National Technical University of Athens
School of Electrical and Computer Engineering
Division of Industrial Electric Devices and Decision Systems

Machine Learning Methods for Recognizing Brain Disorders

Ph.D. Dissertation

Loukas Ilias

Supervisor: Prof. Dimitris Askounis

Athens, December 2024

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Μέθοδοι Μηχανικής Μάθησης για την Αναγνώριση Διαταραχών του Εγκεφάλου

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

του

ΛΟΥΚΑ ΗΛΙΑ

Επιβλέπων Ε.Μ.Π.: Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π.

Αθήνα, Δεκέμβριος 2024

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Ηλεκτρικων Βιομηχανικων Διαταξεων και Συστηματων Αποφασεων

Μέθοδοι Μηχανικής Μάθησης για την Αναγνώριση Διαταραχών του Εγκεφάλου

Διδακτορικη Διατριβη

του

ΛΟΥΚΑ ΗΛΙΑ

Συμβουλευτική Επιτροπή: Δημήτριος Ασκούνης, Καθηγητής ΕΜΠ
Ιωάννης Ψαρράς, Καθηγητής ΕΜΠ
Χρυσόστομος Δούκας, Καθηγητής ΕΜΠ

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την 19η Δεκεμβρίου 2024.

(Υπογραφή)
....
Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π.

(Υπογραφή)
....
Ιωάννης Ψαρράς
Καθηγητής Ε.Μ.Π

(Υπογραφή)
....
Χρυσόστομος Δούκας
Καθηγητής Ε.Μ.Π.

(Υπογραφή)
....
Βασίλειος Ασημακόπουλος
Καθηγητής Ε.Μ.Π.

(Υπογραφή)
....
Γεώργιος Ματσόπουλος
Καθηγητής Ε.Μ.Π

(Υπογραφή)
....
Γρηγόριος Μέντζας
Καθηγητής Ε.Μ.Π.

(Υπογραφή)
....
Ευριπίδης Λουκής
Καθηγητής Πανεπιστημίου Αιγαίου

Αθήνα, Δεκέμβριος 2024

(Υπογραφή)

....

Λουκας Ηλιας
Διδάκτωρ Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν το συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Περίληψη

Οι διαταραχές του εγκεφάλου αποτελούν μία από τις μεγαλύτερες προκλήσεις για την υγεία. Υπολογίζεται ότι περίπου 165 εκατομμύρια άνθρωποι πάσχουν από εγκεφαλική διαταραχή στην Ευρώπη, ενώ 1 στους 3 ανθρώπους θα υποφέρει από εγκεφαλική διαταραχή κάποια στιγμή στη ζωή του. Μερικοί τύποι εγκεφαλικών διαταραχών είναι οι ακόλουθοι: Νόσος Αλτσχάιμερ, διάφοροι τύποι άνοιας, επιληψία, νόσος Πάρκινσον, ψυχικές διαταραχές, κ.ά. Αυτές οι διαταραχές επηρεάζουν τον τρόπο με τον οποίο οι άνθρωποι σκέφτονται, αισθάνονται ή εκτελούν καθημερινές δραστηριότητες. Ωστόσο, εάν αυτές οι διαταραχές διαγνωστούν έγκαιρα και το άτομο λάβει την κατάλληλη φαρμακευτική αγωγή, η εξέλιξή τους μπορεί να καθυστερήσει σημαντικά. Για το λόγο αυτό, η έγκαιρη διάγνωση είναι καθοριστική. Η Τεχνητή Νοημοσύνη (ΤΝ) μετασχηματίζει τον τρόπο με τον οποίο αντιμετωπίζουμε κοινωνικά ζητήματα ενισχύοντας την ευημερία τόσο των ατόμων όσο και των κοινοτήτων. Ο όρος "ΤΝ για το Κοινωνικό Καλό", επίσης γνωστός ως "ΤΝ για το Κοινωνικό Αντίκτυπο", είναι ένα νέο πεδίο έρευνας που στοχεύει στην αντιμετώπιση μερικών από τα πιο σημαντικά κοινωνικά, περιβαλλοντικά και δημόσια υγειονομικά προβλήματα που υπάρχουν σήμερα. Η παρούσα διδακτορική διατριβή έχει ως στόχο να συμβάλει σε αυτό το νέο πεδίο με την ανάπτυξη σύγχρονων μεθόδων μηχανικής μάθησης, για την αναγνώριση τριών μείζονων διαταραχών του εγκεφάλου, συμπεριλαμβανομένης της κατάθλιψης, άνοιας της νόσου Αλτσχάιμερ και επιληψίας.

Η κατάθλιψη συνεπάγεται μεγάλο αριθμό συμπτωμάτων, όπως απώλεια ενδιαφέροντος, θυμό, απαισιοδοξία, αλλαγές στο βάρος, αισθήματα ανικανότητας, σκέψεις αυτοκτονίας και πολλά άλλα. Τα μέσα κοινωνικής δικτύωσης χρησιμοποιούνται σε καθημερινή βάση από ανθρώπους, οι οποίοι εκφράζουν τις σκέψεις και τα συναισθήματά τους συζητώντας με άλλους χρήστες. Οι υπάρχουσες εργασίες χρησιμοποιούν δεδομένα από τα μέσα κοινωνικής δικτύωσης με σκοπό τον εντοπισμό καταθλιπτικών δημοσιεύσεων. Οι εργασίες αυτές χρησιμοποιούν μοντέλα που βασίζονται σε μετασχηματιστές (transformers). Ωστόσο, αυτά τα μοντέλα συχνά δεν μπορούν να συλλάβουν πλούσια τεκμηριωμένη γνώση. Επίσης, η ομιλία είναι ένας αξιόπιστος βιοδείκτης για τη διάγνωση της κατάθλιψης, καθώς οι άνθρωποι με κατάθλιψη παρουσιάζουν μειωμένη παραγωγικότητα λεκτικής δραστηριότητας και "άψυχο"ήχο ομιλίας. Ωστόσο, οι υπάρχουσες μέθοδοι χρησιμοποιούν μονοτροπικά μοντέλα, εφαρμόζουν στρατηγικές early, intermediate, και late fusion για τη συγχώνευση των διαφορετικών τροπικοτήτων, βασίζονται στην εξαγωγή χαρακτηριστικών και εκτελούν τις προσεγγίσεις τους μόνο στην αγγλική γλώσσα. Η άνοια στη νόσο Αλτσχάιμερ χαρακτηρίζεται από απώλεια μνήμης, ενώ επηρεάζει τη γλώσσα και την ομιλία. Προηγούμενες εργασίες χρησιμοποιούν την ομιλία και απομαγνητοφωνήσεις για την αναγνώριση της άνοιας. Ωστόσο, οι προηγούμενες εργασίες επικεντρώνονται απλώς στη βελτίωση της απόδοσης των προτεινόμενων μοντέλων, βασίζονται στην εξαγωγή χαρακτηριστικών, ενώ οι στρατηγικές early και late fusion χρησιμοποιούνται όσον αφορά τις πολυτροπικές προσεγγίσεις, δηλαδή προσεγγίσεις που χρησιμοποιούν τόσο την ομιλία όσο και το απομαγνητοφωνημένο κείμενο. Οι επιληπτικές κρίσεις συνεπάγονται κοινωνικό στίγμα. Οι υπάρχουσες εργασίες βασίζονται στην εξαγωγή χαρακτηριστικών από το ηλεκτροεγκεφαλογράφημα (ΗΕΓ) ή στη διαίρεση των σημάτων ΗΕΓ σε πολλαπλά υποσήματα και στην χρησιμοποίηση τεχνικών majority vote στη συνέχεια.

Αυτή η διδακτορική διατριβή είναι η πρώτη που διερευνά συστηματικά διάφορες μεθόδους για τον εντοπισμό (i) της κατάθλιψης χρησιμοποιώντας αναρτήσεις στα μέσα κοινωνικής δικτύωσης και ομιλία, (ii) ασθενών με άνοια της νόσου Αλτσχάιμερ και πρόβλεψης των βαθμολογιών τους μέσω μίας σύντομης εξέτασης της νοητικής κατάστασης - Βραχεία Κλίμακα Εκτίμησης των Νοητικών Λειτουργιών (Mini Mental State Examination) με χρήση αυθόρμητου λόγου, (iii) επιληψίας μέσω σημάτων ΗΕΓ μονού καναλιού. Οι βασικές συνεισφορές της διατριβής είναι οι εξής: Αρχικά, εισάγονται δύο μέθοδοι για την αναγνώριση της κατάθλιψης. Όσον αφορά την πρώτη μέθοδο, εισάγεται η εργασία της διάγνωσης της κατάθλιψης στα μέσα κοινωνικής δικτύωσης και προτείνεται μια μέθοδος για την ενσωμάτωση εξωτερικών γλωσσικών πληροφοριών σε προεκπαιδευμένα γλωσσικά μοντέλα (π.χ. BERT, MentalBERT). Αναδεικνύεται, έτσι, ότι η ενσωμάτωση γλωσσικών χαρακτηριστικών είναι ευεργετική για την αναγνώριση της κατάθλιψης. Όσον αφορά τη δεύτερη μέθοδο, εισάγεται μία προσέγγιση, η οποία βασίζεται στη χρήση ομιλίας και παραγόμενων από μηχανή (automatic) απομαγνητοφωνημένων κειμένων. Για τον εντοπισμό της άνοιας, βελτιστοποιούνται τα γλωσσικά μοντέλα που βασίζονται σε μετασχηματιστές (transformers) και παρουσιάζονται προσεγγίσεις επεξηγησιμότητας (explainability) και γλωσσικές αναλύσεις για τη διερεύνηση των διαφορών στη γλώσσα μεταξύ υγιών ατόμων και ασθενών με άνοια. Επίσης, εισάγονται μέθοδοι για τη συγχώνευση των διαφορετικών τροπικοτήτων (ομιλία, κείμενο), το καλιμπράρισμα (calibration) των προτεινόμενων μοντέλων με στόχο την αποφυγή δημιουργίας υπερβολικά σίγουρων μοντέλων, την ενίσχυση – βελτίωση των δικτύων αυτοπροσοχής (self – attention) με πληροφορίες σχετικές με τα συμφραζόμενα και την αυτόματη δημιουργία αρχιτεκτονικών Συνελικτικών Νευρωνικών Δικτύων με χρήση τεχνικών αυτόματης αναζήτησης αρχιτεκτονικών νευρωνικού δικτύου. Τέλος, παρουσιάζεται μια πολυτροπική προσέγγιση για την ανίχνευση της επιληψίας αξιοποιώντας μονοκάναλα σήματα ΗΕΓ. Όλα τα πειράματα διεξάγονται σε δημοσίως διαθέσιμα σύνολα δεδομένων.

Αυτή η διδακτορική διατριβή αποτελεί ένα πρώτο, θεμελιώδες βήμα μεταξύ άλλων πρόσφατων προσπαθειών, προς τη βελτίωση της απόδοσης των αυτόματων συστημάτων που στοχεύουν στην αναγνώριση διαφόρων διαταραχών του εγκεφάλου με τη χρήση σύγχρονων τεχνικών βαθιάς μάθησης, προωθεί περαιτέρω την εφαρμογή των νέων τεχνολογιών και ρίχνει φως στα αναδυόμενα πεδία της επεξεργασίας κειμένου, ομιλίας, εικόνας και σήματος.

Λέξεις Κλειδιά

Άνοια της νόσου Αλτσχάιμερ, Επιληψία, Κατάθλιψη, Μέσα Κοινωνικής Δικτύωσης, Ομιλία, Απομαγνητοφωνημένο Κείμενο, Ηλεκτροεγκεφαλογράφημα, Μηχανική Μάθηση

Abstract

Brain disorders represent a significant health challenge. It is estimated that approximately 165 million people suffer from a brain disorder in Europe, while 1 in 3 people will experience such a disorder during their lifetime. Some types of the brain disorders are the following: Alzheimer's disease, dementias, epilepsy, Parkinson's disease, Mental disorders, and more. These disorders affect the way people think, feel, or perform daily activities. However, if these disorders are diagnosed early and the person receives suitable medication, their progression may be delayed. For this reason, early diagnosis is crucial. Artificial Intelligence (AI) holds the promise of transforming how we tackle societal issues and enhancing the welfare of both individuals and communities. "AI for Social Good", also known as "AI for Social Impact" is a new research field aiming to tackle some of the most important social, environmental, and public health challenges that exist today. Another main aim of the "AI for Social Good" is to address the United Nations Sustainable Development Goals (UNSDGs). This PhD thesis aims to contribute to this new field by developing modern machine learning methods, with a particular focus on three major categories (Depression, Alzheimer's Dementia and Epilepsy).

Depression entails a great number of symptoms, including loss of interest, anger, pessimism, changes in weight, feelings of worthlessness, thoughts of suicide, and many more. Social media are used on a daily basis by people, who express their thoughts, feelings by discussing with other users. Prior work employs transformer-based models. However, these models often cannot capture rich factual knowledge. Also, speech is a reliable biomarker for diagnosing depression, since depressed people present decreased verbal activity productivity and "lifeless" sounding speech. However, existing methods employ unimodal models, use early, intermediate, or late fusion strategies to fuse the different modalities, rely on feature extraction, and perform their approaches only in the English language. Alzheimer's dementia is characterized by loss of memory, while it affects language and speech. Previous work utilizes speech and transcripts for recognizing dementia. However, prior work focuses on just improving the performance of proposed models, relies on feature extraction, while early and late fusion strategies are employed in terms of multimodal approaches, i.e., approaches employing both speech and transcripts. Epilepsy and seizures entail social stigma. Existing works rely on extraction of handcrafted features from electroencephalography (EEG) or dividing the EEG signals into multiple sub-signals and exploiting majority vote approaches.

This PhD thesis is the first to systematically investigate various methods for identifying (i) depression by utilizing posts in social media and spontaneous speech, (ii) AD patients and predicting their Mini Mental State Examination scores through spontaneous speech, (iii) epilepsy through single-channel EEG signals. The key contributions of our work are the following: First, we introduce two methods for identifying depression. Regarding the first approach, we present the task of predicting depression in social media and propose a method for injecting external linguistic information into novel pretrained neural language models (e.g. BERT). We show that incorporating linguistic features is beneficial to depression recognition task. In terms of the second approach, we introduce a method which identifies depression based on speech and automatic transcripts. Secondly, for identifying dementia, we fine-tune language models based on transformers and present explainable approaches and linguistic analyses to investigate differences in language between healthy and AD patients. Thirdly, we introduce methods for fusing the different modalities (speech, text), calibrating the proposed models, enhancing the self-attention networks with contextual information, and automatically generating Convolutional Neural Network architectures (Neural Architecture Search). Finally, we present a multimodal approach for detecting epilepsy by exploiting single – channel EEG signals. All experiments are conducted on publicly available datasets.

This PhD thesis represents a first, fundamental step among other recent efforts towards improving the performance of automatic systems aiming at recognizing various brain disorders using modern deep learning techniques. This thesis further advances the application of new technologies and sheds light on the emerging fields of text, speech, image and signal processing.

Keywords

Alzheimer's dementia, Depression, Epilepsy, Social Media, spontaneous speech, Electroencephalogram, deep learning, transformers, explainability, interpretability, multi-task learning, multimodal

Acknowledgements

Με την παρούσα διατριβή κλείνει ένας κύκλος τριών ετών, ο οποίος μου προσέφερε πολύτιμες γνώσεις και εμπειρίες. Θα ήθελα να ευχαριστήσω όλους όσους με στήριξαν και συνεργάστηκαν μαζί μου αυτά τα τρία χρόνια.

Αρχικά, ιδιαίτερες ευχαριστίες οφείλω στον επιβλέποντα Καθ. Δημήτρη Ασκούνη για τη στήριξη, καθοδήγηση και άψογη συνεργασία όλα αυτά τα χρόνια. Ιδιαίτερες ευχαριστίες οφείλω, επίσης, στον Διευθυντή του Εργαστηρίου Συστημάτων Αποφάσεων και Διοίκησης, Καθ. Ιωάννη Ψαρρά, ο οποίος ως μέλος της τριμελούς επιτροπής είχε καθοριστικό ρόλο σε αυτήν την πορεία. Επίσης, θα ήθελα να ευχαριστήσω τον Καθ. Χάρη Δούκα για την τιμή που μου έκανε να είναι μέλος της συμβουλευτικής επιτροπής αλλά και τη συνεργασία που είχαμε στα πλαίσια δημοσιεύσεων.

Παράλληλα με την εκπόνηση της διδακτορικής μου διατριβής, αυτά τα 3 χρόνια συμμετείχα ως ερευνητής σε ευρωπαϊκά ερευνητικά προγράμματα. Για τον λόγο αυτό, ιδιαίτερες ευχαριστίες οφείλω σε όλα τα μέλη της ομάδας του έργου MES-CoBraD για την άψογη συνεργασία που είχαμε όλα αυτά τα χρόνια. Συγκεκριμένα, ευχαριστώ ιδιαίτερα τους Χ. Ντάνο, Γ. Δούκα, Μ. Κοντούλη, Κ. Αλεξάκη, Θ. Πουντρίδη, Γ. Λαδικό, Β. Λάμπρου. Ακόμα, ευχαριστώ όλους τους συναδέλφους με τους οποίους συνεργαστήκαμε όλα τα χρόνια της παρουσίας μου στο Εργαστήριο Συστημάτων Αποφάσεων και Διοίκησης.

Τέλος θα ήθελα να ευχαριστήσω τους γονείς μου, Κωνσταντίνο και Δέσποινα, και τον αδερφό μου Γιάννη, για την στήριξή τους όλα αυτά τα χρόνια, την αμέριστη συμπαράσταση, πίστη, υποστήριξη και υπομονή προς το πρόσωπό μου καθ' όλη τη διάρκεια εκπόνησης της παρούσας διδακτορικής διατριβής.

Εκτεταμένη Περίληψη

1.1 Εισαγωγή

Οι διαταραχές του εγκεφάλου αποτελούν μία από τις μεγαλύτερες προκλήσεις για την υγεία. Υπολογίζεται ότι περίπου 165 εκατομμύρια άνθρωποι πάσχουν από εγκεφαλική διαταραχή στην Ευρώπη, ενώ 1 στους 3 ανθρώπους θα υποφέρει από εγκεφαλική διαταραχή κάποια στιγμή στη ζωή του. Μερικοί τύποι εγκεφαλικών διαταραχών είναι οι ακόλουθοι: Νόσος Αλτσχάιμερ, διάφοροι τύποι άνοιας, επιληψία, Νόσος Πάρκινσον, Ψυχικές διαταραχές και άλλα. Αυτές οι διαταραχές επηρεάζουν τον τρόπο με τον οποίο οι άνθρωποι σκέφτονται, αισθάνονται ή εκτελούν καθημερινές δραστηριότητες. Ωστόσο, εάν αυτές οι διαταραχές διαγνωστούν έγκαιρα και το άτομο λάβει την κατάλληλη φαρμακευτική αγωγή, η εξέλιξή τους μπορεί να καθυστερήσει. Για το λόγο αυτό, η έγκαιρη διάγνωση είναι καθοριστική.

Η Τεχνητή Νοημοσύνη (ΤΝ) μετασχηματίζει τον τρόπο με τον οποίο αντιμετωπίζουμε κοινωνικά ζητήματα ενισχύοντας την ευημερία τόσο των ατόμων όσο και των κοινοτήτων. Ο όρος "ΤΝ για το Κοινωνικό Καλό", επίσης γνωστός ως "ΤΝ για το Κοινωνικό Αντίκτυπο", είναι ένα νέο πεδίο έρευνας που στοχεύει στην αντιμετώπιση μερικών από τα πιο σημαντικά κοινωνικά, περιβαλλοντικά και δημόσια υγειονομικά προβλήματα που υπάρχουν σήμερα. Η παρούσα διδακτορική διατριβή έχει ως στόχο να συμβάλει σε αυτό το νέο πεδίο με την ανάπτυξη σύγχρονων μεθόδων μηχανικής μάθησης, με ιδιαίτερη έμφαση σε τρεις μεγάλες κατηγορίες (Κατάθλιψη, Άνοια της νόσου Αλτσχάιμερ και Επιληψία).

Η κατάθλιψη συνεπάγεται μεγάλο αριθμό συμπτωμάτων, όπως απώλεια ενδιαφέροντος, θυμό, απαισιοδοξία, αλλαγές στο βάρος, αισθήματα ανικανότητας, σκέψεις αυτοκτονίας και πολλά άλλα. Η άνοια στη νόσο Αλτσχάιμερ χαρακτηρίζεται από απώλεια μνήμης, ενώ επηρεάζει τη γλώσσα και την ομιλία. Οι επιληπτικές κρίσεις συνεπάγονται κοινωνικό στίγμα.

1.1.1 Στόχος Διδακτορικής Διατριβής & Συνεισφορές αυτής

Με βάση το περιεχόμενο της έρευνας και τις κλινικές ανάγκες όπως περιγράφονται παραπάνω, ο συνολικός στόχος αυτής της διδακτορικής διατριβής είναι η βελτίωση της ανίχνευσης των διαταραχών του εγκεφάλου χρησιμοποιώντας προηγμένες τεχνικές μηχανικής μάθησης. Ειδικότερα, αυτή η διατριβή παρουσιάζει αυτόματα συστήματα για την αναγνώριση τριών μείζονων διαταραχών του εγκεφάλου, συμπεριλαμβανομένης της κατάθλιψης, της άνοιας της νόσου του Αλτσχάιμερ και της επιληψίας.

Όσον αφορά την κατάθλιψη, η διατριβή εξετάζει δύο μεθόδους αναγνώρισής της μέσω των δημοσιεύσεων στα μέσα κοινωνικής δικτύωσης και της ομιλίας. Όσον αφορά την πρώτη μέθοδο, χρησιμοποιούνται δεδομένα μέσων κοινωνικής δικτύωσης και δημιουργούνται εργαλεία βασισμένα στην επεξεργασία φυσικής γλώσσας για την ανίχνευση καταθλιπτικών αναρτήσεων. Επιπλέον, αυτή η διατριβή αναζητά να βρει διαφορές στη γλώσσα μεταξύ καταθλιπτικών ατόμων και μη-καταθλιπτικών μέσω μιας λεπτομερούς γλωσσολογικής ανάλυσης. Όσον αφορά τη δεύτερη μέθοδο, προτείνεται ένα βαθύ νευρωνικό δίκτυο βασισμένο στα δίκτυα μετασχηματιστών και τις πολυτροπικές μεθόδους συγχώνευσης και εξετάζεται εάν η πρόβλεψη της ηλικίας, του φύλου και του επιπέδου εκπαίδευσης συμβάλλουν στην αύξηση της απόδοσης της αναγνώρισης της κατάθλιψης.

Όσον αφορά την άνοια της νόσου Αλτσχάιμερ, ενθαρρυνόμενοι από το γεγονός ότι άτομα με άνοια παρουσιάζουν ελλείμματα στη γλώσσα και την ομιλία, αυτή η διατριβή χρησιμοποιεί ηχογραφήσεις της αυθόρμητης ομιλίας και δημιουργεί αυτόματα συστήματα βασισμένα στην επεξεργασία φυσικής γλώσσας και την επεξεργασία ήχου. Συγκεκριμένα, προσαρμόζουμε μοντέλα βασισμένα σε μετασχηματιστές, εκμεταλλευόμαστε τεχνικές επεξηγησιμότητας και γλωσσολογικές αναλύσεις και εξερευνούμε ορισμένα γλωσσικά χαρακτηριστικά που είναι χρήσιμα για την ανίχνευση της μείωσης των γνωστικών ικανοτήτων. Αυτή η διατριβή αναζητά, επίσης, να χρησιμοποιήσει πολυτροπικά μοντέλα εκμεταλλευόμενα και την ομιλία και τα απομαγνητοφωνημένα κείμενα αντί να επικεντρωθεί μόνο στα λεκτικά, ακουστικά ή οπτικά χαρακτηριστικά.

Όσον αφορά την επιληψία, ενθαρρυνόμενοι από το γεγονός ότι η παρακολούθηση σημάτων Ηλεκτροεγκεφαλογραφήματος (ΗΕΓ) από νευρολόγους είναι μια κουραστική και ευάλωτη σε λάθη εργασία, παρουσιάζουμε ένα νέο αυτόματο σύστημα βασισμένο σε πολυτροπική μέθοδο για τη διάγνωση της επιληψίας.

Συνολικά, οι κύριες συνεισφορές αυτής της διατριβής είναι οι εξής:

- Προτείνεται μια επεξηγήσιμη προσέγγιση και μια μελέτη γλωσσολογικής ανάλυσης διερευνώντας τα γλωσσικά χαρακτηριστικά των ασθενών με άνοια. Σε αντίθεση με προηγούμενες εργασίες, οι οποίες απλώς εκπαιδεύουν αλγόριθμους μηχανικής μάθησης για την ανίχνευση ασθενών με άνοια, αυτή η διατριβή χρησιμοποιεί μια επεξηγήσιμη προσέγγιση και εισάγει μια γλωσσολογική ανάλυση. Και οι δύο προσεγγίσεις αποκαλύπτουν διαφορές στο λεξιλόγιο μεταξύ υγιών ατόμων και ασθενών με άνοια. Χρησιμοποιούμε την ίδια μελέτη γλωσσολογικής ανάλυσης σε ένα σύνολο δεδομένων που περιέχει καταθλιπτικά κείμενα και βρίσκουμε διαφορές στη γλώσσα μεταξύ υγιών και ανθρώπων με κατάθλιψη.

- Εισάγονται πολυτροπικά μοντέλα. Σε αντίθεση με τις υπάρχουσες ερευνητικές πρωτοβουλίες, οι οποίες χρησιμοποιούν στρατηγικές early, intermediate, late fusion, αυτή η διατριβή εισάγει νέες μεθόδους, προκειμένου να συγχωνεύσει τις διαφορετικές τροπικότητες. Συγκεκριμένα, αυτή η διατριβή επεκτείνει τις προηγούμενες εργασίες αξιοποιώντας μεθόδους, όπως Gated Multimodal Unit, Cross-Modal Attention Layer, Cross-Attention Layer with Gated Self-Attention, Optimal Transport Domain Adaptation methods, κ.ά. Αυτές οι πολυτροπικές προσεγγίσεις υιοθετούνται σε μία σειρά πειραμάτων, όπως ανίχνευση κατάθλιψης μέσω αναρτήσεων στα μέσα κοινωνικής δικτύωσης και αυθόρμητου λόγου, αναγνώριση άνοιας μέσω ομιλίας και απομαγνητοφωνημένου κειμένου, ανίχνευση επιληψίας μέσω ενός καναλιού Ηλεκτροεγκεφαλογραφήματος, και στοχεύουν στην αύξηση της απόδοσης που επιτυγχάνουν τα μονοτροπικά μοντέλα.

- Παρουσίαση μεθόδων βαθμονόμησης (calibration) βαθέων νευρωνικών δικτύων. Προηγούμενες εργασίες αξιολογούν τα βαθιά νευρωνικά δίκτυα με βάση μόνο την απόδοση (performance). Αυτή η διατριβή επεκτείνει τις προηγούμενες εργασίες με αξιοποίηση μεθόδων για τη βαθμονόμηση των μοντέλων και την αξιολόγηση αυτών των μοντέλων αξιοποιώντας μετρήσεις απόδοσης και βαθμονόμησης. Η βαθμονόμηση έχει ως στόχο την αποφυγή δημιουργίας υπερβολικά σίγουρων μοντέλων. Αυτές οι προσεγγίσεις διεξάγονται σε σύνολα δεδομένων που σχετίζονται με την κατάθλιψη και την άνοια.

- Ενσωμάτωση μεθόδου αυτόματης αναζήτησης αρχιτεκτονικής νευρωνικού δικτύου (Neural Architecture Search) σε προτεινόμενα μοντέλα. Σε αντίθεση με προηγούμενες εργασίες, που χρησιμοποιούν σταθερές (fixed) αρχιτεκτονικές, αυτή η διατριβή ενσωματώνει μια προσέγγιση αυτόματης αναζήτησης αρχιτεκτονικής νευρωνικού δικτύου, που ονομάζεται DARTS, σε ένα βαθύ νευρωνικό δίκτυο για την αυτόματη δημιουργία μιας αρχιτεκτονικής συνελικτικών νευρωνικών δικτύων (Convolutional Neural Networks). Με τον τρόπο αυτό, βρίσκουμε τη βέλτιστη αρχιτεκτονική CNN στο δικό μας task.

- Εισάγονται βαθιά νευρωνικά δίκτυα, τα οποία μπορούν να εκπαιδευτούν με τρόπο end-to-end, εξαλείφοντας τη χρονοβόρα διαδικασία εξαγωγής χαρακτηριστικών. Αντίθετα με προηγούμενες ερευνητικές εργασίες που εξάγουν μεγάλο αριθμό χαρακτηριστικών, αξιοποιούν τεχνικές επιλογής χαρακτηριστικών ή μείωσης διαστάσεων και εκπαιδεύουν παραδοσιακούς αλγόριθμους μηχανικής μάθησης, η παρούσα διατριβή στοχεύει στην εξάλειψη της ανάγκης για εξαγωγή χαρακτηριστικών, προτείνοντας βαθιά νευρωνικά δίκτυα και μοντέλα βασισμένα σε μετασχηματιστές.

- Ενίσχυση των δικτύων αυτοπροσοχής με πληροφορίες σχετικές με τα συμφραζόμενα. Αυτή η διατριβή στοχεύει να ενισχύσει το επίπεδο αυτοπροσοχής προσθέτοντας πληροφορίες σχετικές με τα συμφραζόμενα. Συγκεκριμένα, η παρούσα διατριβή παρουσιάζει τρεις στρατηγικές για την κατασκευή ενός διανύσματος, που λαμβάνει υπόψη το περιεχόμενο της πρότασης, σε ένα εκπαιδεύσιμο από άκρο σε άκρο βαθύ νευρωνικό δίκτυο. Αυτή η προσέγγιση πραγματοποιείται σε σύνολα δεδομένων που σχετίζονται με το task της άνοιας Alzheimer.

- Εισάγονται μοντέλα μάθησης πολλαπλών εργασιών. Αυτή η διατριβή προτείνει αρχιτεκτονικές μάθησης πολλαπλών εργασιών για την αναγνώριση της κατάθλιψης και της άνοιας Alzheimer. Αρχικά, η παρούσα διατριβή παρουσιάζει μια προσέγγιση μάθησης πολλαπλών εργασιών για την ταυτόχρονη μοντελοποίηση των εργασιών αναγνώρισης της κατάθλιψης, του επιπέδου εκπαίδευσης, της ηλικίας και του φύλου. Στη συνέχεια, η παρούσα διατριβή εισάγει αρχιτεκτονικές μάθησης πολλαπλών εργασιών με στόχο την πρόβλεψη των εργασιών ανίχνευσης AD και πρόβλεψης MMSE.

1.2 Διάγνωση Κατάθλιψης

Η κατάθλιψη είναι μια σοβαρή διαταραχή της διάθεσης, η οποία επηρεάζει τον τρόπο που οι άνθρωποι αισθάνονται και εκτελούν καθημερινές δραστηριότητες. Οι άνθρωποι χρησιμοποιούν τα μέσα κοινωνικής δικτύωσης για να εκφράσουν τις σκέψεις και τα συναισθήματά τους μέσω αναρτήσεων. Επομένως, τα μέσα κοινωνικής δικτύωσης παρέχουν βοήθεια για την έγκαιρη ανίχνευση ψυχικών καταστάσεων. Εκτός από την αναγνώριση της κατάθλιψης μέσω αναρτήσεων στα μέσα κοινωνικής δικτύωσης, η ομιλία είναι ένας αξιόπιστος βιοδείκτης για τη διάγνωση της κατάθλιψης, καθώς οι καταθλιπτικοί άνθρωποι παρουσιάζουν μειωμένη παραγωγικότητα λεκτικής δραστηριότητας και ομιλία που ακούγεται ¨άψυχη¨.

Σε αυτό το κεφάλαιο, παρουσιάζουμε δύο προσεγγίσεις για την αναγνώριση της κατάθλιψης. Συγκεκριμένα, στην Ενότητα 1.2.1 παρουσιάζουμε μια προσέγγιση για την αναγνώριση της κατάθλιψης μέσω αναρτήσεων στα μέσα κοινωνικής δικτύωσης, ενώ η Ενότητα 1.2.2 παρουσιάζει μια μέθοδο για την αναγνώριση της κατάθλιψης χρησιμοποιώντας αυθόρμητη ομιλία.

1.2.1 Διάγνωση Κατάθλιψης στα Μέσα Κοινωνικής Δικτύωσης

1.2.1.1 Κίνητρο

Η κατάθλιψη[1] συνεπάγεται μεγάλο αριθμό συμπτωμάτων, όπως απώλεια ενδιαφέροντος, θυμός, απαισιοδοξία, αλλαγές στο βάρος, συναισθήματα ανικανότητας, σκέψεις αυτοκτονίας και πολλά άλλα. Σύμφωνα με στον Παγκόσμιο Οργανισμό Υγείας (ΠΟΥ)[2], περίπου 280 εκατομμύρια άνθρωποι στον κόσμο έχουν κατάθλιψη. Κίνα, Ινδία, Ηνωμένες Πολιτείες, Ρωσία, Ινδονησία, και η Νιγηρία είναι μερικές από τις χώρες που παρουσιάζουν τα υψηλότερα ποσοστά κατάθλιψης[3]. Τα άτομα με άγχος και κατάθλιψη χρησιμοποιούν πλατφόρμες μέσων κοινωνικής δικτύωσης, συμπεριλαμβανομένων των X/Twitter και Reddit, και μοιράζονται τις σκέψεις και τα συναισθήματά τους μέσω αναρτήσεων ή σχολίων με άλλους χρήστες. Επομένως, τα μέσα κοινωνικής δικτύωσης αποτελούν μια πολύτιμη πηγή πληροφοριών.

Οι υπάρχουσες ερευνητικές εργασίες χρησιμοποιούν τα δεδομένα των μέσων κοινωνικής δικτύωσης, για να αναγνωρίσουν καταθλιπτικές και αγχωτικές δημοσιεύσεις. Η πλειονότητα αυτών των ερευνητικών εργασιών χρησιμοποιεί εξαγωγή χαρακτηριστικών και εκπαιδεύει ρηχούς αλγόριθμους μηχανικής μάθησης [1, 2]. Η εξαγωγή χαρακτηριστικών αποτελεί μια χρονοβόρα διαδικασία και απαιτεί εξειδίκευση στον τομέα, καθώς οι ερευνητές ενδέχεται να μη βρουν το βέλτιστο σύνολο χαρακτηριστικών για το συγκεκριμένο πρόβλημα. Για την αντιμετώπιση αυτών των περιορισμών, άλλες προσεγγίσεις [3] χρησιμοποιούν βαθιά νευρωνικά δίκτυα, συμπεριλαμβανομένων των συνελικτικών νευρωνικών δικτύων (CNN), BiLSTM, και ούτω καθεξής, ή μοντέλα βασισμένα σε μετασχηματιστές (transformers). Επιπλέον, υπάρχουν ερευνητικές μελέτες που χρησιμοποιούν στρατηγικές late fusion [4]. Ωστόσο, αυτές οι προσεγγίσεις αυξάνουν ουσιαστικά τον χρόνο εκπαίδευσης, αφού πολλά μοντέλα πρέπει

[1] https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide

[2] https://www.who.int/news-room/fact-sheets/detail/depression

[3] https://pulsetms.com/resources/around-world/

να εκπαιδεύονται χωριστά. Επιπλέον, πρόσφατα έγιναν μελέτες που δείχνουν ότι τα μοντέλα που βασίζονται σε μετασχηματιστές δυσκολεύονται ή αποτυγχάνουν να συλλάβουν πλούσια γνώση [5, 6]. Για τον λόγο αυτό, έχουν προταθεί μέθοδοι [7, 8, 9, 10] για τη βελτίωση αυτών των μοντέλων με εξωτερικές πληροφορίες ή πρόσθετες λεπτομέρειες. Επιπλέον, η αξιοπιστία της εμπιστοσύνης ενός μοντέλου ML στις προβλέψεις του, που δηλώνεται ως βαθμονόμηση [11, 12], είναι κρίσιμης σημασίας για εφαρμογές υψηλού κινδύνου, όπως η απόφαση για το αν θα εμπιστευτεί ο γιατρός μια ιατρική διάγνωση - πρόβλεψη μέσω ενός αλγορίθμου μηχανικής μάθησης.

1.2.1.2 Δεδομένα

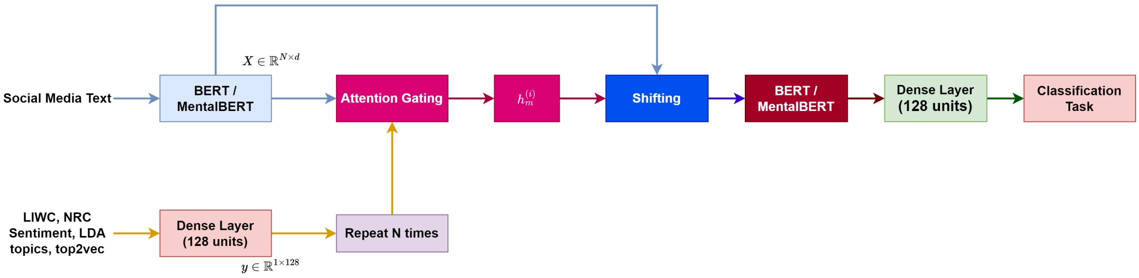
Depression_Mixed. Χρησιμοποιούμε το σύνολο δεδομένων που παρουσιάστηκε στο [13]. Αυτό το σύνολο δεδομένων αποτελείται από 2822 αναρτήσεις. Περιλαμβάνει αναρτήσεις τόσο από το Reddit όσο και από αγγλόφωνα φόρουμ κατάθλιψης [14]. Όσον αφορά τα αγγλόφωνα φόρουμ κατάθλιψης [14], οι συγγραφείς αντλούν δεδομένα από την ομάδα καρκίνου του μαστού. Συγκεκριμένα, συλλέγουν μία μόνο ανάρτηση από ένα όνομα χρήστη, για να αποφευχθούν πολλαπλές εισαγωγές από έναν μόνο χρήστη. Επιπλέον, οι συγγραφείς επιβεβαιώνουν ότι ο συγγραφέας κάθε ανάρτησης ήταν γυναίκα, ενώ απορρίπτουν αναρτήσεις που αναφέρουν ρητά ότι ο συγγραφέας αντιμετωπίζει κατάθλιψη. Για τη δημιουργία συνόλου δεδομένων με καταθλιπτικές αναρτήσεις, οι συγγραφείς στο [13] υιοθετούν ένα πρωτόκολλο παρόμοιο με αυτό του [15, 16] και αναζητούν εκφράσεις όπως ¨Μόλις μου διαγνώστηκε κατάθλιψη' στο subreddit της κατάθλιψης. Όσον αφορά τις μη-καταθλιπτικές αναρτήσεις, οι συγγραφείς συλλέγουν σύνολα αναρτήσεων που ανήκουν στο subreddit συζητήσεων για τον καρκίνο του μαστού, οικογενειακές συμβουλές και φιλίες στο Reddit.

Depression_Severity. Αυτό το σύνολο δεδομένων περιλαμβάνει αναρτήσεις στο Reddit [17] και αναθέτει σε κάθε ανάρτηση ένα επίπεδο σοβαρότητας της κατάθλιψης, δηλαδή ελάχιστο (2587 αναρτήσεις), ελαφρύ (290 αναρτήσεις), μέτριο (394 αναρτήσεις) και σοβαρό είδος κατάθλιψης (282 αναρτήσεις).

1.2.1.3 Προτεινόμενη Μεθοδολογία

Σε αυτή την ενότητα, περιγράφουμε την προτεινόμενη προσέγγισή μας για τον εντοπισμό καταθλιπτικών αναρτήσεων στα μέσα κοινωνικής δικτύωσης. Η προτεινόμενη μέθοδος μας βασίζεται στην εργασία που εισήχθη από τους Rahman et al. [18] και Jin and Aletras [19]. Αντί για διατροπικές αλληλεπιδράσεις, εισάγουμε επιπλέον γλωσσικές πληροφορίες ως εναλλακτικές απόψεις των δεδομένων σε προεκπαιδευμένα γλωσσικά μοντέλα. Η προτεινόμενη αρχιτεκτονική μας απεικονίζεται στο Σχ. 1.1.

- **NRC.** Το NRC Emotion Lexicon είναι μια λίστα αγγλικών λέξεων και οι συσχετίσεις τους με οκτώ βασικά συναισθήματα (θυμός, φόβος, προσμονή, εμπιστοσύνη, έκπληξη, λύπη, χαρά και αηδία) και δύο συναισθήματα (αρνητικό και θετικό) [20]. Κάθε κείμενο αναπαρίσταται ως ένα διάνυσμα 10 διαστάσεων, όπου κάθε στοιχείο είναι η αναλογία των διακριτικών που ανήκουν σε κάθε κατηγορία.



Σχήμα 1.1: Προτεινόμενη Αρχιτεκτονική για Διάγνωση Κατάθλιψης στα Μέσα Κοινωνικής Δικτύωσης

- **LIWC.** Το LIWC είναι μια προσέγγιση βασισμένη σε λεξικό για την καταμέτρηση λέξεων σε γλωσσικές, ψυχολογικές και τοπικές κατηγορίες [21]. Χρησιμοποιούμε το LIWC 2022 [22] για να αναπαραστήσουμε κάθε κείμενο ως διάνυσμα 117 διαστάσεων.

- **LDA topics.** Πριν εκπαιδεύσουμε το μοντέλο LDA, αφαιρούμε τα stop-words και τα σημεία στίξης. Εκμεταλλευόμαστε το LDA (με 25 θέματα) και εξάγουμε 25 πιθανότητες θεμάτων ανά κείμενο [23]. Αυτές οι πιθανότητες περιγράφουν τα θέματα ενδιαφέροντος κάθε κειμένου. Εμπνευσμένοι από τους Liu et al. [24], χρησιμοποιούμε το ακόλουθο διάνυσμα χαρακτηριστικών:

  - **Global Outlier Standard Score (GOSS):** Για να αξιολογήσουμε το ενδιαφέρον του κειμένου i^{th} σε ένα συγκεκριμένο topic k, σε σύγκριση με τα υπόλοιπα κείμενα, χρησιμοποιούμε το GOSS χαρακτηριστικό:

  $$\mu(x_k) = \frac{\sum_{i=1}^{n} x_{ik}}{n} \tag{1.1}$$

  $$GOSS(x_{ik}) = \frac{x_{ik} - \mu(x_k)}{\sqrt{\sum_i \left(x_{ik} - \mu\left(x_k\right)\right)^2}} \tag{1.2}$$

  Επομένως, κάθε κείμενο αναπαρίσταται ως ένα διάνυσμα 25 διαστάσεων.

- **Top2Vec:** Top2Vec [25] είναι ένας αλγόριθμος για τη μοντελοποίηση θεμάτων, ο οποίος εντοπίζει αυτόματα θέματα που υπάρχουν στο κείμενο και δημιουργεί από κοινού ενσωματωμένα διανύσματα θεμάτων, εγγράφων και λέξεων. Μετά την εκπαίδευση του Top2Vec με την εκμετάλλευση του Universal Sentence Encoder, κάθε κείμενο αναπαρίσταται ως διάνυσμα 512-d.

Χρησιμοποιούμε τα εξής προεκπαιδευμένα μοντέλα: BERT [26] και MentalBERT [27].

Αρχικά, δίνουμε ως είσοδο το κείμενο στα προαναφερθέντα μοντέλα. Έστω $C \in \mathcal{R}^{N\times d}$ είναι η έξοδος του μοντέλου, όπου N δηλώνει το μήκος του κειμένου, ενώ d δηλώνει τη διάσταση των μοντέλων. Για χάρη απλότητας έχουμε παραλείψει τη διάσταση που αναφέρεται στο batch size.

Στη συνέχεια, προβάλλουμε τα διανύσματα χαρακτηριστικών σε διαστάσεις ίσες με 128. Επαναλαμβάνουμε το διάνυσμα χαρακτηριστικών N φορές, έτσι ώστε να διασφαλίσουμε ότι το διάνυσμα χαρακτηριστικών και η έξοδος των μοντέλων που βασίζονται σε μετασχηματιστή μπορούν να συνδεθούν κατά γραμμές. Δεδομένης της αναπαράστασης λέξεων $e^{(i)}$, συνενώνουμε το $e^{(i)}$ με διανύσματα χαρακτηριστικών, δηλ. $h_v^{(i)}$.

$$w_v^{(i)} = \sigma \left(W_{hv}[e^{(i)}; h_v^{(i)}] + b_v \right) \tag{1.3}$$

όπου το σ υποδηλώνει τη συνάρτηση ενεργοποίησης σιγμοειδούς, το W_{hv} είναι ένας πίνακας βάρους και το $w_v^{(i)}$ αντιστοιχεί στην πύλη. Το b_v είναι scalar bias.

Στη συνέχεια, υπολογίζουμε ένα διάνυσμα μετατόπισης $h_m^{(i)}$ πολλαπλασιάζοντας τις διανύσματα (embeddings) με την πύλη (gate).

$$h_m^{(i)} = w_v^{(i)} \cdot \left(W_v h_v^{(i)} \right) + b_m^{(i)} \tag{1.4}$$

όπου W_v είναι ένας πίνακας βαρών ανδ $b_m^{(i)}$ είναι bias vector.

Στη συνέχεια, εφαρμόζουμε το στοιχείο Multimodal Shifting με στόχο να μετατοπίσουμε δυναμικά τις αναπαραστάσεις των λέξεων ενσωματώνοντας το διάνυσμα μετατόπισης $h_m^{(i)}$ στην αρχικό διάνυσμα λέξης.

$$e_m^{(i)} = e^{(i)} + \alpha h_m^{(i)} \tag{1.5}$$

$$\alpha = min \left(\frac{||e^{(i)}||_2}{||h_m^{(i)}||_2} \beta, 1 \right) \tag{1.6}$$

, όπου β είναι μια υπερπαράμετρος. Στη συνέχεια, εφαρμόζουμε ένα layer normalization [28] και ένα dropout layer [29] στο $e_m^{(i)}$. Στη συνέχεια, τα συνδυασμένα διανύσματα (embeddings τροφοδοτούνται σε ένα μοντέλο BERT/MentalBERT.

Παίρνουμε την έξοδο του μοντέλου (classification token) [CLS] και το περνάμε μέσα από ένα dense layer που αποτελείται από 128 μονάδες με συνάρτηση ενεργοποίησης ReLU. Τέλος, χρησιμοποιούμε ένα dense layer που αποτελείται είτε από δύο μονάδες (binary classification task) είτε από τέσσερις μονάδες (multiclass classification task).

Ονομάζουμε τα προτεινόμενα μοντέλα μας ως Multimodal BERT (M-BERT) και Multimodal MentalBERT (M-MentalBERT) ακολουθούμενα από τα γλωσσικά χαρακτηριστικά που είναι ενσωματωμένα σε αυτά. Για παράδειγμα, η έγχυση χαρακτηριστικών LIWC σε ένα μοντέλο BERT συμβολίζεται ως M-BERT (LIWC).

1.2.1.4 Model Calibration

Προκειμένου να αποφύγουμε τη δημιουργία υπερβολικά σίγουρων μοντέλων, χρησιμοποιούμε label smoothing [30, 31]. Συγκεκριμένα, η μέθοδος label smoothing βαθμονομεί τα μαθημένα μοντέλα έτσι ώστε η εμπιστοσύνη των προβλέψεών τους να ευθυγραμμίζεται περισσότερο με την ακρίβεια των προβλέψεών τους.

Για ένα δίκτυο εκπαιδευμένο με σκληρούς στόχους, το cross-entropy loss ελαχιστοποιείται μεταξύ των πραγματικών στόχων y_k και των εξόδων του δικτύου p_k, όπως στο $H(y,p) =$

$\sum_{k=1}^{K} -y_k log(p_k)$, όπου y_k είναι ¨1’ για τη σωστή κλάση και ¨0’ για την άλλη. Για ένα δίκτυο εκπαιδευμένο με εξομάλυνση ετικετών, ελαχιστοποιούμε το cross-entropy loss μεταξύ των τροποποιημένων στόχων $y_k^{LS_u}$ και των εξόδων του δικτύου p_k.

$$y_k^{LS_u} = y_k \cdot (1 - \alpha) + \frac{\alpha}{K} \tag{1.7}$$

$$H(y, p) = \sum_{k=1}^{K} -y_k^{LS_u} \cdot \log(p_k) \tag{1.8}$$

, όπου α είναι μία παράμετρος εξομάλυνσης και K είναι ο αριθμός των κλάσεων.

1.2.1.5 Αποτελέσματα

Τα αποτελέσματα της προτεινόμενης μας προσέγγισης αναφέρονται στους Πίνακες 1.1 και 1.2. Ειδικότερα, ο Πίνακας 1.1 παρουσιάζει τις επιδόσεις των προτεινόμενων προσεγγίσεων μας στο σύνολο δεδομένων Depression_Mixed, ενώ ο Πίνακας 1.2 αναφέρει τα αποτελέσματα στο σύνολο δεδομένων Depression_Severity.

Πίνακας 1.1: Performance comparison among proposed models and baselines using the DEPRESSION_MIXED dataset.

	Depression_Mixed					
Μοντέλο	Prec.	Rec.	F1-score	Acc.	ECE	ACE
Baselines						
BERT	91.40	91.40	91.40	-	-	-
MentalBERT	89.27	93.14	91.17	91.15	-	-
Baselines - Proposed Approaches (without label smoothing)						
M-BERT (NRC)	90.56	91.84	91.20	91.15	0.072	0.081
M-BERT (LIWC)	90.98	92.02	91.49	92.04	0.054	0.055
M-BERT (LDA topics)	88.07	95.80	91.77	92.04	0.071	0.071
M-BERT (top2vec)	90.97	92.99	91.97	92.21	0.057	0.069
M-MentalBERT (NRC)	90.65	92.65	91.64	91.86	0.031	0.054
M-MentalBERT (LIWC)	93.49	87.78	90.55	91.50	0.057	0.056
M-MentalBERT (LDA topics)	87.97	93.09	90.46	90.44	0.089	0.086
M-MentalBERT (top2vec)	91.63	93.77	92.69	93.27	0.058	0.054
Proposed Approaches (with label smoothing)						
M-BERT (NRC)	89.82	94.81	92.25	92.39	0.059	0.065
M-BERT (LIWC)	93.06	91.78	92.41	92.21	0.034	0.044
M-BERT (LDA topics)	90.16	92.71	91.42	92.39	0.063	0.067
M-BERT (top2vec)	90.34	94.93	92.58	92.57	0.049	0.056
M-MentalBERT (NRC)	91.44	92.52	91.98	92.74	0.042	0.057
M-MentalBERT (LIWC)	94.96	89.42	92.11	92.57	0.055	0.057
M-MentalBERT (LDA topics)	94.81	90.78	92.75	92.92	0.047	0.049
M-MentalBERT (top2vec)	96.12	90.18	93.06	93.45	0.033	0.043

Σχετικά με το σύνολο δεδομένων Depression_Mixed, συγκρίνουμε αρχικά τις προτεινόμενες προσεγγίσεις μας χωρίς εξομάλυνση ετικέτας με τα μοντέλα BERT και MentalBERT. Παρατηρούμε ότι η ενσωμάτωση γλωσσικών χαρακτηριστικών, εκτός από τα χαρακτηριστικά NRC, στο μοντέλο BERT βελτιώνει το F1-score. Ειδικότερα, παρατηρούμε ότι η ενσωμάτωση χαρακτηριστικών top2vec οδηγεί στο υψηλότερο F1-score και ακρίβεια που ανέρχονται σε 91.97% και 92.21% αντίστοιχα, υπερβαίνοντας την απόδοση του μοντέλου BERT στο F1-score κατά 0.57%. Υποθέτουμε ότι η ενσωμάτωση χαρακτηριστικών top2vec επιτυγχάνει καλύτερη απόδοση από την ενσωμάτωση χαρακτηριστικών που προέρχονται από τα θέματα

Πίνακας 1.2: Performance comparison among proposed models and baselines using the DEPRESSION_SEVERITY dataset.

Μοντέλο	W. Prec.	W. Rec.	W. F1-score	ECE	ACE
Baselines					
BERT	72.99	71.97	71.00	-	-
MentalBERT	73.35	70.81	71.67	-	-
Baselines - Proposed Approaches (without label smoothing)					
M-BERT (NRC)	74.48	70.08	69.96	0.107	0.076
M-BERT (LIWC)	73.77	71.74	72.13	0.110	0.078
M-BERT (LDA topics)	74.25	71.80	71.28	0.114	0.079
M-BERT (top2vec)	72.93	71.97	71.00	0.086	0.071
M-MentalBERT (NRC)	74.43	72.58	69.96	0.097	0.069
M-MentalBERT (LIWC)	72.39	72.53	71.95	0.112	0.075
M-MentalBERT (LDA topics)	73.83	72.58	72.58	0.118	0.078
M-MentalBERT (top2vec)	74.63	72.39	72.06	0.103	0.075
Proposed Approaches (with label smoothing)					
M-BERT (NRC)	74.04	72.84	72.81	0.102	0.074
M-BERT (LIWC)	73.68	72.16	72.37	0.094	0.069
M-BERT (LDA topics)	73.24	71.46	71.42	0.112	0.078
M-BERT (top2vec)	73.36	72.64	72.30	0.113	0.074
M-MentalBERT (NRC)	73.03	71.23	71.46	0.112	0.079
M-MentalBERT (LIWC)	73.21	73.15	72.43	0.099	0.071
M-MentalBERT (LDA topics)	73.74	73.23	73.16	0.111	0.075
M-MentalBERT (top2vec)	73.68	72.70	72.67	0.094	0.071

LDA, δηλαδή τα χαρακτηριστικά GOSS, καθώς το αλγόριθμος top2vec είναι ικανός να αναγνωρίζει αυτόματα τον αριθμό των θεμάτων. Όσον αφορά στο MentalBERT, παρατηρούμε ότι η ενσωμάτωση χαρακτηριστικών top2vec οδηγεί σε F1-score του 92.69%, υπερβαίνοντας το MentalBERT κατά 1.52%. Παρατηρούμε ότι η ολοκλήρωση των χαρακτηριστικών NRC και top2vec βελτιώνει την απόδοση που προκύπτει από το MentalBERT. Όσον αφορά στις προτεινόμενες προσεγγίσεις με εξομάλυνση ετικέτας, παρατηρούμε ότι αυτά τα μοντέλα επιτυγχάνουν καλύτερες επιδόσεις σε σχέση με τα μοντέλα χωρίς εξομάλυνση ετικέτας. Ειδικότερα, παρατηρούμε ότι το M-BERT (top2vec) με εξομάλυνση ετικέτας υπερτερεί στο F1-score και την ακρίβεια από το αντίστοιχο μοντέλο χωρίς εξομάλυνση ετικέτας κατά 0.61% και 0.36% αντίστοιχα. Επίσης, το M-MentalBERT (top2vec) με εξομάλυνση ετικέτας επιτυγχάνει το υψηλότερο F1-score και ακρίβεια ανέρχοντας σε 93.06% και 93.45% αντίστοιχα. Αυτό το μοντέλο υπερτερεί στο F1-score και την ακρίβεια από το αντίστοιχο μοντέλο χωρίς εξομάλυνση ετικέτας κατά 0.37% και 0.18% αντίστοιχα. Εκτός από τη βελτίωση των μετρήσεων απόδοσης, δηλαδή της ακρίβειας, της ανάκλησης, του F1-score και της ακρίβειας, παρατηρούμε ότι τα μοντέλα με εξομάλυνση ετικέτας επιτυγχάνουν καλύτερα αποτελέσματα όσον αφορά τις μετρήσεις βαθμονόμησης, δηλαδή τις μετρήσεις ECE και ACE, σε σύγκριση με τις μετρήσεις που προκύπτουν από τα μοντέλα χωρίς εξομάλυνση ετικέτας. Για παράδειγμα, παρατηρούμε ότι το M-BERT (top2vec) με εξομάλυνση ετικέτας βελτιώνει τις μετρήσεις ECE και ACE που προκύπτουν από το M-BERT (top2vec) χωρίς εξομάλυνση ετικέτας κατά 0.008 και 0.013 αντίστοιχα. Επίσης, το M-MentalBERT (LDA topics) με εξομάλυνση ετικέτας βελτιώνει τις μετρήσεις ECE και ACE που προκύπτουν από το M-MentalBERT (LDA topics) χωρίς εξομάλυνση ετικέτας κατά 0.042 και 0.043 αντίστοιχα.

Όσον αφορά το σύνολο δεδομένων Depression_Severity, συγκρίνουμε αρχικά τις προτεινόμενες προσεγγίσεις μας χωρίς εξομάλυνση ετικέτας με τα μοντέλα BERT και MentalBERT.

Παρατηρούμε ότι η ενσωμάτωση χαρακτηριστικών LIWC και χαρακτηριστικών που εξάγονται με τη μεθοδολογία θέματος LDA, δηλαδή τα χαρακτηριστικά GOSS, στο μοντέλο BERT οδηγεί σε άνοδο της απόδοσης σε σύγκριση με το μοντέλο BERT. Ειδικότερα, το M-BERT (LIWC) υπερτερεί στον αποκριματισμένο F1-score κατά 1.13%. Ταυτόχρονα, η ενσωμάτωση όλων των χαρακτηριστικών, εκτός από τα NRC, σε ένα μοντέλο MentalBERT οδηγεί σε βελτίωση της απόδοσης σε σύγκριση με το μοντέλο MentalBERT. Ειδικότερα, το M-MentalBERT (LDA topics) επιτυγχάνει τον υψηλότερο αποκριματισμένο F1-score που ανέρχεται σε 72.58%, υπερβαίνοντας το MentalBERT κατά 0.91%. Όσον αφορά στα προτεινόμενα μοντέλα με εξομάλυνση ετικέτας, παρατηρούμε μια βελτίωση τόσο στις μετρήσεις απόδοσης όσο και στις μετρήσεις βαθμονόμησης. Ειδικότερα, η ενσωμάτωση χαρακτηριστικών NRC σε ένα μοντέλο BERT επιτυγχάνει έναν αποκριματισμένο F1-score του 72.81%, υπερβαίνοντας το BERT κατά 1.81%, το M-BERT (NRC) χωρίς εξομάλυνση ετικέτας κατά 2.85% και το M-BERT (LIWC) χωρίς εξομάλυνση ετικέτας κατά 0.68%. Επιπλέον, το M-MentalBERT (LDA topics) με εξομάλυνση ετικέτας επιτυγχάνει το υψηλότερο F1-score που ανέρχεται σε 73.16%, υπερβαίνοντας το MentalBERT κατά 1.49% και το M-MentalBERT (LDA topics) χωρίς εξομάλυνση ετικέτας κατά 0.58%. Όσον αφορά στις μετρήσεις βαθμονόμησης, παρατηρούμε ότι και οι δύο μετρήσεις ECE και ACE βελτιώνονται όταν εφαρμόζουμε εξομάλυνση ετικέτας. Για παράδειγμα, το M-BERT (LIWC) με εξομάλυνση ετικέτας επιτυγχάνει μια βαθμονομούμενη βαθμολογία ECE της τάξης του 0.094 και μια βαθμονομούμενη βαθμολογία ACE της τάξης του 0.069, οι οποίες βελτιώνονται κατά 0.016 και 0.009 αντίστοιχα σε σύγκριση με το αντίστοιχο μοντέλο χωρίς εξομάλυνση ετικέτας.

1.2.2 Διάγνωση Κατάθλιψης με Χρήση Ομιλίας

1.2.2.1 Κίνητρο

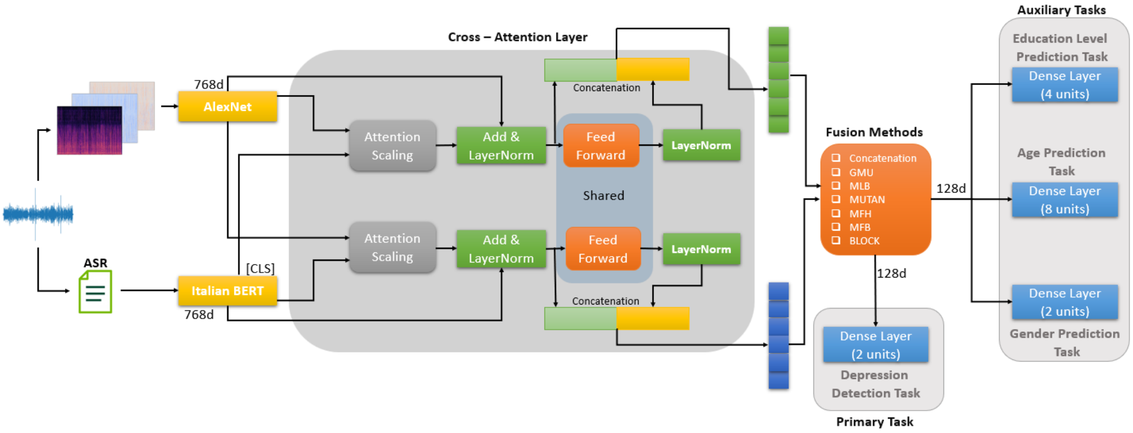
Οι υπάρχουσες ερευνητικές εργασίες βασίζονται στην εξαγωγή χαρακτηριστικών και την εκπαίδευση παραδοσιακών ταξινομητών μηχανικής μάθησης ή προσεγγίσεων βαθιάς μάθησης [32, 33, 34]. Ωστόσο, η εξαγωγή χαρακτηριστικών είναι μια χρονοβόρα διαδικασία που απαιτεί εξειδίκευση στο συγκεκριμένο θέμα. Επιπλέον, η πλειοψηφία των ερευνητικών μελετών χρησιμοποιεί μονοτροπικά μοντέλα για την πρόβλεψη της κατάθλιψης, χρησιμοποιώντας κυρίως την ομιλία [35]. Αν και υπάρχουν μελέτες που χρησιμοποιούν πολυτροπικά μοντέλα, αυτές οι μελέτες εφαρμόζουν στρατηγικές early [36, 37], intermediate [38, 39] ή late fusion [40, 41]. Στη στρατηγική early fusion, οι διανυσματικές αναπαραστάσεις των τροπικοτήτων συνδυάζονται στο επίπεδο εισόδου, ενώ στην intermediate συγχώνευση, οι διανυσματικές αναπαραστάσεις συνδυάζονται κατά την εκπαίδευση, δίνοντας ίση σημασία στις τροπικότητες. Στη στρατηγική late fusion, τα μονοτροπικά μοντέλα εκπαιδεύονται ανεξάρτητα και εφαρμόζεται απόφαση ψηφοφορίας, δηλαδή ψηφοφορία πλειοψηφίας. Επιπλέον, η πλειοψηφία των ερευνητικών εργασιών έχει δοκιμάσει τις προσεγγίσεις τους μόνο στην αγγλική γλώσσα, οπότε το ακουστικό και φωνητικό περιεχόμενο των δεδομένων μπορεί να διαφέρει σε άλλες γλώσσες. Τέλος, καμία υπάρχουσα μελέτη δεν έχει πειραματιστεί με την πρόβλεψη της κατάθλιψης, της ηλικίας, του επιπέδου εκπαίδευσης και του φύλου ταυτόχρονα.

1.2.2.2 Δεδομένα

Χρησιμοποιούμε το Androids corpus [42], το οποίο αποτελείται από δύο εργασίες, συγκεκριμένα την εργασία ανάγνωσης και την εργασία συνέντευξης. Συγκεκριμένα, η εργασία συνέντευξης αποτελείται από 116 δείγματα αυθόρμητης ομιλίας. Όλα τα πειράματα είναι ανεξάρτητα από το άτομο. Τα αρχεία ήχου είναι στην ιταλική γλώσσα. Αυτό το σύνολο δεδομένων περιλαμβάνει πληροφορίες για το φύλο, την ηλικία και το επίπεδο εκπαίδευσης των ατόμων. Οι πληθυσμοί των καταθλιπτικών και μη καταθλιπτικών συμμετεχόντων έχουν την ίδια κατανομή όσον αφορά την ηλικία, το φύλο και το επίπεδο εκπαίδευσης. Χρησιμοποιούμε το whisper large-v3 [43], για να εξάγουμε απομαγνητοφωνήσεις κειμένου παραγόμενες από μηχανή (automatic), καθώς δεν παρέχονται απομαγνητοφωνήσεις κειμένου παραγόμενες από άνθρωπο (manual).

1.2.2.3 Προτεινόμενη Μεθοδολογία

- Single-Task Learning. Σκοπός είναι η πρόβλεψη της κατάθλιψης.
- Multi-Task Learning. Σκοπός είναι η πρόβλεψη της κατάθλιψης, του επιπέδου εκπαίδευσης, της ηλικίας και του φύλου.



Σχήμα 1.2: Προτεινόμενη Αρχιτεκτονική για Διάγνωση Κατάθλιψης με Χρήση Ομιλίας και Απομαγνητοφωνημένου Κειμένου

Επεξεργασία Κειμένου: Χρησιμοποιούμε το Italian BERT[4]. Εξάγουμε το [CLS token] με αναπαράσταση $f^t \in \mathbb{R}^{1\times d}$, όπου $d = 768$.
Επεξεργασία Ομιλίας: Μετατρέπουμε το αρχείο ήχου σε εικόνα τριών καναλιών, log-Mel spectrogram, delta, delta-delta. Περνάμε την κάθε εικόνα σε ένα προεκπαιδευμένο AlexNet [44] μοντέλο. Έστω $f^v \in \mathbb{R}^{1\times d}$, όπου $d = 768$ η έξοδος του μοντέλου.
Επίπεδο Διασταυρούμενης Προσοχής: Με κίνητρο από [45], σχεδιάζουμε ένα επίπεδο διασταυρούμενης προσοχής (cross-attention), το οποίο επιστρέφει ένα ζεύγος βαθμωτών,

[4] https://github.com/dbmdz/berts

ένα για κάθε τροπικότητα. Αυτό το ζεύγος βαθμωτών επιτρέπει την κλιμάκωση των δύο τροπικοτήτων.

Όσον αφορά την τροπικότητα του κειμένου, ορίζουμε $Q_i = FC_q^t(f^v)$, $K_t = FC_k^t\left(f^t\right)$ και $V_t = FC_v^t\left(f^t\right)$. Η τιμή κλίμακας, η οποία αναπαρίσταται ως S_t, μπορεί να υπολογιστεί ως εξής:

$$S_t = sigmoid\left(\frac{Q_i \cdot K_t^T}{\sqrt{d}}\right)$$

Όσον αφορά την τροπικότητα της εικόνας, ορίζουμε $Q_t = FC_q^i\left(f^t\right)$, $K_i = FC_k^i(f^v)$ και $V_i = FC_v^i(f^v)$. Η τιμή κλίμακας, η οποία αναπαρίσταται ως S_i, μπορεί να υπολογιστεί ως εξής:

$$S_i = sigmoid\left(\frac{Q_t \cdot K_i^T}{\sqrt{d}}\right)$$

. Οι έξοδοι του μηχανισμού προσοχής μπορούν να υπολογιστούν ως $S_t \times V_t$ και $S_i \times V_i$. Ορίζουμε $FC_q^t, FC_k^t, FC_v^t, FC_q^i, FC_k^i, FC_v^i \in \mathbb{R}^{d\times d}$. Παρόμοια με [46], χρησιμοποιούμε residual connections ακολουθούμενες από κανονικοποίηση επιπέδου, όπως περιγράφεται στις παρακάτω εξισώσεις:

$$\hat{E}_t = LayerNorm\left(S_t \times V_t + f^t\right)$$

,

$$\hat{E}_i = LayerNorm\left(S_i \times V_i + f^v\right)$$

.

Στη συνέχεια, περνάμε τα $\hat{E}_t$ και $\hat{E}_i$ μέσω δύο κοινών πλήρως συνδεδεμένων δικτύων με συνάρτηση ενεργοποίησης ReLU, ως εξής:

$$\hat{E}_t{}' = LayerNorm\left(FC_m^n\left(ReLU\left(FC_p^q\left(\hat{E}_t\right)\right)\right)\right)$$

,

$$\hat{E}_i{}' = LayerNorm\left(FC_m^n\left(ReLU\left(FC_p^q\left(\hat{E}_i\right)\right)\right)\right)$$

, όπου $FC_p^q \in \mathbb{R}^{d\times 4d}$, $FC_m^n \in \mathbb{R}^{4d\times d}$.

Στη συνέχεια, συνενώνουμε τα $\hat{E}_t$ και $\hat{E}_t{}'$ (ομοίως τα $\hat{E}_i$ και $\hat{E}_i{}'$) σε ένα ενιαίο διάνυσμα, δηλαδή

$$\hat{E}_t{}'' = [\hat{E}_t, \hat{E}_t{}']$$

,

$$\hat{E}_i{}'' = [\hat{E}_i, \hat{E}_i{}']$$

, όπου $\hat{E}_t{}'', \hat{E}_i{}'' \in \mathbb{R}^{2d}$.

Μέθοδοι Συγχώνευσης (Fusion Methods)

- Concatenation
- Gated Multimodal Unit (GMU)
- MUTAN Decomposition

- Multimodal Low-rank Bilinear (MLB) pooling
- MFB
- MFH
- BLOCK

Επίπεδο Εξόδου. Τέλος, ορίζουμε το επίπεδο εξόδου.

1.2.2.4 Αποτελέσματα

Για τον έλεγχο σημαντικότητας, χρησιμοποιούμε το Almost Stochastic Order (ASO) test [47, 48] όπως υλοποιήθηκε από [49]. Συγκεκριμένα, το τεστ ASO καθορίζει αν υπάρχει στοχαστική τάξη [50] μεταξύ δύο μοντέλων, δηλαδή του A και του B. Υπολογίζεται μια βαθμολογία (ϵ_{min}) που αντιπροσωπεύει πόσο μακριά είναι το πρώτο από το να είναι σημαντικά καλύτερο από το δεύτερο. Όταν $\epsilon_{min} = 0$, τότε το A είναι πραγματικά στοχαστικά κυρίαρχο επί του B. Όταν $\epsilon_{min} < 0.5$, το A είναι σχεδόν στοχαστικά κυρίαρχο επί του B. Για $\epsilon_{min} = 0.5$, δεν μπορεί να καθοριστεί τάξη.

Τα αποτελέσματα παρουσιάζονται στον Πίνακα 1.3. Παρατηρούμε ότι η χρήση του *BLOCK* ως μέθοδος συγχώνευσης οδηγεί στο καλύτερο μοντέλο, ξεπερνώντας τις υπόλοιπες προσεγγίσεις στην Ακρίβεια και στο F1-score κατά 1.21-21.99% και 1.32-22.23% αντίστοιχα. Τα πολυτροπικά μοντέλα αποδίδουν καλύτερα από τα μονοτροπικά, επαληθεύοντας την αρχική μας υπόθεση ότι η χρήση πολλαπλών τροπικοτήτων βελτιώνει την απόδοση των αλγορίθμων. Ο μηχανισμός συνένωσης (concatenation) επιτυγχάνει τα χειρότερα αποτελέσματα σε σύγκριση με τις άλλες μεθόδους συγχώνευσης, καθώς αποδίδει ίση σημασία σε κάθε τροπικότητα. Πιστεύουμε ότι το MFB υπερτερεί του MFH, καθώς η μέθοδος MFH αποτελείται από τη σύνδεση δύο MFB μπλοκ, και έτσι φαίνεται να είναι περίπλοκη για το περιορισμένο σύνολο δεδομένων, που χρησιμοποιούμε.

Υποθέτουμε ότι το GMU επιτυγχάνει χαμηλή απόδοση, καθώς ελέγχει τη ροή πληροφοριών χωρίς να καταγράφει τόσο αποτελεσματικά τις αλληλεπιδράσεις μεταξύ των τροπικοτήτων. Παρατηρούμε ότι οι αρχιτεκτονικές single-task learning αποδίδουν καλύτερα από τις αρχιτεκτονικές multi-task learning. Αυτό μπορεί να δικαιολογηθεί από το γεγονός ότι η κατάθλιψη είναι μια ψυχική διαταραχή που μπορεί να συμβεί σε οποιονδήποτε. Υπάρχουν πολλοί λόγοι για την κατάθλιψη, π.χ. αγχωτικά γεγονότα, προσωπικότητα, προβλήματα υγείας (καρκίνος), μοναξιά, κ.λπ. Σύμφωνα με στατιστικό έλεγχο, το καλύτερο μοντέλο μας είναι *σχεδόν* στοχαστικά κυρίαρχο όσον αφορά την ακρίβεια σε σχέση με όλες τις προσεγγίσεις, εκτός από το *Only speech signal*, όπου $\epsilon_{min} = 0$.

Πίνακας 1.3: Συγκριτικός Πίνακας Αξιολόγησης. (∗) σημαίνει ότι $\epsilon_{min} < 0.1$, † σημαίνει ότι $\epsilon_{min} < 0.2$, ‡ σημαίνει ότι $\epsilon_{min} < 0.3$, ∗∗ σημαίνει ότι $\epsilon_{min} < 0.4$ και †† σημαίνει ότι $\epsilon_{min} < 0.5$. Δεν είμαστε σε θέση να πραγματοποιήσουμε στατιστικό έλεγχο με τα αποτελέσματα της μελέτης [42], διότι οι συγγραφείς δεν παρέχουν τα αποτελέσματα που προέκυψαν από τα επιμέρους υποσύνολα.

	Μετρικές Αξιολόγησης				
Αρχιτεκτονική	**Precision**	**Recall**	**F1-score**	**Accuracy**	**Specificity**
Μονοτροπικές Προσεγγίσεις					
Only transcript	$94.72^{\ddagger}$ ±5.38	91.78^{**} ±5.77	$93.04^{\dagger}$ ±3.77	$92.49^{\ddagger}$ ±3.97	93.51^{**} ±6.96
Only Speech signal	80.73^{*} ±12.12	85.70^{*} ±9.57	82.49^{*} ±8.51	80.52^{*} ±8.97	74.21^{*} ±16.87
eGeMAPSv02	79.05^{*} ±13.50	85.46^{*} ±7.92	81.67^{*} ±9.69	80.29^{*} ±10.11	76.64^{*} ±15.26
ComParE_2016	86.03^{*} ±8.92	92.29 ±3.96	88.82^{*} ±5.31	87.97^{*} ±4.93	$84.92^{\dagger}$ ±9.49
Αποτελέσματα Μεθόδων [42]					
BS1	73.50 ±16.10	74.50 ±13.20	73.60 ±13.60	73.30 ±10.60	– –
BS2	85.80 ±3.10	86.10 ±2.70	84.70 ±0.90	83.90 ±1.30	– –
Single - Task Learning					
Concatenation	91.51^{*} ±8.74	93.35 ±5.99	$92.11^{\dagger}$ ±5.54	$91.46^{\dagger}$ ±6.05	$90.91^{\dagger}$ ±10.48
GMU	94.10^{**} ±9.51	93.41 ±6.61	93.38^{**} ±6.25	$92.34^{\ddagger}$ ±7.22	92.33^{**} ±11.91
MLB	95.95 ±7.69	91.82^{**} ±6.31	93.57^{**} ±5.37	92.96^{**} ±5.94	95.33 ±9.71
MUTAN	$93.75^{\ddagger}$ ±8.76	94.46 ±5.57	93.82^{**} ±5.71	92.75^{**} ±6.79	90.78^{**} ±13.07
MFH	95.04^{**} ±6.62	$92.79^{\dagger\dagger}$ ±5.01	93.75^{**} ±4.46	$92.94^{\ddagger}$ ±5.56	91.28^{**} ±17.76
MFB	94.68^{**} ±8.19	93.63 ±4.63	93.95^{**} ±5.32	93.18^{**} ±6.13	92.53^{**} ±10.66
BLOCK	**97.30** ±4.43	**94.52** ±4.52	**95.83** ±3.81	**95.29** ±4.23	**96.42** ±6.04
Multi-Task Learning					
Φύλο, Εκπαίδευση, Ηλικία	96.14 ±5.02	93.24 ±6.95	$94.38^{\dagger\dagger}$ ±3.65	$94.08^{\dagger\dagger}$ ±3.45	96.31 ±4.86
Φύλο, Εκπαίδευση	97.22 ±5.14	$92.28^{\dagger\dagger}$ ±6.82	$94.51^{\dagger\dagger}$ ±4.65	$94.07^{\dagger\dagger}$ ±5.03	95.95 ±9.35
Εκπαίδευση, Ηλικία	94.41^{**} ±7.24	93.63 ±5.97	93.74^{**} ±4.52	$93.62^{\dagger\dagger}$ ±4.48	$93.56^{\dagger\dagger}$ ±8.05
Φύλο, Ηλικία	96.55 ±4.87	$92.51^{\dagger\dagger}$ ±6.09	$94.30^{\dagger\dagger}$ ±3.72	$93.84^{\dagger\dagger}$ ±4.25	94.53 ±13.05
Φύλο	94.61^{**} ±9.28	93.29 ±7.18	93.61^{**} ±6.51	93.20^{**} ±6.81	$93.68^{\dagger\dagger}$ ±10.63
Εκπαίδευση	94.22^{**} ±9.16	93.04 ±7.27	$93.44^{\ddagger}$ ±7.34	93.00^{**} ±7.31	92.03^{**} ±12.41
Ηλικία	94.99^{*} ±7.46	$92.32^{\dagger\dagger}$ ±6.72	$93.34^{\ddagger}$ ±5.09	$92.56^{\ddagger}$ ±5.85	$93.42^{\dagger\dagger}$ ±10.79

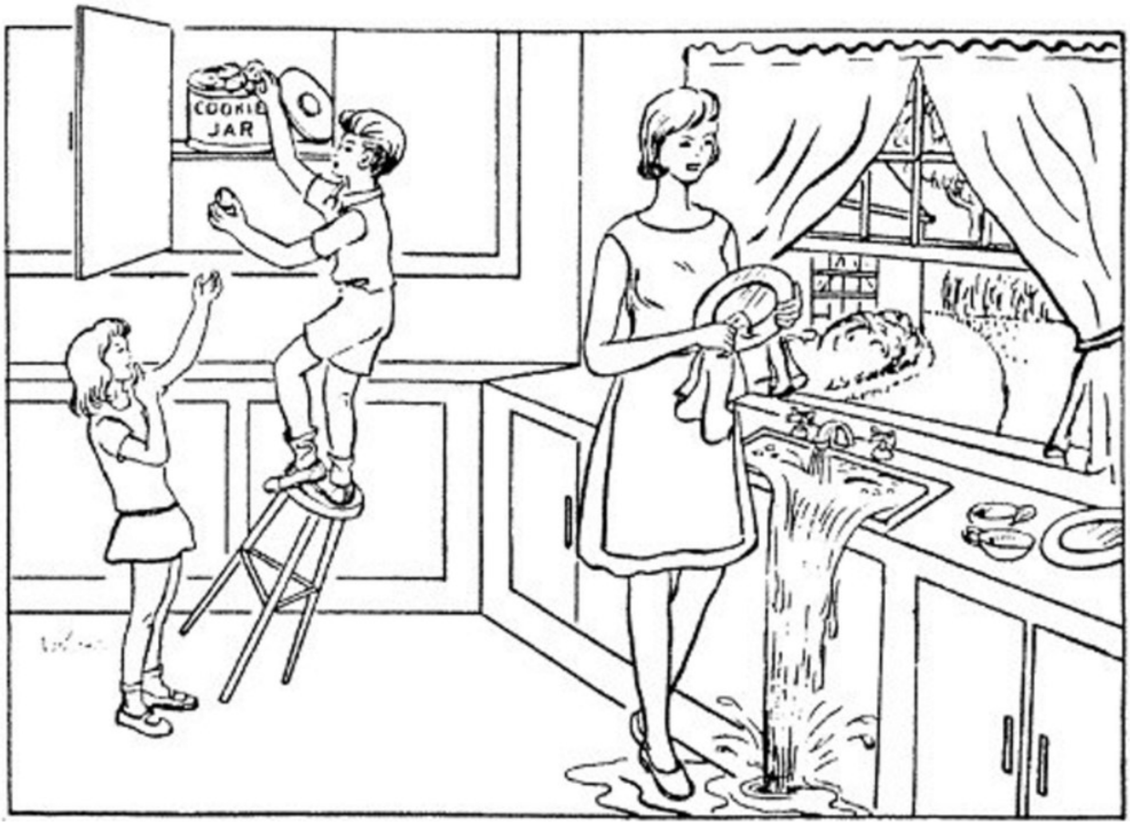
1.3 Διάγνωση Άνοιας

1.3.1 Κίνητρο

Η νόσος Alzheimer (AD) είναι μια νευρολογική διαταραχή, που εξελίσσεται με την πάροδο του χρόνου, και αποτελεί την πιο κοινή αιτία άνοιας. Σύμφωνα με τον ΠΟΥ, περίπου 55 εκατομμύρια άνθρωποι έχουν άνοια παγκοσμίως με πάνω από το 60% να ζει σε χώρες χαμηλού και μεσαίου εισοδήματος [51]. Επιπλέον, η άνοια επηρεάζει την ικανότητα ενός ατόμου να επικοινωνεί. Πιο συγκεκριμένα, τα άτομα με άνοια μπορεί να μην είναι σε θέση να βρουν τις σωστές λέξεις ή να μην μπορούν να βρουν καμία λέξη. Ταυτόχρονα, δεν μπορούν να παραμείνουν συγκεντρωμένοι σε μια συζήτηση και τείνουν να χρησιμοποιούν λέξεις χωρίς νόημα, με αποτέλεσμα να μην μπορούν να επικοινωνήσουν με άλλους ανθρώπους. Σημάδια άνοιας περιλαμβάνουν μεταξύ άλλων: προβλήματα με τη βραχυπρόθεσμη μνήμη, πληρωμή λογαριασμών, προγραμματισμός και προετοιμασία γευμάτων, ραντεβού ή ταξίδια [52]. Αυτό το γεγονός συνεπάγεται σωματικές, ψυχολογικές, κοινωνικές και οικονομικές επιπτώσεις όχι μόνο για τα άτομα που ζουν με άνοια, αλλά και για τους φροντιστές τους, τις οικογένειές τους και την κοινωνία γενικότερα. Λόγω του γεγονότος ότι η άνοια χειροτερεύει με την πάροδο του χρόνου, είναι σημαντικό να διαγνωστεί έγκαιρα.

1.3.2 Δεδομένα

ADReSS Challenge Dataset. Χρησιμοποιούμε το σύνολο δεδομένων ADReSS Challenge [53] για τη διεξαγωγή των πειραμάτων μας. Τα δεδομένα αντιστοιχούν σε προφορικές περιγραφές εικόνων (Σχήμα 1.3) που προέρχονται από τους συμμετέχοντες μέσω της εικόνας κλοπής cookies από την εξέταση αφασίας της Βοστώνης [54]. Επιλέγουμε το συγκεκριμένο σύνολο δεδομένων, καθώς ελαχιστοποιεί πολλά είδη προκαταλήψεων, που θα μπορούσαν να επηρεάσουν την εγκυρότητα των προτεινόμενων προσεγγίσεων κατά τη διαδικασία εκπαίδευσης και αξιολόγησης. Συγκεκριμένα, σε αντίθεση με άλλα σύνολα δεδομένων, το σύνολο δεδομένων ADReSS Challenge αντιστοιχεί στο φύλο και την ηλικία. Επιπλέον, είναι ισορροπημένο, αφού περιλαμβάνει 78 ασθενείς με άνοια και 78 υγιή άτομα. Αυτό που αξίζει επίσης να σημειωθεί είναι το γεγονός ότι το σύνολο δεδομένων ADReSS Challenge έχει επιλεγεί προσεκτικά έτσι ώστε να μετριάζονται κοινές προκαταλήψεις που συχνά παραβλέπονται στις αξιολογήσεις μεθόδων ανίχνευσης άνοιας, συμπεριλαμβανομένων των επαναλαμβανόμενων εμφανίσεων ομιλίας από τον ίδιο συμμετέχοντα και προβλημάτων σε ποιότητα ήχου. Για να είμαστε πιο ακριβείς, οι εγγραφές έχουν βελτιωθεί ακουστικά με σταθερή αφαίρεση θορύβου και έχει εφαρμοστεί κανονικοποίηση της έντασης του ήχου σε όλα τα τμήματα ομιλίας. Το σύνολο δεδομένων έχει χωριστεί από τους διοργανωτές σε ένα σύνολο εκπαίδευσης (train set) και ένα σύνολο δοκιμής (test set). Το train set αποτελείται από 54 ασθενείς με άνοια και 54 υγιείς, ενώ το test set περιλαμβάνει 24 ασθενείς με άνοια και 24 υγιείς.



Σχήμα 1.3: The Cookie Theft picture

1.3.3 Διάγνωση Άνοιας με Χρήση Απομαγνητοφωνημένου Κειμένου

Βελτιστοποιούμε (fine-tune) μοντέλα βασισμένα σε μετασχηματιστές (transformers. Συγκεκριμένα, βελτιστοποιούμε τα εξής μοντέλα: BERT [26], BioBERT [55], BioClinicalBERT [56], ConvBERT [57], RoBERTa [58], ALBERT [59], και XLNet [60].

1.3.3.1 Αποτελέσματα

Τα αποτελέσματα των προτεινόμενων μοντέλων που αναφέρονται παραπάνω αναφέρονται στον Πίνακα 1.4. Επίσης, ο Πίνακας 1.4 παρέχει μια σύγκριση των προτεινόμενων μοντέλων μας με υπάρχουσες ερευνητικές πρωτοβουλίες.

Πίνακας 1.4: Σύγκριση της απόδοσης των προτεινόμενων μοντέλων και εργασιών της βιβλιογραφίας στο ADReSS Challenge test set. Οι τιμές αναπαρίστανται ως: μέσος όρος ± τυπική απόκλιση. Παίρνουμε τον μέσο όρο σε 5 τρεξίματα του μοντέλου.

	Μετρικές Αξιολόγησης				
Αρχιτεκτονική	**Prec.**	**Rec.**	**F1-score**	**Acc.**	**Spec.**
Σύγκριση με μεθόδους της βιβλιογραφίας					
[61]	-	87.50	-	89.58	91.67
[62]	81.82	75.00	78.26	79.17	83.33
[63]	-	-	85.40	85.20	-
[64]	-	-	-	85.00	-
[65]	86.00	79.00	83.00	83.33	88.00
[66]	-	-	-	85.42	-
[67]	94.12	66.67	78.05	81.25	95.83
Προτεινόμενα Μοντέλα					
BERT	87.19 ±3.25	81.66 ±5.00	86.73 ±4.53	87.50 ±4.37	93.33 ±5.65
BioBERT	86.87 ±6.09	78.33 ±4.86	82.11 ±2.83	82.92 ±3.06	87.50 ±6.97
BioClinicalBERT	95.03 ±3.03	76.66 ±4.99	84.72 ±2.74	86.25 ±2.12	95.83 ±2.64
ConvBERT	83.51 ±1.23	79.99 ±4.08	81.65 ±2.06	82.08 ±1.66	84.16 ±1.66
RoBERTa	90.24 ±2.81	76.66 ±4.99	82.81 ±3.52	84.16 ±2.83	91.66 ±2.64
ALBERT	79.15 ±7.89	78.33 ±3.11	78.45 ±3.12	78.33 ±3.86	78.33 ±8.89
XLNet	85.58 ±2.77	68.33 ±6.77	75.75 ±4.05	78.33 ±2.82	88.33 ±3.12

Όσον αφορά τα προτεινόμενα μοντέλα που βασίζονται σε μετασχηματιστές, κάποιος μπορεί εύκολα να παρατηρήσει ότι το BERT επιτυγχάνει την υψηλότερη Ανάκληση, F1-score και Ακρίβεια, με τις μετρικές αυτές να ανέρχονται στο 81.66%, 86.73% και 87.50% αντίστοιχα. Συγκεκριμένα, το BERT υπερτερεί σε σύγκριση με τα άλλα προτεινόμενα μοντέλα που βασίζονται σε μετασχηματιστές στην Ανάκληση κατά 1.67-13.33%, στο F1-score κατά 2.01-10.98%, και στην Ακρίβεια κατά 1.25-9.17%. Το BioClinicalBERT επιτυγχάνει τη δεύτερη υψηλότερη Ακρίβεια και F1-score, με τις μετρικές αυτές να ανέρχονται σε 86.25% και 84.72% αντίστοιχα. Επίσης, το BioClinicalBERT επιτυγχάνει την υψηλότερη Ακρίβεια, που είναι ίση με 95.03%, υπερβαίνοντας τα άλλα μοντέλα που βασίζονται σε μετασχηματιστές κατά 4.79-15.88%. Το RoBERTa επιτυγχάνει παρόμοια αποτελέσματα με το BERT και το BioClinicalBERT με Ακρίβεια και F1-score ίσα με 84.16% και 82.81% αντίστοιχα. Επιπλέον, το BioBERT και το ConvBERT δείχνουν μικρές διαφορές στην Ακρίβεια και το F1-score, με το BioBERT να υπερβαίνει το ConvBERT και στις δύο μετρικές. Συγκεκριμένα, το BioBERT υπερτερεί στο F1-score κατά 0.46% και στην Ακρίβεια κατά 0.84%. Επιπλέον, παρατηρούμε ότι το ALBERT και το XLNet επιτυγχάνουν σκορ Ακρίβειας ίσο με 78.33%, με το ALBERT να υπερτερεί στο F1-score κατά 2.70%.

Σε σύγκριση με τις προηγμένες προσεγγίσεις, κάποιος μπορεί να παρατηρήσει ότι τα προτεινόμενα μοντέλα μας επιτυγχάνουν παρόμοια ή ακόμα και υπερτερούν των προηγούμενων μελετών. Ειδικότερα, το BERT υπερτερεί σε σύγκριση με όλα τα έργα έρευνας, εκτός από το [61], ως προς την Ακρίβεια κατά 2.08-8.33%, το F1-score κατά 1.33-8.68%, και την Ανάκληση κατά 2.66-14.99%.

1.3.3.2 Γλωσσολογική Ανάλυση

Ο κύριος στόχος αυτής της ενότητας είναι να φωτίσει ποια μονογράμματα (unigrams) και μορφές λόγου (pos-tags) συσχετίζονται κυρίως με κάθε κατηγορία ξεχωριστά [68]. Για να διευκολυνθεί αυτό, υπολογίζουμε την συσχέτιση point-biserial μεταξύ κάθε χαρακτηριστικού (μονόγραμμα και μορφή λόγου) σε όλες τις απομαγνητοφωνήσεις κειμένου και της ετικέτας εξόδου - label (0 για τον ομάδα ελέγχου και 1 για την ομάδα άνοιας). Πριν υπολογίσουμε τη συσχέτιση, κανονικοποιούμε τα χαρακτηριστικά έτσι ώστε να αθροίζουν στο 1 σε κάθε κείμενο. Χρησιμοποιούμε τη συσχέτιση point-biserial, αφού αυτή είναι μια συσχέτιση μεταξύ συνεχών και δυαδικών μεταβλητών. Επιστρέφει μια τιμή μεταξύ -1 και 1. Δεδομένου ότι ενδιαφερόμαστε μόνο για τη δύναμη της συσχέτισης, υπολογίζουμε την απόλυτη τιμή, όπου αρνητικές συσχετίσεις αναφέρονται στην ομάδα ελέγχου (ετικέτα 0) και θετικές συσχετίσεις αναφέρονται στην ομάδα άνοιας (ετικέτα 1). Αναφέρουμε τα ευρήματά μας στον Πίνακα 1.5, όπου όλες οι συσχετίσεις είναι σημαντικές στο $p < 0,05$, με διόρθωση Benjamini-Hochberg [69] για πολλαπλές συγκρίσεις.

Όπως εύκολα μπορεί να παρατηρήσει κανείς, τα μέρη του λόγου (pos-tags) που συσχετίζονται με την ομάδα της άνοιας είναι τα ακόλουθα: RB (επιρρήματα), PRP (προσωπική αντωνυμία), VBD (ρήμα σε παρελθόντα χρόνο), και UH (interjection). Από την άλλη πλευρά, οι άνθρωποι στην ομάδα ελέγχου τείνουν να χρησιμοποιούν VBG (ρήμα, γερούνδιο ή μετοχή

Πίνακας 1.5: Χαρακτηριστικά που σχετίζονται με υγιή άτομα και άτομα με άνοια, ταξινομημένα βάσει της συσχέτισης point-biserial. Όλες οι συσχετίσεις είναι σημαντικές στο $p < 0.05$ μετά από διόρθωση Benjamini-Hochberg.

Υγιή Άτομα		**Άνοια**	
Unigrams	**corr.**	**Unigrams**	**corr.**
is	0.364	here	0.310
curtains	0.361	-	-
window	0.301	-	-
are	0.300	-	-
POS	**corr.**	**POS**	**corr.**
VBG	0.285	RB	0.388
DT	0.216	PRP	0.354
NN	0.210	VBD	0.275
-	-	UH	0.242

ενεστώτα), DT (προσδιοριστή), και NN (ουσιαστικό). Αυτά τα ευρήματα μπορούν να δικαιολογηθούν στον Πίνακα 1.6, όπου παρουσιάζουμε τρία παραδείγματα απομαγνητοφωνημένων κειμένων που ανήκουν στην ομάδα ελέγχου και τρία παραδείγματα απομαγνητοφωνημένων κειμένων που ανήκουν στην ομάδα της άνοιας. Συγκεκριμένα, έχουμε αναθέσει χρώματα σε διαφορετικά μέρη του λόγου, έτσι ώστε να γίνουν εύκολα κατανοητές οι διαφορές στα γλωσσικά πρότυπα που χρησιμοποιούνται από κάθε ομάδα στον αναγνώστη. Για να είμαστε πιο ακριβείς, το κόκκινο χρώμα υποδεικνύει το pos-tag VBG, το κίτρινο αναφέρεται στο pos-tag DT, το φούξια στο pos-tag RB, το βερικοκί στο pos-tag PRP, το μπλε στο pos-tag VBD, και το πράσινο στο pos-tag UH.

Παρατηρούμε ότι οι άνθρωποι στην ομάδα της άνοιας τείνουν να χρησιμοποιούν προσωπικές αντωνυμίες (αυτός, αυτή, εγώ, εκείνοι κλπ.) πολύ συχνά, καθώς είναι ανίκανοι να θυμηθούν τους συγκεκριμένους όρους (μαμά, αγόρι κλπ.). Αυτό το εύρημα συμφωνεί με την έρευνα που διεξήχθη από τους [70], όπου οι συγγραφείς αναφέρουν ότι οι προσωπικές αντωνυμίες παρουσιάζουν υψηλή συχνότητα στην ομιλία των ασθενών με Αλτσχάιμερ, καθώς αυτοί οι άνθρωποι δεν μπορούν να βρουν την επιθυμητή λέξη. Για να είμαστε πιο ακριβείς, σε μια συνομιλία οι άνθρωποι πρέπει να θυμούνται τι είπαν κατά τη διάρκεια ολόκληρης της συνομιλίας. Ωστόσο, αυτό δεν είναι εφικτό στους ασθενείς με Αλτσχάιμερ, οι οποίοι παρουσιάζουν ελλείμματα στην εργασιακή μνήμη και έτσι τείνουν να παράγουν άδεια ομιλία (χρήση προσωπικών αντωνυμιών). Από την άλλη πλευρά, οι άνθρωποι στην ομάδα ελέγχου τείνουν να χρησιμοποιούν περισσότερα ουσιαστικά αντί για προσωπικές αντωνυμίες, καθώς είναι σε θέση να διατηρούν διάφορα είδη πληροφοριών.

Επιπλέον, οι ασθενείς με Αλτσχάιμερ τείνουν να χρησιμοποιούν ρήματα στο παρελθόν (ήταν, ξέχασα, έκανα, άρχισαν) αντίθετα με τους ανθρώπους που δεν πάσχουν από άνοια, οι οποίοι χρησιμοποιούν ρήματα στον ενεστώτα. Ένα χαρακτηριστικό παράδειγμα που μπορεί να επισημανθεί στο πέμπτο κείμενο στον Πίνακα 1.6, δηλαδή, ”oh have you heard of that new game that they started to play after christmas ? did you ?”. Ο ασθενής με Αλτσχάιμερ ίσως θυμάται μια προσωπική ιστορία από το παρελθόν που θέλει να διηγηθεί, αντί για την εργασία που του έχει ανατεθεί να εκτελέσει. Συνεπώς, ο ασθενής δεν είναι σε θέση να παραμείνει εστιασμένος στην περιγραφή της εικόνας. Αυτό το εύρημα είναι συμβατό με τις εργασίες [71, 72], όπου οι συγγραφείς αναφέρουν ότι οι ασθενείς με Αλτσχάιμερ παρουσιάζουν δυσκολία στη διατήρηση και τη συνέχιση της ανάπτυξης ενός θέματος και έτσι επιδεικνύουν απρόσμενες αλλαγές θέματος. Επίσης, αυτό το εύρημα αποκαλύπτει διαφορά στη γλώσσα που χρησιμοποιούν οι ασθενείς με Αλτσχάιμερ και οι αφασικοί με αγραμματική άνοια. Συγκεκριμένα, οι ασθενείς με αφασική άνοια τυπικά έχουν προβλήματα στη χρήση του χρόνου παρελθόντος και αντ΄ αυτού βασίζονται σε ρήματα σε παρόντα χρόνο [73].

Επιπλέον, οι ασθενείς με Αλτσχάιμερ τείνουν να χρησιμοποιούν τα pos-tags UH (αχ, ναι, καλά) και RB (ίσως, πιθανώς), καθώς δεν είναι σίγουροι για αυτό που περιγράφουν λόγω της πνευματικής ανασφάλειας. Ταυτόχρονα, το pos-tag UH αποτελεί ένα παράδειγμα άδειας ομιλίας. Συγκεκριμένα, αυτό το pos-tag χρησιμοποιείται ως γέμισμα στην αρχή κάθε εκφώνησης, καθώς οι ασθενείς με Αλτσχάιμερ σκέφτονται τι να πουν.

Πίνακας 1.6: Παραδείγματα απομαγνητοφωνημένων κειμένων με τις ετικέτες τους. Το κόκκινο χρώμα υποδηλώνει το μέρος του λόγου VBG, το κίτρινο αναφέρεται στο μέρος του λόγου DT, το φούξια στο μέρος του λόγου RB, το βερικοκκύ στο μέρος του λόγου PRP, το σκούρο μπλε στο μέρος του λόγου VBD και το πράσινο στο μέρος του λόγου UH.

Απομαγνητοφωνημένο Κείμενο	**Ετικέτα**
” well the girl is watching the boy go into the cookie jar . he has a cookie in his hand . he’s on the stool . the stool is falling . the mother is drying dishes . has a plate in her hand . sink is overflowing . there’s water on the floor . she’s stepping in the water . something that’s going on you said ? the little girl looks like she’s motioning to the boy to be quiet . and I don’t know what else . the woman’s looking out the window . the window’s open . ”	Υγιές Άτομο
” action . alright . a lady’s drying dishes . the boy was standing on a stool but the action is that the stool has slipped and he is falling . and the girl has her hand raised reaching for a cookie . and there’s a lot of action in the sink here . the water is flowing out . she is apparently so daydreaming that she doesn’t realize that the sink is overflowing . any more action ? or is that enough action ? ”	Υγιές Άτομο
” touching lip . raising arm . is that what you mean ? reaching for cookie . handing cookie down . slipping from stool . stool falling over . wiping dishes . water running . water overflowing . breeze . I don’t know if that’s action . stepping out from water . I guess that’s it . ”	Υγιές Άτομο

συνεχίζει στην επόμενη σελίδα

Πίνακας 1.6

Απομαγνητοφωνημένο Κείμενο	**Ετικέτα**
" alright . I see the little boy stealing cookies from the cookie jar . and he gave some to the little girl and she's eating some of the cookies . and I guess this is mama and she's washing the dishes . and she dropped a dish . no she didn't drop a dish . the water that she's washing the dishes with she let run . and it's overflown . that doesn't sound right . did it ? we forgot to turn off the spigot . and so the water is running off onto the floor here . and mom apparently is washing the dishes . and here's this little boy stealing the cookies . he's gonna fall because the four legged stool is gonna fall over with him and the cookie jar . and mama's drying the dishes as usual for mamas if they don't have a husband that dries them or washes them or whatever . let's see now . I guess there's more things I'm sposta see . let's see here now . oh and the water is flowing out of the sink they forgot to turn off whoever's doing the dishwashing . mom apparently here , she forgot to turn off the water and the water is spilling out onto the kitchen floor . and the little girl has pushed over the stool with the boy that was reaching up to get the cookies . either she pushed it over or he fell over with it . you know it excuse me but you know I was ... "	Άνοια
" mhm . oh I see a part of the whole kitchen . is that all the kitchen or isn't it ? oh I can't read ... a lady a mother were in her kitchen . in her kitchen doing some work I suppose . and there's another woman there sharing their pleasures or whatever . oh have you heard of that new game that they started to play after christmas ? did you ? is a . well it looks like ... I'd say this is ... well let's see . it looks like ... oh ... . my wife will beat me by a couple rows of this . that's like the washing machine ? or let me see . I can't ... oh that's the son come from school maybe or something . that's a youngster there . well that's just as though they getting ready to go to school or they're just coming out from school . and right there he's same as back there except for down there in the bottom I think it's ... that's a little . "	Άνοια

συνεχίζει στην επόμενη σελίδα

Πίνακας 1.6

Απομαγνητοφωνημένο Κείμενο	**Ετικέτα**
” yes . the water ? well let’s see . there’s something hasta be where the water goes down over . there’s probably something that’s ... or they don’t have it open or something might have. I don’t know . what ..? when the water goes down what do you call that ? this here . right here . this . what do you call that ? what is that ? what is that ? I don’t know ! that’s what I’m saying . I don’t know what that is . the what ? a pipe . oh water pipe ! oh yeah . okay . well then maybe the water pipe is not broke but there must be things in there . that the water will not go down . I don’t know . huh ? what’s happening to the water ? well the water is going down in the ... I don’t know . what would you call this ? floor ! yeah okay . yeah . well down on this side of the picture . well this thing here is turning over . yeah . no , uhuh . I don’t know what’s going on . well he’s probably getting ... what’s this here ? cocoa jar ? what’s this cocoa ? c o o k i e . I don’t know . I don’t know what ..? huh ? cookie , oh a cookie . oh ! oh okay . mhm . well he’s getting it out . and he’s gonna give it to the girl /. down here . mhm . going on in the picture ? well the boy is giving her the girl the cookie . this probably is broke . so the water will not go down in and it’s coming up and going in here huh . well it looks like she was gonna wash . what they eat with , all that . what do you call that ? what do you call this ? a plate ? oh yeah . what you eat on . is that what you call them a plate ? oh this is a cup ? oh maybe , I don’t know . mhm . okay . ”	Άνοια

1.3.3.3 Επεξηγησιμότητα

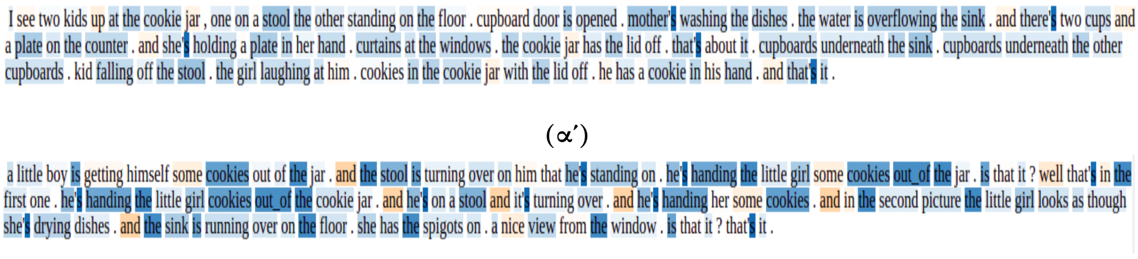
Σε αυτή την ενότητα, χρησιμοποιούμε το LIME [74] (χρησιμοποιώντας 5000 δείγματα) για να εξηγήσουμε τις προβλέψεις που κάνει το καλύτερο μοντέλο μας, δηλαδή το BERT, και να διερευνήσουμε περισσότερο τις διαφορές στη γλώσσα μεταξύ των ασθενών με Αλτσχάιμερ και των μη ασθενών. Πιο συγκεκριμένα, το LIME δημιουργεί τοπικές εξηγήσεις για οποιονδήποτε ταξινομητή μηχανικής μάθησης εισάγοντας ένα ερμηνεύσιμο μοντέλο, το οποίο εκπαιδεύεται σε δεδομένα που παράγονται μέσω της παρατήρησης διαφορών στην απόδοση ταξινόμησης όταν αφαιρούνται λέξεις από το αρχικό κείμενο.

Παραδείγματα εξηγήσεων που δημιουργούνται από το LIME παρουσιάζονται στα Σχήματα 1.4-1.7. Πιο συγκεκριμένα, το Σχήμα 1.4 απεικονίζει δύο απομαγνητοφωνημένα κείμενα, τα οποία αντιστοιχούν σε άνοια. Ωστόσο, το μοντέλο μας τα προβλέπει ως υγιή. Το Σχήμα 1.5 αφορά απομαγνητοφωνημένα κείμενα, τα οποία έχουν προβλεφθεί σωστά από το μοντέλο μας ότι ανήκουν σε ασθενείς με άνοια. Στο Σχήμα 1.6, παρουσιάζονται δύο κείμενα, η πρόβλεψη των οποίων είναι υγιούς ατόμου και η πραγματική ετικέτα είναι επίσης υγιές άτομο. Τέλος, το Σχήμα 1.7 απεικονίζει κείμενα που ταξινομούνται λανθασμένα. Τα απομαγνητοφωνημένα αυτά κείμενα αντιστοιχούν σε υγιή άτομα, ενώ η πρόβλεψη είναι άνοια. Επιπλέον, όπως μπορεί κανείς να παρατηρήσει, κάθε σε κάθε λέξη έχει ανατεθεί ένα χρώμα, είτε μπλε είτε πορτοκαλί. Για να είμαστε πιο ακριβείς, το μπλε χρώμα υποδεικνύει ποιες λέξεις είναι ενδεικτικές της ομάδας ελέγχου, ενώ το πορτοκαλί χρώμα υποδεικνύει λέξεις που χρησιμοποιούνται κυρίως από ασθενείς με άνοια. Όσο πιο έντονα είναι τα χρώματα, τόσο πιο σημαντικές είναι αυτές οι λέξεις προς την τελική ταξινόμηση του απομαγνητοφωνημένου κειμένου.

Όπως είναι εύκολο να παρατηρήσει κανείς στο Σχήμα 1.5, οι λέξεις που ανήκουν στο pos-tag UH, όπως το yeah και το oh, αναγνωρίζονται ως σημαντικές από το μοντέλο μας για την άνοια. Επιπλέον, οι προσωπικές αντωνυμίες (she, they) και τα ρήματα στον παρελθόν (got, had) είναι επίσης ενδεικτικά της νόσου. Επίσης, το μοντέλο μας θεωρεί σημαντική τη λέξη ”here,”, η οποία αντιστοιχεί στο pos-tag RB, ενδεικτικό της κατηγορίας της νόσου. Αυτά τα ευρήματα είναι συμβατά με αυτά που παρουσιάστηκαν στην Ενότητα 1.3.3.2, όπου έχουμε βρει ότι τα μέρη του λόγου PRP, VBD, UH, καθώς και η λέξη ”here” συσχετίζονται σημαντικά με την κατηγορία της νόσου. Επιπλέον, το μοντέλο μας αναγνωρίζει την επανάληψη της λέξης ”and” ως σημαντική για την κατηγορία της νόσου.

Σχετικά με το Σχήμα 1.6, κάποιος μπορεί εύκολα να παρατηρήσει ότι το μοντέλο μας αναγνωρίζει τις λέξεις που ανήκουν στα μέρη του λόγου VBG (putting, drying, blowing, standing, κλπ.), DT (the, a), και NN (cookie, action, stool, κλπ.) ως σημαντικά για την κατηγορία ελέγχου. Ταυτόχρονα, συμφωνώντας με τα ευρήματα της Ενότητας 1.3.3.2, οι λέξεις ”curtain” και ”window” χρησιμοποιούνται κυρίως από υγιείς.

Όσον αφορά στα Σχήματα 1.4 και 1.7, το μοντέλο μας δεν είναι σε θέση να ταξινομήσει σωστά αυτά τα κείμενα. Ένας πιθανός λόγος για τέτοιες λανθασμένες ταξινομήσεις σχετίζεται με το γεγονός ότι αυτά τα κείμενα περιλαμβάνουν pos-tags που είναι ενδεικτικά τόσο της ομάδας ελέγχου όσο και της ομάδας της νόσου. Πιο συγκεκριμένα, στο Σχήμα 1.4, η πλειοψηφία των λέξεων σε κάθε κείμενο ανήκουν στα μέρη του λόγου VBG, NN, και DT, τα οποία αναγνωρίζονται σωστά από το μοντέλο μας ως σημαντικά για την ομάδα ελέγχου. Λέξεις, όπως ”and”, ”him,” και ”well” χρησιμοποιούνται σε χαμηλή συχνότητα. Παρόμοια με το Σχήμα 1.4, στο Σχήμα 1.7, η πλειοψηφία των λέξεων σε κάθε κείμενο ανήκει στα pos-tags που συσχετίζονται σημαντικά με την κατηγορία της νόσου. Αυτό μπορεί να αποδειχθεί στο Σχήμα 1.7γ′, όπου παρατηρούμε τη χρήση λέξεων όπως ”and”, ”yeah,” ”well,” και ”got”.



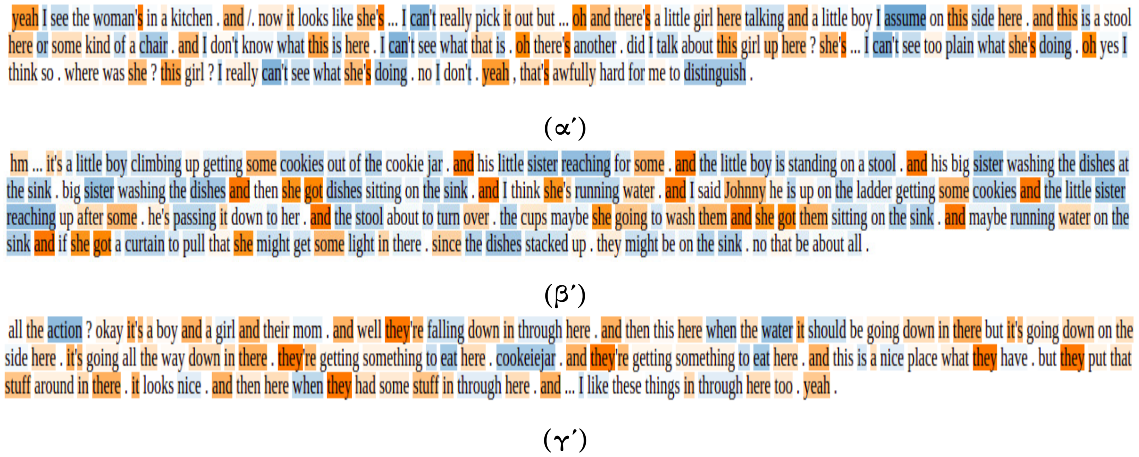
I see two kids up at the cookie jar , one on a stool the other standing on the floor . cupboard door is opened . mother's washing the dishes . the water is overflowing the sink . and there's two cups and a plate on the counter . and she's holding a plate in her hand . curtains at the windows . the cookie jar has the lid off . that's about it . cupboards underneath the sink . cupboards underneath the other cupboards . kid falling off the stool . the girl laughing at him . cookies in the cookie jar with the lid off . he has a cookie in his hand . and that's it .

(α′)

a little boy is getting himself some cookies out of the jar . and the stool is turning over on him that he's standing on . he's handing the little girl some cookies out_of the jar . is that it ? well that's in the first one . he's handing the little girl cookies out_of the cookie jar . and he's on a stool and it's turning over . and he's handing her some cookies . and in the second picture the little girl looks as though she's drying dishes . and the sink is running over on the floor . she has the spigots on . a nice view from the window . is that it ? that's it .

(β′)

Σχήμα 1.4: Label: Dementia, Prediction: Control



yeah I see the woman's in a kitchen . and /. now it looks like she's ... I can't really pick it out but ... oh and there's a little girl here talking and a little boy I assume on this side here . and this is a stool here or some kind of a chair . and I don't know what this is here . I can't see what that is . oh there's another . did I talk about this girl up here ? she's ... I can't see too plain what she's doing . oh yes I think so . where was she ? this girl ? I really can't see what she's doing . no I don't . yeah , that's awfully hard for me to distinguish .

(α′)

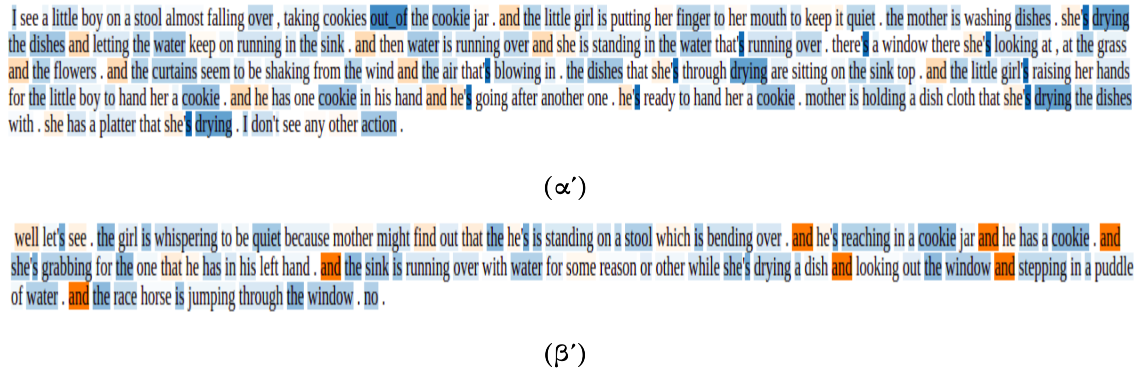
hm ... it's a little boy climbing up getting some cookies out of the cookie jar . and his little sister reaching for some . and the little boy is standing on a stool . and his big sister washing the dishes at the sink . big sister washing the dishes and then she got dishes sitting on the sink . and I think she's running water . and I said Johnny he is up on the ladder getting some cookies and the little sister reaching up after some . he's passing it down to her . and the stool about to turn over . the cups maybe she going to wash them and she got them sitting on the sink . and maybe running water on the sink and if she got a curtain to pull that she might get some light in there . since the dishes stacked up . they might be on the sink . no that be about all .

(β′)

all the action ? okay it's a boy and a girl and their mom . and well they're falling down in through here . and then this here when the water it should be going down in there but it's going down on the side here . it's going all the way down in there . they're getting something to eat here . cookeiejar . and they're getting something to eat here . and this is a nice place what they have . but they put that stuff around in there . it looks nice . and then here when they had some stuff in through here . and ... I like these things in through here too . yeah .

(γ′)

Σχήμα 1.5: Label: Dementia, Prediction: Dementia



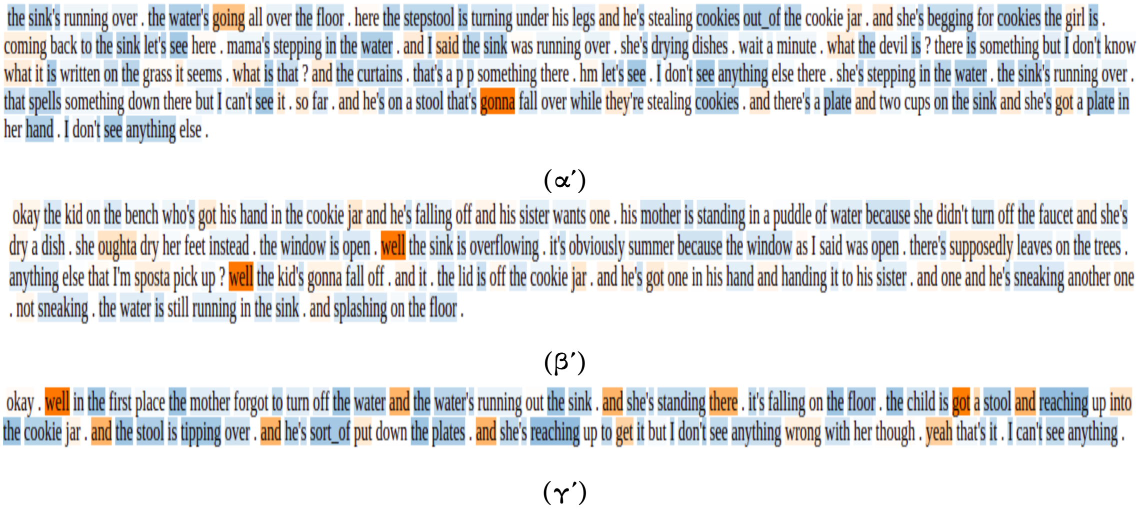
I see a little boy on a stool almost falling over , taking cookies out_of the cookie jar . and the little girl is putting her finger to her mouth to keep it quiet . the mother is washing dishes . she's drying the dishes and letting the water keep on running in the sink . and then water is running over and she is standing in the water that's running over . there's a window there she's looking at , at the grass and the flowers . and the curtains seem to be shaking from the wind and the air that's blowing in . the dishes that she's through drying are sitting on the sink top . and the little girl's raising her hands for the little boy to hand her a cookie . and he has one cookie in his hand and he's going after another one . he's ready to hand her a cookie . mother is holding a dish cloth that she's drying the dishes with . she has a platter that she's drying . I don't see any other action .

(α′)

well let's see . the girl is whispering to be quiet because mother might find out that the he's is standing on a stool which is bending over . and he's reaching in a cookie jar and he has a cookie . and she's grabbing for the one that he has in his left hand . and the sink is running over with water for some reason or other while she's drying a dish and looking out the window and stepping in a puddle of water . and the race horse is jumping through the window . no .

(β′)

Σχήμα 1.6: Label: Control, Prediction: Control



the sink's running over . the water's going all over the floor . here the stepstool is turning under his legs and he's stealing cookies out_of the cookie jar . and she's begging for cookies the girl is . coming back to the sink let's see here . mama's stepping in the water . and I said the sink was running over . she's drying dishes . wait a minute . what the devil is ? there is something but I don't know what it is written on the grass it seems . what is that ? and the curtains . that's a p p something there . hm let's see . I don't see anything else there . she's stepping in the water . the sink's running over . that spells something down there but I can't see it . so far . and he's on a stool that's gonna fall over while they're stealing cookies . and there's a plate and two cups on the sink and she's got a plate in her hand . I don't see anything else .

(α′)

okay the kid on the bench who's got his hand in the cookie jar and he's falling off and his sister wants one . his mother is standing in a puddle of water because she didn't turn off the faucet and she's dry a dish . she oughta dry her feet instead . the window is open . well the sink is overflowing . it's obviously summer because the window as I said was open . there's supposedly leaves on the trees . anything else that I'm sposta pick up ? well the kid's gonna fall off . and it . the lid is off the cookie jar . and he's got one in his hand and handing it to his sister . and one and he's sneaking another one . not sneaking . the water is still running in the sink . and splashing on the floor .

(β′)

okay . well in the first place the mother forgot to turn off the water and the water's running out the sink . and she's standing there . it's falling on the floor . the child is got a stool and reaching up into the cookie jar . and the stool is tipping over . and he's sort_of put down the plates . and she's reaching up to get it but I don't see anything wrong with her though . yeah that's it . I can't see anything .

(γ′)

Σχήμα 1.7: Label: Control, Prediction: Dementia

1.3.4 Πολυτροπικά Μοντέλα για Διάγνωση Άνοιας με χρήση Ομιλίας και Απομαγνητοφωνημένου Κειμένου

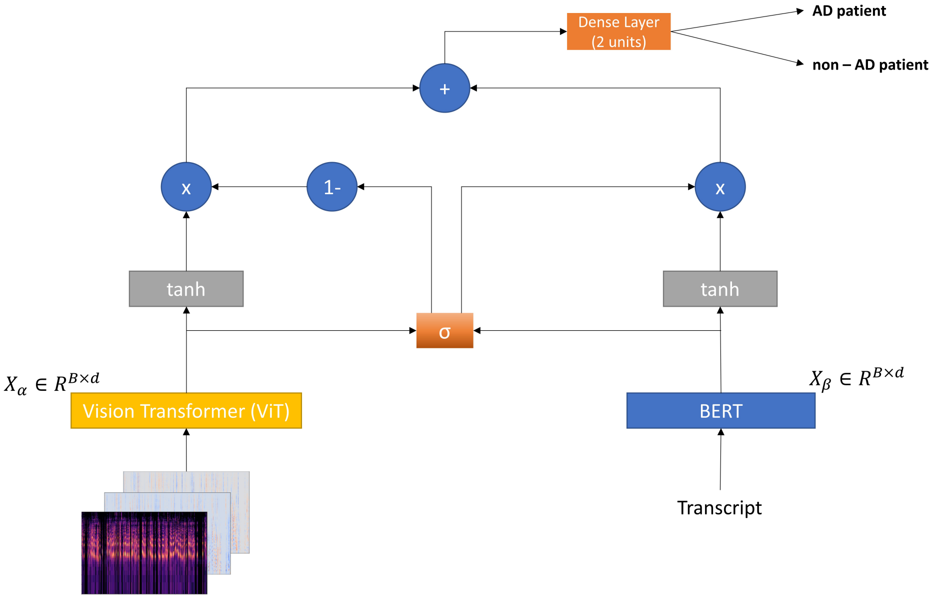
Στην προηγούμενη ενότητα, χρησιμοποιήσαμε μόνο το απομαγνητοφωνημένο κείμενο για τη διάγνωση της άνοιας. Στην ενότητα αυτή, θα χρησιμοποιήσουμε και το απομαγνητοφωνημένο κείμενο και τον ήχο.

Σε αντίθεση με υπάρχουσες εργασίες, που χρησιμοποιούν στρατηγικές early και late fusion αμελώντας έτσι τις αλληλεπιδράσεις διαφορετικών τροπικοτήτων μεταξύ τους, η διατριβή αυτή προτείνει μεθόδους αποτελεσματικού συνδυασμού των διαφορετικών τροπικοτήτων.

Στις Εικόνες 1.8 - 1.14, παρουσιάζουμε τις προτεινόμενες μεθόδους συγχώνευσης των διαφορετικών τροπικοτήτων.

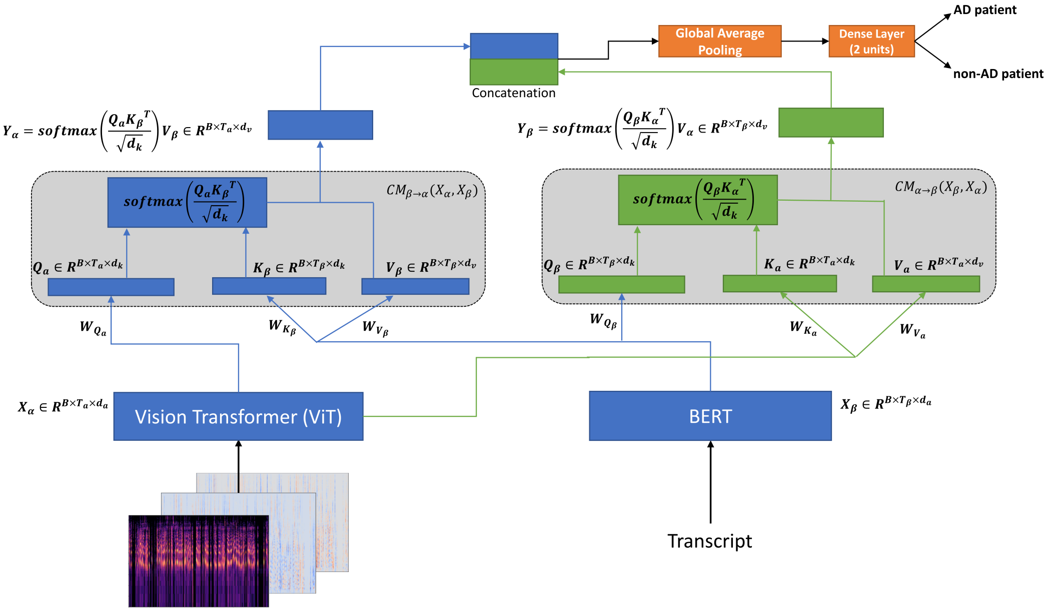
Ως είσοδο σε όλα τα νευρωνικά δίνεται το απομαγνητοφωνημένο κείμενο, ενώ το αρχείο ήχου μετατρέπεται σε εικόνα 3 καναλιών, log-Mel spectrogram, delta, delta-delta. Στη συνέχεια, χρησιμοποιούμε πολυτροπικές μεθόδους, οι οποίες περιγράφονται παρακάτω:

BERT + ViT + Gated Multimodal Unit Η προτεινόμενη αρχιτεκτονική απεικονίζεται στην Εικόνα 1.8. Χρησιμοποιούμε ως μέθοδο συγχώνευσης των διαφορετικών τροπικοτήτων το Gated Multimodal Unit [75], προκειμένου να ελέγξουμε τη συνεισφορά της κάθε τροπικότητας ως προς την τελική έξοδο/ταξινόμηση.



Σχήμα 1.8: BERT + ViT + Gated Multimodal Unit

BERT + ViT + Crossmodal Attention Η προτεινόμενη αρχιτεκτονική απεικονίζεται στην Εικόνα 1.9. Χρησιμοποιούμε ως μέθοδο συγχώνευσης των διαφορετικών τροπικοτήτων το crossmodal attention [76, 77, 78]. Συγκεκριμένα, ο μηχανισμός cross-attention διακρίνεται σε δύο επίπεδα προσοχής, ένα από τα κείμενα X_β προς τα οπτικά χαρακτηριστικά X_α και ένα από τα οπτικά προς τα κειμενικά χαρακτηριστικά. Στη συνέχεια, υπολογίζουμε το scaled dot attention, όπως προτάθηκε στο [46] και δίνεται από την εξίσωση: $(\alpha = softmax(QK^T/\sqrt{d_{proj}})V)$ με την αναπαράσταση του κειμένου ως query (Q), και την αναπαράσταση της εικόνας ως key (K) και value (V), και αντίστροφα.

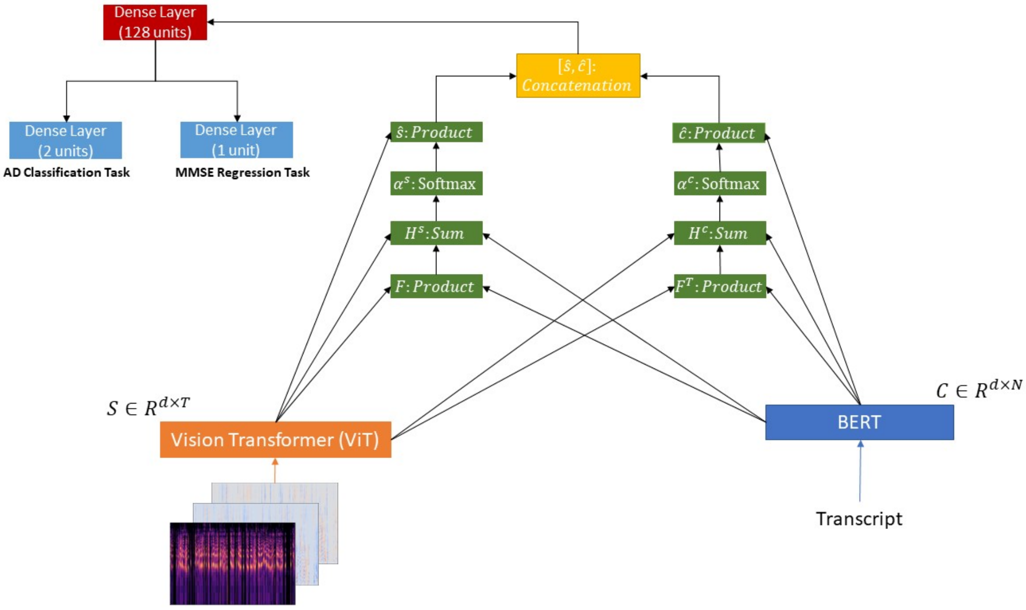

AD patient
non-AD patient
Global Average Pooling
Dense Layer (2 units)
Concatenation
$Y_\alpha = softmax\left(\frac{Q_\alpha K_\beta{}^T}{\sqrt{d_k}}\right)V_\beta \in R^{B\times T_\alpha\times d_v}$
$Y_\beta = softmax\left(\frac{Q_\beta K_\alpha{}^T}{\sqrt{d_k}}\right)V_\alpha \in R^{B\times T_\beta\times d_v}$
$softmax\left(\frac{Q_\alpha K_\beta{}^T}{\sqrt{d_k}}\right)$
$CM_{\beta\rightarrow\alpha}(X_\alpha, X_\beta)$
$softmax\left(\frac{Q_\beta K_\alpha{}^T}{\sqrt{d_k}}\right)$
$CM_{\alpha\rightarrow\beta}(X_\beta, X_\alpha)$
$Q_\alpha \in R^{B\times T_\alpha\times d_k}$
$K_\beta \in R^{B\times T_\beta\times d_k}$
$V_\beta \in R^{B\times T_\beta\times d_v}$
$Q_\beta \in R^{B\times T_\beta\times d_k}$
$K_\alpha \in R^{B\times T_\alpha\times d_k}$
$V_\alpha \in R^{B\times T_\alpha\times d_v}$
W_{Q_α}
W_{K_β}
W_{V_β}
W_{Q_β}
W_{K_α}
W_{V_α}
$X_\alpha \in R^{B\times T_\alpha\times d_a}$
Vision Transformer (ViT)
BERT
$X_\beta \in R^{B\times T_\beta\times d_a}$
Transcript

Σχήμα 1.9: BERT + ViT + Crossmodal Attention

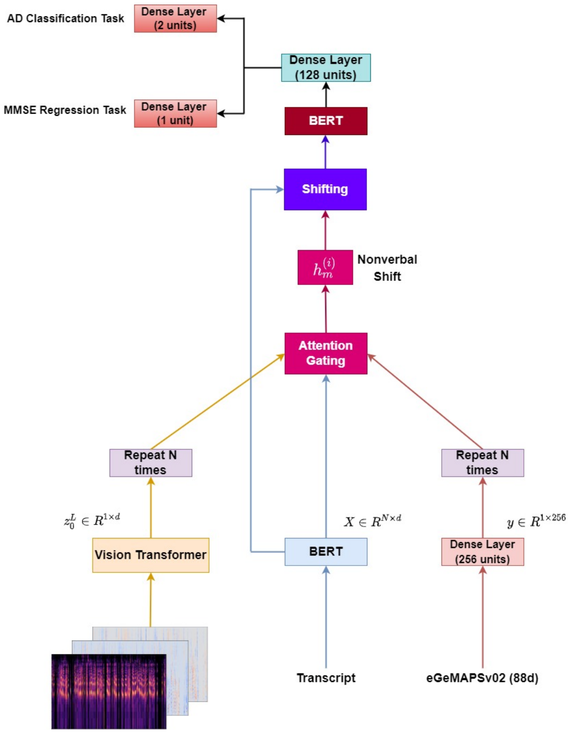
BERT + ViT + Co-Attention Η προτεινόμενη αρχιτεκτονική απεικονίζεται στην Εικόνα 1.10. Ως μέθοδος συγχώνευσης των γλωσσικών και ακουστικών αναπαραστάσεων χρησιμοποιείται ο μηχανισμός co-attention [79, 80]. Ο μηχανισμός αυτός χρησιμοποιείται στις αναπαραστάσεις του κειμένου και της εικόνας και βοηθάει στη μάθηση των βαρών προσοχής των απομαγνητοφωνημένων κειμένων και τμημάτων της εικόνας ταυτόχρονα.

Multimodal BERT Η προτεινόμενη αρχιτεκτονική απεικονίζεται στην Εικόνα 1.11. Χρησιμοποιούμε μία μέθοδο, η οποία εισάγει ακουστική και οπτική πληροφορία στο γλωσσικό μοντέλο BERT [18, 81, 19, 82].

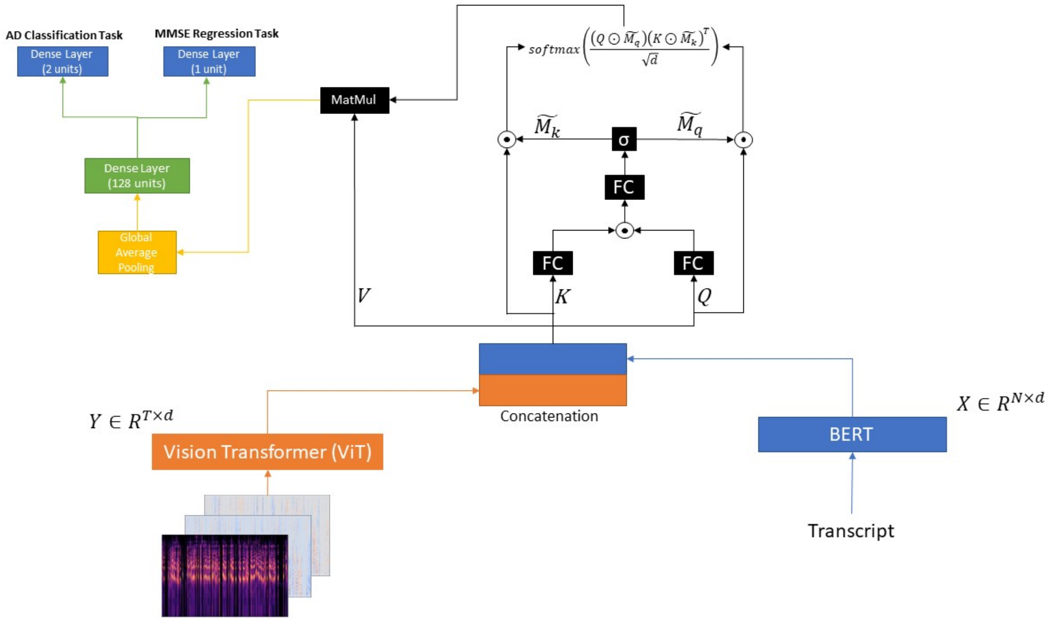
BERT + ViT + Gated Self-Attention Η προτεινόμενη αρχιτεκτονική απεικονίζεται στην Εικόνα 1.12. Ως μέθοδος συγχώνευσης των διαφορετικών τροπικοτήτων χρησιμοποιείται το Gated Self-Attention [83]. Συγκεκριμένα, συνενώνουμε κατά γραμμές τις αναπαραστάσεις του κειμένου και της εικόνας και χρησιμοποιούμε έναν μηχανισμό αυτο-προσοχής, που περιέχει ένα μοντέλο πύλης gated model.



Σχήμα 1.10: BERT + ViT + Co-Attention



Σχήμα 1.11: Multimodal BERT - eGeMAPS + ViT



Σχήμα 1.12: BERT + ViT + Gated Self-Attention

1.3.4.1 Καλιμπράρισμα (Calibration)

Υιοθετούμε τη μέθοδο που αναφέρθηκε στην Ενότητα 1.2.1.4. Η προτεινόμενη αρχιτεκτονική απεικονίζεται στην Εικόνα 1.13. Συγκεκριμένα, αφού έχουμε λάβει τις αναπαραστάσεις του κειμένου και της εικόνας, ακολουθούμε την ακόλουθη διαδικασία:

- Βέλτιστος Πυρήνας Μεταφοράς (Optimal Transport Kernel): Για να εξασφαλίσουμε ότι το μήκος της ακολουθίας των διανυσμάτων που προκύπτουν από το BERT και το DeiT είναι το ίδιο, εκμεταλλευόμαστε έναν Βέλτιστο Πυρήνα Μεταφοράς (OTK).

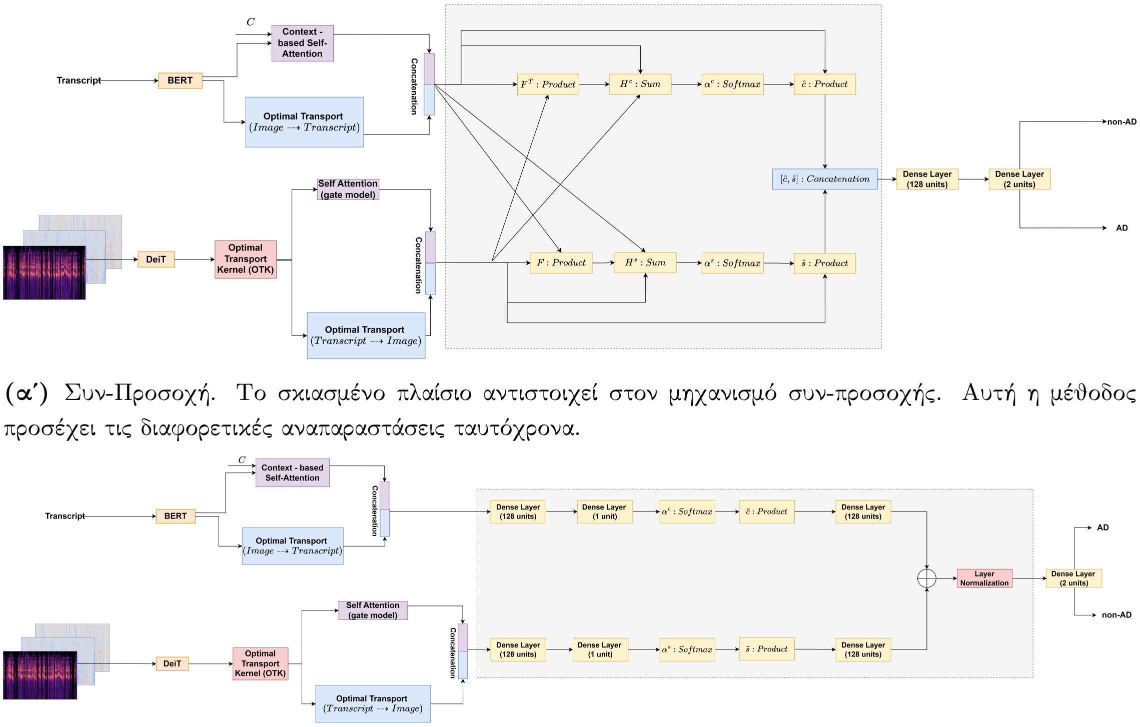
- Αναπαράσταση κειμένου: Περνάμε την κειμενική αναπαράσταση μέσω ενός ενισχυμένου επιπέδου αυτο-προσοχής με πληροφορίες περιεχομένου. Εκμεταλλευόμαστε τρεις κύριες μεθόδους για την παροχή πλαισίου (contextualization), συμπεριλαμβανομένου του καθολικού περιεχομένου (global context), του βαθέος περιεχομένου (deep context) και του βαθέος-καθολικού περιεχομένου.

- Αναπαράσταση Εικόνας: Περνάμε την εικονική αναπαράσταση μέσω ενός μηχανισμού αυτο-προσοχής με ένα νέο μοντέλο πύλης για τη μοντελοποίηση των εσωτερικών ενδοτροπικών αλληλεπιδράσεων.

- Βέλτιστη Μεταφορά (Optimal Transport): Χρησιμοποιούμε μεθόδους βέλτιστης μεταφοράς για την καταγραφή των δια-τροπικών αλληλεπιδράσεων.

- Πολυτροπικές Μέθοδοι: Στη συνέχεια, προτείνουμε δύο μεθόδους βασισμένες σε μηχανισμό προσοχής για τη συγχώνευση των χαρακτηριστικών αυτο-προσοχής και συν-προσοχής.

- Καλιμπράρισμα: Τέλος, για την αποτροπή δημιουργίας μοντέλων με υπερβολική αυτοπεποίθηση, χρησιμοποιούμε label smoothing.



(α΄) Συν-Προσοχή. Το σκιασμένο πλαίσιο αντιστοιχεί στον μηχανισμό συν-προσοχής. Αυτή η μέθοδος προσέχει τις διαφορετικές αναπαραστάσεις ταυτόχρονα.

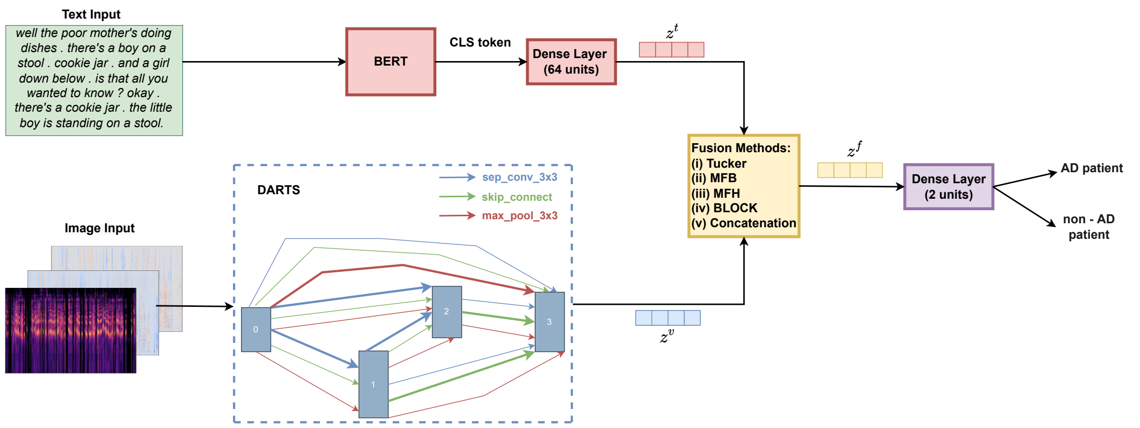
(β΄) Το σκιασμένο πλαίσιο δείχνει τη μέθοδο συγχώνευσης. Αυτή η μέθοδος χρησιμοποιεί δύο ανεξάρτητα μοντέλα προσοχής. Τα χαρακτηριστικά συγχωνεύονται μέσω μιας λειτουργίας πρόσθεσης, ενώ το layer normalization χρησιμοποιείται για τη σταθεροποίηση της εκπαίδευσης.

Σχήμα 1.13: Προτεινόμενες Αρχιτεκτονικές - Optimal Transport - Calibration

1.3.4.2 Αυτόματη Αναζήτηση Αρχιτεκτονικής Νευρωνικού Δικτύου

Η προτεινόμενη αρχιτεκτονική απεικονίζεται στην Εικόνα 1.14. Αποτελείται από τα παρακάτω τμήματα:

- Απομαγνητοφωνημένο Κείμενο: BERT
- Αναζήτηση Ομιλίας-Νευρωνικής Αρχιτεκτονικής χρησιμοποιώντας τη μέθοδο DARTS [84]. .
  - Πολυτροπικές Μέθοδοι.
  - Tucker Decomposition [85]
  - Multimodal Factorized Bilinear Pooling (MFB) [86]
  - Multimodal Factorized High-order pooling (MFH) [86]
  - BLOCK [87]
  - Συνένωση (Concatenation)



Σχήμα 1.14: Μοντέλο που χρησιμοποιεί αυτόματη αναζήτηση αρχιτεκτονικής νευρωνικού δικτύου και πολυτροπικές μεθόδους.

1.3.5 Αποτελέσματα

Στον Πίνακα 1.7 παρατηρούμε τα αποτελέσματα των προτεινόμενων αρχιτεκτονικών. Επίσης, πραγματοποιείται σύγκριση των αρχιτεκτονικών αυτών με υπάρχουσες μεθόδους της βιβλιογραφίας.

Όσον αφορά τα προτεινόμενα μοντέλα, παρατηρούμε ότι το DARTS + BERT + BLOCK είναι το καλύτερο μοντέλο σε απόδοση πετυχαίνοντας ποσοστά Accuracy και F1-score ίσα με 92.08% και 91.94% αντίστοιχα. Πιστεύουμε ότι το μοντέλο αυτό είναι το καλύτερο, επειδή περιέχει μηχανισμό αυτόματης αναζήτησης αρχιτεκτονικής νευρωνικού δικτύου. Συγχρόνως, χρησιμοποιείται ως μέθοδος συγχώνευσης το BLOCK. Το δεύτερο σε απόδοση μοντέλο είναι το Attention-based fusion - Optimal Transport, το οποίο εξασφαλίζει απόδοση σε Accuracy ίση με 91.25%. Ο μηχανισμός calibration, που συμπεριλαμβάνεται σε αυτό το μοντέλο, συμβάλλει στην απόδοση αυτή. Στην απόδοση αυτή συμβάλλουν επίσης και οι μέθοδοι συγχώνευσης των τροπικοτήτων. Το μοντέλο BERT + ViT + Gated Self-Attention πετυχαίνει την τρίτη μεγαλύτερη απόδοση σε Accuracy, καθώς η μέθοδος αυτή ¨πιάνει’ όλες τις αλληλεπιδράσεις μεταξύ κειμένου και ήχου. Παρατηρούμε ότι το μοντέλο BERT + ViT + Gated Multimodal Unit πετυχαίνει τη δεύτερη χειρότερη απόδοση. Πιστεύουμε ότι αυτό οφείλεται στο γεγονός ότι η μέθοδος Gated Multimodal Unit ελέγχει τη ροή πληροφορίας προς την έξοδο καθορίζοντας ποια τροπικότητα είναι περισσότερη σημαντική χωρίς να ¨πιάνει’ τις αλληλεπιδράσεις μεταξύ των διαφορετικών τροπικοτήτων. Η χειρότερη απόδοση σε Accuracy με ποσοστό ίσο με 80.83%

Όσον αφορά τη σύγκριση των μοντέλων μας με υπάρχουσες μεθόδους της βιβλιογραφίας, παρατηρούμε τα εξής:

- Συγκριτικά με τις πολυτροπικές μεθόδους, παρατηρούμε ότι το καλύτερο μοντέλο μας, δηλαδή το DARTS + BERT + BLOCK, τις ξεπερνά σε Accuracy κατά 2.50 - 17.08%. Αυτό συμβαίνει, επειδή οι εργασίες αυτές χρησιμοποιούν μεθόδους early & late fusion ή συγχωνεύουν τις αναπαραστάσεις διαφορετικών τροπικοτήτων κατά τη διάρκεια της εκπαίδευσης. Επομένως, δεν ¨πιάνουν’ τις αλληλεπιδράσεις των διαφορετικών τροπικοτήτων.

- Συγκριτικά με τις μεθόδους, που χρησιμοποιούν ως είσοδο μόνο μία τροπικότητα (κείμενο ή ήχο), παρατηρούμε ότι το μοντέλο μας πετυχαίνει καλύτερη απόδοση.
  - Όσον αφορά τις μεθόδους που χρησιμοποιούν μόνο το απομαγνητοφωνημένο κείμενο (BERT), το μοντέλο μας πετυχαίνει καλύτερη απόδοση σε Accuracy και F1-score κατά 4.58% και 5.21% αντίστοιχα.
  - Όσον αφορά τις πολυτροπικές μεθόδους, η αρχιτεκτονική μας υπερβαίνει τις εργασίες αυτές ως προς το Accuracy κατά 2.5 - 17.08%.

Πίνακας 1.7: Σύγκριση απόδοσης μεταξύ των προτεινόμενων μοντέλων και των υπάρχουσων μεθόδων στο σύνολο δεδομένων του ADReSS Challenge. Οι αναφερόμενες τιμές είναι ο μέσος όρος ± η τυπική απόκλιση. Τα αποτελέσματα είναι μέσος όρος από πέντε εκτελέσεις. Τα καλύτερα αποτελέσματα ανά μετρική αξιολόγησης παρουσιάζονται με έντονα γράμματα.

	Μετρικές Αξιολόγησης				
Αρχιτεκτονική	**Precision**	**Recall**	**F1-score**	**Accuracy**	**Specificity**
Unimodal state-of-the-art approaches (only transcripts)					
BERT	87.19 ±3.25	81.66 ±5.00	86.73 ±4.53	87.50 ±4.37	93.33 ±5.65
Unimodal state-of-the-art approaches (only Speech)					
DARTS	70.04 ±3.84	89.99 ±2.04	76.09 ±0.87	72.92 ±2.28	62.3 ±7.05
AT-LSTM (x-vector) [88]	66.00	69.00	67.00	67.00	-
ECAPA-TDNN [89]	-	-	-	66.70	-
SiameseNet [63]	-	-	70.80	70.80	-
x-vectors_SRE [67]	54.17	54.17	54.17	54.17	54.17
Acoustic+Silence [90]	70.00	58.00	63.00	66.70	75.00
YAMNet [91]	64.40±3.93	73.40±8.82	68.60±4.84	66.20±4.79	59.20±7.73
Majority vote (Acoustic) [64]	-	-	-	65.00	-
Audio (Fusion) [92]	-	83.33	-	81.25	79.17
DemCNN [93]	62.50	62.50	62.50	62.50	62.50
CNN-LSTM (MFCC) [94]	82.00	38.00	51.00	64.58	92.00
Multimodal state-of-the-art approaches (speech and transcripts)					
Audio + Text (Fusion) [92]	-	87.50	-	89.58	91.67
Fusion Maj. (3-best) [63]	-	-	85.40	85.20	-
Fusion of system [67]	94.12	66.67	78.05	81.25	**95.83**
GFI,NUW,Duration,Character 4-grams, Suffixes,POS tag,UD [95]	-	-	-	77.08	-
Acoustic & Transcript [90]	70.00	88.00	78.00	75.00	83.00
Dual BERT [91]	83.04 ±3.97	83.33 ±5.89	82.92 ±1.86	82.92 ±1.56	82.50 ±5.53
Majority vote (NLP + Acoustic) [64]	-	-	-	83.00	-
Προτεινόμενες Αρχιτεκτονικές					
BERT + ViT + Gated Multimodal Unit	80.92	**91.67**	85.92	85.00	78.33
BERT + ViT + Crossmodal Attention	86.13	**91.67**	88.69	88.33	85.00
Multimodal BERT	76.57	89.17	82.28	80.83	72.50
BERT + ViT + Co-Attention	92.83	81.67	86.81	87.50	93.33
BERT + ViT + Gated Self-Attention	90.87	89.17	89.94	90.00	90.83
BERT + ViT + Gated Multimodal Unit	89.16	85.00	86.73	87.08	89.16
Co-Attention – Optimal Transport	93.57	84.16	88.53	89.16	**94.16**
Attention - based fusion – Optimal Transport	93.08	89.17	91.06	91.25	93.33
DARTS+BERT+Tucker Decomposition	89.16	85.00	86.73	87.08	89.16
DARTS+BERT+MFB	91.29	88.29	89.80	89.58	91.66
DARTS+BERT+MFH	**94.46**	86.66	88.31	88.74	**94.16**
DARTS+BERT+BLOCK	94.09	91.66	**91.94**	**92.08**	**94.16**
DARTS+BERT+Concatenation	86.68	90.83	88.65	88.33	85.83

1.4 Διάγνωση Επιληψίας

1.4.1 Κίνητρο

Η επιληψία είναι μια νευρολογική νόσος, η οποία επηρεάζει άτομα όλων των ηλικιών. Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας (ΠΟΥ), περίπου 50 εκατομμύρια άνθρωποι έχουν επιληψία παγκοσμίως, γεγονός που την καθιστά μια από τις περισσότερο κοινές νευρολογικές παθήσεις [96]. Η επιληψία έχει αρνητικό αντίκτυπο στην καθημερινή ζωή των ανθρώπων κυρίως λόγω των διακρίσεων και του στίγματος που περιβάλλει την ίδια την ασθένεια. Ωστόσο, ο ΠΟΥ αναφέρει ότι έως και το 70% των ανθρώπων που ζουν με επιληψία θα μπορούσαν να ζήσουν χωρίς επιληπτικές κρίσεις, εάν οι άνθρωποι διαγνωστούν έγκαιρα και λάβουν την κατάλληλη θεραπεία. Επομένως, η έγκαιρη διάγνωση της επιληψίας είναι σημαντική για την παροχή καλύτερης ποιότητας ζωής στους επιληπτικούς ασθενείς.

Υπάρχει ένας σημαντικός αριθμός μελετών που προτείνουν μεθόδους για την ανίχνευση επιληπτικών κρίσεων. Η πλειοψηφία αυτών των μελετών εξάγει χαρακτηριστικά τόσο του τομέα χρόνου όσο και του τομέα συχνότητας από το ηλεκτροεγκεφαλογράφημα (ΗΕΓ). Για παράδειγμα, οι συγγραφείς εφαρμόζουν το Discrete Wavelet Transform (DWT) [97, 98] για την αποσύνθεση των σημάτων ΗΕΓ σε υποζώνες και στη συνέχεια την εξαγωγή χαρακτηριστικών από κάθε υποζώνη. Αφού εξάγουν μεγάλο αριθμό χαρακτηριστικών, οι συγγραφείς συνήθως εκμεταλλεύονται την επιλογή χαρακτηριστικών feature selection ή τεχνικές μείωσης διαστάσεων (dimensionality reduction techniques) για την εύρεση του καλύτερου υποσυνόλου χαρακτηριστικών ή τη μείωση της διάστασης του διανύσματος χαρακτηριστικών αντίστοιχα. Το τελευταίο βήμα από τις προτεινόμενες μεθόδους περιλαμβάνει το σύνολο των παραδοσιακών ταξινομητών μηχανικής μάθησης, π.χ. Logistic Regression (LR), Support Vector Machines (SVMs), Random Forests (RF), Decision Trees, κ.λπ. Αυτές οι μέθοδοι είναι χρονοβόρες, καθώς απαιτούν κάποιο επίπεδο τεχνογνωσίας για την εξαγωγή των καλύτερων αντιπροσωπευτικών χαρακτηριστικών. Μόνο μερικές μελέτες [99, 100, 101, 102] έχουν εκμεταλλευτεί βαθιά νευρωνικά δίκτυα, δηλ. CNNs, LSTMs ή BiLSTMs για την ανίχνευση και πρόβλεψη της επιληψίας. Ωστόσο, οι περισσότερες από αυτές τις μεθόδους εξακολουθούν να βασίζονται στην εξαγωγή χαρακτηριστικών [100, 101, 99]. Ένας άλλος περιορισμός είναι το γεγονός ότι οι υπάρχουσες εργασίες χωρίζουν τα σήματα ΗΕΓ σε τμήματα και προτείνουν majority vote προσεγγίσεις [102]. Έτσι, πρέπει να εκπαιδεύονται πολλαπλά μοντέλα αυξάνοντας σημαντικά τον υπολογιστικό χρόνο. Ταυτόχρονα, τα περισσότερα μοντέλα CNN δεν είναι σε θέση να μοντελοποιήσουν αποτελεσματικά τις χρονικές εξαρτήσεις μεταξύ των δεδομένων ΗΕΓ. Αν και τα LSTM και τα BiLSTM μπορούν να συλλάβουν τις χρονικές εξαρτήσεις στα δεδομένα ΗΕΓ, συνήθως έχουν υψηλή πολυπλοκότητα μοντέλου.

1.4.2 Δεδομένα

EEG Database of the University of Bonn. Αυτό το σύνολο δεδομένων [103] αποτελείται από πέντε υποσύνολα, τα οποία συμβολίζονται ως A, B, C, D και E. Κάθε υποσύνολο περιέχει 100 τμήματα ΗΕΓ ενός καναλιού διάρκειας 23,6 δευτερολέπτων. Η συχνότητα δειγματοληψίας είναι ίση με 173,61 Hz. Έτσι, κάθε τμήμα ΗΕΓ αποτελείται από 4097 δείγματα. Τα σετ A και B έχουν συλλεχθεί από πέντε υγιείς εθελοντές με τα μάτια τους ανοιχτά και κλειστά αντίστοιχα. Τα σετ C και D έχουν συλλεχθεί κατά τη διάρκεια της ενδιάμεσης κατάστασης (διάστημα χωρίς επιληπτικές κρίσεις). Συγκεκριμένα, τμήματα στο σύνολο D έχουν καταγραφεί από τον σχηματισμό του ιππόκαμπου που προσδιορίζεται ως επιληπτογόνος ζώνη, ενώ τα σήματα στο σύνολο δεδομένων C έχουν καταγραφεί από τον ιππόκαμπο σχηματισμό του αντίθετου ημισφαιρίου του εγκεφάλου. Το σύνολο δεδομένων E περιέχει τμήματα από δραστηριότητα επιληπτικής κρίσης. Εφαρμόστηκε ένα ζωνοπερατό φίλτρο στα σήματα ΗΕΓ με χαμηλές και υψηλές συχνότητες αποκοπής 0,53 Hz και 40 Hz αντίστοιχα. Όλα αυτά τα τμήματα έχουν επιθεωρηθεί χειροκίνητα από έναν ειδικό λόγω της μυϊκής δραστηριότητας και των κινήσεων των ματιών.

Παρακάτω, θα πραγματοποιήσουμε τα πειράματά μας χρησιμοποιώντας την εξής κατηγοριοποίηση: **AB (υγιή άτομα) - CD (interictal) - E (ictal)**.

1.4.3 Μεθοδολογία

Σε αυτήν την ενότητα, περιγράφουμε την αρχιτεκτονική που εισαγάγαμε για την ανίχνευση της επιληψίας χρησιμοποιώντας σήματα ΗΕΓ και φασματογράμματα STFT. Η προτεινόμενη αρχιτεκτονική απεικονίζεται στην Εικόνα 1.15.

- Σήμα ΗΕΓ: Όπως φαίνεται στο Σχ. 1.15, υλοποιούμε δύο κλάδους CNN με διαφορετικά μεγέθη πυρήνα για την επεξεργασία των σημάτων ΗΕΓ. Η επιλογή αυτών των δύο κλάδων CNN με μικρά και μεγάλα μεγέθη φίλτρου είναι εμπνευσμένα από τους [104, 105], όπου οι συγγραφείς αναφέρουν ότι το μικρό φίλτρο είναι σε θέση καταγράφει χρονικές πληροφορίες, ενώ το μεγαλύτερο φίλτρο είναι ικανό για τη σύλληψη πληροφοριών συχνότητας. Κάθε κλάδος αποτελείται από τρία συνελικτικά στρώματα και δύο στρώματα max-pooling, όπου κάθε συνελικτικό στρώμα περιλαμβάνει ένα επίπεδο κανονικοποίησης [106] και μια συνάρτηση ενεργοποίησης ReLU. Όπως μπορεί κανείς να παρατηρήσει από το Σχ. 1.15, το πρώτο συνελικτικό μπλοκ κάθε κλάδου δείχνει το μέγεθος του φίλτρου, τον αριθμό των φίλτρων και το μέγεθος του διασκελισμού (stride size). Τα επόμενα δύο συνελικτικά μπλοκ του κάθε κλάδου δείχνουν το μέγεθος του φίλτρου και τον αριθμό των φίλτρων. Το μέγεθος του διασκελισμού είναι ίσο με 1. Κάθε μπλοκ max-pooling δείχνει το μέγεθός του και το μέγεθος του διασκελισμού. Για τη μείωση της υπερπροσαρμογής, εφαρμόζουμε dropout layer με συντελεστή 0,5 μετά το πρώτο μπλοκ max-pool κάθε κλάδου και μετά τη συνένωση και των δύο κλάδων. Τέλος, χρησιμοποιούμε ένα flatten layer, οπότε η έξοδος έχει διάσταση 1d. Έστω το αποτέλεσμα αυτού του τμήματος της αρχιτεκτονικής: f^t.

- Αναπαράσταση εικόνας: Εφαρμόζουμε τον μετασχηματισμό Fourier μικρού χρόνου (STFT) στα ακατέργαστα σήματα ΕΕΓ. Μετά τον υπολογισμό των απόλυτων τιμών του φασματογράμματος STFT (μέτρο STFT), υπολογίζουμε το φασματογράμμα σε κλίμακα db, το δέλτα και το δέλτα-δέλτα. Έτσι, κατασκευάζουμε μια εικόνα που αποτελείται από τρία

Πίνακας 1.8: Απόδοση του προτεινόμενου μοντέλου μέσω της μεθόδου cross-validation (AB - CD - E). Τα αποτελέσματα είναι στη μορφή: μέσος όρος ± τυπική απόκλιση.

Μοντέλο	**Μετρικές Αξιολόγησης**									
	Precision			**Recall**			**F1-score**			**Accuracy**
	AB	CD	E	AB	CD	E	AB	CD	E	
Προτεινόμενο Μοντέλο	97.14 ±3.10	97.16 ±3.75	97.18 ±4.31	97.99 ±2.45	96.49 ±2.29	96.00 ±6.63	97.52 ±1.91	96.77 ±1.88	96.41 ±4.09	97.00 ±1.84

κανάλια, δηλαδή φασματογράμμα σε κλίμακα db, δέλτα και δέλτα-δέλτα. Κάθε εικόνα κλιμακώνεται σε [0,1]. Κάθε εικόνα μετασχηματίζεται σε 224 × 224 pixels. Οι τιμές των εικόνων όλων κανονικοποιούνται.

Όπως φαίνεται στο Σχήμα 1.15, κάθε εικόνα δίνεται σε ένα προεκπαιδευμένο μοντέλο EfficientNet-B7, ακολουθούμενο από ένα επίπεδο απόρριψης με ποσοστό 0,5. Επίσης, αφαιρούμε το τελευταίο επίπεδο του EfficientNet που χρησιμοποιείται για την κατηγοριοποίηση. Έτσι, το προεκπαιδευμένο μοντέλο EfficientNet-B7 ενεργεί ως εξαγωγέας χαρακτηριστικών. Έστω η έξοδος είναι: f^v.

- Gated Multimodal Unit: Εφαρμόζουμε το Gated Multimodal Unit [75], προκειμένου να αναθέσουμε περισσότερη σημασία στη σχετική τροπικότητα αγνοώντας τις μη σχετικές πληροφορίες. Δεδομένων των f^t και f^v όπως υπολογίστηκαν παραπάνω, υπολογίζουμε την έξοδο αυτής της πολυτροπικής μεθόδου h.

- Επίπεδο εξόδου: Μεταβιβάζεται η πολυτροπική αναπαράσταση h σε ένα dropout layer με ρυθμό 0,5 ακολουθούμενο από ένα πυκνό στρώμα, που δίνει το τελικό αποτέλεσμα. Ο αριθμός των μονάδων στο πυκνό στρώμα εξαρτάται από κάθε περίπτωση που εξετάζεται για ταξινόμηση και μπορεί να είναι είτε δύο (δυαδική ταξινόμηση) είτε τρεις μονάδες (multiclass classification).

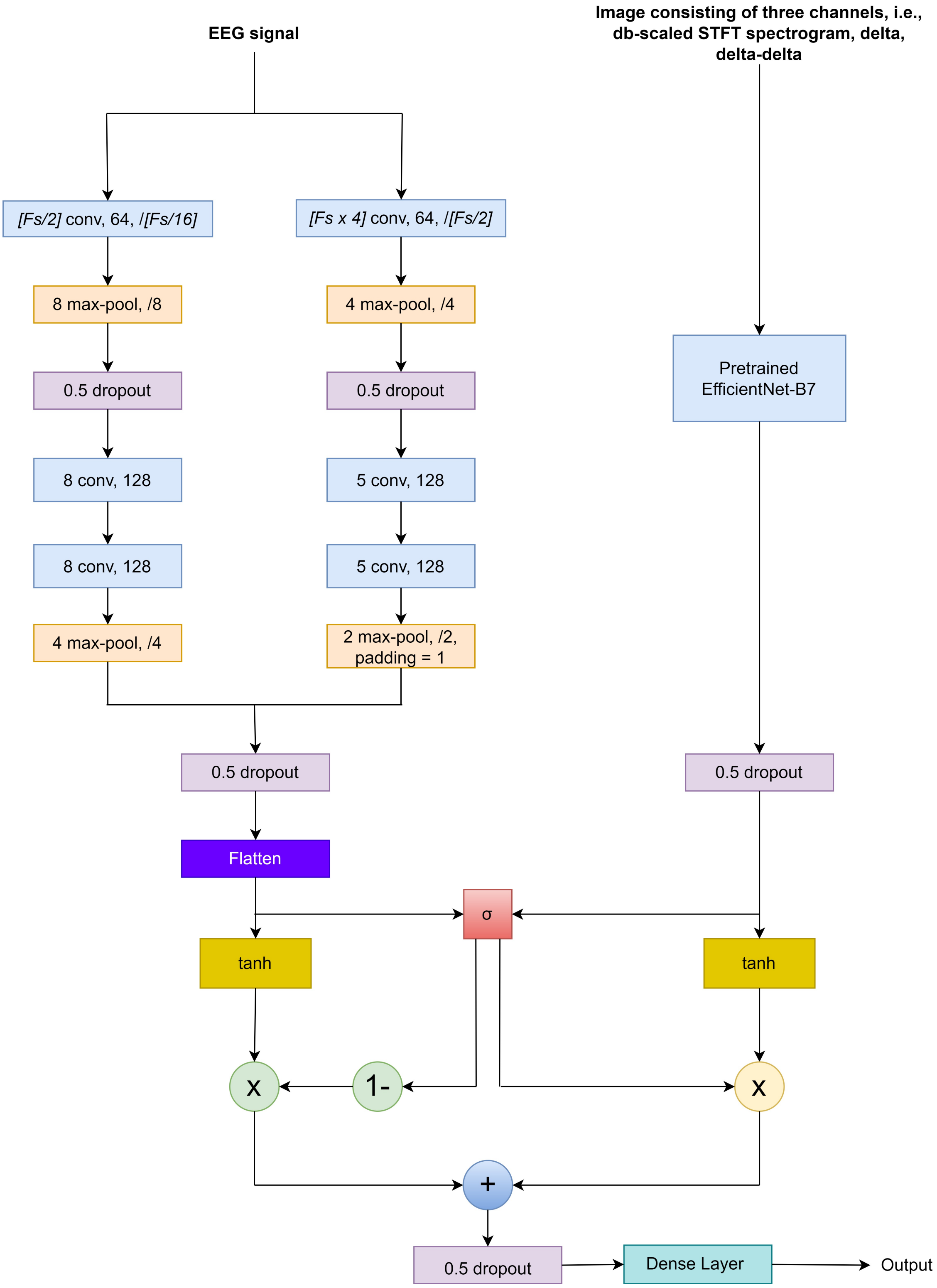
1.4.4 Αποτελέσματα

Όσον αφορά την περίπτωση (AB–CD–E), όπως παρατηρείται στον Πίνακα 1.8, το μοντέλο μας επιτυγχάνει ένα βαθμό ακρίβειας που ανέρχεται στο 97.00%. F1-score που ισούνται με 97.52%, 96.77% και 96.41% επιτυγχάνονται για τις κλάσεις AB (healthy), CD (interictal) και E (ictal) αντίστοιχα.

Μπορεί κανείς να παρατηρήσει από τον Πίνακα 1.9 ότι το μοντέλο μας υπερτερεί 15 ερευνητικών πρωτοβουλιών σε ακρίβεια κατά 0.50-17.00%.

1.4.5 Ablation Study

Σε αυτήν την ενότητα, εκτελούμε μια σειρά από πειράματα, για να εξετάσουμε την αποτελεσματικότητα και την ευρωστία της προτεινόμενης αρχιτεκτονικής που περιγράφεται στην


EEG signal
Image consisting of three channels, i.e., db-scaled STFT spectrogram, delta, delta-delta
[Fs/2] conv, 64, /[Fs/16]
[Fs x 4] conv, 64, /[Fs/2]
8 max-pool, /8
4 max-pool, /4
0.5 dropout
0.5 dropout
8 conv, 128
5 conv, 128
8 conv, 128
5 conv, 128
4 max-pool, /4
2 max-pool, /2, padding = 1
Pretrained EfficientNet-B7
0.5 dropout
0.5 dropout
Flatten
σ
tanh
tanh
x
1-
x
+
0.5 dropout
Dense Layer
Output

Σχήμα 1.15: Προτεινόμενη Αρχιτεκτονική - Επιληψία

Πίνακας 1.9: Σύγκριση απόδοσης μεταξύ του προτεινόμενου πολυτροπικού μοντέλου και υπάρχουσων εργασιών (AB - CD - E). Οι καταγεγραμμένες τιμές είναι ο μέσος όρος ± η τυπική απόκλιση. Τα καλύτερα αποτελέσματα εμφανίζονται με έντονους χαρακτήρες.

	Μετρική Αξιολόγησης
Αρχιτεκτονική	**Accuracy**
State-of-the-art approaches	
Novel RF [107]	96.70
EMD, higher order moments, ANN [108]	80.00
BiLSTM [99]	88.00
DWT + Kmeans + MLPNN [98]	95.60
CNN [109]	96.97
Random Forest [110]	87.00
Matrix Determinant and MLP [111]	96.50
EMD and SVM [112]	93.00
dual-tree complex wavelet transform domain [113]	96.28
statistical dual-tree complex wavelet transform domain [114]	83.50
ANN, hierarchical multi-class SVM with new kernel [115]	95.00
Random Forest, wavelets [116]	95.84
CNN [117]	88.67
OPF [118]	89.20
Symlets wavelets, statistical mean energy std and PCA, GBM-GSO, RF, SVM [119]	96.50
Προτεινόμενη Αρχιτεκτονική	
	97.00 ±1.84

Ενότητα 1.4.3. Τα αποτελέσματα των πειραμάτων αναφέρονται στον Πίνακα 1.10.

Πρώτα, εξετάζουμε την αποτελεσματικότητα της πολυτροπικής μεθόδου - GMU. Συγκεκριμένα, αφαιρούμε τη GMU και συνενώνουμε (Concatenation) τις αναπαραστάσεις h^t και h^v. Το παραγόμενο διάνυσμα περνάει σε ένα dropout layer με ποσοστό 0.5, ακολουθούμενο από ένα dense layer (με δύο ή τρεις μονάδες), το οποίο δίνει την τελική πρόβλεψη. Όσον αφορά την Περίπτωση (AB - CD - E), μπορεί κάποιος να παρατηρήσει από τους Πίνακες 1.10 και 1.8 ότι η αφαίρεση της GMU οδηγεί σε μείωση της Ακρίβειας κατά 0.80%.

Στη συνέχεια, εξετάζουμε τα αποτελέσματα του μέρους της αρχιτεκτονικής που αντιστοιχεί στην εικόνα. Αφαιρούμε τόσο το μέρος αναπαράστασης της εικόνας όσο και την πολυτροπική μέθοδο GMU και πειραματιζόμαστε με τον εντοπισμό επιληπτικών κρίσεων χρησιμοποιώντας

Πίνακας 1.10: Ablation Study (AB - CD - E). Reported values are mean ± standard deviation.

Μοντέλο	**Μετρικές Αξιολόγησης**									
	Precision			**Recall**			**F1-score**			**Accuracy**
	AB	CD	E	AB	CD	E	AB	CD	E	
(AB - CD - E)										
Concatenation	95.86	98.09	94.45	99.00	94.50	94.00	97.34	96.15	93.98	96.20
	±4.17	±3.11	±6.09	±2.00	±4.72	±6.63	±2.23	±2.43	±4.34	±2.27
Χρήση μόνο	97.44	92.90	96.51	93.00	95.50	99.00	95.08	94.06	97.61	95.20
του ΗΕΓ ως είσοδο	±3.40	±5.16	±5.67	±5.57	±6.10	±2.99	±3.75	±4.62	±3.14	±3.59
Αφαίρεση	93.33	96.68	97.09	97.50	94.50	90.00	95.11	95.41	92.86	94.80
του αριστερού κλάδου του CNN	±8.31	±4.24	±4.45	±4.03	±5.68	±11.83	±5.05	±3.31	±7.19	±4.12
Αφαίρεση	94.56	97.05	93.65	97.99	94.50	90.00	96.14	95.63	91.31	95.00
του δεξί κλάδου του CNN	±4.93	±2.41	±8.88	±2.45	±5.22	±10.00	±2.32	±2.44	±6.95	±2.41

μόνο τους δύο κλάδους του CNN. Όσον αφορά την Περίπτωση (AB - CD - E), μπορεί κανείς να παρατηρήσει από τους Πίνακες 1.10 και 1.8 μια μείωση της Ακρίβειας κατά 1.80%.

Στη συνέχεια, εξετάζουμε την αποτελεσματικότητα του κλάδου της αρχιτεκτονικής CNN με το μικρό φίλτρο. Για να το κάνουμε αυτό, αφαιρούμε αυτόν τον κλάδο και το σήμα του ΗΕΓ διέρχεται μόνο μέσω του κλάδου με το μεγαλύτερο φίλτρο. Κάποιος μπορεί να παρατηρήσει από τους Πίνακες 1.10 και 1.8 μια μείωση της Ακρίβειας κατά 2.20%.

Τέλος, εξετάζουμε την αποτελεσματικότητα του κλάδου της αρχιτεκτονικής CNN με το μεγάλο φίλτρο. Για να το πετύχουμε, αφαιρούμε αυτόν τον κλάδο και το σήμα ΗΕΓ διέρχεται μόνο μέσω του κλάδου με το μικρό φίλτρο. Παρατηρώντας τους Πίνακες 1.10 και 1.8, βλέπουμε μια μείωση της Ακρίβειας κατά 2.00%.

1.5 Επίλογος και Μελλοντικές Επεκτάσεις

Σε αυτή τη διδακτορική διατριβή ερευνήσαμε τις πιο πρόσφατες μεθόδους μηχανικής μάθησης για (i) την αναγνώριση της κατάθλιψης με χρήση αναρτήσεων στα κοινωνικά μέσα δικτύωσης και τον αυθόρμητο λόγο, (ii) την ανίχνευση ασθενών με άνοια της νόσου Αλτσχάιμερ και την πρόβλεψη των σκορ MMSE μέσω αυθόρμητης ομιλίας, και (iii) την αναγνώριση επιληπτικών ασθενών μέσω σημάτων ΗΕΓ ενός καναλιού.

Παρακάτω, παρουσιάζονται ιδέες για μελλοντική επέκταση:

- **Ερμηνεύσιμα πολυτροπικά μοντέλα βαθιάς μάθησης.** Ο γιατρός πρέπει να ενημερώνεται γιατί ο αλγόριθμος μηχανικής μάθησης έφτασε σε μια συγκεκριμένη απόφαση. Συγκεκριμένα, οι μέθοδοι GRAD-CAM και Integrated Gradients είναι δύο τεχνικές ερμηνευσιμότητας που μπορούν να εφαρμοστούν για την εξήγηση των αποτελεσμάτων οποιουδήποτε αλγορίθμου μηχανικής μάθησης.

- **Έλλειψη ετικετών (labels).** Η συλλογή μεγάλων συνόλων δεδομένων που συνοδεύονται με ετικέτες για την εκπαίδευση των αλγορίθμων τεχνητής νοημοσύνης / μηχανικής μάθησης είναι κρίσιμης σημασίας. Για αυτό το λόγο, σχεδιάζουμε να εφαρμόσουμε προσεγγίσεις αυτο-επιβλεπόμενης μάθησης (self-supervised learning) στο μέλλον για να αντιμετωπίσουμε την ανάγκη απόκτησης μεγάλων συνόλων δεδομένων.

- **Ανίχνευση της Mild Cognitive Impairement κατάστασης.** Στο μέλλον, στοχεύουμε στην εφαρμογή των προτεινόμενων προσεγγίσεων στο σύνολο δεδομένων VAS που προτάθηκε στο [120, 121]. Αυτό το σύνολο δεδομένων περιλαμβάνει ασθενείς με Αλτσχάιμερ, μη-Αλτσχάιμερ και άτομα με ήπια προσβολή της γνώσης (MCI). Η ανίχνευση των ατόμων με MCI αποτελεί πρόκληση και έχει αποδειχθεί ότι είναι κρίσιμης σημασίας. Συγκεκριμένα, η πρόοδος της νόσου μπορεί να καθυστερήσει σημαντικά με την έγκαιρη ανίχνευση των ατόμων σε κατάσταση MCI.

- **Προβλήματα απορρήτου - Ομοσπονδιακή Μάθηση (Federated Learning).** Η επεξεργασία δεδομένων υγείας συνεπάγεται προβλήματα απορρήτου. Για να είμαστε πιο ακριβείς, η πλειονότητα των υπαρχουσών προσεγγίσεων βασίζεται σε κεντρικές ρυθμίσεις, όπου τα δεδομένα συγκεντρώνονται σε έναν κεντρικό εξυπηρετητή. Αντίθετα, η ομοσπονδιακή μάθηση αντιμετωπίζει αυτό το πρόβλημα διανέμοντας τη διαδικασία εκπαίδευσης σε συσκευές των τελικών χρηστών.

- **Επαύξηση Δεδομένων.** Τα παραγωγικά αντιπαλικά δίκτυα (Generative Adversarial Networks) μπορούν επίσης να αξιοποιηθούν για τη δημιουργία σημάτων, δηλαδή ομιλίας, ΗΕΓ, κ.ά. Συγκεκριμένα, τα βαθιά νευρωνικά δίκτυα μπορούν να εκπαιδευτούν με δεδομένα που έχουν δημιουργηθεί τεχνητά, ενώ η απόδοσή τους μπορεί να δοκιμαστεί σε πραγματικά δεδομένα.

- **Εφαρμογή των μεθόδων μας σε άλλες διαταραχές του εγκεφάλου.** Οι προσεγγίσεις που προτείναμε μπορούν να εφαρμοστούν και σε άλλες νόσους. Για παράδειγμα, η έρευνα έχει δείξει ότι η νόσος του Πάρκινσον επηρεάζει την ομιλία, επομένως η νόσος του Πάρκινσον μπορεί να ανιχνευθεί μέσω της ομιλίας και των απομαγνητοφωνημένων κειμένων.

- **Χρήση πολυκαναλικών δεδομένων ΗΕΓ.** Στο μέλλον σκοπεύουμε να χρησιμοποιήσουμε πολυκάναλα σήματα ΗΕΓ [122, 123].

- **Πολυγλωσσικές προσεγγίσεις.** Σχεδιάζουμε να εφαρμόσουμε τις προσεγγίσεις που προτείναμε σε ένα πολυγλωσσικό πλαίσιο. Συγκεκριμένα, στοχεύουμε στην εκπαίδευση των μοντέλων μας σε μία γλώσσα και στην αξιολόγηση της απόδοσής τους σε μια άλλη γλώσσα. Για παράδειγμα, κάποιος μπορεί να εκμεταλλευτεί το σύνολο δεδομένων MADReSS Challenge [124]. Μπορούν να εκπαιδευτούν μοντέλα βασισμένα σε δεδομένα ομιλίας στα αγγλικά και να αξιολογηθεί η απόδοσή τους σε δεδομένα ομιλίας στα ελληνικά.

- **Απόσταξη Γνώσης (Knowledge Distillation).** Για να αντιμετωπίσουμε την ανάγκη δημιουργίας μεγάλων μοντέλων, τα οποία συνεπάγονται προβλήματα υπολογιστικής φύσεως, στοχεύουμε στην εκμετάλλευση προσεγγίσεων Απόσταξης Γνώσης [125, 126]. Με αυτόν τον τρόπο, ένα μεγάλο νευρωνικό δίκτυο συμπιέζεται σε ένα μικρότερο και πιο απλό, χωρίς να μειώνεται η απόδοσή του.

- **Προσαρμογείς (Adapters).** Σε αυτή τη διατριβή, βελτιστοποιήσαμε μερικά προεκπαιδευμένα μοντέλα βασισμένα σε μετασχηματιστές. Ωστόσο, κατά τη βελτιστοποίηση χάνεται κάποια πληροφορία, αφού χρησιμοποιούνται μόνο δεδομένα που είναι συγκεκριμένα στην εκάστοτε εργασία για την ενημέρωση των παραμέτρων των μοντέλων. Αυτό το φαινόμενο είναι γνωστό ως catastrophic forgetting [127]. Επομένως, στο μέλλον, σχεδιάζουμε να χρησιμοποιήσουμε προσαρμογείς [128, 129].

- **Παρακολούθηση της Εξέλιξης των Διαταραχών του Εγκεφάλου με την Πάροδο του Χρόνου.** Επειδή η κατάθλιψη και η άνοια τύπου Αλτσχάιμερ εξελίσσονται με την πάροδο του χρόνου, είναι σημαντικό να διαγνωστούν έγκαιρα. Η παρακολούθηση της πορείας της νόσου κατά μήκος του χρόνου έχει μεγάλη σημασία στις μέρες μας. Για παράδειγμα, ένα από τα tasks στο πλαίσιο του συνόλου δεδομένων ADReSSo είναι η πρόβλεψη εξέλιξης της νόσου, όπου μπορεί κανείς να δημιουργήσει ένα μοντέλο για να προβλέψει τις αλλαγές στην γνωστική κατάσταση με την πάροδο του χρόνου.

Contents

List of Figures

List of Tables

Chapter 2

Introduction

2.1 Brain Disorders - Artificial Intelligence for Social Good

Brain disorders are one of the greatest challenges to health. It is estimated that approximately 165 million people suffer from a brain disorder in Europe, while 1 in 3 people will suffer from a brain disorder at some point in their lives. Some types of brain disorders include Alzheimer's disease, various types of dementia, epilepsy, Parkinson's disease, mental disorders, and others. These disorders affect the way people think, feel, or perform everyday activities. However, if these disorders are diagnosed early and the individual receives appropriate medication, their progression can be delayed. For this reason, timely diagnosis is crucial.

Artificial Intelligence (AI) is transforming the way we address social issues by enhancing the well-being of both individuals and communities. The term "AI for Social Good," also known as "AI for Social Impact," is a new field of research aimed at addressing some of the most significant social, environmental, and public health problems existing today. This doctoral dissertation aims to contribute to this new field by developing modern machine learning methods for improving the recognition of brain disorders, with particular emphasis on three major categories (Depression, Alzheimer's Dementia, and Epilepsy).

Depression involves a large number of symptoms, such as loss of interest, anger, pessimism, changes in weight, feelings of helplessness, suicidal thoughts, and many others. Alzheimer's dementia is characterized by memory loss and affects language and speech. Epileptic seizures involve social stigma.

2.2 Depression

Depression rates have presented a surge due to the covid-19 pandemic[1]. Depression entails a great number of symptoms, including loss of interest, anger, pessimism, changes in weight, feelings of worthlessness, thoughts of suicide, and many more. According to

[1] https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide

the World Health Organization (WHO)[2], around 280 million people in the world have depression. Recent surveys[3] indicate that global rates of depression are rising. China, India, the United States, Russia, Indonesia, Nigeria are some of the countries presenting the highest rates of depression.

People with depression use social media platforms, including X/Twitter and Reddit, and share their thoughts, emotions, feelings, etc. through posts or comments with other users. Therefore, social media constitute a valuable form of information, where linguistic patterns of depressed people can be investigated.

Research has shown that speech constitutes a reliable biomarker for detecting depression [130]. Specifically, people with depression present anomalities in speech, including lower speech rates, less pitch variability and more self-referential speech. Depression affects language also [131]. For instance, depressed people use first-person singular pronouns, negative thinking, and self-focus. Therefore, employing both speech and transcripts in a multimodal setting is a hot research topic nowadays.

2.2.1 Cognitive Tools

Below, we mention one cognitive tool, which is used for recognizing depression.

2.2.1.1 Patient Health Questionnaire-9 (PHQ-9)

The Patient Health Questionnaire-9[4] is a multipurpose instrument for screening, diagnosing, monitoring and measuring the severity of depression. It consists of 9 questions. Each participant is asked to answer to a variety of questions pertinent to sleep problems, tiredness, little energy, limited concentration, poor appetite or overeating, and more. A score is computed based on the answers given by the participant. A score lower than 5 indicates minimal depression, a score lower than 10 denotes mild depression, a score lower than 15 denotes moderate depression, a score lower than 20 means moderately severe depression, and a score ranging from 20 to 27 denotes severe depression. Findings of the study introduced in [132] state that PHQ-9 is useful in clinics specializing in psychiatry.

2.3 Dementia

Alzheimer's disease is the most common form of dementia and may contribute to 60-70% of cases. According to the WHO, approximately 55 million people suffer from dementia nowadays, while this number is going to present a surge in the upcoming years reaching up to 78 million and 139 million people in 2030 and 2050 respectively [51]. Due to the fact that Alzheimer's disease is a neurodegenerative disease, meaning that the symptoms become worse over time, the early diagnosis seems to be imperative for promoting

[2] https://www.who.int/news-room/fact-sheets/detail/depression

[3] https://pulsetms.com/resources/around-world/

[4] https://www.hiv.uw.edu/page/mental-health-screening/phq-9

early and optimal management. In addition, dementia is inextricably linked with difficulties in speech, since dementia affects how a person can use language and communicate [133, 134, 135]. For this reason, current research works have moved their interest towards Alzheimer's dementia (AD) identification from spontaneous speech, in order to save money and time.

2.3.1 Stages of Dementia

As mentioned above, dementia is progressive meaning that the symptoms become worse with time, usually over several years. There are three stages of dementia, which are clinically identified, namely the early stage, the middle stage, and the late stage [136]. These stages are described in detail below.

Early Stage. This stage is also known as mild stage, since the symptoms are relatively mild and not always easy to notice. The early stage of dementia has a duration of approximately two years, where the person suffers from daily problems, including memory problems, difficulty in planning or making complex decisions, difficulties in language and communication, changes in mood or emotion, visual-perceptual difficulties, and poor orientation [137].

Middle Stage. This stage of dementia is also known as moderate stage, since the symptoms become more noticeable [138]. Memory and thinking skills, communication abilities, problem with the orientation, and symptoms of apathy, depression, and anxiety will get worse in this stage. At the same time, patients may suffer from delusions [139] and hallucinations. In terms of the behavioural changes during this stage, AD patients experience symptoms, including screaming or shouting, disturbed sleep patterns, repetitive behaviour, losing inhibitions, and many more.

Later Stage. This stage of dementia is also known as severe stage, since the person will need full-time care and support with daily living and personal care [140]. In this stage, the language difficulties will become severe, where the person's spoken language may eventually be reduced to only a few words or lost altogether. In addition, people with dementia often think they are at an earlier period of their life, widely known as time shifting.

2.3.2 Cognitive Tools

Each cognitive tool is used to screen for dementia. The participant performs several tasks and based on the answers to each task, a score is calculated at the end of the test. Then, the examiner compares this score with the cut-off values recommended by each test and decides the degree of cognitive impairement of the person. Below we describe some cognitive tools.

2.3.2.1 Mini Mental State Examination

The Mini-Mental State Examination score is a 30-point questionnaire, which was proposed by Folstein et al. [141]. It is used to screen for dementia. Administration of the test takes between 5 and 10 minutes. The maximum score for the MMSE is 30 points. According to [62], there are four groups of cognitive severity: healthy (MMSE score ≥ 25), mild dementia (MMSE score of 21–24), moderate dementia (MMSE score of 10–20), and severe dementia (MMSE score ≤ 9). However, the MMSE entails some drawbacks, including its sensitivity to progressive changes occuring with severe Alzheimer's disease and its inability to distinguish patients with mild Alzheimer's disease from healthy patients. In addition, according to [142] the MMSE should not be used clinically unless the person has at least a grade-eight education and is fluent in English.

2.3.2.2 Montreal Cognitive Assessment

The Montreal Cognitive Assessment (MoCA) is used for predicting dementia in people with mild cognitive impairment [143]. The MoCA checks different types of cognitive or thinking abilities, including orientation, short-term memory, language, abstraction, animal naming, attention, and many more. Similar to MMSE, the scores on the MoCA range from 0 to 30, where a score of 26 and higher is considered normal. Also, compared to MMSE, MoCA is better at detecting mild disease. However, the most appropriate cut-off point is not clearly agreed [144].

2.3.2.3 Addenbrooke's Cognitive Examination

The Addenbrooke's Cognitive Examination (ACE) [145] and its subsequent versions (Addenbrooke's Cognitive Examination-Revised, ACE-R [146] and Addenbrooke's Cognitive Examination III, ACE-III [147]) are neuropsychological tests used to identify cognitive impairment in conditions such as dementia. This test was developed for improving the screening performance of the MMSE. It is scored out of 100, with a higher score denoting better cognitive function and with the recommended cut-off scores accounting for 88 and 83. Regarding the current version of the test, i.e., ACE-R, it consists of 19 activities which test five cognitive domains: attention, memory, fluency, language and visuospatial processing.

2.3.2.4 Boston Naming Test

The Boston Naming Test (BNT) was introduced by [148] and is a widely used neuropsychological assessment tool to measure confrontational word retrieval in individuals with aphasia or other language disturbance caused by stroke, Alzheimer's disease, or other dementing disorder. This test comprises 60 pictures which are presented to the patient one at a time and the patient is asked to name each picture. In case of an error response, there are two types of cues, namely the stimulus cue and the phonemic cue. A stimulus cue is presented when the subject clearly misperceives the picture or indicates a lack of recognition of the picture. A phonemic cue is presented after each error response, including following a stimulus cue. The BNT is recommended as a supplement to the Boston Diagnostic Aphasia Examination [149].

2.3.2.5 Wechsler Adult Intelligence Scale

The Wechsler Adult Intelligence Scale (WAIS) was introduced by [150] and is the most common intelligence quotient (IQ) test, measuring intelligence and cognitive abilities in adults. The current version of the test is the WAIS-IV, which is composed of 10 core subtests and five supplemental subtests. There are four index scores representing major components of intelligence, namely the Verbal Comprehension Index (VCI), Perceptual Reasoning Index (PRI), Working Memory Index (WMI), and Processing Speed Index (PSI).

2.3.2.6 Wechsler Memory Scale

The Wechsler Memory Scale (WMS) is a neuropsychological test developed to measure different memory functions [151]. Anyone ages 16 to 90 is eligible to take this test. The current version is the fourth edition (WMS-IV) and was designed to be used with the WAIS-IV. A person's performance is reported as five Index Scores: Auditory Memory, Visual Memory, Visual Working Memory, Immediate Memory, and Delayed Memory.

2.3.2.7 Alzheimer's Disease Assessment Scale (ADAS)

The Alzheimer's Disease Assessment Scale (ADAS) was developed to evaluate cognitive and behavioral dysfunctions characteristic of Alzheimer's disease [152]. It consists of cognitive (ADAS-Cog [153]) and noncognitive (ADAS-Noncog). The ADAS-Cog consists of 11 parts and takes approximately 30 minutes to administer. The original version of ADAS-Cog consists of 11 items, including Word Recall task, naming omjects and fingers, following commands, orientation, spoken language, etc. Scores of the ADAS-Cog range from 0 to 70, where a score of 70 represents the most severe impairment and 0 represents the least impairment. The greater the dysfunction, the greater the score.

2.3.2.8 General Practitioner assessment of Cognition (GPCOG)

The GPCOG is a screening instrument rather than a diagnostic test [154, 155]. The participant has to perform some tasks, including: remember a name and address and recall it in a few minutes, state today's date, make a clock drawing with all of the numbers drawn correctly on the face of the clock, describe something specific that has happened in the news in the last week, etc. Scores of the GPCOG range from 0 to 9, where a score of 9 indicates no significant impairement. A score between 5 and 8 indicates that informant interview must be conducted, while a score less than 4 indicates cognitive impairement.

During the informant interview, the test administrator asks a caregiver or family member if the patient has more difficulty than they used to five to ten years ago with some tasks. This test is free, brief, and the education has little to no effect on the accuracy.

2.4 Epilepsy

Epilepsy is a neurological disease, which affects people of all ages. According to the World Health Organization (WHO), approximately 50 million people have epilepsy worldwide, rendering it one of the most common neurological diseases [96]. Epilepsy has a negative impact in peoples' everyday life mainly due to the discrimination and stigma surrounding the disease itself. However, the WHO states that up to 70% of people living with epilepsy could live seizure-free, if people are diagnosed early and receive the proper treatment. Therefore, the early diagnosis of the epilepsy is important for providing a better quality of life to epileptic patients. Electroencephalogram (EEG) is used by neurologists for diagnosing epilepsy. However, manually reviewing and analyzing EEG signals by neurologists is a task requiring significant amount of time, while it is prone to errors as well. Thus, the need for an automatic system is crucial.

2.5 Motivation and Research Questions

2.5.1 Motivation

Depression. Existing research initiatives exploit social media data for identifying depressive posts. The majority of these research works [1, 2] employ feature extraction approaches and train shallow Machine Learning (ML) algorithms. Employing feature extraction approaches constitutes a tedious procedure and demands domain expertise, since the authors may not find the optimal feature set for the specific problem. At the same time, the train of shallow ML algorithms does not yield optimal performance and does not generalize well to new data. For addressing these limitations, other approaches [3] use deep neural networks, including Convolutional Neural Networks (CNNs), bidirectional long short-term memory (BiLSTM), and so on, or transformer-based networks. In addition, there are researches employing ensemble strategies [4]. However, these approaches increase substantially the training time, since multiple models must be trained separately. In addition, recently there have been studies [5, 6] showing that transformer-based models struggle or fail to capture rich knowledge. For this reason, there have been proposed methods for enhancing these models with external information or additional modalities [7, 8, 9, 10]. However, existing research initiatives in the task of depression detection through social media have not exploited any of these approaches yet. In addition, the reliability of a machine learning model's confidence in its predictions, denoted as calibration [11, 12], is critical for high risk applications, such as deciding whether to trust a medical diagnosis prediction [156, 157, 158]. Although methods regarding the confidence of models' predictions have been introduced in many studies, including suicide risk assessment [159], sleep stage classification [160], and so on, no prior work for depression detection has taken into account the level of confidence of models' predictions, creating in this way overconfident models.

Existing research works use spontaneous speech and rely on the extraction of handcrafted features and the train of traditional machine learning classifiers or deep learning approaches [32, 33, 34]. However, extracting features is a timely procedure requiring expertise on the specific topic. Additionally, the majority of research studies uses unimodal approaches for predicting depression using mainly speech [35]. Although there are studies employing multimodal models, these studies employ early [36, 37], intermediate [38, 39], or late fusion [40, 41] strategies. In the early fusion strategy, representation vectors of the modalities are concatenated at the input level, while in the intermediate fusion, the representation vectors are concatenated during training, thus equal importance is assigned to the modalities. In the late fusion strategy, unimodal models are trained independently and decision voting is applied, i.e., majority voting. The inter-modal interactions cannot be captured through these approaches. In addition, the majority of research works have tested their approaches only in English language, thus the acoustic and phonetic content of data might differ in other languages. Finally, to the best of our knowledge, no study has experimented with predicting depression, age, education level, and gender at the same time.

Alzheimer's Dementia. Several research works have been conducted with regard to the identification of AD patients using speech and transcripts. The majority of them have employed feature extraction techniques [161, 162, 163, 164, 165], in order to train traditional Machine Learning (ML) algorithms, such as Logistic Regression, k-NN, Random Forest, etc. However, feature extraction constitutes a time-consuming procedure achieving poor classification results and often demands some level of domain expertise. Recently, researchers introduce deep learning architectures [166, 167], such as CNNs and BiLSTMs, so as to improve the classification results. Despite the success of transformer-based models in several domains, their potential has not been investigated to a high degree in the task of dementia identification from transcripts, where research works [61] having proposed them, use their outputs as features to train shallow machine learning algorithms. Concurrently, all research works except one [91], train machine learning models, in order to distinguish AD patients from non-AD patients, without taking into account the severity of dementia via Mini-Mental State Exam (MMSE) scores. At the same time, to the best of our knowledge, the research works that have proposed deep learning models based on transformer networks have focused their interest only on improving the classification results obtained by CNNs, BiLSTMs etc. instead of exploring possible explainability techniques. Specifically, due to the fact that deep learning models are considered black boxes, it is important to propose ways of making them interpretable, since it is imperative for a clinician to be informed why the specific deep neural network classified a person as AD patient or not. To the best of our knowledge, only one work [168] has experimented with interpreting its proposed deep learning model (CNN-LSTM model) in the field of dementia detection using transcripts. In terms of the proposed multimodal approaches, the majority of them have introduced label fusion and majority-voting or average approaches [169, 170, 61]. Specifically, regarding the AD classification task they train several textual and acoustic models and they make the final prediction of the given transcript based on the class, which received the most votes by the individual models. With regards to the MMSE regression task, they simply average the predictions of the individual models. Concurrently, they extract a large number of features corresponding to the textual and acoustic modalities and some of them train traditional machine learning algorithms, such as Logistic Regression, XGBoost, etc. Thus, it is evident that these approaches are not time-efficient, since a lot of models must be trained and tested separately. At the same time, these approaches do not exploit the interaction between the two modalities. Moreover, research initiatives introducing multimodal models use the add and concatenation operation treating in this way equally the two modalities [91]. Another limitation of this approach has to do with the fact that one modality may override the other one with a negative impact on the classification performance. In terms of the textual modality recent studies have shown that Self-Attention layers treat the input sequence as a bag-of-word tokens and each token individually performs attention over the bag-of-word tokens. Consequently, the contextual information is not taken into account in the calculation of dependencies between elements. In addition, the reliability of a machine learning model's confidence in its predictions, denoted as calibration [11, 12], is critical for high risk applications, such as deciding whether to trust a medical diagnosis prediction [156, 157, 158]. However, no prior work has taken into account the calibration of the models, creating in this way overconfident models. According to [171], modern neural networks are not well-calibrated, while they are overconfident at the same time.

Epilepsy. There have been a number of studies proposing methods for detecting epileptic seizures. The majority of these studies first extract both time-domain and frequency domain features from the electroencephalogram (EEG) signals. For instance, the authors apply the Discrete Wavelet Transform (DWT) [97, 98] for decomposing the EEG signals into sub-bands and then extract features from each sub-band. After having extracted a large number of features, the authors usually exploit feature selection or dimensionality reduction techniques for finding the best subset of features or reducing the dimension of the feature vector respectively. The last step of the proposed methods includes the train of traditional machine learning classifiers, i.e., Logistic Regression (LR), Support Vector Machines (SVMs), Random Forests (RF), Decision Trees, etc. However, these methods are time-consuming, since they demand some level of domain expertise for extracting the best representative features. Only a few number of studies [99, 100, 101, 102] have exploited deep neural networks, i.e., CNNs, LSTMs, or BiLSTMs in the task of epilepsy detection and prediction. However, most of these methods still rely on handcrafted features [100, 101, 99]. Another limitation is the fact that existing research works split the EEG signals into segments and propose majority-voting approaches [102]. Thus, they have to train multiple models separately increasing substantially the computation time. At the same time, most of the CNN models are not able to model effectively the temporal dependencies among the EEG data. Although LSTMs and BiLSTMs can capture the temporal dependencies in EEG data, they usually have high model complexity.

2.5.2 Research Questions

Considering the aforementioned limitations and the adaptation of deep learning models in the Natural Language Processing (NLP), Speech Processing, Signal Processing, and Computer Vision (CV) domains, this thesis seeks to answer seven main research questions:

- **RQ1:** Do transformer-based networks, i.e., BERT, ALBERT, etc. achieve better performance than traditional techniques, i.e., LSTMs, CNNs, etc.?
- **RQ2:** Can we provide explanations, which will show how our models reach their decisions? Especially in health-related tasks, it is very important for a clinician to be informed why the deep neural network classified a person as an AD patient or a non-AD one. At the same time, according to the European Union General Data Protection Regulation (GDPR) [172] each person has the right to the explanation. Also, can we propose interpretable models, which will achieve comparable performance to existing research initiatives?
- **RQ3:** Can we propose multi-task learning models, consisting of primary and auxiliary tasks, to explore if the axiliary tasks help the primary one in improving its performance?
- **RQ4:** How can we combine the representation vectors of the different modalities (multimodal approaches) effectively?
- **RQ5:** Instead of creating fixed deep neural networks, can we create automatically architectures which will perform best for our specific task?
- **RQ6:** How can we improve self-attention networks through capturing the richness of context?
- **RQ7:** How can we prevent deep learning models from becoming too overconfident?

2.6 Thesis Contributions

Based on the research context and clinical need as detailed above, the overall, high-level aim of this Ph.D. thesis is to improve the detection and monitoring of brain disorders by exploiting advanced machine learning techniques. Specifically, this thesis presents automatic systems for recognizing three major brain disorders, including depression, Alzheimer's dementia, and epilepsy.

In terms of depression, we present to approaches. In terms of the first approach, we utilize social media data and create tools based on natural language processing to detect depressive posts. Additionally, this thesis seeks to find differences in language between depressed people and non-depressed ones through a detailed linguistic analysis. In terms of the second approach, we utilize spontaneous speech and create tools based on both natural language processing and speech processing to detect depression.

Regarding Alzheimer's dementia, motivated by the fact that people with Alzheimer's dementia present deficits in language and speech, this thesis utilizes recordings of spontaneous speech and creates automatic systems based on natural language processing and speech processing. Specifically, we fine-tune transformer-based models, exploit explainability techniques and linguistic analyses and explore some linguistic features which are useful for detecting cognitive decline. This thesis seeks also to use multimodal models exploiting both speech and transcripts instead of just focusing on lexical, acoustic, or visual features alone.

With regards with Epilepsy, motivated by the fact that the manual review of EEG signals by neurologists is a laborious task, we present a new automatic system based on a multimodal method for diagnosing epilepsy.

Overall, the main contributions of this thesis are the following:

- **Introducing deep neural networks, which can be trained in an end-to-end trainable manner eliminating the timely procedure of feature extraction.** Contrary to prior research works extracting a large number of features, exploiting feature selection or dimensionality reduction techniques, and training shallow machine learning algorithms, this thesis aims to eliminate the need of feature extraction by proposing deep neural networks and transformer-based models.

- **An explainable approach and a linguistic analysis study is proposed.** Contrary to prior works, which simply train ML algorithms for detecting AD patients, this thesis extends prior work by employing an explainable approach and introducing a linguistic analysis. Both approaches reveal the linguistic patterns used by AD patients, i.e., pos-tags. Differences in language between AD patients and non-AD ones are also revealed. We use the same linguistic analysis on a depression dataset and reveal differences in language between depressed people and non-depressed ones.

- **Introducing multi-task learning models.** This thesis proposes multi-task learning architectures for identifying depression and Alzheimer's dementia. Firstly, this thesis presents a multi-task learning approach for jointly modelling the depression, education level, age, and gender identification tasks. Secondly, this thesis introduces multi-task learning architectures aiming to predict the AD detection and MMSE recognition tasks.

- **Multimodal Fusion methods are introduced for capturing the inter- and intra-modal interactions.** Contrary to existing research initiatives, which exploit early, intermediate, and late fusion strategies, this thesis introduces new methods for fusing the different modalities. These methods aim to capture the inter- and intra-modal interactions, while increasing the performance at the same time. Therefore, this thesis extends prior work by exploiting fusion methods, including Gated Multimodal Unit, Cross-Modal Attention Layer, Cross-Attention Layer incorporating a gating model, Cross - Attention Scaling Layer, Multimodal Shifting Gate, Optimal Transport Domain Adaptation, and many more. These multimodal approaches are adopted in a series of experiments, i.e., depression detection through social media posts and spontaneous speech, dementia identification through speech and transcripts, epilepsy detection via single-channel EEG signals, and aim to increase the performance achieved by the unimodal ones.

- **Incorporating a Neural Architecture Search approach into a deep neural network.** In contrast with prior work, which exploit fixed architectures, this thesis incorporates a NAS approach, called DARTS, into a deep neural network for generating automatically a CNN architecture. This CNN architecture fits best for this specific task.

- **Enhancing self-attention networks with contextual information.** This thesis aims to enhance the self-attention layer by adding contextual information. Specifically, this thesis presents three strategies for constructing a contextual vector into an end-to-end trainable deep neural network. This approach is conducted on datasets related to the Alzheimer's dementia task.

- **Presenting methods for calibrating deep neural networks.** Prior works evaluate deep neural networks based only on the performance by reporting accuracy, precision, recall, and more metrics. This thesis extends prior work by exploiting methods for calibrating the introduced models and evaluating these models by exploiting both performance and calibration metrics. These approaches are conducted on datasets related to depression and alzheimer's dementia.

2.7 Thesis Outline

The rest of this thesis is organized as follows:

Chapter 3: Literature Review. Review of the literature for studies that proposed systems *(i)* for identifying depressive posts in social media and recognizing depression via spontaneous speech, *(ii)* either for classifying people into AD patients and non-AD ones or for predicting the Mini-Mental State Exam scores. Specifically, the approaches have been divided into unimodal, i.e., approaches which use either only speech or transcripts, and multimodal, i.e., approaches which exploit both speech and transcripts, and *(iii)* for detecting epileptic patients using EEG recordings. Also, we provide a list of datasets for each case.

Chapter 4: Methods for Recognizing Depression through Social Media posts and Spontaneous Speech. In this chapter, we present two methods aiming to recognize depression. The first approach injects linguistic information into transformer-based models for identifying depressive posts in social media. The proposed approach is evaluated based on both the performance and the calibration. The second approach utilizes both speech and automatic transcripts into a multimodal deep neural network.

Chapter 5: Explainable Identification of Dementia from Transcripts using Transformer Networks. In this chapter, we propose models in both single-task and multi-task learning settings by utilizing only transcripts for detecting AD patients. Also, we perform a detailed linguistic analysis and explainability techniques, which shed light on the main differences in language between AD patients and Healthy Control group.

Chapter 6: Detecting Dementia from Speech and Transcripts Using Transformers. In this chapter, we introduce two methods for fusing the two modalities (speech and transcripts).

Chapter 7: Multimodal Deep Learning Models for Detecting Dementia and Predicting Mini-Mental State Examination scores from Speech and Transcripts. This chapter presents three deep neural networks, which exploit both speech and transcripts. The proposed models are trained both for the AD Classification task and the MMSE Regression task.

Chapter 8: Context-Aware Attention Layers coupled with Optimal Transport Domain Adaptation and Multimodal Fusion methods for recognizing dementia. In this chapter, we present approaches for enhancing the self-attention mechanisms with contextual information, calibrating the proposed models, and fusing the different modalities. Both manual and automatic transcripts are exploited.

Chapter 9: Neural Architecture Search with Multimodal Fusion Methods for Recognizing Dementia. In this chapter, we present a deep neural network, which incorporates a Neural Architecture Search method for generating automatically a CNN architecture and multimodal fusion methods.

Chapter 10: Multimodal Detection of Epilepsy with Deep Neural Networks. This chapter introduces a multimodal deep neural network for detecting epilepsy through single-channel EEG signals.

Chapter 11: Conclusions and Future Work. This chapter concludes the work proposed in this thesis, presents some limitations of this thesis, and provides some suggestions for future research directions.

2.8 Supporting Publications

All the materials presented in this thesis are built on the publications considered by various international conferences and journals, as follows:

- **L. Ilias**, S. Mouzakitis and D. Askounis, “Calibration of Transformer-Based Models for Identifying Stress and Depression in Social Media,” in *IEEE Transactions on*

Computational Social Systems, vol. 11, no. 2, pp. 1979-1990, April 2024 [173] **(Chapter 4)**.

- **L. Ilias** and D. Askounis, "A Cross-Attention Layer coupled with Multimodal Fusion Methods for Recognizing Depression from Spontaneous Speech," *Proc. Interspeech 2024*, 2024, pp. 912-916 [174] **(Chapter 4)**.
- **L. Ilias** and D. Askounis, "Explainable Identification of Dementia From Transcripts Using Transformer Networks," in *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 4153-4164, Aug. 2022 [175] **(Chapter 5)**.
- **L. Ilias**, D. Askounis, and J. Psarras,"Detecting dementia from speech and transcripts using transformers," *Computer Speech & Language*, vol. 79, p. 101485, 2023 [176] **(Chapter 6)**.
- **L. Ilias** and D. Askounis, "Multimodal deep learning models for detecting dementia from speech and transcripts," *Frontiers in Aging Neuroscience*, vol. 14, 2022 [177] **(Chapter 7)**.
- **L. Ilias** and D. Askounis, "Context-aware attention layers coupled with optimal transport domain adaptation and multimodal fusion methods for recognizing dementia from spontaneous speech," *Knowledge-Based Systems*, vol. 277, p. 110834, 2023 [178] **(Chapter 8)**.
- M. Chatzianastasis*, **L. Ilias***, D. Askounis and M. Vazirgiannis, "Neural Architecture Search with Multimodal Fusion Methods for Diagnosing Dementia," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1-5 [179] **(Chapter 9)**.
  * *The first two authors contributed equally.*
- **L. Ilias**, D. Askounis and J. Psarras, "Multimodal detection of epilepsy with deep neural networks," *Expert Systems With Applications*, vol. 213, p. 119010, 2023 [180] **(Chapter 10)**.

Chapter 3

Literature Review

3.1 State-of-the-art analysis of Machine Learning Methods used for the recognition of Depression in Social Media

Some studies have focused on the extraction of features and then the train of shallow machine learning classifiers. For instance, Tadesse et al. [1] extracted n-grams via the tf-idf approach, LIWC features, and LDA topics. Then, they trained LR, SVM, Random Forest (RF), AdaBoost, and Multilayer Perceptron (MLP). Results showed that the bigram features trained on an SVM classifier achieved 80.00% accuracy, while the best accuracy accounting for 91.00% was achieved by exploiting the MLP classifier with all the features, i.e., LIWC, LDA, and bigrams. Liu and Shi [181] extracted a set of textual features, namely part-of-speech, emotional words, personal pronouns, polarity, and so on, and a set of features indicating the posting behaviour of the user, i.e., posting habits and time. Next, feature selection techniques were applied, including recursive elimination, mutual information, extreme random tree. Finally, naive bayes, k-nearest neighbor, regularized logistic regression, and support vector machine were used as base learners, and a simple logistic regression algorithm was used as a combination strategy to build a stacking model. Nguyen et al. [182] extracted a set of features, including LDA topics, LIWC features, affective features by using the affective norms for english words (ANEW) lexicon, and mood labels. The authors trained a LASSO regression classifier for detecting depressive posts and analyzing the importance of each feature. The authors applied also statistical tests and found significant differences between depressive and non-depressive posts. Tsugawa et al. [183] extracted features and trained an SVM classifier to detect depression in Twitter. Specifically, the authors extracted the frequency of words used in tweets, ratio of tweet topics found by LDA, ratio of positive and negative words, and many more. Pirina and Çöltekin [13] collected several corpora and trained an SVM classifier using character and word n-grams. Doc2vec and tf-idf features were extracted and given as input to AdaBoost, LR, RF, and SVM for identifying the severity of depression.

Recently, deep learning approaches have introduced, since they obtain better performance than the traditional ML algorithms and do not often require the tedious procedure of feature extraction. For example, Wani et al. [184] represented words as word2vec and tf-idf approach and trained a deep neural network consisting of CNNs and LSTMs. Kim et al. [185] collected a dataset consisting of posts written by people, who suffer from mental disorders, including depression, anxiety, bipolar, borderline personality disorder, schizophrenia, and autism. This study developed six binary classification models for detecting mental disorders, i.e., depression vs. non-depression, and so on. Specifically, the authors utilized the tf-idf approach and trained an XGBoost classifier. Next, the authors used the word2vec and trained a CNN model. Naseem et al. [17] reformulated depression identification as an ordinal classification problem, where they used four depression severity levels. The authors introduced a deep neural network consisting of Text Graph Convolutional Network, BiLSTM, and Attention layer. A similar approach was proposed by Ghosh and Anwar [186], where the authors extracted features and trained LSTMs for estimating the depression intensity levels. A hybrid deep neural network consisting of CNN and BiLSTM was introduced by Kour and Gupta [187]. Zogan et al. [188] introduced the first dataset including posts from users with and without depression during COVID-19 and presented a new hierarchical convolutional neural network. An emotion-based attention network model was proposed by Ren et al. [189], where the authors extracted the positive and negative words and passed through two separate BiLSTM layers followed by Attention layers.

Ensemble strategies have also been explored in the literature. This means that multiple models are trained separately and the final decision is taken usually by a majority voting approach. For instance, an ensemble strategy was introduced by Ansari et al. [4]. Firstly, the authors exploited some sentiment lexicons, including AFINN, NRC, SenticNet, and multi-perspective question answering (MPQA), extracted features, and applied Principal Component Analysis for reducing the dimensionality of the feature set. A Logistic Regression classifier was trained using the respective feature set. Next, the authors trained an LSTM neural network coupled with an attention mechanism. Finally, the authors combined the predictions of these two approaches via an ensemble method. Also, an ensemble approach was proposed by Trotzek et al. [190]. Firstly, the authors trained a Logistic Regression classifier using as input user-level linguistic metadata. Specifically, the authors extracted LIWC features, length of the text, four readability scores, and so on. Next, the authors trained a CNN model. Finally, the authors combined the outputs of these approaches via a late fusion strategy, i.e., by averaging the predictions of the classifiers. Figuerêdo et al. [3] designed a CNN along with early and late fusion strategies. Specifically, the authors exploited fastText and GloVe embeddings. In the early fusion approach, multiple word embeddings were concatenated and passed to the CNN model. In the late fusion strategy, a majority-vote approach was performed based on the predictions of multiple CNN models. The CNN model comprised a simple convolution layer, max-pooling, fully connected layers, and Concatenated Rectified Linear Units as the activation function.

Explainable approaches have also been introduced. Souza et al. [191] introduced a stacking ensemble neural network, which addresses a multilabel classification task. Specifically, the proposed architecture consists of two levels. In the first level, binary base classifiers were trained with two distinct roles, i.e, expert and differentiating. The expert base classifiers were used for differentiating between users belonging to the control group and those diagnosed with anxiety, depression, or comorbidity. The differentiating base models aimed at distinguishing between two target conditions, e.g., anxiety vs. depression. In the second level, a meta-classifier uses the base models' outputs to learn a mapping function that manages the multi-label problem of assigning control or diagnosed labels. The authors used LSTMs and CNNs. Finally, this study explored Shapley additive explanations (SHAP) metrics for identifying the influential classification features. Zogan et al. [192] proposed also an explainable approach, where textual, behavioural, temporal, and semantic aspect features from social media were exploited. An hierarchical attention network was used in terms of explainable purposes. An hierarchical attention network was also used by Uban et al. [193], where the authors extracted a feature set consisting of content, style, LIWC, and emotions/sentiment features. An interpretable approach was proposed by Song et al. [194], where the authors introduced the Feature Attention Network. The Feature Attention Network consists of four feature networks, each of which analyzes posts based on an established theory related to depression and a post-level attention on top of the networks. However, this method did not attain satisfactory results.

Recently, transformer-based models have been applied in the task of depression detection in social media. Specifically, Boinepelli et al. [195] introduced a method for finding the subset of posts that would be a good representation of all the posts made by the user. Firstly, they employed BERT and computed the embeddings for all posts made by the user. Next, they used a clustering and ranking algorithm. After finding the representative posts per user, the authors added domain specific elements by exploiting RoBERTa. Finally, the authors experimented with two ways for diagnosing depression, i.e., by either employing a majority-vote approach or training a hierarchical attention network. Anantharaman et al. [196] fine-tuned a BERT model for classifying the signs of depression into three labels namely "not depressed", "moderately depressed", and "severely depressed". Similarly, Nilsson and Kovács [197] exploited a BERT model and used abstractive summarization techniques for data augmentation. Zogan et al. [198] presented an abstractive-extractive automatic text summarization model based on BERT, k-Means clustering, and bidirectional auto-regressive transformers (BART). Then, they proposed a deep learning framework, which combines user behaviour and user post history or user activity.

Multimodal approaches combining both text and images have also been proposed. For instance, a multimodal approach was introduced by Ghosh et al. [199] for detecting depression in Twitter. Specifically, the authors utilized the user's description and profile image. The authors used the IBM Watson NaturalLanguageUnderstanding tool and extracted sentiment and emotion information for all user descriptions along with the possible categories (at most 3) that the description may belong to. Next, the authors designed a neural network consisting of BiGRU, Attention layers, Convolution layers, and dense layers. The authors used GloVe embeddings. The proposed architecture can predict whether the user suffers from depression or not as well as predict the sadness, joy, fear, disgust, and anger score. Li et al. [200] exploited text, pictures, and auxiliary information (post time, dictionary, social information) and used attention mechanism within and between the modalities at the same time. The authors exploited TextCNN, ResNet-18, and fully connected layers for extracting representation vectors of text, images, and auxiliary information respectively. A multimodal approach was proposed by Cheng and Chen [201], where the authors exploited texts, images, posting time, and the time interval between the posts in Instagram. Shen et al. [202] collected multimodal datasets and extracted six depression-oriented feature groups, namely social network, user profile, visual, emotional, topic-level, and domain-specific features. Gui et al. [203] combined texts and images and proposed a new cooperative multi-agent reinforcement learning method.

Multitask approaches have been introduced. A multitask approach was introduced by Zhou et al. [204]. Specifically, the authors proposed a hierarchical attention network consisting of BiGRU layers and integrated LDA topics. The main task was the identification of depression, i.e., binary classification task, while the auxiliary task was the prediction of the domain category of the post, i.e., multiclass classification task. Both multitask and multimodal approaches were introduced by Wang et al. [205]. The authors extracted a total of ten features from text, social behaviour, and pictures. XLNet, BiGRU coupled with Attention layers, and Dense layers were used. The authors in [206] presented two approaches based on a multi-task learning framework. Depression detection corresponded to the primary task, while stress detection corresponded to the auxiliary task. Experiments showed that the proposed approach improved single-task learning and transfer learning strategies.

3.1.1 Literature Review Findings

Existing research initiatives rely on the feature extraction process and the train of shallow machine classifiers targeting at diagnosing mental disorders in social media. This fact demands domain expertise and does not generalize well to new data. Other existing approaches train CNNs, BiLSTMs, or employ hybrid models and ensemble strategies. Recently, transformer-based models have been used also. Only few works have experimented with injecting linguistic, including emotion, features into deep neural networks. These approaches employ multi-task learning models, fine-tuning, or multimodal approaches. All these approaches employing transformer-based models usually fine-tune these models. None of these approaches have used modifications of BERT aiming to enhance its performance by injecting into it external knowledge. Also, no prior work has taken into account model calibration creating in this way overconfident models.

3.1.2 Datasets

3.1.2.1 Depression_Mixed

This dataset [13] consists of 1482 non-depressive posts and 1340 depressive posts. These posts have been written by users in Reddit and English depression forums.

3.1.2.2 Depression_Severity

This dataset includes posts in Reddit [17] and assigns each post to a severity level, i.e., minimal (2587 posts), mild (290 posts), moderate (394 posts), and severe form of depression (282 posts).

3.2 State-of-the-art analysis of Machine Learning Methods used for the recognition of Depression through Spontaneous Speech

3.2.1 Early Fusion

The study in [37] constructs a graph based on question-answering pairs. Specifically, a Graph Attention Network is trained. In terms of the multimodal fusion, the authors employ an early fusion approach. A multitask learning framework is adopted, which predicts the level of depression severity (regression) and classifies the subject as depressive or non-depressive. A similar approach is introduced by [36], where the authors employ an early fusion approach and concatenate the representation vectors of audio, visual, and textual modalities. A multi-task learning framework is trained for classifying the level of disorder and predicting the disorder score.

3.2.2 Intermediate Fusion

The study in [207] converts speech signals into spectrogram and uses a VGG16 pretrained model followed by Gated Convolutional Neural Networks and one LSTM layer. The authors pass the BERT embeddings into CNN layers followed by LSTM layers. The representation vectors of the two modalities are concatenated for predicting the Patient Health Questionnaire (PHQ) score. In [32], the authors use articulatory coordination features (ACFs) derived from vocal tract variables. A staircase regression approach is used, where an ensemble of models is trained on multiple partitions of the same training data set. A hierarchical attention network (HAN) is used for extracting textual representation. Additional features representing the prosodic information are extracted. The abovementioned feature representations are concatenated for estimating the depression severity score. In [38], speech signals are represented as log-Mel spectrograms and fed into temporal CNNs, while text is passed through the encoder part of the transformer. Representation vectors of these two modalities are concatenated for predicting whether the individual has depression or not. Ref. [39] adopts a similar approach. DeepSpectrum features are obtained from speech signals and fed into Temporal Convolutional Networks (TCNs) followed by Attention and Dense Layers. The authors feed the word2vec embeddings into a transformer encoder. Finally, the audio and textual vectors are concatenated into a single vector. Ref. [208] proposes a multimodal neural network consisting of two branches of LSTMs for extracting textual and acoustic representations. These two representations are concatenated in one single vector. The authors in [209] concatenate the audio and transcript representations during training. The authors in [210] pass the MFCC features through CNN layers, while the visual and textual modalities are passed through dense layers. The two representations are concatenated into one feature vector.

3.2.3 Late Fusion

The authors in [41] use audio, videos, and transcripts and combine the respective representations via a late fusion approach, namely adaptive nonlinear judge classifier. A majority vote approach is adopted by [40].

3.2.4 Other approaches

In [211], the authors use sentence embeddings, log-Mel spectrograms, and facial expressions and employ ConvBiLSTMs. They fuse the representation vectors by using an attention layer and state that the proposed approach outperforms late fusion strategies. A different approach is proposed by [212], where feed-forward highway layers with gating units are used for controlling the information flow of the different modalities. This approach is compared with early and late fusion strategies. Results suggest that the proposed approach yields the highest results.

3.2.5 Literature Review Findings

Existing research works rely on the feature extraction approach, which is a time-consuming procedure, demands a level of domain expertise and does not generalize well to new data. In terms of multimodal approaches, early, intermediate, and late fusion strategies are employed, which cannot capture the inter-modal interactions. Additionally, the majority of studies are performing their experiments on the english language, thus limiting the generalization to other languages. Finally, no study has experimented with multi-task learning approaches for exploring if the education level, age, and gender aid in the depression detection task.

3.2.6 Datasets

3.2.6.1 Androids Corpus

The Androids corpus [42] consists of two tasks, namely the reading and interview task. Specifically, the interview task consists of 116 spontaneous speech samples. All experiments are person independent. Audio files are in Italian language. This dataset includes information about the gender, age, and education level of the individuals. The populations of depressed and non-depressed participants have the same distribution in terms of age, gender, and education. Manual transcripts are not provided.

3.2.6.2 Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ)

This database includes clinical interviews for supporting the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder [213]. 189 clinical interviews are included. PHQ-9 is used for annotating depressed patients.

3.2.6.3 Emotional Audio-Textual Depression Corpus (EATD-Corpus)

The emotional audio-textual depression corpus [214] includes audio recordings and transcripts of 162 chinese student volunteers. Specifically, this dataset includes 30 depressed volunteers and 132 non-depressed volunteers. Each volunteer is asked to complete an SDS questionnaire.

3.2.6.4 Depression and Anxiety Crowdsourced Corpus (DEPAC)

The depression and anxiety crowdsourced corpus [215] includes 2,674 audio samples collected from 571 subjects. Firstly, each participant is asked to provide some demographic information, i.e., age, gender, education level. Then, each participant is asked to perform five speech tasks, phoneme pronunciation, phonemic fluency test, picture description, semantic fluency test and prompted narrative task. Each participant is asked to complete two assessment tools, including Patient Health Questionnaire (PHQ-9) and Generalized Anxiety Disorder - 7 (GAD-7).

3.2.6.5 Multimodal Open Dataset for Mental-Disorder Analysis (MODMA)

The multimodal open dataset for mental-disorder analysis [216] consists of audio recordings, EEG signals, and questionnaires. It includes 24 depressed patients and 29 healthy controls. In terms of the questionnaires, all the participants have completed depression assessment questionnaires, including PHQ-9 and GAD-7 (generalized anxiety disorder-7), and a psychiatric evaluation. Each participant is asked to perform an interview, reading, and picture description task.

3.3 State-of-the-art analysis of Machine Learning Methods used in dementia from spontaneous speech

3.3.1 Unimodal Approaches

Meghanani et al. [94] used the ADReSS Challenge Dataset and proposed three deep learning models to detect AD patients using only speech data. Firstly, they converted the audio files into Log-Mel spectrograms and MFCCs along with their delta and delta-delta, in order to create an image consisting of three channels. Next, they divided the images into non-overlapping segments of 224 frames and passed each frame through five convolution layers followed by LSTM layers. In the second proposed model, they replaced the five convolution layers with a pretrained ResNet18 model. Finally, they trained a model consisting of BiLSTMs and CNN layers. Results showed that Log-Mel spectrograms and MFCCs are effective for the AD detection problem. One limitation of this study is that the authors employed only one image-based pretrained model, i.e., ResNet18.

Gauder et al. [217] used the ADReSSo Challenge Dataset and extracted a set of features from speech, namely eGeMAPS [218], trill [219], allosaurus [220], and wav2vec2 [221], where each feature vector was fed into two convolution layers. Then, the outputs of the convolution layers were concatenated and were passed through a global average pooling layer followed by a dense layer, in order to get the final output. Results from an ablation study showed that trill and wav2vec2 constituted the best features. The main limitations of this study are the feature extraction process and the concatenation of the feature representations.

Balagopalan and Novikova [222] used the ADReSSo Challenge Dataset and introduced three approaches to differentiate AD from non-AD patients by extracting 168 acoustic features from the speech audio files, computing the embeddings of the audio files using wav2vec2, and finally combining the aforementioned approaches by simply concatenating the two representations. Results showed that a Support Vector Machine trained on the acoustic features yielded the highest precision, whereas the SVM classifier trained on the concatenation of the embeddings achieved the highest accuracy, recall, and F1-score. The limitation of this study lies on the feature extraction process, the train of traditional machine learning classifiers, and the usage of the concatenation operation, where the same importance is assigned to the features.

Ref. [93] used the ADReSS Challenge Dataset and introduced two approaches targeting at diagnosing dementia only from speech. Firstly, after employing VGGish [223], they used the features extracted via VGGish and trained shallow machine learning algorithms to detect AD patients. Next, they proposed a convolutional neural network for speech classification, namely DemCNN, and claimed that DemCNN outperformed the other approaches. The main limitation of this research work is the train of shallow machine learning classifiers using the VGGish features, which increase the training time.

The authors in [224] proposed a feature extraction approach. Specifically, they extracted 54 acoustic features, including duration, intensity, shimmer, MFCCs, etc. Finally, they trained the LIBSVM with a radial basis kernel function. The limitation of this study lies on the feature extraction process and the train of only one traditional machine learning classifier. In addition, the authors have not applied feature selection or dimensionality reduction techniques.

Research works [225, 226] used the DementiaBank Dataset and exploited a set of acoustic features along with shallow machine learning classifiers. More specifically in

[225], the authors extracted a set of 121 features, including the fundamental frequency, the frequency alteration from cycle to cycle, the F0 amplitude variability, features assessing the voice quality, spectral features, etc. The authors expanded this feature set with some statistical sub-features, i.e., min, max, mean, etc. and thus increased the number of features to 811. After employing feature selection techniques, the authors applied two classification algorithms, namely SVM and Stochastic Gradient Descent for classifying subjects into AD, non-AD patients, and Mild Cognitive Impairment (MCI) groups in a cross-visit framework. In [226], the authors extracted a set of features, including the emobase, ComParE [227], eGeMAPS, and MRCG functionals [228] and performed three experiments, namely segment level classification, majority vote classification, and active data representation. The authors exploited many classifiers, including Decision Trees, k-Nearest Neighbours, Linear Discriminant Analysis (LDA), Random Forests, and Support Vector Machines. The limitations of these studies lie on the tedious procedure of feature extraction, which demands domain expertise. Also, both studies train shallow machine learning classifiers.

Bertini et al. [229] used the DementiaBank Dataset and employed an autoencoder used in the audio data domain called *auDeep* [230] and passed the encoded representation (latent vector) to a multilayer perceptron, in order to detect AD patients. Results showed significant improvements over state-of-the-art approaches. The main limitation of this study is the way the speech signal is represented as image. Specifically, the speech signal is converted to a log-Mel spectrogram. On the contrary, the addition of delta and delta-delta features as channels of the image adds more information, since these features add dynamic information to the static cepstral features.

The authors in [231] introduced the Open Voice Brain Model (OVBM), which uses 16 biomarkers. Audio files were converted into MFCCs. The ResNet has been used by eight biomarkers for feature representation. Finally, the authors have applied Graph Neural Networks (GNNs) and have extracted a personalized subject saliency map. The limitation of this study lies on the way the speech signal is represented as an image. Specifically, the authors convert the speech signal only to MFCCs. On the contrary, the addition of delta and delta-delta features as channels of the image adds more information, since these features add dynamic information to the static cepstral features. In addition, the authors train multiple models increasing in this way both the training time and computational resources.

Li et al. [89] extracted a set of acoustic and a set of linguistic features for categorizing people into AD patients and non-AD ones. Regarding the acoustic features, they extracted the ComParE feature set and x-vectors. In terms of the linguistic features, they exploited CLAN [232] for extracting the Linguistic feature set. They also extracted tf-idf and BERT features. Next, they employed feature selection, i.e., Pearson's Correlation, and dimensionality reduction, i.e., Principal Component Analysis (PCA), techniques. Finally, they trained three classifiers, namely Linear Discriminant Analysis, Support Vector Machine, and LSTM coupled with an attention mechanism. They used both manual and automatic transcripts. Findings revealed that linguistic features achieved better performance than the acoustic ones. Also, the authors stated that linguistic features extracted from automatic transcripts achieved similar performance with the one obtained by using manual transcripts.

Pan et al. [233] introduced Sinc-CLA, which consists of SincNet [234], Convolutional Layers, Long Short-Term Memory layers, and an attention layer. They used this architecture as a task-driven feature extractor, where they passed the outpits of the attention layer and the dense layer in LR and SVM classifiers. The authors extracted also ComParE and IS10 feature sets and trained LR and SVM classifiers. They conducted their experiments both at chunk and recording-level. Results showed that the task-driven features yielded superior performance compared with IS10 and ComParE. Moreover, the authors performed an analysis of the learned SincNet filters and stated that low-frequency information is critical for classifying Mild Cognitive Impairement (MCI) and Neurodegenerative Disorders (ND) from Healthy Control (HC).

Ref. [235] used only transcripts and introduced three deep neural networks. Firstly, the authors trained a Convolutional Neural Network. Secondly, the authors trained an architecture consisting of CNN and Bidirectional LSTM layers. Finally, the authors introduced an architecture, namely SDDNN, where they passed the representation vectors of the transcripts through: (a) CNN, (b) CNN + BiLSTM, and (c) BiLSTM coupled with an attention mechanism, and concatenated the obtained representation vectors. The authors experimented with both GloVe embeddings and randomly initialized embeddings. Findings showed that SDDNN using GloVe embeddings achieved the best evaluation results.

Wankerl et al. [236] combined two perplexity estimates, namely one from a model trained on transcripts of speech produced by healthy controls and the other from a model trained on transcripts from patients with dementia. An AUC score of 0.83 was achieved by using n-gram Language Models (LMs) in a participant-level leave-one-out-cross validation (LOOCV) evaluation across the DementiaBank dataset. Fritsch et al. [237] further improved performance of this approach by substituting a neural LM (a LSTM model) for the n-gram LM, and report an improved AUC of 0.92. However, it is currently unclear as to whether this level of accuracy is due to dementia-specific linguistic markers, or a result of markers of other significant differences between the case and control group such as age ($\bar{x} = 71.4$ vs. 63) and years of education ($\bar{x} = 12.1$ vs. 14.3) [54]. The work proposed by [238] investigated why these approaches are effective by interrogating neural LMs trained on participants with and without dementia using synthetic narratives previously developed to simulate progressive semantic dementia by manipulating lexical frequency. Findings suggested that the “two perplexities” approach is successful at distinguishing between cases and controls in the DementiaBank corpus because of its ability to capture specifically linguistic manifestations of the disease.

Reference [239] exploited only transcripts and employed a feature extraction process. Specifically, the authors extracted n-gram and lexicosyntactic features, including stopwords ratio, word count, quantity of expressed propositions to the total spoken words, etc.

Next, the authors proposed feature selection techniques, including the student's t-test and the Kolmogorov-Smirnov test. For dealing with the imbalanced dataset, a subsampling technique was adopted, where the authors performed a random selection of a subset of the majority with a matching size to that of the minority. Several machine learning algorithms were trained, including Gaussian Naive Bayes, SVM, and Multilayer Perceptron Neural Networks. The authors evaluated their proposed approaches on five classification tasks, namely AD vs. HC, MCI vs. HC, MCI vs. AD, HC vs. Possible AD (PoAD), and AD vs. PoAD. Results showed that early stages of dementia can be efficiently diagnosed through linguistic patterns and deficits. In addition, the authors stated the superiority of their approaches over state-of-the-art ones.

In [240], the authors introduced a stacked fusion model. Firstly, the authors extracted lexicosyntactics and character n-gram features. Next, they applied feature selection techniques, namely Pearson's correlation and mutual information. After that, they trained and evaluated several classifiers, including Random Forest, Extreme Gradient Boosting, Linear discriminant analysis, Support Vector Machine, Gaussian Naïve Bayes, Logistic Regression, and Multi-layer Perceptron, where they returned the best n classifiers. Finally, the predictions of the best n classifiers were used as input to a Meta-Classifier. Findings suggested the effectiveness of ensemble methods for AD diagnosis.

The authors in [241] used the DementiaBank dataset and translated the transcripts into the Nepali language. Next, the authors used CountVectorizer, tf-idf, Word2Vec, and fastText. They trained both shallow and deep learning classifiers. Regarding the shallow machine learning algorithms, they used Decision Trees, k-Nearest Neighbours, Support Vector Machines, Naive Bayes, Random Forests, AdaBoost, and XGBoost. In terms of the deep learning models, they experimented with CNNs, BiLSTMs, Attention Layers, and their combinations. Findings showed that the deep learning models performed better than the traditional machine learning classifiers.

Nasreen et al. [164] extracted two feature sets, namely disfluency and interactional features, and performed an in-depth statistical analysis in an attempt to investigate the differences between AD and non-AD subjects in terms of these features. Findings show that these two groups of people present significant differences. Then, they exploited shallow machine learning algorithms using the aforementioned feature sets to distinguish AD from non-AD patients and obtained an accuracy of 0.90 when providing both feature sets as input to the SVM classifier.

Al-Hameed et al. [242] used a longitudinal dataset to study the natural deterioration of AD patients across three visits. More specifically, they used only acoustic features and employed feature selection techniques to predict MMSE scores and distinguish people with AD from people with Mild Cognitive Impairment (MCI) and healthy control (HC). A similar approach was proposed by [165], who extracted features only from transcripts in order to detect AD patients. Findings suggest that word entropy, phone entropy, and rate of pauses in utterances achieve competitive performance when they are given as input to a Decision Tree classifier. Haider et al. [243] introduced three approaches, namely segment-level, majority-voting on segments, and the novel active data representation (ADR), for identifying AD patients using only acoustic features. They claimed that ADR outperformed the other two approaches due to its ability to encode acoustic information of a full audio recording into a single feature vector for model training.

The authors in [65] introduced some approaches to detect AD patients and predict the MMSE scores using only text data. Specifically, the authors proposed a Convolutional Neural Network (CNN) and fastText-based classifiers. Regarding the AD classification task, they fitted 21 models and the outputs were combined by a majority voting scheme for final classification. In terms of the MMSE regression task, the outputs of these bootstrap models were averaged for calculating the final MMSE score.

Research works [244, 245] employed a hierarchical attention neural network to detect AD patients. More specifically, the authors in [244] evaluated their proposed model in both manual and automatic transcripts and found that a hierarchical neural network achieves an improvement in F1-score in comparison to other deep learning models. In [245], the authors tried to interpret the decisions made by the proposed model by visualizing words and sentences and performing statistical analyses. However, they were not able to explain why their model pays attention to some specific words more than others.

Authors in [246] proposed a multi-task learning framework (Sinc-CLA), so as to predict age and MMSE scores (both considered as regression tasks) and used only speech as input for their proposed network. Concurrently, they introduced shallow networks with input i-vectors and x-vectors both in single and multi-task learning frameworks. They claimed that using x-vectors in a multi-task learning framework yields the best results in terms of the estimation of both age and MMSE scores.

The research work proposed by [247] employed unimodal approaches by using only either speech or text to classify subjects into AD patients or non-AD ones. For the text modality, the authors extracted embeddings by using fastText, BERT, LIWC, and CLAN. For the acoustic modality, the authors extracted i-vectors and x-vectors. For both modalities, they employed dimensionality reduction techniques and trained shallow machine learning classifiers and neural networks (CNNs and LSTMs). The authors claimed that the Support Vector Machine and the Random Forest Classifiers trained on BERT embeddings achieved the highest accuracy. One limitation of this study is the fact that the authors used BERT embeddings as features for training additional algorithms. They did not experiment with extracting the [CLS] token and passing it to a dense layer for performing the classification.

Karlekar et al. [168] applied three deep neural networks based on CNNs, LSTM-RNNs, and their conjunction to distinguish AD patients from non-AD ones utilizing only transcripts. Next, they proposed explainability techniques by applying automatic cluster pattern analysis and first derivative saliency heat maps, in order to uncover differences in language between AD patients and healthy control groups. The main limitation of this paper is the fact that the authors did not experiment with language models based on transformers, i.e., BERT, RoBERTa, and so on.

Similarly, the work proposed by [248] extracted seventeen features from transcripts for detecting AD patients. Specifically, the authors extracted the rate of pauses in utterances, filler sounds, number of no answers, part-of-speech tags, intelligibility of speech, diversity and complexity of the words, and many more. Next, they trained Support Vector Machines, Linear Discriminant Analysis, and Decision Trees. Results indicated that 90% prediction accuracy can be obtained using only phone entropy, silence rate per utterance, and word entropy with a Decision Tree classifier. The limitation of this paper lies on the feature extraction process, which is a time-consuming and tedious procedure. Additionally, the optimal feature set may not be found, since some level of domain expertise is required.

An augmented adversarial self - supervised learning method was proposed by [249]. Specifically, the introduced approach was based on contrastive predictive encoding. For dealing with the imbalanced dataset, i.e., limited number of speech samples corresponding to AD patients, the authors applied three augmentation schemes, including speed based augmentation, tempo based augmentation, and tremolo based augmentation. Findings indicated that the proposed methods improved the performance for AD detection to a large margin compared to other models.

3.3.2 Multimodal Approaches

Several approaches have been introduced which fuse the representation vectors or features of the different modalities at the input level. This strategy is known as an early fusion approach and does not capture effectively the inter-modal interactions. Edwards et al. [250] proposed a multimodal (audio and text) and multiscale (word and phoneme levels) approach. For the acoustic modality, the authors extracted features using the OpenSMILE toolkit, applied feature selection techniques, and trained shallow machine learning classifiers, including SVM, latent discriminant analysis (LDA), and LR. In terms of the language models, the authors trained a Random Forest Classifier on Word2Vec and GloVe embeddings. Also, they trained a FastText classifier from scratch. In addition, pretrained embeddings obtained by Sent2Vec, RoBERTa, ELECTRA, and so on were fine-tuned with the FastText classifier. The authors transcribed the segmented text into phoneme written pronunciation using CMUDict and stated that the FastText classifier was the best performing model trained on the phoneme representation. Results also showed that the combination of phonemes and audio yielded to the highest accuracy accounting for 79.17%. Martinc and Pollak [95] proposed also an early fusion approach. The authors extracted a large number of features corresponding to the textual and acoustic modality. They fused the feature sets via an early fusion method. Finally, they trained four machine learning classifiers, namely XGBoost, Random Forest, SVM, and Logistic Regression. Findings showed that the logistic regression and SVMs were proved to be better than XGBoost and Random Forest. Also, the authors stated that the readability features led to a surge in the classification performance. In terms of the audio features, the duration was the best performing one. Pompili et al. [67] proposed an early fusion approach for fusing the modalities of speech and transcript. Specifically, for the text modality, the authors employed the BERT model first and then trained three deep neural models on top of the BERT embeddings, namely (i) a Global Maximum pooling, (ii) a bidirectional LSTM-RNNs provided with an attention module, and (iii) the second model augmented with part-of-speech (POS) embeddings. For the audio modality, the authors extracted the x-vectors. Finally, the authors merged the feature sets corresponding to the two different modalities and trained a Support Vector Machine classifier. Results showed that the fusion of the two modalities increased the performance obtained by unimodal approaches exploiting only speech or text. Ref. [251] extracted three sets of features, namely lexicosyntactic, acoustic, and semantic features. In terms of the lexicosyntactic features, the authors extracted the proportion of POS-tags, average sentiment valence of all words in a transcript, and many more. Regarding the acoustic features, MFCCs, fundamental frequency, statistics related to zero-crossing rate, etc. were exploited. With regards to the semantic features, the authors extracted proportions of various information content units used in the picture. Next, they performed feature selection by using the ANOVA and trained four machine learning classifiers, including SVM, neural network, RF, and NB. Results showed that SVM outperformed the other approaches in the multimodal framework. The limitation of this study lies on the way the features from different modalities are combined. More specifically, the authors apply an early fusion strategy, where they fuse the features at the input level. This approach does not capture the inter- and intra-modal interactions. In addition, another limitation is the feature extraction procedure. In [252], an early fusion approach was proposed. Specifically, the authors extracted a set of acoustic features, i.e., articulation, prosody, i-vectors, and x-vectors, and a set of linguistic features, including word2vec, BERT, and BERT-Base trained with the Spanish Unannotated Corpora (BETO) embeddings. The authors concatenated these sets of features and trained a Radial Basis Function-Support Vector Machine. The main limitation of this paper is the early-fusion approach. [253] compared the performance of traditional machine learning classifiers with the performance obtained by pre-trained transformer models, namely BERT. More specifically, the authors extracted a large number of features, i.e., lexicosyntactic, semantic, and acoustic features and applied feature selection by choosing top-k number of features, based on ANOVA F-value between label and features. Four conventional machine learning models, namely Support Vector Machine, Neural Network, Random Forest, and Naive Bayes, were trained with the respective sets of features. Next, the authors trained a BERT model and stated that BERT outperformed the feature-based approaches in terms of all the evaluation metrics. [254] introduced some approaches to predict MMSE scores using textual and acoustic features. More specifically, the authors extracted lexicosyntactic features weighted via tf-idf, psycholinguistic features, discourse-based features, and acoustic features (MFCCs). The authors trained a Support Vector Regressor for predicting the MMSE scores. Results indicated that a selection of verbal and non-verbal cues achieved the lowest RMSE score. The authors in [247, 64] introduced approaches based on multimodal data (both linguistic and acoustic features) to detect AD patients (binary classification task) and predict MMSE score (regression task). More specifically, the authors in [247] exploited dimensionality reduction techniques followed by machine learning classifiers and stated that Logistic Regression (LR) with language features was their best performing model in terms of classifying AD and non-AD patients. With regards to estimating the MMSE score, they claimed that a Random Forest classifier with language features achieves the lowest RMSE and R^2 scores. The combination of linguistic and acoustic features did not perform well on both tasks. In [64], the authors trained both shallow and deep learning models (LSTM and CNN) on a feature set consisting of acoustic features (i-vectors, x-vectors) and text features (word vectors, BERT embeddings, LIWC features, and CLAN features) to detect AD patients. They found that the top-performing classification models were the Support Vector Machine (SVM) and Random Forest classifiers trained on BERT embeddings, which both achieved an accuracy of 85.4% on the test set. Regarding the regression task, they claimed that the gradient boosting regression model using BERT embeddings outperformed all the other introduced architectures.

Other approaches employ late-fusion strategies. This means that multiple models, i.e., acoustic and language, are trained separately and the final result/prediction is often taken after a majority vote approach. In this way, the inter-modal interactions are not captured. The authors in [92] proposed a majority-level approach for classifying AD patients using the audio and textual modalities. In terms of the textual modality, the authors extracted handcrafted textual features and deep textual embeddings of transcripts. For the extraction of deep textual embeddings, they used BERT, RoBERTa, and distilled versions of BERT and RoBERTa. Next, they exploited feature aggregation techniques and classified the subject as AD or non-AD patient by training either a Logistic Regression (LR) or a Support Vector Machine (SVM) classifier. In terms of the audio modality, the authors extracted handcrafted acoustic features, i.e., ComParE, COVAREP, etc. and deep acoustic embeddings, i.e., YAMNet, VGGish, etc. Similarly to the textual modality, they used feature aggregation techniques and trained a LR and SVM classifier. Results indicated that the majority-level approach of text models yielded the highest evaluation results, while the fusion of textual and acoustic modalities led to a degredation in performance. Shah et al. [64] introduced a weighted majority-vote ensemble meta-algorithm for classification utilizing the modalities of speech and transcripts. For the textual modality, the authors extracted language and fluency features, including the type-token ratio, the number of verbs per utterance, etc. and n-gram features. For the acoustic modality, the authors extracted four feature sets using the OpenSMILE v2.1 toolkit. After that, the authors applied dimensionality reduction techniques, i.e., Principal Component Analysis, and feature selection techniques, i.e., ANOVA F-values. Finally, shallow machine learning classifiers were trained. Best results were obtained by using only the textual modality, while the majority vote approach by combining textual and acoustic modalities led to a decrease in the classification performance. Sarawgi et al. [170] trained acoustic and language models separately and proposed three kinds of ensemble modules for classification. Specifically, the authors experimented with hard ensemble, meaning that a majority vote was taken between the predictions of the three individual models. A soft ensemble was also proposed, where a weighted sum of the class probabilities was computed for final decision, in order to leverage the confidence of the predictions. Also, a learnt ensemble was exploited, where a logistic regression classifier was trained using class probabilities as inputs. Results showed that the hard ensemble approach yielded the best results. Mittal et al. [255] proposed a late fusion strategy using the modalities of speech and transcripts. Firstly, they trained separately acoustic and language models. For the acoustic modality, the authors trained a VGGish model with log-mel spectrograms. For the textual modality, the authors concatenated the representation obtained by BERT, Sentence-BERT, and fastText-CNN. Finally, the probabilities calculated by the audio and text-based model were combined in a weighted manner, and a threshold was fixed for classifying the persons into AD and healthy control. Pappagari et al. [256] trained acoustic and language models separately and used the output scores as inputs to a Logistic Regression classifier for obtaining the final prediction. For the language models, the authors used automatic speech recognition models for transcribing the recordings and employed a BERT model. For the acoustic modality, the authors used x-vectors for classifying subjects into AD patients and non-AD ones. Also, they extracted eGeMAPS, VGGish, prosody features, etc. and trained Logistic Regression and XGBoost classifiers. The authors stated that the combination of the different models and the BERT model trained on automatic transcripts achieved equal accuracy on the test set. Similarly, the authors in [90] trained also acoustic and language models separately. In terms of the acoustic models, the authors extracted the x-vectors and trained a Probabilistic Linear Discriminant Analysis classifier. For the textual modality, the authors employed a BERT model. For fusing the two modalities, the authors employed the scores from the whole training subset to train a final fusion GBR model that was used to perform the fusion of scores coming from the acoustic and transcript-based models for the challenge evaluation. Results showed that the proposed approach was the best performing one. The authors in [257] introduced three speech-based systems and two text-based systems for diagnosing dementia from spontaneous speech. Also, they proposed methods for fusing the different modalities. In terms of the speech based systems, the authors extracted i-vectors, x-vectors, and rhythmic features and trained an SVM and a Linear Discriminant Analysis (LDA) classifier. Regarding the text-based models, the authors fine-tuned a BERT model and trained an SVM classifier using linguistic features. Finally, the authors exploited three fusion strategies based on late fusion approach. Therefore, the main limitation of this study is the late fusion approach for fusing the different modalities. Ref. [63] used the ADReSS Challenge Dataset and introduced neural network architectures which use language and acoustic features. Regarding the multimodal approach, the authors fuse the predictions of the three best performing models using a majority vote approach and show that label fusion outperforms the neural networks using either only speech or transcripts. The limitation of this study lies on the usage of a late fusion strategy, i.e., majority vote approach. In this way, multiple models must be trained separately increasing the training time. Also, the inter-modal interactions cannot be captured. [61] proposed several acoustic and language individual models. Specifically, they extracted both handcrafted features and embeddings via BERT, RoBERTa, VGGish, YAMNet, etc. After applying feature aggregation techniques, they trained and tested a Logistic Regression and Support Vector Machine classifier for differentiating AD from non-AD patients. For fusing the two modalities, the authors applied a majority voting based label fusion strategy, where each model made a decision on whether it considered the subject to be healthy or suffering from Alzheimer's dementia. Results showed that the multimodal fusion did not achieve better performance than the unimodal models. Regarding the MMSE regression task, the authors used SVR and PLSR and fused the two modalities by applying average-based fusion. A similar approach was conducted by [66], where the authors extracted a set of acoustic features, i.e., Prosody, Voice Quality, ComParE, IS10-Paling, etc., and a set of linguistic features using transformer-based networks, including BERT, RoBERTa, and their distilled versions. They categorized people into AD patients or not by training a Support Vector Machine (SVM) and a Logistic Regression (LR) classifier. The authors used label fusion from the top performing models and stated that the label fusion of the 10 best performing textual models achieved an accuracy of 85.42%. For predicting the MMSE scores, the authors used support vector machines based regression (SVR) and a partial least squares regressor (PLSR). They achieved a Root Mean Squared Error (RMSE) score equal to 4.30 by averaging the predictions of the MMSE scores from the top-10 performing models. Research works [169] extracted a set of acoustic and linguistic features using the ADReSSo Challenge Dataset. Next, they concatenated these sets of features and trained a Logistic Regression classifier. They also proposed three label fusion approaches, namely majority voting, average fusion, and weighted average fusion, based on the predictions of several neural networks. The limitations of this study are related to the early and late fusion strategies introduced for detecting AD patients. In [258], the authors introduced an approach, which accounts for temporal aspects of both linguistic and acoustic features. In terms of the acoustic features, the authors exploited the eGeMAPS feature set, while they used GloVE embeddings with regards to the language features. Next, the Active Data Representation [226] with some modifications was employed. The authors used a Random Forest Classifier for performing their experiments. The authors performed a series of experiments and stated that the majority vote approach yielded the best result. The method for fusing the two modalities, i.e., late fusion strategy, constitutes the main limitation of this study.

There are also approaches, which add or concatenate the representation vectors of different modalities during training. However, in this way, the inherent correlations between the different modalities are not captured. On the contrary, equal importance is assigned to the different modalities. Research work [259] employed also a bi-modal model consisting of Dense, GRU, CNN, BiLSTM, and attention layers. The authors fused the two modalities by concatenating their respective representations. Results on the ADReSS

Challenge Dataset showed an improvement of evaluation results of the multimodal approach over unimodal architectures. The usage of the concatenation operation for fusing the two modalities constitutes a limitation of this study. Also, the feature extraction process proposed by the authors, constitutes another limitation. Zhu et al. [91] proposed both unimodal and multimodal approaches. Regarding unimodal models, they employed first MobileNet and YamnNet to discriminate between AD patients and non-AD ones. They converted audio files into MFCC features, duplicated the MFCC feature map twice and made the MFCC feature map as a (p, t, 3)-matrix, in order to match with the module input of the proposed architectures. They used also BERT and Speech BERT. In terms of the multimodal models, the authors exploited Speech BERT, YamnNet, Longformer, and BERT. After extracting the representations of audio and transcripts, they used the add and concatenation operation to fuse these two modalities. Results on the ADReSS Challenge Dataset showed that the concatenation operation of the representations extracted via BERT and Speech BERT outperformed the unimodal models. The limitations of this study are the following: (i) the way the speech signal is represented as an image. More specifically, this study duplicates the MFCC feature map twice and makes the MFCC feature map as a (p, t, 3)-matrix. On the contrary, the delta and delta-delta features can be used for adding more information [260, 261]. (ii) In terms of the multimodal models, the authors fuse the different modalities via an add and concatenation operation. These methods do not capture the inherent correlations between the two modalities. The authors in [262] proposed both unimodal and multimodal approaches. Regarding unimodal approaches using speech data, the authors extracted acoustic features and trained four shallow machine learning classifiers. For the language modality, the authors trained a BERT model. In terms of the multimodal approach, the authors simply concatenated the representations obtained by BERT and acoustic modality. Results on the test set indicated that the fusion approach achieved lower performance than the unimodal one using the textual modality. Koo et al. [263] used the ADReSS Challenge Dataset and proposed a deep learning model consisting of BiLSTMs, CNNs, and self-attention mechanism and exploited both textual, i.e., transformer-based models, psycholinguistic, repetitiveness, and lexical complexity features, and acoustic features, i.e., openSMILE and VGGish features. Specifically, they passed each modality through a self attention layer, where key, value, and query corresponded to one single modality. However, the authors concatenated the outputs of the attention layer, which correspond to the two different modalities, and passed them through a CNN layer. The main limitations of this study are pertinent to the feature extraction process and the concatenation of the representation vectors of the two modalities into one vector.

A different approach was proposed by [62]. More specifically, the authors extracted textual and acoustic features and passed them through two different branches of BiLSTM layers. A gating mechanism consisting of highway networks was proposed for fusing the two modalities. However, the authors did not experiment with replacing the proposed fusion method with a concatenation operation via an ablation study. Thus, this fusion method cannot guarantee performance improvement. Similarly, [264] used BERT instead of BiLSTM for extracting the text representation and stated that the BiLSTM performed better than BERT due to the fewer parameters used.

3.3.3 Other Multimodal Tasks

Villegas et al. [265] introduced multimodal approaches for inferring the political ideology of an ad sponsor and identifying whether the sponsor is an official political party of a third-party organization. The authors employed BERT and EfficientNet [266] for extracting textual and visual representations respectively. They concatenated these two representations and passed the resulting vector to an output layer for binary classification. Results suggested that the combination of both modalities led to a surge in the classification performance.

Villegas and Aletras [77] proposed multimodal approaches for the task of point-of-interest type prediction. Specifically, the authors exploited BERT and Xception [267] for extracting text and visual representations respectively. Next, they introduced three different architectures for fusing the two modalities. First, they exploited the Gated Multimodal Unit introduced by [75]. Secondly, inspired by [76], they proposed a model for modeling the cross-modal interactions. Finally, the authors introduced an architecture, which includes the gated multimodal mechanism and the cross-attention layers on the top of the gated multimodal mechanism. Findings suggested that the proposed architecture yielded new state-of-the-art results outperforming significantly the previous text-only models.

Gu et al. [268] presented a deep multimodal network with both feature attention and modality attention to classify utterance-level speech data. The authors used the modalities of audio signal and text data as input to the deep neural network. In terms of the modality fusion approach proposed, it consisted of three main parts, namely the modality attention module, the weighted operation, and the decision making module. Findings showed that the multimodal system achieved state-of-the-art performance and was tolerant to noisy data indicating in this way its generalizability.

Pan et al. [269] proposed a multimodal architecture for detecting sarcasm in Twitter. More specifically, the authors exploited the ResNet-152 model and obtained a visual representation. Regarding the textual modality, they used a pretrained BERT model. After obtaining embeddings for the input sequence and the hashtags included in the sequence, the authors passed the corresponding embeddings through encoders of the transformer. For modeling the cross-modal interactions, an additional encoder was used, where the visual representation corresponded to the key and value, while the sequence representation corresponded to the query. In addition, an intra-modality attention approach was used, which gets as input the sequence and the hashtag representations. The outputs obtained were concatenated and passed to an output layer for the final prediction. Findings stated that the proposed architecture achieved state-of-the-art results.

Inspired by the transformer model in machine translation [46], the authors in [270] presented some multimodal approaches for the task of visual question answering. More specifically, the authors employed a self-attention and a guided-attention unit for capturing the intra- and inter-modal interactions respectively. Next, they obtained a Modular Co-Attention layer, which constitutes the modular composition of the self-attention and guided-attention units. Finally, the authors proposed a deep Modular Co-Attention Network consisting of cascaded Modular Co-Attention layers. Results indicated that the introduced approach surpassed the existing co-attention models.

Zadeh et al. [271] introduced a novel model, termed Tensor Fusion Network, for the task of multimodal sentiment analysis. The authors used visual, language, and acoustic modalities. For capturing the intra-modal interactions, the authors proposed three Modality Embedding Subnetworks. For capturing the inter-modal interactions, the Tensor Fusion layer has been used. Finally, the authors employed the Sentiment Inference Subnetwork, which is conditioned on the output of the Tensor Fusion layer and performs sentiment inference. Results indicated a surge in performance in comparison with existing research initiatives.

Cai et al. [272] presented a multimodal approach for sarcasm detection in Twitter. The authors used the modalities of text features, image features, and image attributes. After extracting image features and attributes, the authors leveraged attribute features and BiLSTM layers for extracting the text features. Next, the authors employed a representation fusion approach for reconstructing the features of the three modalities. Finally, they proposed a modality fusion approach motivated by [268]. Results showed the effectiveness of the proposed architecture and the usefulness of the three modalities.

A different approach was proposed by [273], where the authors utilized optimal transport for capturing the cross-modal interactions and self attention mechanisms for capturing the intra-modal correspondence. Specifically, they exploited three different modalities, namely visual, language, and acoustic modalities. After utilizing self-attention and optimal transport methods, they used the multimodal attention fusion method introduced by [268]. Experiments conducted towards the sarcasm and humor detection tasks demonstrated valuable advantages over existing research initiatives.

Yu et al. [83] introduced an approach for capturing both the inter- and intra-modal interactions for the visual question answering and the visual grounding tasks using the modalities of text and image. Specifically, after obtaining text and visual representations, they passed these two representations through a unified attention block. The authors proposed also a variation of the self-attention mechanism by introducing a novel gating model. Findings showed the effectiveness of the proposed approach on five datasets.

3.3.4 Literature Review Findings

From the aforementioned research works, it is evident that despite the negative consequences dementia has in people's everyday life, little work has been done so far towards its identification. More specifically, most researchers introduce feature extraction approaches from audio and transcripts and train ML algorithms, such as SVM, LR, etc. Because of the fact that feature extraction constitutes a time-consuming procedure and does not generalize well to new AD patients, researchers have started exploiting deep learning methods, such as CNNs and LSTMs, which obtain low performances. However, despite the fact that pretrained transformer models achieve new state-of-the-art results in several domains, including the biomedical one, their potential has been mainly used as embeddings for training shallow ML algorithms, such as SVM or LR. Concurrently, little has been done regarding the interpretability of the proposed deep learning models as well as the main differences observed in the language between AD patients and non-AD patients.

In terms of the architectures using only speech data, it is evident that current research works [224, 217, 222, 93, 225, 226, 247] have been focused mainly on acoustic feature extraction and then the usage of shallow machine learning algorithms, i.e., SVM, LR, RF etc., or CNNs and BiLSTMs. The study in [94], which has converted audio files into images of three channels, namely log-Mel spectrograms (and MFCCs), their delta, and delta-delta, has exploited only one pretrained model of the domain of computer vision, i.e., ResNet18. In addition, the study introduced in [91] has converted the audio files into MFCC features, has duplicated the MFCC feature map twice and has made the MFCC feature map as a *(p, t, 3)*-matrix. Next, this study has employed YAMNet, MobileNet, and Speech BERT. However, the limitation of this study lies on the way images are created. On the contrary, delta and delta-delta coefficients are used for recognizing speech better, since the dynamics of the power spectrum, i.e., trajectories of MFCCs over time, are understood better.

Regarding the multimodal models, the majority of the research works have either concatenated or added the representations corresponding to the two different modalities [91, 262, 259, 263]. However, the concatenation operation assigns equal importance to each modality and it neglects the inter- and intra-modal interactions. Other research works have trained several language and acoustic models separately and then use majority voting for the final classification of the people as AD patients or non-AD patients [63, 170, 92, 258, 169]. Late fusion approaches have been also proposed including [257, 256, 90, 64, 255, 66, 169]. However, these approaches increase substantially the computation time, while the inter-modal interactions are not captured. In addition, there are studies [252, 251, 169, 67, 95, 250] proposing early fusion approaches, meaning that the features corresponding to the different modalities are concatenated at the input level. None of these works capture the inter- and intra-modal interactions.

3.3.5 Datasets

3.3.5.1 DementiaBank Pitt Corpus

The DementiaBank English Pitt Corpus [54] consists of participants with probable and possible Alzheimer's Disease, people with other dementia diagnoses, and healthy people. Regarding the study eligibility criteria, the age of the participant must be over 44 years old, while the person must have at least seven years of education. Also, the person should be able to read and write english fluently before dementia onset, should not have history of major nervous system disorders, such as cerebral trauma, stroke, etc. In addition, the person must not receive any neuroleptic or other medication affecting central nervous system functions, must have an initial MMSE score greater than 10, must be able to give informed consent, and have an informant (for patients only). Subjects in this study made up to five visits. This dataset includes four tasks, which are described below:

- Cookie: Description of the Cookie Theft picture, which is illustrated in Fig. 3.1.
- Fluency: Responses to the Word Fluency task for the dementia group only.
- Recall: Responses to the Story Recall task for the Dementia group only.
- Sentence: Responses to the Sentence Construction task for the dementia group only.

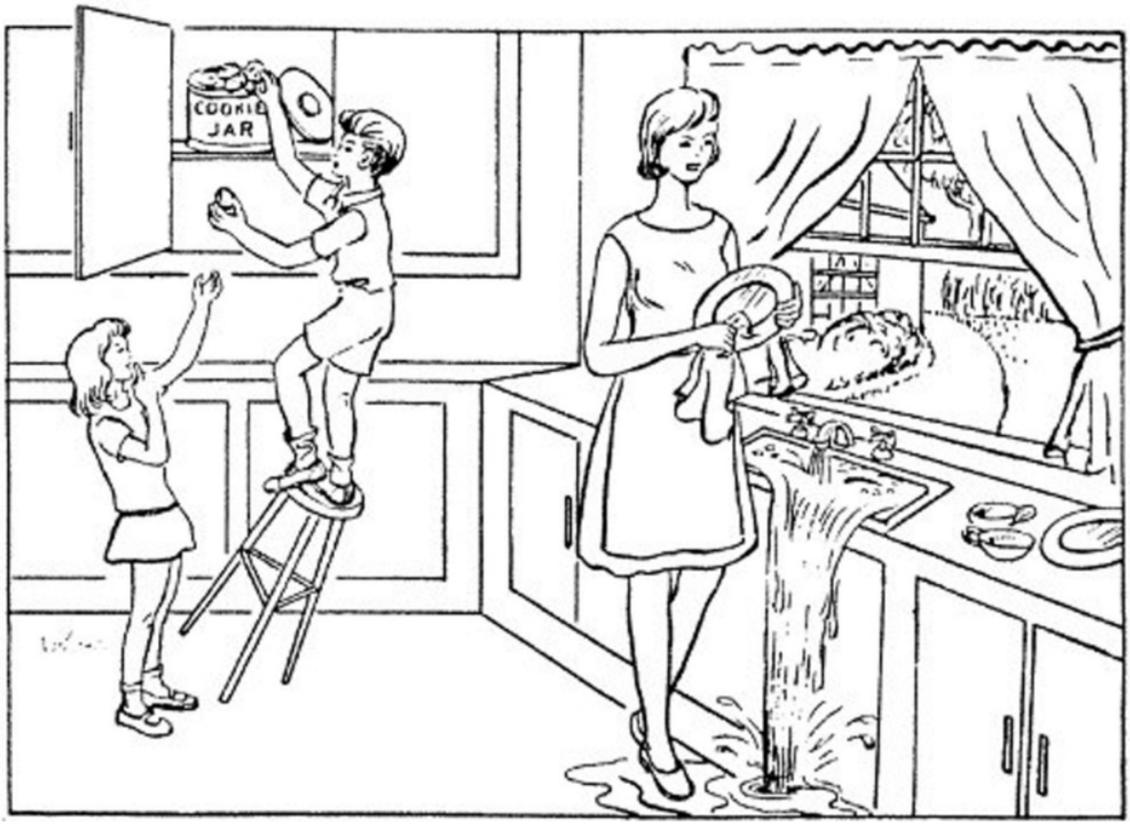


Figure 3.1: The Cookie Theft picture

3.3.5.2 ADReSS Challenge Dataset

In contrast to other datasets, the ADReSS Challenge dataset [53] is matched for gender and age, so as to minimize the risk of bias in the prediction tasks. Moreover, it has been selected in such a way so as to mitigate biases often overlooked in evaluations of AD detection methods, including repeated occurrences of speech from the same participant (common in longitudinal datasets) and variations in audio quality. It consists of speech recordings along with their associative transcripts and includes 78 non-AD and 78 AD subjects. In addition, the dataset includes the MMSE scores for each subject except one. We report the mean and standard deviation of the MMSE scores for the two main groups, i.e., AD patients and non-AD ones, in Table 3.1. Each participant (PAR) has been assigned by the interviewer (INV) to describe the Cookie Theft picture from the Boston Diagnostic Aphasia Exam [54]. Due to the fact that the transcripts are annotated using the CHAT coding system [274], the python library PyLangAcq [275] is used for having access to the dataset. The ADReSS Challenge dataset has been divided into a train and a test set. The train set consists of 54 AD patients and 54 non-AD ones, while the test set consists of 24 AD patients and 24 non-AD ones.

Table 3.1: Mean and standard deviation of the MMSE scores for the two main groups (AD and non-AD patients).

	MMSE	
	mean	**standard deviation**
AD	17.79	5.48
non-AD	29.01	1.17

3.3.5.3 ADReSSo Challenge Dataset

The ADReSSo Challenge Dataset [276] includes two datasets described below:

- a dataset consisting of speech recordings of Alzheimer's patients performing a category (semantic) fluency task [277] at their baseline visit, for prediction of cognitive decline over a two year period, and
- a dataset consisting of healthy people and AD patients describing the Cookie Theft picture.

Similarly to the ADReSS Challenge dataset, the ADReSSo Challenge dataset has been carefully selected to mitigate several kinds of biases. However, it does not include manual transcripts. It includes only speech recordings.

3.3.5.4 B-SHARP Dataset

B-SHARP dataset [278] includes 185 normal controls and 141 MCI patients. Each subject has been examined with multiple cognitive tests, including the Montreal Cognitive Assessment and the Boston Naming Test. In addition, each person speaks about three topics, which are described below:

- Q1: daily activity

- Q2: room environment
- Q3: Description of the Circus Procession picture, which is illustrated in Fig. 3.2.

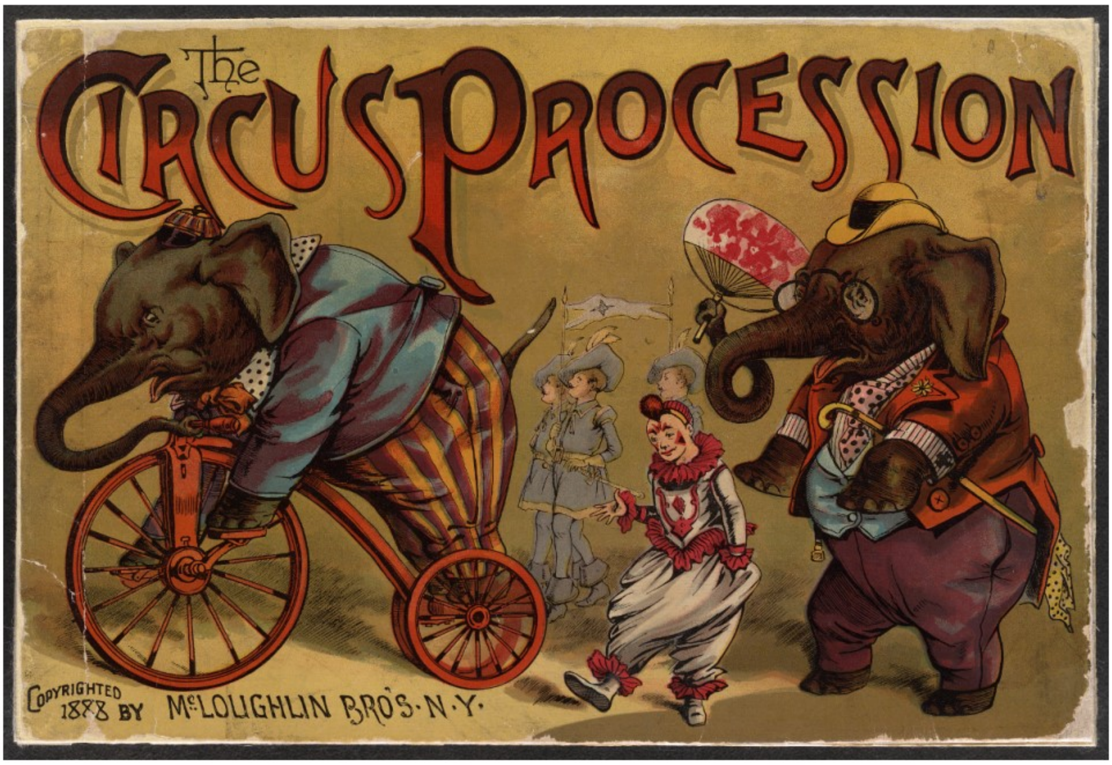


Figure 3.2: The Circus Procession picture

The B-SHARP study is still growing and is not publicly available yet.

3.3.5.5 Longitudinal Multimodal Dataset

This dataset includes data from three 4-week phases [279]. Each subject is examined with multiple cognitive tools at the start and end of each phase, including the Mini Mental State Examination score and the Addenbrooke's Cognitive Examination-III. This dataset is still growing. Until now, the dataset includes 22 people, 14 people with dementia or MCI and 8 age matched controls. Each person is given a tablet application, which shows four pictures every day, each one of them representing a topic from the 50's, 60's, and 70's. Also, three questions are shown to the person, in order to help him/her perform the task. Then, each person can choose one picture and is able to record the conversation, type or write the thoughts.

Regarding the eligibility criteria, the age of the participants ranges from 65 to 80 years at the time of consent. They need to have lived in the United Kingdom during the 50's, 60's, or 70's, and they must be able to use the provided tablet application. Also they need to be in contact with a carer or a family member.

This dataset is not publicly available yet.

3.3.5.6 Carolinas Conversations Collection

The Carolinas Conversations Collection [280] includes a digital archive of transcribed audio and video recordings of people over 65 years of age in natural conversations about health and is supported by the National Libraries of Medicine.

The Carolinas Conversations Collections consists of two cohorts. Specifically, Cohort 1 includes over 200 consented conversations with 125 older men and women of multiple ethnicities, with any of 12 chronic medical conditions recorded twice a year. Cohort 2 includes over 400 naturally occurring conversations with 125 persons with Alzheimer's disease in a longitudinal set of persons.

3.3.5.7 Intelligent Virtual Agent (IVA)

The IVA dataset was collected at the University of Sheffield's Department of Neurology at the Royal Hallamshire Hospital in the UK in a real clinical setting [281, 282]. The IVA acts as a neurologist, i.e., virtual doctor, who asks several questions similar to those asked in real assessment situations. The IVA runs in a laptop, where two cameras are also used for capturing the participants' movements. The participants are asked a total of 10 conversational questions and encouraged to take part in two 1-minute verbal fluency tests.

The participants are grouped into four categories, namely functional memory disorder (FMD), neurodegenerative disorder (ND), MCI, and HC.

3.4 State-of-the-art analysis of Machine Learning Methods used in epilepsy detection/prediction from EEG signals

In the research work proposed by [116], the authors applied a five-level Discrete Wavelet Transform (DWT) to the EEG signals for decomposing them into sub-bands. Then, the authors extracted features from each sub-band, namely the energy and entropy of the coefficients, as well as the standard deviation, variance, and mean of the absolute values of the coefficients. The resulting feature vector was used for training a Random Forest Classifier. Results showed the robustness of the proposed method.

A different approach was proposed by [283], where the authors exploited a novel feature called successive decomposition index (SDI) for the automated seizure detection task. They trained a Support Vector Machine (SVM) classifier and stated that experiments on three EEG databases demonstrated the robustness of this approach. Also, findings suggested that the successive decomposition index was computationally more efficient in comparison to methods proposing wavelet decomposition and feature extraction from each sub-band.

Ref. [284] trained a convolutional neural network (CNN) to distinguish ictal, preictal, and interictal segments for epileptic seizure detection. As input to the CNN, the authors experimented with raw EEG data in the time domain and data in the frequency domain by applying Fast Fourier Transform to the EEG signals. Results suggested that the frequency domain signals achieved higher evaluation results than the ones achieved by the time domain signals.

In [285], the authors proposed an approach to optimize the parameters of the SVM classifier by using the Genetic Algorithm (GA) and the particle swarm optimization (PSO). Firstly, the authors applied discrete wavelet transform (DWT) to decompose the EEG signal into sub-bands and extracted a set of statistical features. Next, they used these features for training the SVM classifier along with GA and PSO. Findings stated that the PSO-based approach outperformed the GA.

The authors in [286] introduced an approach, where the deep neural network is hybridized with a novel Adaptive Haar Wavelet-based Binary version of Grasshopper Optimization (AHW-BGOA). This method can be used both for hyperparameter optimization and the selection of the most informative features, which are capable of enhancing the classification performance. Regarding the process of feature extraction, the authors decomposed the signal into sub-bands via the DWT and extracted a set of features, including non-linear features, hurst exponent, and entropy-based features.

A straightforward approach was introduced by [287], where the authors trained a neural network consisting of BiLSTM layers and stated that this model can predict seizure episodes reaching F1-score up to 88.00%.

A similar approach was proposed by [288], where the authors trained a neural network consisting of two LSTM layers and stated that the proposed model can attain high evaluation results. A similar approach was conducted by [289], where the authors trained an architecture consisting of two BiLSTM layers for detecting and predicting epileptic seizures.

Reference [290] adopted a deep convolutional neural network consisting of 23 layers including the input layer. The proposed deep neural network is able to detect abnormal EEG signals with an accuracy of 79.34% without requiring the tedious procedure of feature extraction.

Similarly, the authors in [117] adopted a deep convolutional neural network consisting of 13 layers for detecting normal, preictal, and seizure classes.

In [291], the authors proposed a long short-term memory (LSTM) network for classifying epileptic EEG signals. First, the authors applied DWT for decomposing the EEG signal into sub-bands and extracted a set of statistical features from each sub-band. Next, for reducing the number of features, the authors exploited feature selection and dimensionality reduction techniques. Regarding the feature selection, they employed the correlation coefficient and the P-value analysis. In terms with the dimensionality reduction, they exploited the principal component analysis (PCA). Finally, they trained the LSTM neural network and found that three features were sufficient for building an effective model for epilepsy. Concurrently, results suggested that the proposed approach outperformed traditional machine learning algorithms, including SVM, Logistic Regression (LR), etc.

Research work [102] proposed an ensemble approach to detect abnormal EEG signals. More specifically, the authors split the EEG signal into sub-signals using fixed-size overlapping windows and passed them through a deep neural network consisting of three convolutional layers followed by two fully-connected layers. Finally, the authors classified the EEG signal via a majority-voting approach of the prediction of each sub-signal. Results indicated that the proposed approach outperformed state-of-the-art systems.

Reference [292] adopted a convolutional neural network (CNN), which is capable of extracting spectral, spatial, and temporal features from EEG data and predicting abnormal EEG signals. Also, the authors exploited visualization techniques, where the clinician can see what the CNN learns. Results suggested the robustness of the introduced approach both for patient-specific and cross-patient tasks.

A different approach was proposed by [293], where the authors exploited autoencoders for predicting seizures. More specifically, the authors trained an autoencoder, where both the encoder and decoder consist of convolutional neural networks. The latent vector was passed either through a multilayer perceptron (MLP) or a BiLSTM neural network for classifying the given EEG signal into ictal or interictal. Results indicated the superiority of the BiLSTM over the MLP.

In [294], the authors introduced a deep learning approach for epilepsy detection. First, they used a five level DWT for decomposing the EEG signal into sub-bands, and eliminated the D1 and D2 coefficients. The rest of the available coefficients were used as input to a CNN for classification. Results stated that the proposed approach is comparable to the current state-of-the-art.

An interpretable approach was introduced by [295]. More specifically, the authors exploited the tiny visual geometry group CNN architecture [296] for epilepsy detection from EEG signals. In addition, the authors exploited the Gradient-weighted Class Activation Mapping method for interpreting the decisions made by the proposed network and stated that the model was able to learn sensible features associated with well-known epilepsy markers.

In [101], the authors introduced a deep learning approach to detect epileptic seizures. More specifically, the authors, applied Discrete Cosine Transform (DCT) to the EEG signals and extracted the hurst exponent and ARMA features. Finally, they trained an LSTM network and claimed that the proposed approach improved the binary classification accuracy by 2% in comparison with the previous SVM classifier.

The work proposed by [100] extracted a set of features from segments of EEG signals, including time and frequency domain features, between EEG channels cross-correlation, and graph theoretical features. Then, the authors trained an LSTM neural network and stated that the introduced approach presents a surge in the performance compared to the one obtained by traditional machine learning algorithms and convolutional neural networks.

A different approach was adopted by [297]. Firstly, the authors applied some preprocessing steps for the noise removal, including the empirical mode decomposition and the bandpass filter. For dealing with the imbalanced dataset, they exploited Generative Adversarial Neural Networks generating in this way synthetic EEG segments of preictal states. Next, they extracted a set of handcrafted features from Intrinsic Mode Functions (IMFs) and a set of automated features. Regarding the automated features, they converted the EEG signals to STFT spectrograms and used them as input to a CNN architecture. Then, they concatenated the set of handcrafted and automated features and applied feature selection techniques including Pearson Correlation Coefficient and Particle Swarm Optimization (PSO). Finally, the resulting feature set was used for training an ensemble classifier, which combines the output of SVM, CNN, and LSTM using Model-Agnostic Meta-Learning (MAML). Findings suggested that the proposed approach performs better than existing research works in terms of sensitivity, specificity, and average anticipation time.

In [298], a stacking ensemble approach based model was introduced for predicting epileptic seizures. Findings stated that the stacking ensemble approach achieved higher evaluation results than the ones achieved by the base Deep Neural Network (DNN).

A deep learning approach was also proposed by [299]. More specifically, the authors proposed a deep neural network consisting of an attention mechanism, a BiLSTM layer, a time-distributed fully connected layer, a pooling layer, a fully connected layer, and a softmax layer. According to the authors, the attention mechanism is able to capture the spatial features, while the BiLSTM layer is capable of extracting the discriminating temporal features. Results indicated that the introduced approach performed well against current state-of-the-art methods.

3.4.1 Literature Review Findings

From the aforementioned research works, it is evident that the majority of the research works have employed feature extraction techniques for training shallow machine learning classifiers or deep neural networks. Specifically, most of them apply the DWT to decompose the EEG signal into multiple sub-bands. However, the limitation of DWT is that one should select carefully the number of levels of decomposition and the mother wavelet, increasing in this way the computational time [283]. Another limitation of feature extraction is the fact that it demands some level of domain expertise rendering it a time-consuming procedure.

3.4.2 Datasets

3.4.2.1 EEG Database of the University of Bonn

This dataset [103] consists of five subsets, denoted as A, B, C, D, and E. Each subset contains 100 single channel EEG segments of 23.6 second duration. The sampling frequency is equal to 173.61 Hz. Thus, each EEG segment consists of 4097 samples. Sets A and B have been collected from five healthy volunteers having their eyes open and closed respectively. Sets C and D have been collected during interictal state (seizure-free interval). Specifically, segments in set D have been recorded from the hippocampal formation identified as epileptogenic zone, while the signals in dataset C have been recorded from hippocampal formation of opposite hemisphere of the brain. The dataset E contains segments from seizure activity (ictal state). A band-pass filter was applied to the EEG signals with 0.53 Hz and 40 Hz low and high cutoff frequencies respectively. All these segments have been manually inspected by an expert due to the muscle activity and eye movements.

3.4.2.2 Temple University EEG corpus

It includes a variety of corpora, which are publicly available [300]. Specifically, the TUH Abnormal EEG Corpus is available, where EEGs have been annotated as normal or abnormal. The TUH EEG Artifact Corpus includes annotations of 5 different artifacts, while the TUH EEG Epilepsy Corpus contains subjects with and without epilepsy. The TUH EEG Events Corpus contains annotations of EEG segments belonging into 6 classes, including artifacts, spikes and sharp waves, and more. The TUH EEG Seizure Corpus is also available, which provides information about the start and stop time of seizures as well as the seizure type. Finally, the TUH EEG Slowing Corpus is provided, which includes annotations of slowing events.

3.4.2.3 CHB-MIT Scalp EEG Database

This database includes 22 pediatric subjects [301, 302]. Specifically, 5 males of ages 3-22 and 17 females of ages 1.5-19 are included. There are 9-42 .edf files per subject, while each .edf file contains 23-26 channels. All signals have a sampling frequency of 256Hz.

3.4.2.4 Siena Scalp EEG Database

This database includes 14 epileptic patients (9 males and 5 females) [303]. A sampling frequency of 512 Hz has been used. The start and end time of seizures is provided, while three types of seizures are annotated, i.e., focal onset with and without impaired awareness, and focal to bilateral tonic-clonic (FBTC). This dataset can be used for the task of seizure prediction.

3.4.2.5 A dataset of neonatal EEG recordings with seizures annotations

This dataset comprises multi-channel EEG recordings from 79 neonates, where 39 of them have been diagnosed with neonatal seizures [304]. A sampling rate of 256 Hz has been used. Butterworth high-pass filtering has also been applied. This dataset can be used for the task of seizure detection.

Chapter 4

Methods for Recognizing Depression through Social Media posts and Spontaneous Speech

4.1 Introduction

As mentioned in Section 2.2, depression is a serious mood disorder, which affects the way people feel and perform daily activities. People use social media for expressing their thoughts and feelings through posts. Therefore, social media provide assistance for the early detection of mental health conditions. Apart from recognizing depression via social media posts, speech is a reliable biomarker for diagnosing depression, since depressed people present decreased verbal activity productivity and "lifeless" sounding speech.

In this chapter, we present two approaches for recognizing depression. Specifically, in Section 4.2 we present an approach for identifying depression through social media posts, while Section 4.3 introduces a method for recognizing depression by using spontaneous speech.

4.2 Calibration of Transformer-based Models for Identifying Depression in Social Media

Existing research initiatives exploit social media data for identifying depressive posts. The majority of these research works [1] employ feature extraction approaches and train shallow Machine Learning (ML) algorithms. Employing feature extraction approaches constitutes a tedious procedure and demands domain expertise, since the authors may not find the optimal feature set for the specific problem. At the same time, the train of shallow ML algorithms does not yield optimal performance and does not generalize well to new data. For addressing these limitations, other approaches [3] use deep neural networks, including Convolutional Neural Networks (CNNs), bidirectional long short-term memory

(BiLSTM), and so on, or transformer-based networks. In addition, there are researches employing ensemble strategies [4]. However, these approaches increase substantially the training time, since multiple models must be trained separately. In addition, recently there have been studies [5, 6] showing that transformer-based models struggle or fail to capture rich knowledge. For this reason, there have been proposed methods for enhancing these models with external information or additional modalities [7, 8, 9, 10]. However, existing research initiatives in the task of depression detection through social media have not exploited any of these approaches yet. In addition, the reliability of a machine learning model's confidence in its predictions, denoted as calibration [11, 12], is critical for high risk applications, such as deciding whether to trust a medical diagnosis prediction [156, 157, 158]. Although methods regarding the confidence of models' predictions have been introduced in many studies, including suicide risk assessment [159], sleep stage classification [160], and so on, no prior work for depression detection has taken into account the level of confidence of models' predictions, creating in this way overconfident models.

To tackle the aforementioned limitations, in this section, we propose a method, which injects extra linguistic information into transformer-based models, namely BERT and MentalBERT. Firstly, we extract various linguistic features, including NRC Sentiment Lexicon, features derived by Latent Dirichlet Allocation (LDA) topics, Top2vec, and Linguistic Inquiry and Word Count (LIWC) features. Regarding the LDA topic-based features, this is the first study in terms of the task of depression detection via social media texts utilizing the Global Outlier Standard Score (GOSS) [24], which captures the text's interest on a specific topic in comparison with other texts. After passing each text through a transformer-based model, we project the linguistic information to the same dimensionality with the outputs of the transformer models. Next, we concatenate the representations obtained by BERT (or MentalBERT) and linguistic information and apply a Multimodal Adaptation Gate [18], where an attention gating mechanism is used for controlling the importance of each representation. Similarly to [19], we modify M-BERT [18] by replacing the multimodal information with linguistic information. Finally, a shifting component is exploited for calculating the new combined embeddings. The new combined embeddings are passed through a BERT (or MentalBERT) model, where the classification [CLS] token is fed to Dense layers for getting the final prediction. In addition, for preventing models becoming too overconfident, we use label smoothing. According to Müller et al. [31], label smoothing has been used successfully to improve the accuracy of deep learning models across a range of tasks, while at the same time it implicitly calibrates learned models so that the confidences of their predictions are more aligned with the accuracies of their predictions. We use metrics for assessing both the performance and the calibration of our model. We also demonstrate the efficiency of label smoothing in both calibrating and enhancing the performance of our model. We test our proposed approaches on three publicly available datasets, which differentiate *(i)* depressive from non-depressive posts, and *(ii)* posts indicating the severity of depression, namely minimal, mild, moderate, and severe. We demonstrate the robustness of our model and advantages over state-of-the-art approaches. Finally, we conduct an extensive linguistic analysis and show differences in linguistic patterns between depressive posts and non-depressive ones.

The contributions of this section can be summarized as follows:

- We introduce a method, which injects linguistic features into transformer-based neural models.
- We perform model calibration by using label smoothing. We evaluate the calibration of our approaches by using two metrics. To the best of our knowledge, this is the first study exploiting label smoothing and utilizing calibration metrics.
- We contribute to the existing literature by performing a detailed linguistic analysis, which reveals significant differences in language between depressive and non-depressive posts.

4.2.1 Methodology

4.2.1.1 Architecture

In this section, we describe our proposed approach for detecting depressive posts in social media. Our proposed method is based on the work introduced by Rahman et al. [18], and Jin and Aletras [19]. Instead of cross-modal interactions, we inject extra linguistic information as alternative views of the data into pretrained language models. Our proposed architecture is illustrated in Fig. 4.1.

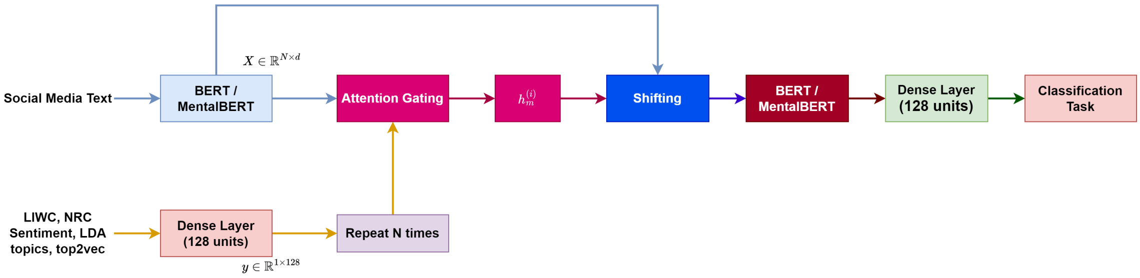


Figure 4.1: Our Proposed Architecture

Specifically, we use the following feature vectors:

- **NRC.** The NRC Emotion Lexicon is a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive) [20]. Each text is represented as a 10-d vector, where each element is the proportion of tokens belonging to each category.
- **LIWC.** LIWC is a dictionary-based approach to count words in linguistic, psychological, and topical categories [21]. We use LIWC 2022 [22] to represent each text as a 117-d vector.

- **LDA topics.** Before training the LDA model, we remove stop words and punctuation. We exploit LDA (with 25 topics) and extract 25 topic probabilities per text [23]. These probabilities describe the topics of interest of each text. Inspired by Liu et al. [24], we use the following feature vector:
  - **Global Outlier Standard Score (GOSS):** For evaluating the i^{th} text's interest on a certain topic k, compared to the rest of the texts, we use the GOSS feature:
$$\mu(x_k) = \frac{\sum_{i=1}^{n} x_{ik}}{n} \tag{4.1}$$
$$GOSS(x_{ik}) = \frac{x_{ik} - \mu(x_k)}{\sqrt{\sum_i \left(x_{ik} - \mu\left(x_k\right)\right)^2}} \tag{4.2}$$

  Therefore, each text is represented as a 25-d vector.

- **Top2Vec:** Top2Vec [25] is an algorithm for topic modelling, which automatically detects topics present in text and generates jointly embedded topic, document and word vectors. After training Top2Vec by exploiting the Universal Sentence Encoder, each text is represented as a 512-d vector.

We experiment with the following pretrained models: BERT [26] and MentalBERT [27].

First, we pass each text through the aforementioned transformer-based models. Let $C \in \mathcal{R}^{N \times d}$ be the output of the transformer-based models, where N denotes the sequence length, while d denotes the dimensionality of the models. We have omitted the dimension corresponding to the batch size for the sake of simplicity.

Then, we project the feature vectors to dimensionality equal to 128. We repeat the feature vector N times, so as to ensure that the feature vector and the output of the transformer-based models can be concatenated. Given the word representation $e^{(i)}$, we concatenate $e^{(i)}$ with feature vectors, i.e., $h_v^{(i)}$.

$$w_v^{(i)} = \sigma\left(W_{hv}[e^{(i)}; h_v^{(i)}] + b_v\right) \tag{4.3}$$

where σ denotes the sigmoid activation function, W_{hv} is a weight matrix, and $w_v^{(i)}$ corresponds to the gate. b_v is the scalar bias.

Next, we calculate a shift vector $h_m^{(i)}$ by multiplying the embeddings with the gate.

$$h_m^{(i)} = w_v^{(i)} \cdot \left(W_v h_v^{(i)}\right) + b_m^{(i)} \tag{4.4}$$

where W_v is a weight matrix and $b_m^{(i)}$ is the bias vector.

Next, we apply the Multimodal Shifting component aiming to dynamically shift the word representations by integrating the shift vector $h_m^{(i)}$ into the original word embedding.

$$e_m^{(i)} = e^{(i)} + \alpha h_m^{(i)} \tag{4.5}$$

$$\alpha = min\left(\frac{||e^{(i)}||_2}{||h_m^{(i)}||_2}\beta, 1\right) \tag{4.6}$$

, where β is a hyperparameter. Then, we apply a layer normalization [28] and dropout layer [29] to $e_m^{(i)}$. Next, the combined embeddings are fed to a BERT/MentalBERT model.

We get the classification [CLS] token of this model and pass it through a Dense layer consisting of 128 units with a ReLU activation function. Finally, we use a dense layer consisting of either two units (binary classification task) or four units (multiclass classification task).

We denote our proposed models as Multimodal BERT (M-BERT) and Multimodal MentalBERT (M-MentalBERT) followed by the linguistic features which are integrated into them. For example, the injection of LIWC features into a BERT model is denoted as M-BERT (LIWC).

4.2.1.2 Model Calibration

To prevent the model becoming too overconfident, we use label smoothing [30, 31]. Specifically, label smoothing calibrates learned models so that the confidences of their predictions are more aligned with the accuracies of their predictions.

For a network trained with hard targets, the cross-entropy loss is minimized between the true targets y_k and the network's outputs p_k, as in $H(y,p) = \sum_{k=1}^{K} -y_k log(p_k)$, where y_k is "1" for the correct class and "0" for the other. For a network trained with label smoothing, we minimize instead the cross-entropy between the modified targets $y_k^{LS_u}$ and the network's outputs p_k.

$$y_k^{LS_u} = y_k \cdot (1-\alpha) + \frac{\alpha}{K} \tag{4.7}$$

$$H(y,p) = \sum_{k=1}^{K} -y_k^{LS_u} \cdot \log(p_k) \tag{4.8}$$

, where α is the smoothing parameter and K is the number of classes.

4.2.2 Experiments

4.2.2.1 Datasets

Depression_Mixed. We use the dataset described in Section 3.1.2.1.
Depression_Severity. We use the dataset described in Section 3.1.2.2.

4.2.2.2 Experimental Setup

We use the Adam optimizer with a learning rate of 0.001. We apply *StepLR* with a step size of 5 and a gamma of 0.1. We use batch size of 8. With regards to *Depression_Mixed* dataset, we split the dataset into a train and a test set ($80\% - 20\%$) similar to Ansari et al. [4]. Regarding *Depression_Severity* dataset, we use 5-fold stratified cross-validation, since the study [17] has also exploited cross-validation. All train sets are divided into a train and a validation set. Regarding *Depression_Severity* dataset, we apply *EarlyStopping* with a patience of 7 epochs based on the validation loss. In terms of the *Depression_Mixed* dataset, we train our introduced model for a maximum of 30 epochs, choose the epoch with the smallest validation loss, and test the model on the test set. We set β of Eq. 4.6 equal to 0.0001.[1] We choose α of Eq. 4.7 equal to 0.001. We use the Python library, namely Transformers [305], for BERT and MentalBERT. Specifically, we use the BERT base uncased version and the MentalBERT base uncased version. We use PyTorch [306] for performing our experiments. All experiments are trained on a single Tesla P100-PCIE-16GB GPU.

4.2.2.3 Evaluation Metrics

Performance In terms of the binary classification task, i.e., 0 for non-depressive and 1 for depressive texts, we use Precision, Recall, F1-score, and Accuracy to evaluate the performance of our proposed approach. We use these metrics similar to Wani et al. [184].

Regarding multiclass classification task reported on Depression_Severity dataset, we use Weighted Precision, Weighted Recall, and Weighted F1-score. We use these metrics similar to Mishra et al. [307].

Calibration We evaluate the calibration of our model using the metrics proposed by relevant literature [308, 309, 171]. Specifically, we use the metrics mentioned below:

- **Expected Calibration Error (ECE).** The calibration error is the difference between the fraction of predictions in the bin that are correct (accuracy) and the mean of the probabilities in the bin (confidence). First, we divide the predictions into M equally spaced bins (size $1/M$).

$$acc(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} 1(\hat{y}_i = y_i) \tag{4.1}$$

$$conf(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i \tag{4.2}$$

  , where y_i and $\hat{y}_i$ are the true and predicted labels for the sample i and $\hat{p}_i$ is the confidence (predicted probability value) for sample i.

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} |acc(B_m) - conf(B_m)| \tag{4.3}$$

[1] We experimented with values of β, including 0.01 and 0.001, but setting β equal to 0.0001 yielded the best results.

, where N is the total number of data points and B_m is the group of samples whose predicted probability values falls into the interval $I_m = \left(\frac{m-1}{M}, \frac{m}{M}\right]$.

Perfectly calibrated models have an ECE of 0.

- **Adaptive Calibration Error (ACE).** Adaptive Calibration Error uses an adaptive scheme which spaces the bin intervals so that each contains an equal number of predictions.

$$ACE = \frac{1}{KR}\sum_{k=1}^{K}\sum_{r=1}^{R}|acc(r,k) - conf(r,k)| \tag{4.4}$$

, where $acc(r, k)$ and $conf(r, k)$ are the accuracy and confidence of adaptive calibration range r for class label k, respectively; and N is the total number of data points. Calibration range r is defined by the $[N/R]$th index of the sorted and thresholded predictions.

4.2.2.4 Baselines

We use the following baselines as comparisons with our proposed approaches:

- **BERT, MentalBERT:** We fine-tune these pretrained language models in order to explore whether our method of injecting linguistic information to pretrained models leads to performance improvement.

  In terms of Depression_Mixed dataset, we report the performance of BERT obtained by Yang et al. [310]. We finetune MentalBERT and report its performance on this dataset.

  Regarding Depression_Severity dataset, we finetune BERT and MentalBERT and report their performances.

  We do not report calibration metrics for these models, since our goal in this case is to compare only the performances of these models with our proposed approaches.

- **Proposed Approaches (without label smoothing):** We train the proposed models introduced in Section 4.2.1 without label smoothing. We explore whether label smoothing leads to performance improvement and better calibration of our models.

4.2.3 Results

The results of our proposed approach are reported in Tables 4.1 and 4.2. Specifically, Table 4.1 reports the performances of our proposed approaches on the Depression_Mixed dataset, while Table 4.2 reports the results on the Depression_Severity dataset.

Regarding the Depression_Mixed dataset, we first compare our proposed approaches without label smoothing with the BERT and MentalBERT models. We observe that the

Table 4.1: Performance comparison among proposed models and baselines using the DEPRESSION_MIXED dataset

Model	Prec.	Rec.	F1-score	Acc.	ECE	ACE
Baselines						
BERT	91.40	91.40	91.40	-	-	-
MentalBERT	89.27	93.14	91.17	91.15	-	-
Baselines - Proposed Approaches (without label smoothing)						
M-BERT (NRC)	90.56	91.84	91.20	91.15	0.072	0.081
M-BERT (LIWC)	90.98	92.02	91.49	92.04	0.054	0.055
M-BERT (LDA topics)	88.07	95.80	91.77	92.04	0.071	0.071
M-BERT (top2vec)	90.97	92.99	91.97	92.21	0.057	0.069
M-MentalBERT (NRC)	90.65	92.65	91.64	91.86	0.031	0.054
M-MentalBERT (LIWC)	93.49	87.78	90.55	91.50	0.057	0.056
M-MentalBERT (LDA topics)	87.97	93.09	90.46	90.44	0.089	0.086
M-MentalBERT (top2vec)	91.63	93.77	92.69	93.27	0.058	0.054
Proposed Approaches (with label smoothing)						
M-BERT (NRC)	89.82	94.81	92.25	92.39	0.059	0.065
M-BERT (LIWC)	93.06	91.78	92.41	92.21	0.034	0.044
M-BERT (LDA topics)	90.16	92.71	91.42	92.39	0.063	0.067
M-BERT (top2vec)	90.34	94.93	92.58	92.57	0.049	0.056
M-MentalBERT (NRC)	91.44	92.52	91.98	92.74	0.042	0.057
M-MentalBERT (LIWC)	94.96	89.42	92.11	92.57	0.055	0.057
M-MentalBERT (LDA topics)	94.81	90.78	92.75	92.92	0.047	0.049
M-MentalBERT (top2vec)	96.12	90.18	93.06	93.45	0.033	0.043

Table 4.2: Performance comparison among proposed models and baselines using the DEPRESSION_SEVERITY dataset.

Model	W. Prec.	W. Rec.	W. F1-score	ECE	ACE
Baselines					
BERT	72.99	71.97	71.00	-	-
MentalBERT	73.35	70.81	71.67	-	-
Baselines - Proposed Approaches (without label smoothing)					
M-BERT (NRC)	74.48	70.08	69.96	0.107	0.076
M-BERT (LIWC)	73.77	71.74	72.13	0.110	0.078
M-BERT (LDA topics)	74.25	71.80	71.28	0.114	0.079
M-BERT (top2vec)	72.93	71.97	71.00	0.086	0.071
M-MentalBERT (NRC)	74.43	72.58	69.96	0.097	0.069
M-MentalBERT (LIWC)	72.39	72.53	71.95	0.112	0.075
M-MentalBERT (LDA topics)	73.83	72.58	72.58	0.118	0.078
M-MentalBERT (top2vec)	74.63	72.39	72.06	0.103	0.075
Proposed Approaches (with label smoothing)					
M-BERT (NRC)	74.04	72.84	72.81	0.102	0.074
M-BERT (LIWC)	73.68	72.16	72.37	0.094	0.069
M-BERT (LDA topics)	73.24	71.46	71.42	0.112	0.078
M-BERT (top2vec)	73.36	72.64	72.30	0.113	0.074
M-MentalBERT (NRC)	73.03	71.23	71.46	0.112	0.079
M-MentalBERT (LIWC)	73.21	73.15	72.43	0.099	0.071
M-MentalBERT (LDA topics)	73.74	73.23	73.16	0.111	0.075
M-MentalBERT (top2vec)	73.68	72.70	72.67	0.094	0.071

injection of linguistic features, except for NRC features, into the BERT model improves the F1-score. Specifically, we observe that the injection of top2vec features yields the highest F1-score and Accuracy accounting for 91.97% and 92.21% respectively, surpassing the performance of the BERT model in F1-score by 0.57%. We speculate that the injection of top2vec features obtains better performance than the injection of features derived by LDA topics, i.e., GOSS features, since the top2vec algorithm is capable of identifying the number of topics automatically. In terms of MentalBERT, we observe that the injection of top2vec features obtains an F1-score of 92.69% surpassing MentalBERT by 1.52%. We observe that the integration of NRC and top2vec features improves the performance obtained by MentalBERT. Regarding the proposed approaches with label smoothing, we observe that these models attain better performances than the ones obtained by the models without label smoothing. Specifically, we observe that M-BERT (top2vec) with label smoothing surpasses the respective model without label smoothing in F1-score and Accuracy by 0.61% and 0.36% respectively. Similarly, M-MentalBERT (top2vec) with label smoothing obtains the highest F1-score and Accuracy accounting for 93.06% and 93.45% respectively. This model surpasses the respective model without label smoothing in F1-score and Accuracy by 0.37% and 0.18%. Except for the improvement of the performance metrics, i.e., Precision, Recall, F1-score, and Accuracy, we observe that the models with label smoothing obtain better results in terms of the calibration metrics, i.e., ECE and ACE, than the ones obtained by the models without label smoothing. For example, we observe that M-BERT (top2vec) with label smoothing improves the ECE and ACE scores obtained by M-BERT (top2vec) without label smoothing by 0.008 and 0.013 respectively. Similarly, M-MentalBERT (LDA topics) with label smoothing improves the ECE and ACE scores obtained by M-MentalBERT (LDA topics) without label smoothing by 0.042 and 0.043 respectively.

With regards with the Depression_Severity dataset, we first compare our proposed approaches without label smoothing with the BERT and MentalBERT models. We observe that the integration of LIWC features and features extracted by LDA topic modelling, i.e., GOSS features, into the BERT model leads to a performance surge in comparison with the BERT model. Specifically, M-BERT (LIWC) outperforms BERT in weighted F1-score by 1.13%. At the same time, the integration of all the features, except NRC, to a MentalBERT model yields to a performance improvement compared to the MentalBERT model. Specifically, M-MentalBERT (LDA topics) attains the highest weighted F1-score accounting for 72.58% surpassing MentalBERT by 0.91%. When it comes to proposed models with label smoothing, we observe an improvement in both the performance metrics and calibration ones. More specifically, the integration of NRC features to a BERT model obtains a weighted F1-score of 72.81% outpeforming BERT by 1.81%, M-BERT (NRC) without label smoothing by 2.85%, and M-BERT (LIWC) without label smoothing by 0.68%. In addition, M-MentalBERT (LDA topics) with label smoothing obtains the highest F1-score accounting for 73.16% surpassing MentalBERT by 1.49% and M-MentalBERT (LDA topics) without label smoothing by 0.58%. In terms of the calibration metrics, we observe that both ECE and ACE scores are improved when we apply label smoothing. For example, M-BERT (LIWC) with label smoothing obtains an ECE score of 0.094 and an ACE score of 0.069, which are improved by 0.016 and 0.009 respectively compared to the respective model without label smoothing.

4.2.4 Linguistic Analysis

We finally perform an analysis on the Depression_Mixed dataset to uncover the peculiarities of depression. Specifically, we seek to find the correlations of LIWC features

Table 4.3: LIWC Features associated with depressive and non-depressive posts, sorted by point-biserial correlation. All correlations are significant at $p < 0.05$ after Benjamini-Hochberg correction.

Depression_Mixed			
Non-Depressive		**Depressive**	
LIWC	**corr.**	**LIWC**	**corr.**
Tone	0.3156	health	0.4108
Clout	0.3022	mental health	0.3674
Social referents	0.2914	physical	0.3603
shehe	0.2634	emo_sad	0.3506
we	0.2415	tone_neg	0.3274
social	0.2401	1st person singular	0.2974
male references	0.2199	Authentic	0.2961
affiliation	0.1960	cognition	0.2957
female references	0.1923	emo_neg	0.2843
conversation	0.1794	cognitive processes	0.2601
netspeak	0.1741	feeling	0.2507
culture	0.1732	focuspresent	0.2139
allpunc	0.1667	insight	0.2138
family	0.1589	emotion	0.2120
technology	0.1524	negations	0.2116
exclam	0.1519	verb	0.2076
analytic	0.1507	linguistic	0.1893
period	0.1458	death	0.1842
drives	0.1168	function	0.1834
OtherP	0.1156	all-or-none	0.1748
number	0.1075	dic	0.1708
assent	0.1013	affect	0.1609
tone_pos	0.0998	adverb	0.1588
leisure	0.0980	illness	0.1584
Social behavior	0.0937	emo_anx	0.1458
communication	0.0920	auxverb	0.1442
lifestyle	0.0917	discrepancy	0.1373
friend	0.0812	apostro	0.1256
curiosity	0.0792	want	0.1146
you	0.0774	achieve	0.1136
determiners	0.0758	pronoun	0.1092
politic	0.0751	lack	0.1054
relig	0.0673	differ	0.1004
focusfuture	0.0666	prepositions	0.0951
visual	0.0660	risk	0.0928
motion	0.0647	allure	0.0923
money	0.0626	causation	0.0916
ethnicity	0.0591	tentative	0.0880
article	0.0493	time	0.0811
emo_pos	0.0480	personal pronouns	0.0801
home	0.0420	impersonal pronoun	0.0789
food	0.0401	perception	0.0747
words per sentence (WPS)	0.0391	swear	0.0697
Nonfluencies	0.0385	substances	0.0686
-	-	memory	0.0673
-	-	BigWords	0.0588
-	-	adj	0.0575
-	-	certitude	0.0547
-	-	wellness	0.0457
-	-	moral	0.0433
-	-	conflict	0.0411
-	-	acquire	0.0394
-	-	QMark	0.0384

with depressive and non-depressive texts. First, we normalize LIWC features, so as to ensure that they sum up to 1 across each post. Next, we use the point-biserial correlation between each LIWC category and the label of the post. The output of the point-biserial correlation is a number ranging from -1 to 1. Positive correlations mean that the specific LIWC category is correlated with the depressive class (label 1), while negative correlations mean that the specific LIWC category is correlated with the non-depressive class (label

0). We consider the absolute values of the correlations. Results are reported in Table 4.3. All the correlations are significant at $p < 0.05$ with Benjamini-Hochberg correction [69] for multiple comparisons.

In terms of the *Depression_Mixed* dataset, we observe that the control group tends to use words with positive tone and emotion, i.e., good, well, happy, hope, and so on. In addition, healthy control group discusses topics of the everyday life, including lifestyle (work, home, school), culture (car, phone), politics (govern, congress), family, and friends (boyfriend, girlfriend, dude). Also, these people make plans for the future, thus use words indicating focus on the future (correlation equal to 0.0666). However, it must be noted that this is a very weak correlation. On the other hand, people with depression focus on the present and do not make plans for the future. They discuss about negative topics, including death, illnesses, mental health, and substances. This can be justified by the fact that people with depression often have tendencies to suicide and believe that they cannot achieve anything. In addition, they use swear words, i.e., shit, fuck, damn, since they think that everything goes wrong in their life. Also, their posts are full of sadness, anxiety, and negative tone.

4.2.5 Discussion

Our study contributes to the literature by introducing the first approach of integrating extra linguistic information into pretrained language models based on transformers, namely BERT and MentalBERT. Specifically, we adapt M-BERT [18] by replacing multimodal information with linguistic information. To be more precise, we extract NRC, LIWC, features derived by LDA topics, and top2vec features. We apply a Multimodal Adaptation Gate and exploit also a Shifting component for creating new combined embeddings which are given as input to BERT (and MentalBERT) models. In addition, motivated by the fact that in real-world decision making systems, classification networks must not only be accurate, but also should indicate when they are likely to be incorrect, we apply label smoothing and evaluate our proposed approaches both in terms of classification and calibration.

Therefore, our study is different from the state-of-the-art approaches described in Section 3.1, since:

- Prior works having proposed multimodal, multitask, ensemble strategies in conjunction with transformer-based models, have just fine-tuned these pretrained transformer-based models instead of using some modifications of them. Thus, this study is the first attempt to inject extra knowledge into BERT (and MentalBERT), in order to enhance its performance.

- All the prior works evaluate only the classification performance of their approaches neglecting the confidence of the prediction. To tackle this, this is the first study in the task of depression detection through social media posts utilizing label smoothing and evaluating both the classification performance and the calibration of the models.

- Finally, this is the first study utilizing features derived by LDA topics, namely the Global Outlier Standard Score, which captures the text's interest compared to other texts.

From the results of this study, we found that:

- *Finding 1:* The integration of linguistic features into transformer-based models yields to an increase in the classification performance. However, it is worth noting that in some cases this improvement is limited. For instance, the integration of LIWC features into the MentalBERT model with label smoothing obtains better performance than MentalBERT in F1-score by 3.36% and better performance than M-MentalBERT (LIWC) without label smoothing in F1-score by 0.63%. However, we believe that even a small improvement can make a difference.

- *Finding 2:* Label smoothing improves both the performance and the calibration of the proposed approaches. The calibration of the proposed approaches is measured via two metrics, namely Expected Calibration Error and Adaptive Calibration Error.

- *Finding 3:* Findings from a linguistic analysis reveal that people in depressive conditions use words belonging to specific LIWC categories more frequently than others.

There are several limitations related to this study.

- *Hyperparameter Tuning:* Due to limited access to GPU resources, we were not able to perform hyperparameter tuning. On the contrary, we tried some combinations of parameters. We believe that the adoption of the hyperparameter tuning procedure through the access to GPU resources would increase further the classification performance.

- *Explainability:* The present study is not accompanied with explainability techniques, i.e., Integrated Gradients [311], and so on. Therefore, we aim to apply explainability techniques in the future.

- Due to limited access to GPU resources and similarly to prior work [184, 4, 186], we were not able to perform multiple runs for testing for statistical significance.

4.3 A Cross-Attention Layer coupled with Multimodal Fusion Methods for Recognizing Depression from Spontaneous Speech

Existing research works rely on the extraction of handcrafted features and the train of traditional machine learning classifiers or deep learning approaches [32, 33, 34]. However, extracting features is a timely procedure requiring expertise on the specific topic. Additionally, the majority of research studies uses unimodal approaches for predicting depression using mainly speech [35]. Although there are studies employing multimodal models, these studies employ early [36, 37], intermediate [38, 39], or late fusion [40, 41] strategies. In the early fusion strategy, representation vectors of the modalities are concatenated at the input level, while in the intermediate fusion, the representation vectors are concatenated during training, thus equal importance is assigned to the modalities. In the late fusion strategy, unimodal models are trained independently and decision voting is applied, i.e., majority voting. The inter-modal interactions cannot be captured through these approaches. In addition, the majority of research works have tested their approaches only in English language, thus the acoustic and phonetic content of data might differ in other languages. Finally, to the best of our knowledge, no study has experimented with predicting depression, age, education level, and gender at the same time.

To tackle these limitations, we present a new method for detecting depression from spontaneous speech in the Italian language. Specifically, we feed each transcript into a pretrained Italian BERT model. Each speech signal is transformed into an image of three channels, namely log-Mel spectrogram, delta, and delta-delta. Each image is passed through a pretrained AlexNet [44] model. Next, the textual and image representations are passed through a cross-attention scaling layer. Finally, we employ and compare a variety of multimodal fusion methods, including Multimodal Factorized Bilinear Pooling (MFB), Multimodal Factorized High-order pooling (MFH) [86], and more, for fusing the outputs of the cross-attention scaling layer and predicting depression. Additionally, we introduce multi-task learning (MTL) architectures to explore if gender, age, and education level as auxiliary tasks help the primary task (depression recognition). Results demonstrate the effectiveness of the proposed approach via an extensive ablation study, as well as multiple advantages over state-of-the-art approaches.

The main contributions of this section can be summarized as follows:

- We introduce a method which includes a cross-attention layer and multimodal fusion approaches.
- We perform multi-task learning experiments to explore whether the prediction of gender, age, and education level lead towards the increase of depression detection's performance.
- We compare our approaches with competitive baselines, including shallow machine learning classifiers and deep learning.
- We perform an extensive ablation study to verify the effectiveness of the proposed approach.

4.3.1 Proposed Methodology

In this section, we describe our proposed methodology for recognizing depression from spontaneous speech. Fig. 4.2 illustrates our proposed architecture.

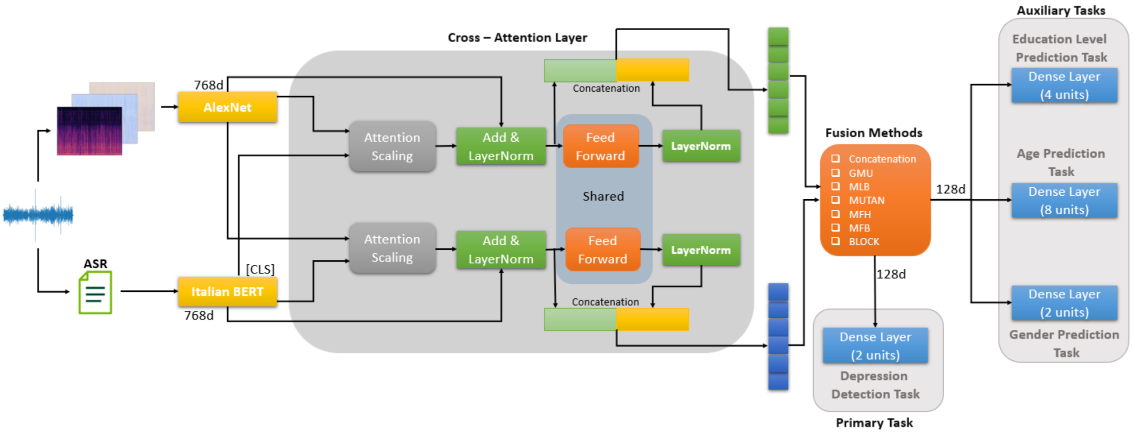


Figure 4.2: Our Proposed Methodology

4.3.1.1 Single - Task Learning

Text Processing: Since data are in Italian language, we employ Italian BERT[2]. Firstly, each transcript is passed through the Italian BERT tokenizer, where input_ids and attention mask are returned. Transcripts are padded to a maximum length of 512 tokens, while transcripts with number of tokens greater than 512 are truncated. Next, the input_ids and attention mask are fed to the Italian BERT model. Let $f^t \in \mathbb{R}^{1 \times d}$, corresponding to the [CLS] token, be the transcript representation, where $d = 768$.

Speech Processing: We use the Python library librosa [312] for converting the speech signals into images consisting of three channels, namely log-Mel spectrogram, delta, and delta-delta. We use 224 Mel bands, hop length equal to 512, and a Hanning window. Each image is resized to (224×224) pixels. We pass each image through a pretrained *AlexNet* [44] model. Let $f^v \in \mathbb{R}^{1 \times d}$ be the image representation, where $d = 768$.

Cross-Attention Layer: Motivated by [45], we design a cross-attention layer, which returns a pair of scalars, one for each modality. This pair of scalars allows for scaling the two modalities with respect to each other. One modality is used as a query for the attention of the other.

In terms of the textual modality, let $Q_i = FC_q^t\left(f^v\right)$, $K_t = FC_k^t\left(f^t\right)$, and $V_t = FC_v^t\left(f^t\right)$. The scaling value, denoted as S_t can be calculated as follows:

$$S_t = sigmoid\left(\frac{Q_i \cdot K_t^T}{\sqrt{d}}\right)$$

. In terms of the image modality, let $Q_t = FC_q^i\left(f^t\right)$, $K_i = FC_k^i\left(f^v\right)$, and $V_i = FC_v^i\left(f^v\right)$. The scaling value, denoted as S_i can be calculated as follows:

$$S_i = sigmoid\left(\frac{Q_t \cdot K_i^T}{\sqrt{d}}\right)$$

. The outputs of the attention mechanism can be calculated as $S_t \times V_t$ and $S_i \times V_i$. Note that $FC_q^t, FC_k^t, FC_v^t, FC_q^i, FC_k^i, FC_v^i \in \mathbb{R}^{d \times d}$.

[2] https://github.com/dbmdz/berts

Similar to [46], we use residual connections followed by layer normalization, as described via the equations below:

$$\hat{E_t} = LayerNorm\left(S_t \times V_t + f^t\right)$$

,

$$\hat{E_i} = LayerNorm\left(S_i \times V_i + f^v\right)$$

.

Next, we pass $\hat{E_t}$ and $\hat{E_i}$ through two shared fully connected feed-forward networks with a ReLU activation function in between, as follows:

$$\hat{E_t}' = LayerNorm\left(FC_m^n\left(ReLU\left(FC_p^q\left(\hat{E_t}\right)\right)\right)\right)$$

,

$$\hat{E_i}' = LayerNorm\left(FC_m^n\left(ReLU\left(FC_p^q\left(\hat{E_i}\right)\right)\right)\right)$$

, where $FC_p^q \in \mathbb{R}^{d \times 4d}$, $FC_m^n \in \mathbb{R}^{4d \times d}$.

Next, we concatenate $\hat{E_t}$ and $\hat{E_t}'$ (similarly $\hat{E_i}$ and $\hat{E_i}'$) into one single vector, i.e.,

$$\hat{E_t}'' = [\hat{E_t}, \hat{E_t}']$$

,

$$\hat{E_i}'' = [\hat{E_i}, \hat{E_i}']$$

, where $\hat{E_t}'', \hat{E_i}'' \in \mathbb{R}^{2d}$.

Fusion Methods: Next, we employ a variety of fusion methods, which are described in detail below, so as to fuse $\hat{E_t}''$ and $\hat{E_i}''$ in one single vector:

- Concatenation: We concatenate $\hat{E_t}''$ and $\hat{E_i}''$ into one single vector, i.e., $z \in \mathbb{R}^{4d}$. We use a dropout layer with a rate of 0.4. We use a dense layer of 128 units.

- Gated Multimodal Unit (GMU): We adopt the method introduced in [75], which controls the information flow of the two modalities towards the final classification. The equations govering the GMU are described as follows: $h^t = \tanh\left(W^t \hat{E_t}'' + b^t\right), h^v = \tanh\left(W^v \hat{E_i}'' + b^v\right), z = \sigma(W^z[\hat{E_t}''; \hat{E_i}''] + b^z), h = z * h^t + (1 - z) * h^v$, where $W^t, W^v, W^z \in \mathbb{R}^{128}$ denote the learnable parameters, and [.;.] the concatenation operation. h is the output of the GMU.

- MUTAN decomposition [85]

- Multimodal Low-rank Bilinear (MLB) pooling [313]

- MFB [86]

- MFH [86]: It is based on cascading two MFB blocks.

- BLOCK [87]: This method is based on the block-term tensor decomposition [314] and combines the strengths of the Candecomp/PARAFAC (CP) [315] and Tucker decompositions.

The output of the aforementioned fusion methods corresponds to a vector with dimensionality accounting for 128.

Output Layer: Finally, we use a dense layer consisting of two units, which gives the final prediction. The cross-entropy loss function is minimized.

4.3.1.2 Multi - Task Learning

According to research, gender [316], age[3], and education level [317] are linked with depression. In this section, we design a multi-task learning framework consisting of a primary task, i.e., depression detection (binary classification), and auxiliary tasks, i.e., gender recognition (binary classification), estimation of education level (multiclass classification), and age prediction (multiclass classification). In this approach, we explore if the auxiliary tasks help the primary task in increasing its performance. As illustrated in Fig. 4.2, in terms of gender recognition, we add a dense layer consisting of two units. In terms of education level recognition, we add a dense layer consisting of four units. Regarding the age prediction, we define the following age groups: [19,25],[26,32],[33,39],[40,46],[47,53],[54,60],[61,67],[68,71]. Thus, we add a dense layer consisting of 8 units.

All the tasks are learnt simultaneously and updated by the following loss function:

$$L = (1 - \alpha - \beta - \gamma) \cdot L_{depression} + \alpha \cdot L_{gender} + \beta \cdot L_{education} + \gamma \cdot L_{age}$$

, where $L_{depression}, L_{gender}$, $L_{education}$, and L_{age} correspond to the cross-entropy loss function. α, β, γ are hyperparameters denoting the importance we place to each task.

4.3.2 Experiments

4.3.2.1 Dataset

We use the Androids corpus, which is described in Section 3.2.6.1, for performing our experiments. We use data of the interview task. Due to the fact that manual transcripts are not provided, we use whisper large-v3 [43], in order to produce automatic transcripts.

4.3.2.2 Baselines.

We compare our approaches with the following baselines:

- *Only transcript:* We use a pretrained Italian BERT model and a learning rate of 1e-5.

[3] https://www.nhs.uk/mental-health/conditions/depression-in-adults/causes

- *Only Speech signal:* Each speech signal is represented as an image and fed into a pretrained AlexNet model. A learning rate of 1e-5 is employed.
- *BS1* [42]: This approach segments the audio signal into analysis windows of 25ms length and extracts features per window. SVM classifier is trained.
- *BS2* [42]: After calculating the feature sets per analysis windows as above, this approach segments the speech signal into frames of length equal to 128 and passes each frame through an LSTM layer. A majority vote approach is adopted.
- eGeMAPSv02 features (functional): This method trains a SVM classifier. We use the openSMILE Python toolkit [318].
- ComParE_2016 features (functional): This method trains a SVM classifier. We use the openSMILE Python toolkit [318].

4.3.2.3 Experimental Setup

In [42], the split of the participants into subsets to be used for a 5-fold setup is provided. In our study, we repeat the experiments four times and report the average and standard deviation over four runs. For Italian BERT and AlexNet, the learning rate is set to 1e-5, while for the rest layers, the learning rate is set to 1e-4. We train our models for 40 epochs with a batch size of 4. In terms of the MTL setting, we set $\alpha = \beta = \gamma = 0.1$. We use PyTorch for performing our experiments. All experiments are performed on a single Tesla P100-PCIE-16GB GPU with the running time ranging from 1 hour to 1.5 hours. For significance testing, we use the Almost Stochastic Order (ASO) test [47, 48] as implemented by [49]. Specifically, the ASO test determines whether a stochastic order [50] exists between two models, i.e., A and B. A score (ϵ_{min}) is calculated representing how far the first is from being significantly better than the second. When $\epsilon_{min} = 0$, then A is truly stochastically dominant over B. When $\epsilon_{min} < 0.5$, A is almost stochastically dominant over B. For $\epsilon_{min} = 0.5$, no order can be determined.

Evaluation Metrics. Precision, Recall, F1-score, Accuracy, and Specificity are used to evaluate the performance of the introduced approaches.

4.3.3 Results

Results are reported in Table 4.4. We observe that the usage of *BLOCK* as fusion method leads to the best performing model outperforming the rest approaches in Accuracy and F1-score by 1.21-21.99% and 1.32-22.23% respectively. Multimodal models perform better than unimodal ones verifying our initial hypothesis that the usage of multiple modalities improves detection performance. The concatenation mechanism achieves the worst results compared with the other fusion methods, since it assigns equal importance to each individual modality. We believe that MFB outperforms MFH, since the MFH method is developed by cascading two MFB blocks, thus appears to be complex for our limited dataset. We hypothesize that GMU achieves a poor performance, since it controls the information flow without capturing so effectively the cross-modal interactions. We observe that single-task learning settings perform better than multi-task learning ones. This can be justified by the fact that depression is a mental disorder, which can happen to anyone. There are many causes of depression, e.g. stressful events, personality, health issues (cancer), loneliness, etc. According to statistical test, our best performing model is almost stochastically dominant in terms of accuracy over all the approaches, except for *Only speech signal*, where $\epsilon_{min} = 0$. We are not able to perform statistical tests with [42],

Table 4.4: Performance comparison among proposed models and baselines. Reported values are mean ± standard deviation. Results are averaged across four runs (5-fold setting). $(*)$ means that $\epsilon_{min} < 0.1$, $\dagger$ means that $\epsilon_{min} < 0.2$, $\ddagger$ means that $\epsilon_{min} < 0.3$, $**$ means that $\epsilon_{min} < 0.4$, and $\dagger\dagger$ means that $\epsilon_{min} < 0.5$. We are not able to perform statistical test regarding baselines in [42], since the authors have not provided the results obtained over individual folds.

	Evaluation metrics				
Architecture	**Precision**	**Recall**	**F1-score**	**Accuracy**	**Specificity**
Unimodal approaches					
Only transcript	$94.72^{\ddagger}$ ±5.38	91.78^{**} ±5.77	$93.04^{\dagger}$ ±3.77	$92.49^{\ddagger}$ ±3.97	93.51^{**} ±6.96
Only Speech signal	80.73^{*} ±12.12	85.70^{*} ±9.57	82.49^{*} ±8.51	80.52^{*} ±8.97	74.21^{*} ±16.87
eGeMAPSv02	79.05^{*} ±13.50	85.46^{*} ±7.92	81.67^{*} ±9.69	80.29^{*} ±10.11	76.64^{*} ±15.26
ComParE_2016	86.03^{*} ±8.92	92.29 ±3.96	88.82^{*} ±5.31	87.97^{*} ±4.93	$84.92^{\dagger}$ ±9.49
Baselines reported in [42]					
BS1	73.50 ±16.10	74.50 ±13.20	73.60 ±13.60	73.30 ±10.60	– –
BS2	85.80 ±3.10	86.10 ±2.70	84.70 ±0.90	83.90 ±1.30	– –
Single - Task Learning					
Concatenation	91.51^{*} ±8.74	93.35 ±5.99	$92.11^{\dagger}$ ±5.54	$91.46^{\dagger}$ ±6.05	$90.91^{\dagger}$ ±10.48
GMU	94.10^{**} ±9.51	93.41 ±6.61	93.38^{**} ±6.25	$92.34^{\ddagger}$ ±7.22	92.33^{**} ±11.91
MLB	95.95 ±7.69	91.82^{**} ±6.31	93.57^{**} ±5.37	92.96^{**} ±5.94	95.33 ±9.71
MUTAN	$93.75^{\ddagger}$ ±8.76	94.46 ±5.57	93.82^{**} ±5.71	92.75^{**} ±6.79	90.78^{**} ±13.07
MFH	95.04^{**} ±6.62	$92.79^{\dagger\dagger}$ ±5.01	93.75^{**} ±4.46	$92.94^{\ddagger}$ ±5.56	91.28^{**} ±17.76
MFB	94.68^{**} ±8.19	93.63 ±4.63	93.95^{**} ±5.32	93.18^{**} ±6.13	92.53^{**} ±10.66
BLOCK	**97.30** ±4.43	**94.52** ±4.52	**95.83** ±3.81	**95.29** ±4.23	**96.42** ±6.04
Multi-Task Learning					
Gender, Education, Age	96.14 ±5.02	93.24 ±6.95	$94.38^{\dagger\dagger}$ ±3.65	$94.08^{\dagger\dagger}$ ±3.45	96.31 ±4.86
Gender, Education	97.22 ±5.14	$92.28^{\dagger\dagger}$ ±6.82	$94.51^{\dagger\dagger}$ ±4.65	$94.07^{\dagger\dagger}$ ±5.03	95.95 ±9.35
Education, Age	94.41^{**} ±7.24	93.63 ±5.97	93.74^{**} ±4.52	$93.62^{\dagger\dagger}$ ±4.48	$93.56^{\dagger\dagger}$ ±8.05
Gender, Age	96.55 ±4.87	$92.51^{\dagger\dagger}$ ±6.09	$94.30^{\dagger\dagger}$ ±3.72	$93.84^{\dagger\dagger}$ ±4.25	94.53 ±13.05
Gender	94.61^{**} ±9.28	93.29 ±7.18	93.61^{**} ±6.51	93.20^{**} ±6.81	$93.68^{\dagger\dagger}$ ±10.63
Education	94.22^{**} ±9.16	93.04 ±7.27	$93.44^{\ddagger}$ ±7.34	93.00^{**} ±7.31	92.03^{**} ±12.41
Age	94.99^{*} ±7.46	$92.32^{\dagger\dagger}$ ±6.72	$93.34^{\ddagger}$ ±5.09	$92.56^{\ddagger}$ ±5.85	$93.42^{\dagger\dagger}$ ±10.79

since the results obtained over individual folds are not available.

4.3.4 Ablation Study

In this section, we perform a series of ablation experiments to explore the effectiveness of the best performing architecture. Results are reported in Table 4.5. Firstly, we experiment with removing both the cross-attention layer and the fusion methods. Results show that a decrease of Accuracy ($\epsilon_{min} = 0.25$) and F1-score by 3.14% and 2.59% ($\epsilon_{min} = 0.27$) respectively. Secondly, we remove the cross-attention layer and pass the outputs of Italian BERT and AlexNet through the fusion methods. Findings suggest that Accuracy and F1-score drop by 3.19% ($\epsilon_{min} = 0.16$) and 3.01% ($\epsilon_{min} = 0.16$). Thirdly, we replace the shared layer with two non-shared ones and observe that Accuracy presents a decrease accounting for 2.36% ($\epsilon_{min} = 0.31$), while F1-score is decreased by 2.26% ($\epsilon_{min} = 0.26$). Next, we remove the concatenation mechanisms in the cross-attention layer and pass the outputs of LayerNorm through the fusion methods. Findings suggest that Accuracy and F1-score are decreased by 1.68% ($\epsilon_{min} = 0.45$) and 1.69% ($\epsilon_{min} = 0.38$) respectively. Finally, we remove the shared layer followed by LayerNorm and thus pass the outputs of Add & LayerNorm directly through fusion methods. Results show that Accuracy drops by 1.29% ($\epsilon_{min} = 0.45$).

Table 4.5: Ablation Study. (∗) means that $\epsilon_{min} < 0.1$, † means that $\epsilon_{min} < 0.2$, ‡ means that $\epsilon_{min} < 0.3$, ∗∗ means that $\epsilon_{min} < 0.4$, and †† means that $\epsilon_{min} < 0.5$.

	Evaluation metrics				
Architecture	**Precision**	**Recall**	**F1-score**	**Accuracy**	**Specificity**
– Cross-Attention and Fusion Methods	92.77‡ ±11.29	94.66 ±5.41	93.24‡ ±6.87	92.15‡ ±8.15	90.35‡ ±15.75
– Cross-Attention	92.99† ±8.22	93.16 ±5.68	92.82† ±5.19	92.10† ±5.47	92.08‡ ±9.44
Not shared	96.21 ±5.51	91.66** ±8.01	93.57‡ ±4.80	92.93‡ ±5.56	94.69 ±7.75
– Concatenation in Cross-Attention Layer	95.49** ±7.73	93.33 ±5.39	94.14** ±4.72	93.61†† ±5.19	95.03 ±8.88
– Shared feed forward and LayerNorm	94.36** ±8.21	95.52 ±4.88	94.60 ±4.45	94.00†† ±5.04	92.29** ±11.01
Proposed Approach	**97.30** ±4.43	**94.52** ±4.52	**95.83** ±3.81	**95.29** ±4.23	**96.42** ±6.04

4.4 Summary

In this chapter, we presented two methods for detecting depression by utilizing social media posts and spontaneous speech.

Firstly, we introduced a method for identifying depression in social media text by injecting linguistic information into transformer-based models. Also, it is the first study exploiting label smoothing, in order to ensure that our model is calibrated. We evaluated our proposed methods on two publicly available datasets, which include two depression detection datasets (binary classification and multiclass classification - severity of depression). Findings suggested that transformer-based networks combined with linguistic information lead to performance improvement in comparison with transformer-based networks. Also, applying label smoothing yielded both to the performance improvement and better calibration of the proposed models. Specifically, in terms of the Depression_Mixed dataset, we found that the injection of top2vec features into BERT and MentalBERT models along with label smoothing obtained the highest F1-score and Accuracy. With regards to the Depression_Severity dataset, findings showed that the injection of NRC features into the BERT model and the integration of features derived by LDA topics, namely GOSS features, into the MentalBERT model yielded the highest weighted F1-scores. We also conducted a linguistic analysis and showed that depressive posts are full of sadness, anxiety, and negative tone.

Secondly, we presented the first study utilizing a cross-attention scaling layer and multimodal fusion methods in a single neural network for detecting depression from spontaneous speech in the Italian language through speech and automatic transcripts. This is also the first study experimenting with a multi-task learning setting to investigate if the prediction of gender, age, and education level as auxiliary tasks aid the depression detection task (primary task) in increasing its performance. Results showed that our introduced approach improves competitive baselines in Accuracy by 1.21-21.99% and in F1-score by 1.32-22.23%. Results also showed that the introduced single-task learning model outperforms the multitask learning ones. Finally, we performed an ablation study, where we removed several parts of the proposed architecture and observe differences in performance. Findings showed degradation in performance in terms of Accuracy by 1.29-3.19%.

Chapter 5

Explainable Identification of Dementia from Transcripts using Transformer Networks

5.1 Introduction

Several research works have been conducted with regard to the identification of AD patients using speech and transcripts. The majority of them have employed feature extraction techniques [161, 162, 163, 164, 165], in order to train traditional Machine Learning (ML) algorithms, such as Logistic Regression, k-NN, Random Forest, etc. However, feature extraction constitutes a time-consuming procedure achieving poor classification results and often demands some level of domain expertise. Recently, researchers introduce deep learning architectures [166, 167], such as CNNs and BiLSTMs, so as to improve the classification results. Despite the success of transformer-based models in several domains, their potential has not been investigated to a high degree in the task of dementia identification from transcripts, where research works [61] having proposed them, use their outputs as features to train shallow machine learning algorithms. Concurrently, all research works except one [91], train machine learning models, in order to distinguish AD patients from non-AD patients, without taking into account the severity of dementia via Mini-Mental State Exam (MMSE) scores. Motivated by this limitation, in this chapter, we propose two multi-task learning models minimizing the loss of both dementia identification and its severity.

At the same time, to the best of our knowledge, the research works that have proposed deep learning models based on transformer networks have focused their interest only on improving the classification results obtained by CNNs, BiLSTMs etc. instead of exploring possible explainability techniques. Specifically, due to the fact that deep learning models are considered black boxes, it is important to propose ways of making them interpretable, since it is imperative for a clinician to be informed why the specific deep neural network classified a person as AD patient or not. To the best of our knowledge, only one work [168] has experimented with interpreting its proposed deep learning model (CNN-LSTM model) in the field of dementia detection using transcripts. In order to tackle this limitation, our contribution is twofold. First, we propose an interpretable neural network architecture. Next, we extend prior work and employ LIME [74], a model agnostic framework for interpretability, aiming to explain the predictions made by our best performing model. Concurrently, we propose an in-depth analysis of the language patterns used between AD and non-AD patients aiming to shed more light on the main differences observed in the vocabulary that may distinguish people suffering from dementia from healthy people.

The contributions of this chapter can be summarized as follows:

- We employ several transformer-based models, pretrained in biomedical and general corpora, and compare their performances.
- We propose an interpretable method based on the siamese neural networks along with a co-attention mechanism, so as to detect AD patients.
- We introduce two models in a multi-task learning framework, where the one task is the identification of dementia and the second one is the detection of MMSE score (severity of dementia). We model the MMSE detection task as a multiclass classification task instead of a regression task.
- We perform a thorough linguistic analysis regarding the differences in language between control and dementia groups.
- We employ LIME, in order to explain the predictions of our best performing model.

5.2 Dataset

We use the ADReSS Challenge Dataset described in Section 3.3.5.2 for conducting our experiments.

5.3 Problem Statement

In this section, the problem statement used in this chapter is presented. More specifically, it can be divided into two problems, namely the Single-Task Learning (STL) Problem and the Multi-Task Learning (MTL) Problem, which are presented in detail in Sections 5.3.1 and 5.3.2 respectively.

5.3.1 Single-Task Learning Problem

Let a dataset $\mathcal{S}_{n\times 2} = \begin{bmatrix} s_1, label_1 \\ s_2, label_2 \\ \vdots \\ s_n, label_n \end{bmatrix}$ consist of a set of transcriptions belonging to the dementia group, $d \subset \mathcal{S}$, and a set of transcriptions belonging to the control group, $c \subset \mathcal{S}$. Furthermore, $label_i \in \{0,1\}, 1 \leq i \leq n$, where *0* denotes that $s_i \in c$, while *1* denotes that $s_i \in d$. The task is to identify if a transcription $s_i \in \mathcal{S}$, belongs to a person suffering from dementia, i.e., $s_i \in d$, or not, i.e., $s_i \in c$.

5.3.2 Multi-Task Learning Problem

Let a dataset $\mathcal{S}_{n\times 3} = \begin{bmatrix} s_1, label_1, mmse_1 \\ s_2, label_2, mmse_2 \\ \vdots \\ s_n, label_n, mmse_n \end{bmatrix}$ consist of a set of transcriptions belonging to the dementia group, $d \subset \mathcal{S}$, and a set of transcriptions belonging to the control group, $c \subset \mathcal{S}$. Furthermore, $label_i \in \{0,1\}, 1 \leq i \leq n$, where *0* denotes that $s_i \in c$, while *1* denotes that $s_i \in d$. Moreover, $mmse_i$ indicates the MMSE scores. The tasks here are to identify ***(i)*** if a transcription $s_i \in \mathcal{S}$, belongs to a person suffering from dementia, i.e., $s_i \in d$, or not, i.e., $s_i \in c$, as well as ***(ii)*** to identify the MMSE scores of each person.

5.4 Predictive Models

In this section, we describe the models used for detecting AD patients. Specifically, Section 5.4.1 refers to the models employed in the single-task learning setting, whereas in Section 5.4.2 we refer to the models used for jointly learning to identify AD patients and detect the severity of dementia.

5.4.1 Single-Task Learning

5.4.1.1 Transformer-based models

We exploit the following transformer-based networks in our experiments: **BERT** [26], **BioBERT** [55], **BioClinicalBERT** [56], **ConvBERT** [57], **RoBERTa** [58], **ALBERT** [59], and **XLNet** [60].

Regarding our experiments, we pass each transcription through each pretrained model mentioned above. The output of each model is passed through a Global Average Pooling layer followed by two dense layers. The first dense layer consists of 128 units with a ReLU activation function and the second one has one unit with a sigmoid activation function to give the final output.

5.4.1.2 Transformer-based models with Co-Attention Mechanism

In this section, we present an interpretable method to differentiate AD from non-AD patients. First, we split each transcription s in the dataset into two statements of equal length (s_1 & s_2). In this way, we have to categorize a pair of statements (s_1 & s_2) into dementia or control group. To do this, we pass s_1 and s_2 through the transformer-based models mentioned in Section 5.4.1.1, i.e., BERT, BioBERT, BioClinicalBERT, ConvBERT, RoBERTa, ALBERT, and XLNet. These models can be considered as siamese in our experiments, since we make them share the same weights. Then, we implement a co-attention mechanism introduced by [319] and adopted in several studies, including [320, 321], over the two embeddings of the two statements (outputs of the transformer-based models), in order to render the entire architecture interpretable.

Formally, let $C \in \mathbb{R}^{d \times N}$ and $S \in \mathbb{R}^{d \times T}$ be the outputs of each model mentioned above, i.e., BERT, BioBERT, BioClinicalBERT, ConvBERT, RoBERTa, ALBERT, and XLNet, where d denotes the hidden size of the model. We have omitted the first dimension, which corresponds to the batch size. Following the methodology proposed by [319], the affinity matrix $F \in \mathbb{R}^{N \times T}$ is calculated using the equation presented below:

$$F = \tanh\left(C^T W_l S\right) \tag{5.1}$$

where $W_l \in \mathbb{R}^{d \times d}$ is a matrix of learnable parameters. Next, this affinity matrix is considered as a feature and we learn to predict the attention maps for both statements via the following,

$$H^s = \tanh\left(W_s S + \left(W_c C\right) F\right) \tag{5.2}$$

$$H^c = \tanh\left(W_c C + \left(W_s S\right) F^T\right) \tag{5.3}$$

where $W_s, W_c \in \mathbb{R}^{k \times d}$ are matrices of learnable parameters. The attention probabilities for each word in both statements are calculated through the softmax function as follows,

$$a^s = softmax\left(w_{hs}^T H^s\right) \tag{5.4}$$

$$a^c = softmax\left(w_{hc}^T H^c\right) \tag{5.5}$$

where $a_s \in \mathbb{R}^{1 \times T}$ and $a_c \in \mathbb{R}^{1 \times N}$. $W_{hs}, W_{hc} \in \mathbb{R}^{k \times 1}$ are the weight parameters. Based on the above attention weights, the attention vectors for each statement are obtained by calculating the weighted sum of the features from each statement. Formally,

$$\hat{s} = \sum_{i=1}^{T} a_i^s s^i, \quad \hat{c} = \sum_{j=1}^{N} a_j^c c^j \tag{5.6}$$

where $\hat{s} \in \mathbb{R}^{1 \times d}$ and $\hat{c} \in \mathbb{R}^{1 \times d}$.

Finally, these two vectors are concatenated, i.e.,

$$p = [\hat{s}, \hat{c}] \tag{5.7}$$

where $p \in \mathbb{R}^{1 \times 2d}$ and we pass the vector p to a dense layer with 128 units and a ReLU activation function followed by a dense layer consisting of one unit with a sigmoid activation function.

5.4.2 Multi-Task Learning

In this section we propose two architectures based on multi-task learning [322] and adopt the methodology followed by [323] & [82]. To be more precise, we employ a multi-task learning framework consisting of a primary and an auxiliary task. The identification of dementia constitutes the primary task, while the prediction of the MMSE score constitutes the auxiliary one. Our main objective is to explore whether the MMSE score helps in classifying groups into dementia or control. The introduced architectures are trained on the two tasks and updated at the same time with a joint loss:

$$L = (1 - \alpha) \, L_{dementia} + \alpha L_{MMSE} \tag{5.1}$$

,where $L_{dementia}$ and L_{MMSE} are the losses of dementia identification and MMSE prediction tasks respectively. α is a hyperparameter that controls the importance we place on each task. We mention below the MTL architectures developed.

MTL-BERT (Multiclass) We pass each transcription through a BERT model (which constitutes our best performing STL model). The output of the BERT model is passed through two separate dense layers, so as to identify dementia and predict the MMSE score. For identifying dementia, we use a dense layer with 2 units and a softmax activation function and minimize the cross-entropy loss function. Regarding the estimation of the MMSE score, in contrast with previous research works, we convert the MMSE regression task into a multiclass classification task. More specifically, according to [62], we can create 4 groups of cognitive severity: **healthy** (MMSE score ≥ 25), **mild dementia** (MMSE score of 21–24), **moderate dementia** (MMSE score of 10–20), and **severe dementia** (MMSE score ≤ 9). Thus, for classifying transcriptions into one of these 4 groups, we use a dense layer of 4 units with a softmax activation function and minimize the cross-entropy loss function.

MTL-BERT-DE (Multiclass) Similarly to [82], we pass each transcription into a BERT model. The output of the BERT model is passed through two separate BERT encoders, i.e, double encoders, which are followed by dense layers so as to identify dementia and classify MMSE score into one of the four classes mentioned above. For identifying dementia, we use a dense layer with 2 units and a softmax activation function and minimize the cross-entropy loss function. For classifying the MMSE score, we use a dense layer with 4 units and a softmax activation function and minimize the cross-entropy loss function.

5.5 Experiments

All experiments are conducted on a single Tesla P100-PCIE-16GB GPU.

5.5.1 Single-Task Learning

Comparison with state-of-the-art approaches We compare our introduced models with the following research works, since these research works propose single-task learning models and test their proposed approaches on the ADReSS Challenge test set: **(1)** Text [61], **(2)** LSTM with Gating (Acoustic + Lexical + Dis) [62], **(3)** Fusion Maj. (3-best) [63], **(4)** Logistic Regression (NLP) [64], **(5)** fastText, bi + trigram [65], **(6)** Attempt 5 [66], and **(7)** Fusion of system [67].

Experimental Setup Firstly, we divide the train set provided by the Challenge into a train and a validation set (65%-35%). Next, we train the proposed architectures five times and test them using the test set provided by the Challenge. Specifically, we freeze the weights of each pretrained model (BERT, BioBERT, BioClinicalBERT, ConvBERT, RoBERTa, ALBERT, and XLNet) and update the weights of the rest layers. In this way, these pretrained models act as fixed feature extractors. We train the proposed architectures using Adam optimizer with a learning rate of 1e-4. We apply *EarlyStopping* and stop training, if the validation loss has stopped decreasing for 9 consecutive epochs. We also apply *ReduceLROnPlateau*, where we reduce the learning rate by a factor of 0.2, if the validation loss has stopped decreasing for 3 consecutive epochs. When this training procedure stops, we unfreeze the weights of the pretrained models and train the entire deep learning architectures using Adam optimizer with a learning rate of 1e-5. We apply *EarlyStopping* with a patience of 3 based on the validation loss. In terms of models with a co-attention mechanism, we start training the proposed architectures using Adam optimizer with a learning rate of 1e-3 and follow the same methodology. We also apply dropout after the co-attention mechanism with a rate of 0.4. For **BERT**, we have used the base-uncased model, for **BioBERT** we have used BioBERT v1.1 (+PubMed), for **ConvBERT** we have used the base model, for **RoBERTa** we have employed the base model, for **ALBERT** we have used the base-v1 model, and for **XLNet** we have used the base model. For these pretrained models, we have used the Transformers library [305].[1]

Evaluation Metrics We evaluate our results using Accuracy, Precision, Recall, F1-score, and Specificity. All these metrics have been calculated using the dementia class as the positive one.

[1] For BioClinicalBERT we have used the model in: `https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT`

5.5.2 Multi-Task Learning

Comparison with state-of-the-art approaches For the primary task (AD Classification task), we compare our introduced models with BERT base [91], since this research work proposes a multi-task learning model and tests its proposed approach on the ADReSS Challenge test set.

Experimental Setup Firstly, we divide the train set provided by the Challenge into a train and a validation set (65%-35%). Next, we train the proposed architectures five times and test them using the test set provided by the Challenge. We use the Adam optimizer with a learning rate of 1e-6. We apply *EarlyStopping* and stop training, if the validation loss has stopped decreasing for 8 consecutive epochs. Regarding MTL-BERT-DE (Multiclass), we freeze the weights of the shared BERT model. Moreover, because of the class imbalance of the MMSE categories, we apply balanced class weights to the loss function (L_{MMSE}). We set α of (5.1) equal to 0.1. [2]

Evaluation Metrics For the primary task (AD Classification task), we evaluate our results using Accuracy, Precision, Recall, F1-score, and Specificity. All these metrics have been calculated using the dementia class as the positive one.

For the auxiliary task (MMSE Classification task), we evaluate our results using the average weighted Precision, average weighted Recall, and average weighted F1-score.

5.6 Results

5.6.1 Single-Task Learning Experiments

The results of the proposed models mentioned in Section 5.4.1 are reported in Table 5.1. Also, Table 5.1 provides a comparison of our introduced models with existing research initiatives.

Regarding our proposed transformer-based models, one can easily observe that BERT obtains the highest Recall, F1-score, and Accuracy accounting for 81.66%, 86.73%, and 87.50% respectively. Specifically, BERT outperforms the other introduced transformer-based models in Recall by 1.67-13.33%, in F1-score by 2.01-10.98%, and in Accuracy by 1.25-9.17%. BioClinicalBERT achieves the second highest Accuracy and F1-score accounting for 86.25% and 84.72% respectively. Also, BioClinicalBERT obtains the highest Precision score equal to 95.03% surpassing the other transformer-based models by 4.79-15.88%. RoBERTa achieves comparable results to BERT and BioClinicalBERT yielding an Accuracy and F1-score of 84.16% and 82.81% respectively. In addition, BioBERT and ConvBERT demonstrate slight differences in Accuracy and F1-score, with BioBERT surpassing ConvBERT in both metrics. Specifically, BioBERT surpasses ConvBERT in F1-score by 0.46% and in Accuracy by 0.84%. Moreover, we observe that ALBERT and

[2] We used also the experimental setup of Section 5.5.1. However, lower evaluation results were achieved.

XLNet achieve Accuracy scores equal to 78.33%, with ALBERT surpassing XLNet in F1-score by 2.70%.

Regarding our proposed transformer-based models with a co-attention mechanism, they achieve lower performance than the proposed transformer-based models except for ConvBERT + Co-Attention, ALBERT + Co-Attention, and XLNet + Co-Attention. More specifically, ConvBERT + Co-Attention presents a slight surge of 0.42% in Accuracy in comparison with ConvBERT, ALBERT+Co-Attention presents an increase in Accuracy by 1.67% in comparison with ALBERT, and XLNet + Co-Attention demonstrates a slight increase of 0.42% in Accuracy in comparison with XLNet. BERT+Co-Attention attains the highest F1-score and Accuracy accounting for 83.85% and 83.75% respectively. BERT+Co-Attention outperforms the other models in terms of F1-score by 1.42-7.43%, and in terms of Accuracy by 1.25-5.00%. ConvBERT + Co-Attention and BioClinicalBERT + Co-Attention demonstrate slight differences in F1-score and Accuracy, with ConvBERT + Co-Attention surpassing BioClinicalBERT + Co-Attention in F1-score by 0.44% and in Accuracy by 0.42%. BioBERT + Co-Attention and ALBERT + Co-Attention achieve almost equal F1-score results, with BioBERT + Co-Attention attaining a higher Accuracy score than ALBERT + Co-Attention by 1.66%. RoBERTa + Co-Attention and XLNet + Co-Attention demonstrate low performances attaining an Accuracy of 79.16% and 78.75% respectively.

Overall, BERT constitutes our best performing model, since it outperforms all the other introduced models in F1-score and Accuracy. Although there are models surpassing BERT in Precision and Recall, BERT outperforms all of them in F1-score, which constitutes the weighted average of Precision and Recall. In addition, there are models that outperform BERT in Specificity. However, high specificity and low recall means that the model cannot diagnose the AD patients pretty well and consequently AD patients are misdiagnosed as non-AD ones.

In comparison with the state-of-the-art approaches, one can observe that our proposed models achieve comparable performance to or outperform previous studies. More specifically, BERT outperforms all the research works, except [61], in terms of Accuracy by 2.08-8.33%, in F1-score by 1.33-8.68%, and in Recall by 2.66-14.99%. Moreover, BERT + Co-Attention surpasses [62, 65, 67] in Accuracy by 2.50%, 0.42%, and 4.58% respectively. Also, it surpasses [62, 65, 67] in Recall by 17.49%, 5.16%, and 9.16% respectively. BERT+Co-Attention outperforms [62, 65, 67] in F1-score by 5.80%, 0.85%, and 5.59% respectively.

Table 5.1: Performance comparison among proposed STL models and state-of-the-art approaches on the ADReSS Challenge test set. Reported values are mean ± standard deviation. Results are averaged across five runs.

	Evaluation metrics				
Architecture	**Prec.**	**Rec.**	**F1-score**	**Acc.**	**Spec.**
Comparison with state-of-the-art approaches					
[61]	-	87.50	-	89.58	91.67
[62]	81.82	75.00	78.26	79.17	83.33
[63]	-	-	85.40	85.20	-
[64]	-	-	-	85.00	-
[65]	86.00	79.00	83.00	83.33	88.00
[66]	-	-	-	85.42	-
[67]	94.12	66.67	78.05	81.25	95.83
Proposed Transformer-based models					
BERT	87.19	81.66	86.73	87.50	93.33
	±3.25	±5.00	±4.53	±4.37	±5.65
BioBERT	86.87	78.33	82.11	82.92	87.50
	±6.09	±4.86	±2.83	±3.06	±6.97
BioClinicalBERT	95.03	76.66	84.72	86.25	95.83
	±3.03	±4.99	±2.74	±2.12	±2.64
ConvBERT	83.51	79.99	81.65	82.08	84.16
	±1.23	±4.08	±2.06	±1.66	±1.66
RoBERTa	90.24	76.66	82.81	84.16	91.66
	±2.81	±4.99	±3.52	±2.83	±2.64
ALBERT	79.15	78.33	78.45	78.33	78.33
	±7.89	±3.11	±3.12	±3.86	±8.89
XLNet	85.58	68.33	75.75	78.33	88.33
	±2.77	±6.77	±4.05	±2.82	±3.12
Proposed Transformer-based models with co-attention mechanism					
BERT	83.67	84.16	83.85	83.75	83.33
Co-Attention	±3.36	±1.66	±1.09	±1.56	±4.56
BioBERT	85.41	76.66	80.72	81.66	86.66
Co-Attention	±4.91	±3.33	±3.16	±3.06	±4.86
BioClinicalBERT	82.60	81.66	81.99	82.08	82.50
Co-Attention	±3.60	±4.25	±2.11	±2.12	±4.86
ConvBERT	83.78	81.66	82.43	82.50	83.33
Co-Attention	±6.13	±4.24	±2.37	±3.12	±8.74
RoBERTa	79.39	79.16	79.06	79.16	79.16
Co-Attention	±2.26	±6.45	±2.15	±1.32	±4.56

ALBERT	77.94	84.16	80.77	80.00	75.83
Co-Attention	±3.20	±4.86	±1.68	±1.66	±5.53
XLNet	85.63	69.16	76.42	78.75	88.33
Co-Attention	±3.45	±5.00	±3.75	±3.06	±3.12

5.6.2 Multi-Task Learning Experiments

5.6.2.1 Primary Task

The results of the introduced models described in Section 5.4.2 are reported in Table 5.2. Also, Table 5.2 provides a comparison of our introduced approaches with state-of-the-art approaches.

With regards to our introduced models, one can easily observe that MTL-BERT (Multiclass) outperforms MTL-BERT-DE (Multiclass) in terms of all the evaluation metrics except Recall. Specifically, MTL-BERT (Multiclass) surpasses MTL-BERT-DE (Multiclass) in Precision by 3.40%, in F1-score by 0.88%, in Accuracy by 1.25%, and in Specificity by 4.16%. Although MTL-BERT-DE (Multiclass) surpasses MTL-BERT (Multiclass) in Recall by 1.67%, MTL-BERT (Multiclass) obtains a higher F1-score, which constitutes the weighted average of Precision and Recall. Therefore, MTL-BERT (Multiclass) constitutes our best performing model in the MTL framework.

In comparison to the research work [91], as one can easily observe, both our introduced models attain a higher Accuracy score. To be more precise, MTL-BERT (Multiclass) outperforms BERT base [91] in Accuracy by 5.42%. In addition, MTL-BERT-DE (Multiclass) surpasses the research work [91] in Accuracy by 4.17%. These differences in performance are attributable to the fact that we adopt a different training procedure than the one adopted by [91], we consider the MMSE task as a multiclass classification task instead of a regression task, as well as to the different architectures proposed.

5.6.2.2 Auxiliary Task

The results of the introduced models mentioned in Section 5.4.2 for the auxiliary task (MMSE Classification task) are reported in Table 5.3.

As one can easily observe, MTL-BERT (Multiclass) obtains an average weighted Precision of 73.62% surpassing MTL-BERT-DE (Multiclass) by 3.12%. However, MTL-BERT-DE (Multiclass) outperforms MTL-BERT (Multiclass) in average weighted Recall and average weighted F1-score by 1.26% and 3.82% respectively.

Table 5.2: Performance comparison among proposed MTL models and state-of-the-art approaches on the ADReSS Challenge test set for the primary task (AD Classification Task). Reported values are mean ± standard deviation. Results are averaged across five runs.

	Evaluation metrics				
Architecture	**Prec.**	**Rec.**	**F1-score**	**Acc.**	**Spec.**
Comparison with state-of-the-art approaches					
[91]	-	-	-	80.83	-
	-	-	-	±1.56	-
Proposed Multi-task learning models					
MTL-BERT	88.59	83.33	85.84	86.25	89.16
(Multiclass)	±3.05	±2.64	±2.12	±2.13	±3.33
MTL-BERT-DE	85.19	85.00	84.96	85.00	85.00
(Multiclass)	±3.46	±5.00	±2.60	±2.43	±4.25

Table 5.3: Results of the proposed MTL models on the ADReSS Challenge test set for the auxiliary task (MMSE Classification Task). Reported values are mean ± standard deviation. Results are averaged across five runs.

	Evaluation metrics		
Architecture	**Avg. W. Prec.**	**Avg. W. Rec.**	**Avg. W. F1-score**
Proposed Multi-task learning models			
MTL-BERT	73.62	69.16	64.75
(Multiclass)	±2.95	±4.04	±3.50
MTL-BERT-DE	70.50	70.42	68.57
(Multiclass)	±5.59	±3.06	±2.04

5.7 Analysis of the Language used in Control and Dementia groups

We finally perform an extensive analysis to uncover some unique characteristics, which discriminate the AD patients from the non-AD ones, and understand the predictions made by our best performing model as well as its limits.

5.7.1 Text Statistics

We first extract some statistics, namely the syllable count, the lexicon count, the difficult words, and the sentence count, using the `TEXTSTAT` library in Python, in order to understand better the differences in language used between control and dementia groups. More specifically, the syllable count refers to the number of syllables, the lexicon count to the number of words, and the sentence count to the number of sentences present in the given text. With regards to the difficult words, they refer to the number of polysyllabic words with a Syllable Count > 2 that are not included in the list of words of common usage in English [324]. After extracting these statistics per transcript, we calculate the mean and standard deviation for both control and dementia groups. We test for statistical significance using an independent t-test for each metric between control and dementia groups and adjust the p-values using Benjamini-Hochberg correction [69]. As one can easily observe in Table 5.4, the control group presents a significantly higher number of syllables, lexicon, and difficult words than the dementia group.

Table 5.4: mean ± standard deviation metrics per transcript. † indicates statistical significance between transcripts of control and dementia groups. All differences are significant at $p < 0.05$ after Benjamini-Hochberg correction.

	Transcript	
Metric	Control	Dementia
Syllable Count†	151.63 ± 79.98	119.95 ± 71.18
Lexicon Count†	107.49 ± 62.02	86.08 ± 54.10
Difficult Words†	10.58 ± 3.64	6.38 ± 3.53
Sentence Count	1.67 ± 1.03	1.92 ± 1.62

5.7.2 Vocabulary Uniqueness

In order to understand the vocabulary similarities and differences between control and dementia groups, we adopt the methodology proposed by [325]. Formally, let $\mathcal{P}$ and $\mathcal{C}$ be the sets of unique words included in the control group and dementia group respectively. Next, we calculate the Jaccard's index given by (5.1), in order to measure the similarity between finite sample sets. More specifically, the Jaccard's index is a number between 0 and 1, where 1 indicates that the two sets, namely $\mathcal{P}$ and $\mathcal{C}$, have the same elements, while 0 indicates that the two sets are completely different.

$$J(P, C) = |P \cap C| / |P \cup C| \tag{5.1}$$

As observed in Table 5.5, the Jaccard's index between the control and dementia groups is equal to 0.4049, which indicates that people with dementia tend to use a different vocabulary than those in the control group.

Table 5.5: Jaccard's Index between transcripts of control and dementia group

Jaccard's Index between transcripts	**Result**
$J(\mathcal{P}=$ control, $\mathcal{C}=$dementia$)$	0.4049

5.7.3 Word Usage

Apart from finding the vocabulary similarities and differences, it is imperative that patterns of word usage be investigated. Thus, following the methodology introduced in [325], the main objective of this section is to explore the differences between the two classes (control and dementia) with regard to the probability of using specific words more than others. Formally, let D_1 and D_2 be two documents, where D_1 includes all the transcriptions of the control group, whereas D_2 consists of transcriptions of the dementia group. Moreover, we define S as the entire corpus consisting of D_1 and D_2. Now we can define the probability of a word w_i in the document D_1 in a collection of documents S given by (5.1):

$$P(w_i|D_1, S) = (1 - \alpha_D)P(w_i|D_1) + \alpha_D P(w_i|S) \tag{5.1}$$

Similarly, we can define the probability of a word w_i in the document D_2 in a collection of documents S given by (5.2):

$$P(w_i|D_2, S) = (1 - \alpha_D)P(w_i|D_2) + \alpha_D P(w_i|S) \tag{5.2}$$

We employ the Jelinek-Mercer smoothing method and consider that $\alpha_D \in [0, 1]$. More specifically, α_D is a parameter that controls the probability of words included only in one document (D_1 or D_2). In our experiments, we set α_D equal to 0.2.

Moreover, we define $P(w_i|S) = \frac{s_{w_i}}{|S|}$, where s_{w_i} denotes the number of times a word w_i is included in the collection, whereas $|S|$ is the total number of words occurrences in the collection. Similarly, $P(w_i|D_1) = \frac{d_{w_i}}{|D_1|}$, where d_{w_i} denotes the number of times a word w_i is presented in the document D_1, whereas $|D_1|$ is the total number of words occurrences in the document D_1. The same methodology has been adopted for calculating the $P(w_i|D_2)$.

After having calculated the two distributions, i.e., $P(w_i|D_1, S)$ and $P(w_i|D_2, S)$, we exploit the Kullback-Leibler (KL) divergence, in order to measure the difference of these two distributions. KL-divergence is always greater than zero and is given by (5.3). The larger it gets, the more different the two distributions are.

$$KL(P||C) = \sum_x P(x) log \frac{P(x)}{C(x)} \tag{5.3}$$

As one can easily observe in Table 5.6, the KL divergence between control and dementia groups is high indicating that these two groups present differences regarding the probability of using some words more than others. Our findings agree with the ones in [325], where the authors state that there are clear differences in terms of language use between positive (depression and self-harm) and control group, where the values of KL-divergence range from 0.18 to 0.21.

Table 5.6: Kullback-Leibler divergence

KL divergence	Result
KL(Control || Dementia)	0.2047
KL(Dementia || Control)	0.2161

5.7.4 Linguistic Feature Analysis

Following the method introduced by [68], the main objective of this section is to shed light on which unigrams and pos-tags are mostly correlated with each class separately. To facilitate this, we compute the point-biserial correlation between each feature (unigram and pos-tag) across all the transcriptions and a binary label (0 for the control and 1 for the dementia group). Before computing the correlation, we normalize features so that they sum up to 1 across each transcription. We use the point-biserial correlation, since it is a correlation used between continuous and binary variables. It returns a value between -1 and 1. Since we are only interested in the strength of the correlation, we compute the absolute value, where negative correlations refer to the control group (label 0) and positive correlations refer to the dementia one (label 1). We report our findings in Table 5.7, where all correlations are significant at $p < 0.05$, with Benjamini-Hochberg correction [69] for multiple comparisons.

As one can easily observe, the pos-tags associated with the dementia group are the following: RB (adverbs), PRP (personal pronoun), VBD (verb in past tense), and UH (interjection). On the other hand, people in the control group tend to use VBG (verb, gerund, or present participle), DT (determiner), and NN (noun). These findings can be justified in Table 5.8, where we present three examples of transcripts belonging to the control group and three examples of transcripts belonging to the dementia one. More specifically, we have assigned colours to different pos-tags, so as to render the differences in the language patterns used by each group easily understandable to the reader. To be more precise, red colour indicates the VBG pos-tag, yellow refers to the DT pos-tag, fuchsia to the RB pos-tag, apricot to the PRP pos-tag, navy blue to the VBD pos-tag, and the pine green to the UH pos-tag.

We observe that people in the dementia group tend to use personal pronouns (he, she, I, them etc.) very often, since they are unable to remember the specific terms (mom, boy, etc.). This finding agrees with the research conducted by [70], where the authors state that personal pronouns present a high frequency in the speech of AD patients, since these people cannot find the target word. To be more precise, in a conversation people have to remember what they have said during the entire conversation. However, this is not possible in AD patients, who present working memory impairment and thus tend to produce empty conversational speech (use of personal pronouns). On the other hand, people in the control group tend to use more nouns instead of personal pronouns, since they are able to maintain various kinds of information.

Moreover, AD patients tend to use verbs in the past tense (were, forgot, did, started) in contrast to people who are not suffering from dementia and use verbs in the present participle. One typical example that can illustrate this difference can be seen in the fifth transcription in Table 5.8, i.e., *"oh have you heard of that new game that they started to play after christmas ? did you"*. The AD patient perhaps remembers a personal story from the past that wants to narrate, instead of the task he has been assigned to conduct. Therefore, the patient is not able to stay focused on describing the picture. This finding is consistent with [71, 72], where the authors state that AD patients present difficulty in maintaining and continuing the development of a topic and thus demonstrate unexpected topic shifts. Also, this finding reveals a difference in language used by the AD patients and the agrammatic aphasics. Specifically, patients with agrammatic aphasia typically have problems using past tense inflection and instead rely on infinitive or present tense verb forms [73].

In addition, AD patients tend to use the UH (oh, yeah, well) and the RB (maybe, probably) pos-tags, since they are not certain of what they are describing due to the cognitive impairment. Concurrently, the UH pos-tag constitutes an example of empty speech. More specifically, this pos-tag is used as filler at the beginning of each utterance, since AD patients are thinking of what to say.

Table 5.7: Features associated with control and dementia subjects, sorted by point-biserial correlation. All correlations are significant at $p < 0.05$ after Benjamini-Hochberg correction.

Control		**Dementia**	
Unigrams	**corr.**	**Unigrams**	**corr.**
is	0.364	here	0.310
curtains	0.361	-	-
window	0.301	-	-
are	0.300	-	-
POS	**corr.**	**POS**	**corr.**
VBG	0.285	RB	0.388
DT	0.216	PRP	0.354
NN	0.210	VBD	0.275
-	-	UH	0.242

Table 5.8: Examples of transcripts along with their labels. red colour indicates the VBG pos-tag, yellow refers to the DT pos-tag, fuchsia to the RB pos-tag, apricot to the PRP pos-tag, navy blue to the VBD pos-tag, and the pine green to the UH pos-tag.

Transcript	**Label**
” well the girl is watching the boy go into the cookie jar . he has a cookie in his hand . he’s on the stool . the stool is falling . the mother is drying dishes . has a plate in her hand . sink is overflowing . there’s water on the floor . she’s stepping in the water . something that’s going on you said ? the little girl looks like she’s motioning to the boy to be quiet . and I don’t know what else . the woman’s looking out the window . the window’s open . ”	Control
” action . alright . a lady’s drying dishes . the boy was standing on a stool but the action is that the stool has slipped and he is falling . and the girl has her hand raised reaching for a cookie . and there’s a lot of action in the sink here . the water is flowing out . she is apparently so daydreaming that she doesn’t realize that the sink is overflowing . any more action ? or is that enough action ? ”	Control
” touching lip . raising arm . is that what you mean ? reaching for cookie . handing cookie down . slipping from stool . stool falling over . wiping dishes . water running . water overflowing . breeze . I don’t know if that’s action . stepping out from water . I guess that’s it . ”	Control
” alright . I see the little boy stealing cookies from the cookie jar . and he gave some to the little girl and she’s eating some of the cookies . and I guess this is mama and she’s washing the dishes . and she dropped a dish . no she didn’t drop a dish . the water that she’s washing the dishes with she let run . and it’s overflown . that doesn’t sound right . did it ? we forgot to turn off the spigot . and so the water is running off onto the floor here . and mom apparently is washing the dishes . and here’s this little boy stealing the cookies . he’s gonna fall because the four legged stool is gonna fall over with him and the cookie jar . and mama’s drying the dishes as usual for mamas if they don’t have a husband that dries them or washes them or whatever . let’s see now . I guess there’s more things I’m sposta see . let’s see here now . oh and the water is flowing out of the sink they forgot to turn off whoever’s doing the dishwashing . mom apparently here , she forgot to turn off the water and the water is spilling out onto the kitchen floor . and the little girl has pushed over the stool with the boy that was reaching up to get the cookies . either she pushed it over or he fell over with it . you know it excuse me but you know I was ... ”	Dementia

Continued on next page

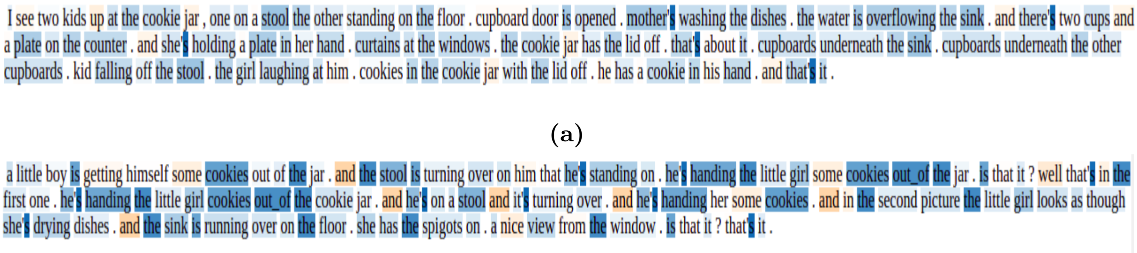
Table 5.8 – continued from previous page

Transcript	Label
” mhm . oh I see a part of the whole kitchen . is that all the kitchen or isn’t it ? oh I can’t read ... a lady a mother were in her kitchen . in her kitchen doing some work I suppose . and there’s another woman there sharing their pleasures or whatever . oh have you heard of that new game that they started to play after christmas ? did you ? is a . well it looks like ... I’d say this is ... well let’s see . it looks like ... oh ... . my wife will beat me by a couple rows of this . that’s like the washing machine ? or let me see . I can’t ... oh that’s the son come from school maybe or something . that’s a youngster there . well that’s just as though they getting ready to go to school or they’re just coming out from school . and right there he’s same as back there except for down there in the bottom I think it’s ... that’s a little . ”	Dementia
” yes . the water ? well let’s see . there’s something hasta be where the water goes down over . there’s probably something that’s ... or they don’t have it open or something might have. I don’t know . what ..? when the water goes down what do you call that ? this here . right here . this . what do you call that ? what is that ? what is that ? I don’t know ! that’s what I’m saying . I don’t know what that is . the what ? a pipe . oh water pipe ! oh yeah . okay . well then maybe the water pipe is not broke but there must be things in there . that the water will not go down . I don’t know . huh ? what’s happening to the water ? well the water is going down in the ... I don’t know . what would you call this ? floor ! yeah okay . yeah . well down on this side of the picture . well this thing here is turning over . yeah . no , uhuh . I don’t know what’s going on . well he’s probably getting ... what’s this here ? cocoa jar ? what’s this cocoa ? c o o k i e . I don’t know . I don’t know what ..? huh ? cookie , oh a cookie . oh ! oh okay . mhm . well he’s getting it out . and he’s gonna give it to the girl /. down here . mhm . going on in the picture ? well the boy is giving her the girl the cookie . this probably is broke . so the water will not go down in and it’s coming up and going in here huh . well it looks like she was gonna wash . what they eat with , all that . what do you call that ? what do you call this ? a plate ? oh yeah . what you eat on . is that what you call them a plate ? oh this is a cup ? oh maybe , I don’t know . mhm . okay . ”	Dementia

5.7.5 Explainability - Error Analysis

In this section, we employ LIME [74] (using 5000 samples) to explain the predictions made by our best performing model, namely BERT, and shed more light regarding the differences in language between AD and non-AD patients. More specifically, LIME generates local explanations for any machine learning classifier by introducing an interpretable model, which is trained on data generated through observing differences in the classification performance when removing tokens from the input string.

Examples of explanations generated by LIME are illustrated in Figs. 5.1-5.4. More specifically, Fig. 5.1 illustrates two transcripts, whose ground-truth label is dementia, while our model predicts them as belonging to non-AD patients. Fig. 5.2 refers to transcripts with both ground-truth label and prediction corresponding to dementia. In Fig. 5.3, two transcripts are presented, whose prediction is control and true label is control too. Finally, Fig. 5.4 illustrates transcripts, which are misclassified. The ground-truth is control, whereas the prediction is dementia. Moreover, as one can observe, each token has been assigned a colour, either blue or orange. To be more precise, the blue colour indicates which tokens are indicative of the control group, whilst the orange colour indicates tokens, which are used mainly by AD patients. The more intense the colours are, the more important these tokens are towards the final classification of the transcript.



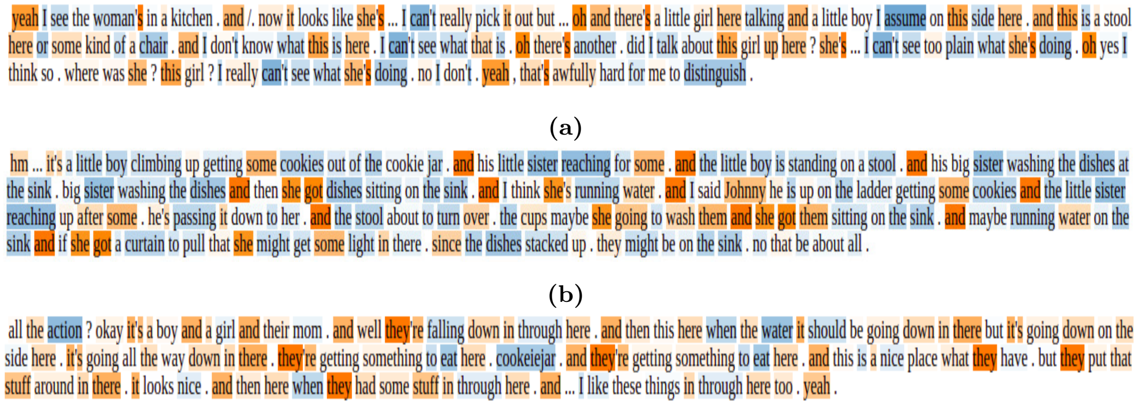
I see two kids up at the cookie jar , one on a stool the other standing on the floor . cupboard door is opened . mother's washing the dishes . the water is overflowing the sink . and there's two cups and a plate on the counter . and she's holding a plate in her hand . curtains at the windows . the cookie jar has the lid off . that's about it . cupboards underneath the sink . cupboards underneath the other cupboards . kid falling off the stool . the girl laughing at him . cookies in the cookie jar with the lid off . he has a cookie in his hand . and that's it .

(a)

a little boy is getting himself some cookies out of the jar . and the stool is turning over on him that he's standing on . he's handing the little girl some cookies out_of the jar . is that it ? well that's in the first one . he's handing the little girl cookies out_of the cookie jar . and he's on a stool and it's turning over . and he's handing her some cookies . and in the second picture the little girl looks as though she's drying dishes . and the sink is running over on the floor . she has the spigots on . a nice view from the window . is that it ? that's it .

(b)

Figure 5.1: Label: Dementia, Prediction: Control



yeah I see the woman's in a kitchen . and /. now it looks like she's ... I can't really pick it out but ... oh and there's a little girl here talking and a little boy I assume on this side here . and this is a stool here or some kind of a chair . and I don't know what this is here . I can't see what that is . oh there's another . did I talk about this girl up here ? she's ... I can't see too plain what she's doing . oh yes I think so . where was she ? this girl ? I really can't see what she's doing . no I don't . yeah , that's awfully hard for me to distinguish .

(a)

hm ... it's a little boy climbing up getting some cookies out of the cookie jar . and his little sister reaching for some . and the little boy is standing on a stool . and his big sister washing the dishes at the sink . big sister washing the dishes and then she got dishes sitting on the sink . and I think she's running water . and I said Johnny he is up on the ladder getting some cookies and the little sister reaching up after some . he's passing it down to her . and the stool about to turn over . the cups maybe she going to wash them and she got them sitting on the sink . and maybe running water on the sink and if she got a curtain to pull that she might get some light in there . since the dishes stacked up . they might be on the sink . no that be about all .

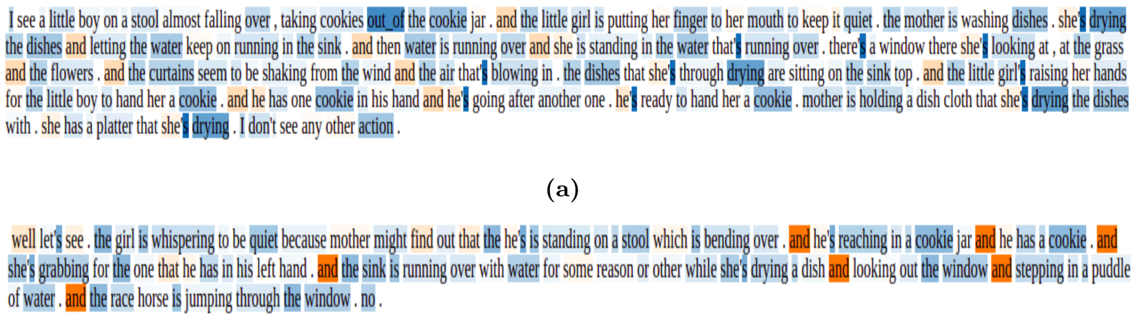
(b)

all the action ? okay it's a boy and a girl and their mom . and well they're falling down in through here . and then this here when the water it should be going down in there but it's going down on the side here . it's going all the way down in there . they're getting something to eat here . cookeiejar . and they're getting something to eat here . and this is a nice place what they have . but they put that stuff around in there . it looks nice . and then here when they had some stuff in through here . and ... I like these things in through here too . yeah .

(c)

Figure 5.2: Label: Dementia, Prediction: Dementia

As one can easily observe in Fig. 5.2, tokens belonging to the UH pos-tag, such as yeah and oh, are identified as important for the dementia class by our best performing model. Moreover, personal pronouns (she, they) and verbs in the past tense (got, had) are also indicative of dementia. Also, our model considers the token "here", which corresponds to the RB pos-tag, indicative of the dementia class. These findings are consistent with



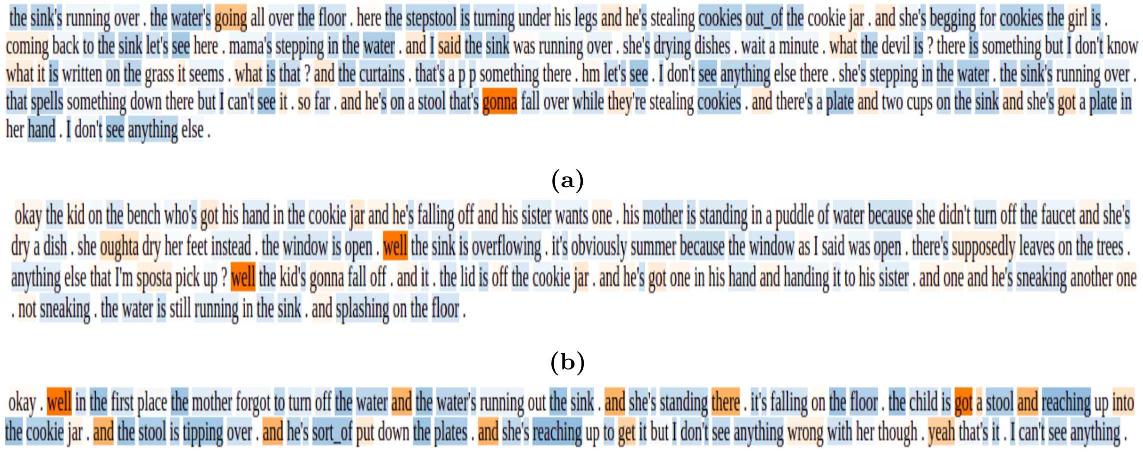
I see a little boy on a stool almost falling over , taking cookies out_of the cookie jar . and the little girl is putting her finger to her mouth to keep it quiet . the mother is washing dishes . she's drying the dishes and letting the water keep on running in the sink . and then water is running over and she is standing in the water that's running over . there's a window there she's looking at , at the grass and the flowers . and the curtains seem to be shaking from the wind and the air that's blowing in . the dishes that she's through drying are sitting on the sink top . and the little girl's raising her hands for the little boy to hand her a cookie . and he has one cookie in his hand and he's going after another one . he's ready to hand her a cookie . mother is holding a dish cloth that she's drying the dishes with . she has a platter that she's drying . I don't see any other action .

(a)

well let's see . the girl is whispering to be quiet because mother might find out that the he's is standing on a stool which is bending over . and he's reaching in a cookie jar and he has a cookie . and she's grabbing for the one that he has in his left hand . and the sink is running over with water for some reason or other while she's drying a dish and looking out the window and stepping in a puddle of water . and the race horse is jumping through the window . no .

(b)

Figure 5.3: Label: Control, Prediction: Control



the sink's running over . the water's going all over the floor . here the stepstool is turning under his legs and he's stealing cookies out_of the cookie jar . and she's begging for cookies the girl is . coming back to the sink let's see here . mama's stepping in the water . and I said the sink was running over . she's drying dishes . wait a minute . what the devil is ? there is something but I don't know what it is written on the grass it seems . what is that ? and the curtains . that's a p p something there . hm let's see . I don't see anything else there . she's stepping in the water . the sink's running over . that spells something down there but I can't see it . so far . and he's on a stool that's gonna fall over while they're stealing cookies . and there's a plate and two cups on the sink and she's got a plate in her hand . I don't see anything else .

(a)

okay the kid on the bench who's got his hand in the cookie jar and he's falling off and his sister wants one . his mother is standing in a puddle of water because she didn't turn off the faucet and she's dry a dish . she oughta dry her feet instead . the window is open . well the sink is overflowing . it's obviously summer because the window as I said was open . there's supposedly leaves on the trees . anything else that I'm sposta pick up ? well the kid's gonna fall off . and it . the lid is off the cookie jar . and he's got one in his hand and handing it to his sister . and one and he's sneaking another one . not sneaking . the water is still running in the sink . and splashing on the floor .

(b)

okay . well in the first place the mother forgot to turn off the water and the water's running out the sink . and she's standing there . it's falling on the floor . the child is got a stool and reaching up into the cookie jar . and the stool is tipping over . and he's sort_of put down the plates . and she's reaching up to get it but I don't see anything wrong with her though . yeah that's it . I can't see anything .

(c)

Figure 5.4: Label: Control, Prediction: Dementia

the ones in Section 5.7.4, where we have found that PRP, VBD, UH pos-tags as well as the unigram ”here” are significantly correlated with the dementia class. In addition, our model identifies the repetition of token ”and” as important for the dementia class. This finding agrees with previous research works [168], where the word ”and” indicates a short answer and burst of speech.

Regarding Fig. 5.3, one can easily observe that our model identifies tokens belonging to the VBG (putting, drying, blowing, standing, etc.), DT (the, a), and NN (cookie, action, stool, etc.) pos-tags as significant for the control class. Concurrently, in consistence with the findings in Section 5.7.4, the unigrams ”curtain” and ”window” are used mainly by non-AD patients.

With regards to Figs. 5.1 and 5.4, our model is not able to classify these transcripts correctly. One possible reason for such misclassifications has to do with the fact that these transcripts include pos-tags which are indicative of both the control and the dementia class. To be more precise, in Fig. 5.1, the majority of tokens in both transcripts belong to the VBG, NN, and DT pos-tags, which are correctly identified by our model as significant for the control group. Words, like "and", "him", and "well" are used in a low frequency. Similarly to Fig. 5.1, in Fig. 5.4, the majority of tokens in each transcript belong to the pos-tags which are significantly correlated with the dementia class. This can be illustrated in Fig. 5.4c, where we observe the usage of words, like "and", "yeah", "well" & "got".

5.8 Summary

In this chapter, we introduced both single-task and multi-task learning models. Regarding single-task learning models, we employed several transformer-based networks and compared their performances. Results showed that BERT achieved the highest classification performance with accuracy accounting for 87.50%. Concurrently, we introduced siamese networks coupled with a co-attention mechanism which can detect AD patients with an accuracy up to 83.75%. In terms of the multi-task learning setting, it consisted of two tasks, the primary and the auxiliary one. The primary task was the identification of dementia (binary classification), whereas the auxiliary task was the categorization of the severity of dementia into one of the four categories -healthy, mild/moderate/severe dementia- (multiclass classification). Specifically, we proposed two multi-task learning models. Results showed that our model achieves competitive results in the MTL framework reaching accuracy up to 86.25% on the detection of AD patients. Next, we performed an in-depth linguistic analysis, in order to understand better the differences in language between AD and non-AD patients. Finally, we employed LIME, in order to shed light on how our best performing model works. Findings suggest that AD patients tend to use personal pronouns, interjection, adverbs, verbs in the past tense, and the token "and" at the beginning of utterances in a high frequency. On the contrary, healthy people use verbs in present participle or gerund, nouns as well as determiners.

In this chapter, we concentrated on the usage of linguistic information, i.e., transcripts, for recognizing Alzheimer's dementia, thus neglecting the acoustic modality. Moving forward to the next chapter, we will introduce unimodal (acoustic) and multimodal (linguistic and acoustic) methods for identifying AD patients.

Chapter 6

Detecting Dementia from Speech and Transcripts Using Transformers

6.1 Introduction

In Chapter 5, we utilized only transcripts and used transformer-based models along with explainable approaches for identifying AD patients. However, speech contains valuable information. Existing research works using audio data to categorize people into AD and non-AD patients use mainly acoustic features extracted from speech, such as eGeMAPS [218], duration of speech etc. After having extracted the respective feature sets, they train traditional machine learning classifiers, such as Support Vector Machines (SVM), Decision Trees (DT) etc. However, feature extraction constitutes a time-consuming procedure, does not generalize well to data from new patients, and often demands some level of domain expertise. Log-Mel Spectrograms and Mel-frequency cepstral coefficients (MFCCs) are being used extensively in heart sound classification [326], emotion recognition [327], depression detection [328], etc. In addition, pretrained models on the domain of computer vision, including AlexNet, MnasNet, EfficientNet, VGG, etc., have been exploited extensively in many tasks, including Alzheimer's disease detection through MRIs [329], detection of epileptic seizures using EEG signals [330, 331], facial emotion recognition [332], analysis of online political advertisements [265], heart sound classification [326], voice pathology diagnosis [333], etc. Thus, the representation of speech signal as an image constitutes a motivation for exploiting image-based models. However, limited research has considered speech in such a way [94, 91, 231]. Therefore, in this chapter, we convert each audio file into an image consisting of three channels, namely log-Mel spectrograms (and MFCCs), their delta, and delta-delta. Contrary to [91, 231], we use the delta and delta-delta features for adding more information [260, 261]. Next, we employ many pretrained models, including AlexNet, VGG16, DenseNet, EfficientNet, Vision Transformer, etc. and compare their performances. Our main motivation is to find the best model for extracting acoustic features and exploiting it in the multimodal setting.

Moreover, another limitation of the existing research works lies in the usage of multimodal models. To be more precise, research works train first acoustic and language models separately and then use the majority voting approach for classifying people into AD and non-AD patients [63, 170, 92]. This fact increases substantially the training time and does not take into account the inter- and intra-modal interactions. Other research works add or concatenate acoustic and language representations during training [91, 262, 259]. This approach may decrease the performance of the multimodal models in comparison with the unimodal ones, since the different modalities are treated equally. In addition, there are studies, which concatenate the features from different modalities at the input level (early fusion approaches) [169, 67, 95]. Little work has been done in terms of exploiting techniques to control the influence of each modality towards the final classification and capturing the inter- and intra-modal interactions. Specifically, the authors in [62, 264] used feed-forward highway layers with a gating mechanism. However, the authors did not experiment with replacing the gating mechanism with a simple concatenation operation. Thus, the addition of the introduced gating mechanism cannot guarantee increase in the performance. To tackle this limitation, in this chapter, we propose new methods, which can be trained in an end-to-end trainable way, to combine the representations of the different modalities. Firstly, we convert each audio file into an image consisting of three channels, namely log-Mel spectrograms (and MFCCs), their delta, and delta-delta. We pass these images through a Vision Transformer, which is the best performing model among the proposed pretrained models, i.e., AlexNet, VGG16, DenseNet, EfficientNet, etc. Each transcript is passed through a BERT model. Next, we propose a Gated Multimodal Unit in order to assign more importance to the most relevant modality while suppressing irrelevant information. In addition, we introduce crossmodal attention so as to model crossmodal interactions.

The contributions of this chapter can be summarized as follows:

- We propose multimodal deep learning models to detect AD patients from speech and transcripts. We also introduce a multimodal gate mechanism, so as to control the influence of each modality towards the final classification.
- We introduce the crossmodal attention and show that crossmodal models outperform the multimodal ones.

6.2 Dataset

We use the ADReSS Challenge Dataset described in Section 3.3.5.2 for conducting our experiments.

6.3 Proposed Predictive Models using only Speech

In this section, we describe the models used for detecting AD patients using only speech. Our main motivation of exploiting these pretrained models is to find the best performing one and exploit it in the multimodal setting, which will be discussed in detail in Section 6.4. Firstly, we use the Python library librosa [334] for converting the audio files into Log-Mel spectrograms (and MFCCs), their delta, and delta-delta. We extract Log-Mel spectrograms with 224 Mel bands, window length equal to 2048, hop length equal to 1024, and a Hanning window. For extracting MFCCs, we use 40 MFCCs, a Hanning window, window length equal to 2048, and a hop length of 512. We employ the following pretrained models: **GoogLeNet (Inception v1)** [335], **ResNet50** [336], **WideResNet-50-2** [337], **AlexNet** [44], **SqueezeNet1_0** [338], **DenseNet-201** [339], **MobileNetV2**[340], **MnasNet1_0** [341], **ResNeXt-50 32×4d** [342], **VGG16** [343], **EfficientNet-B2**[1] [344], and **Vision Transformer** [345].

For all the models, we add a classification layer with two units at the top of the models. Regarding the Vision Transformer, the output of the Vision Transformer (z_0^L) serving as the image representation is passed through a dense layer with two units in order to get the final output.

6.3.1 Experiments

All experiments are conducted on a single Tesla P100-PCIE-16GB GPU.

Experimental Setup Firstly, we divide the train set provided by the Challenge into a train and a validation set (65-35%). All models have been trained with an Adam optimizer and a learning rate of 1e-5. We train the proposed architectures five times. We apply *ReduceLROnPlateau*, where we reduce the learning rate by a factor of 0.1, if the validation loss has stopped decreasing for three consecutive epochs. Also, we apply *EarlyStopping* and stop training if the validation loss has stopped decreasing for six consecutive epochs. We minimize the cross-entropy loss function. We test the proposed models using the ADReSS Challenge test set. We average the results obtained by the five repetitions. All models have been created using the PyTorch library [346]. We have used the Transformers library [305] for exploiting the Vision Transformer[2,3].

Evaluation Metrics Accuracy, Precision, Recall, F1-Score, and Specificity have been used for evaluating the results of the introduced architectures. These metrics have been computed by regarding the dementia class as the positive one.

[1] We experimented with EfficientNet-B0 to B7, but EfficientNet-B2 was the best performing model.

[2] google/vit-base-patch16-224-in21k

[3] We also use the ViTFeatureExtractor.

6.3.2 Results

The results of the proposed models mentioned in Section 6.3, which receive as input either log-Mel Spectrograms or MFCCs, are reported in Table 6.1.

In terms of the proposed models with log-Mel Spectrograms as input, as one can easily observe, the Vision Transformer constitutes our best performing model outperforming the other pretrained models in terms of all the evaluation metrics except specificity. To be more precise, Vision Transformer surpasses the other models in accuracy by 2.08-14.58%, in precision by 1.64-11.22%, in recall by 5.00-39.17%, and in F1-score by 2.85-21.85%. The second best performing model is the AlexNet achieving accuracy and F1-score equal to 62.92% and 66.91% respectively. VGG16 constitutes the third best model achieving F1-score and Accuracy equal to 65.55% and 61.25% respectively. The other pretrained models achieve almost equal accuracy results ranging from 53.33% to 59.16% except for DenseNet-201, which performs very poorly with the accuracy accounting for 50.42%.

In terms of the proposed models with MFCCs as input, we observe that the Vision Transformer constitutes the best performing model attaining an Accuracy score of 63.33% and an F1-score of 60.30%. Specifically, it surpasses the other models in Accuracy by 0.41-9.17%, in F1-score by 0.10-6.24%, and in Precision by 0.13-12.85%. AlexNet is the second best performing model achieving an Accuracy of 62.92%, while it surpasses the other models in Accuracy by 2.93-8.76%. MnasNet1_0, GoogleNet, and VGG16 achieve almost equal accuracy scores ranging from 59.17% to 59.99% with the MnasNet1_0 achieving the highest Accuracy score. Next, SqueezeNet1_0 and DenseNet-201 yield equal accuracy scores accounting for 58.75%, with SqueezeNet1_0 outperforming DenseNet-201 in F1-score by 0.72%. MobileNetV2 achieves an Accuracy score of 57.92% followed by EfficientNet-B2, whose accuracy accounts for 57.08. EfficientNet-B2 yields the highest Recall equal to 65.00%, surpassing the other models by 5.00-10.84%. ResNeXt-50 32×4d achieves the worst accuracy score accounting for 54.16%.

In both cases, i.e., log-Mel spectrograms and MFCCs, we observe that Vision Transformer constitutes our best performing model. This can be justified by the fact that all the other pretrained models are based on Convolutional Neural Networks (CNNs). On the contrary, the Vision Transformer does not imply any convolution layer. Specifically, the image is split in patches and is fed to the Vision Transformer network, which exploits the concept of the self-attention mechanism introduced in [347]. Therefore, we believe that the difference in performance is attributable to the transformer encoder, which consists of multi-head self-attention and is implemented in the Vision Transformer.

Table 6.1: Performance comparison among proposed models (using only speech) on the ADReSS Challenge test set. Reported values are mean ± standard deviation. Results are averaged across five runs. Best results per evaluation metric and method are in bold.

	Evaluation metrics				
Architecture	**Precision**	**Recall**	**F1-score**	**Accuracy**	**Specificity**

log-Mel Spectrogram					
GoogLeNet (Inception v1)	57.01	70.00	60.92	57.08	44.17
	±4.70	±19.08	±7.43	±4.86	±24.80
ResNet50	58.93	41.66	47.91	55.00	**68.33**
	±9.31	±6.97	±3.61	±4.86	±14.58
WideResNet-50-2	52.99	64.16	57.70	53.75	43.33
	±1.95	±10.74	±5.39	±2.43	±8.58
AlexNet	60.07	75.83	66.91	62.92	50.00
	±2.60	±9.28	±5.35	±4.04	±2.64
SqueezeNet1_0	57.13	74.16	64.52	59.16	44.16
	±2.61	±1.66	±2.18	±3.12	±4.99
DenseNet-201	50.49	70.00	58.46	50.42	30.83
	±3.58	±7.64	±3.81	±4.82	±10.74
MobileNetV2	54.92	73.33	62.69	56.66	40.00
	±1.51	±7.73	±3.70	±2.43	±4.25
MnasNet1_0	56.66	70.00	59.84	55.83	41.66
	±6.08	±22.88	±7.42	±4.45	±28.99
ResNeXt-50 32 × 4d	53.69	64.16	58.09	53.75	43.33
	±3.99	±6.24	±2.06	±4.45	±13.59
VGG16	58.89	74.16	65.55	61.25	48.33
	±1.18	±6.66	±3.27	±2.12	±3.33
EfficientNet-B2	54.16	58.33	55.46	53.33	48.33
	±6.44	±7.91	±3.48	±5.53	±15.72
Vision Transformer (ViT)	**61.71**	**80.83**	**69.76**	**65.00**	49.16
	±2.93	±6.24	±1.61	±2.76	±10.34
MFCCs					
GoogLeNet (Inception v1)	60.77	55.00	57.49	59.58	64.17
	±3.84	±6.66	±4.19	±3.39	±6.77
ResNet50	56.38	59.16	57.30	56.25	53.33
	±3.83	±8.50	±3.45	±3.49	±12.47
WideResNet-50-2	55.87	54.16	54.06	55.00	55.83
	±3.86	±11.79	±4.51	±2.12	±14.34
AlexNet	65.88	55.00	59.53	62.92	**70.83**
	±5.94	±7.64	±5.13	±4.04	±8.33
SqueezeNet1_0	58.82	59.16	58.55	58.75	58.33
	±2.89	±9.65	±5.13	±2.76	±8.33
DenseNet-201	59.40	56.66	57.83	58.75	60.83
	±3.31	±4.25	±2.22	±2.43	±6.77
MobileNetV2	57.76	57.50	57.22	57.92	58.33
	±2.44	±11.61	±6.38	±3.58	±5.89

MnasNet1_0	63.42 ±10.27	56.66 ±14.81	57.63 ±9.56	59.99 ±6.77	63.33 ±17.95
ResNeXt-50 32 × 4d	53.16 ±3.19	60.00 ±14.09	55.88 ±8.32	54.16 ±3.73	48.33 ±6.77
VGG16	59.20 ±2.75	60.00 ±3.33	59.49 ±1.61	59.17 ±2.12	58.33 ±5.89
EfficientNet-B2	56.40 ±5.89	**65.00** ±7.26	60.20 ±5.51	57.08 ±5.98	49.16 ±10.34
Vision Transformer (ViT)	**66.01** ±3.36	55.83 ±4.25	**60.30** ±1.89	**63.33** ±1.66	**70.83** ±5.89

6.4 Proposed Predictive Models using Speech and Transcripts

In this section, we describe the models used for detecting AD patients using transcripts along with their audio files. We have exploited the python library PyLangAcq [275] for having access to the manual transcripts, since the dataset has been created using the CHAT [274] coding system. For processing the audio files, we use the same procedure mentioned in Section 6.3. We mention below the proposed models used in our experiments.

BERT + ViT In this model we pass each transcription through a pretrained BERT model [347, 26] and get the output of the BERT model (CLS token). Regarding the audio files, we convert them into Log-Mel spectrograms (and MFCCs), their delta, and delta-delta for constructing an image consisting of three channels and pass the image through the ViT. We exploit the Vision Transformer, since it constitutes the best performing model as discussed in Section 6.3.2. The output of the ViT (z_0^L) is concatenated with the output of the BERT and then the resulting vector is passed through a dense layer with 512 units and a ReLU activation function followed by a dense layer consisting of two units to get the final output. The proposed model is illustrated in Fig. 6.1.

BERT + ViT + Gated Multimodal Unit In this model we pass each transcription through a pretrained BERT model and get the output of the BERT model (CLS token). Regarding the audio files, we convert them into Log-Mel spectrograms (and MFCCs), their delta, and delta-delta for constructing an image consisting of three channels and pass the image through the ViT. We exploit the Vision Transformer, since it constitutes the best performing model as discussed in Section 6.3.2. We get the output of the ViT (z_0^L). Next, we employ the Gated Multimodal Unit (GMU) introduced by [75], in order to control the contribution of each modality towards the final classification. The equations governing the GMU are described below:

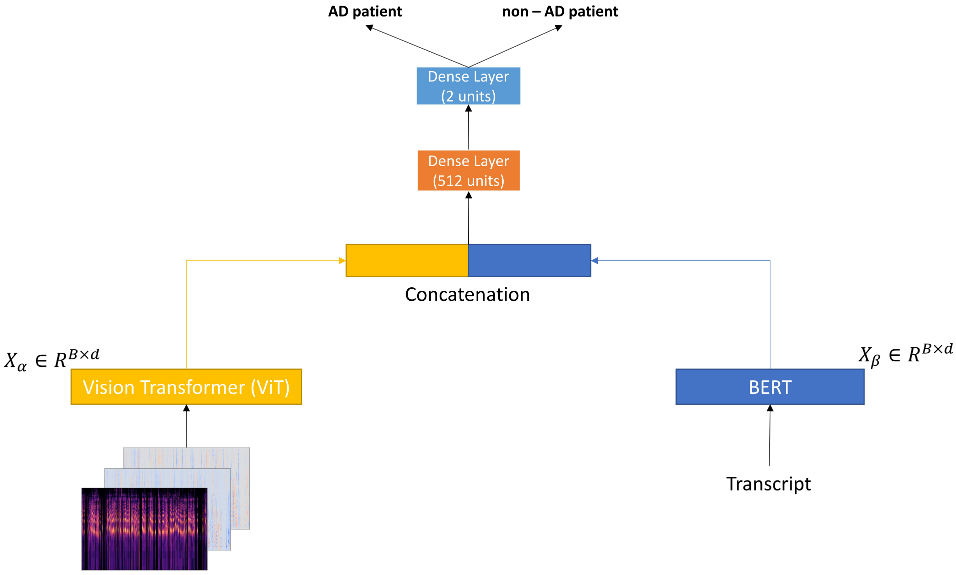


Figure 6.1: BERT + ViT

$$h^t = \tanh\left(W^t f^t + b^t\right) \tag{6.1}$$
$$h^v = \tanh\left(W^v f^v + b^v\right) \tag{6.2}$$
$$z = \sigma(W^z[f^t; f^v] + b^z) \tag{6.3}$$
$$h = z * h^t + (1 - z) * h^v \tag{6.4}$$
$$\Theta = \{W^t, W^v, W^z\} \tag{6.5}$$

where f^t and f^v denote the text and image representations respectively, Θ the parameters to be learned, and $[.;.]$ the concatenation operation. Specifically, $W^t \in \mathcal{R}^{128}, W^v \in \mathcal{R}^{128}, W^z \in \mathcal{R}^{128}$.

The output h of the gated multimodal unit is passed through a dense layer consisting of two units.

The proposed model is illustrated in Fig. 6.2.

BERT + ViT + Crossmodal Attention Similar to the previous models, we pass each transcription through a BERT model, and each image through a ViT model. We exploit the Vision Transformer, since it constitutes the best performing model as discussed in Section 6.3.2. The image representation can be denoted as $X_\alpha \in \mathbb{R}^{B,T_\alpha,d_\alpha}$, while the text representation can be represented as $X_\beta \in \mathbb{R}^{B,T_\beta,d_\alpha}$, where B constitutes the batch size, $T_{(.)}$ the sequence length, and d_α the feature dimension. Next, we employ the crossmodal attention [76, 77, 78]. Specifically, we employ two crossmodal attentions, one from text to image representations and another one from image to text representations. Formally, the crossmodal attention from text to image representation is given by the equations below.

Specifically, we define the queries, keys, and values as:

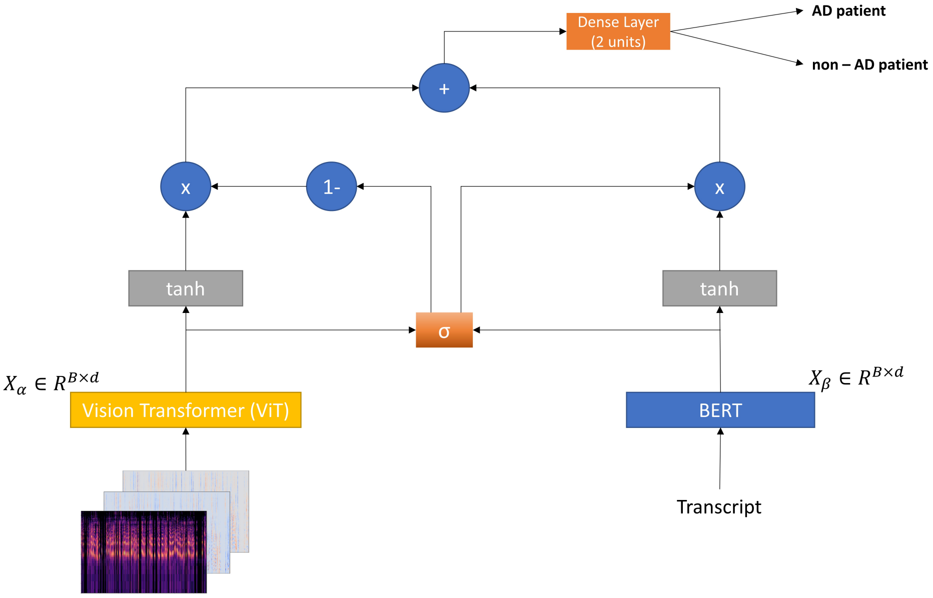


Figure 6.2: BERT + ViT + Gated Multimodal Unit

$$Q_\alpha = X_\alpha W_{Q_\alpha}, K_\beta = X_\beta W_{K_\beta}, V_\beta = X_\beta W_{V_\beta} \tag{6.6}$$

, where $W_{Q_\alpha} \in \mathcal{R}^{d_\alpha \times d_k}, W_{K_\beta} \in \mathcal{R}^{d_\alpha \times d_k}$, and $W_{V_\beta} \in \mathcal{R}^{d_\alpha \times d_v}$ are learnable parameters. Therefore,

$$Q_\alpha \in \mathcal{R}^{B\times T_\alpha \times d_k}, K_\beta \in \mathcal{R}^{B\times T_\beta \times d_k}, V_\beta \in \mathcal{R}^{B\times T_\beta \times d_v} \tag{6.7}$$

The latent adaptation from β to α is presented as the crossmodal attention, given by the equations below:

$$\begin{aligned} Y_\alpha &= CM_{\beta\rightarrow\alpha}(X_\alpha, X_\beta) \\ &= softmax\left(\frac{Q_\alpha K_\beta^T}{\sqrt{d_k}}\right) V_\beta \\ &= softmax\left(\frac{X_\alpha W_{Q_\alpha} W_{K_\beta}^T X_\beta^T}{\sqrt{d_k}}\right) X_\beta W_{V_\beta} \end{aligned} \tag{6.8}$$

The scaled (by $\sqrt{d_k}$) softmax is a score matrix, where the (i,j)-th entry measures the attention given by the i-th time step of modality α to the j-th time step of modality β. The i-th time step of Y_α is a weighted summary of V_β, with the weight determined by i-th row in softmax$(\cdot)$.

Similarly, the crossmodal attention from image to text representation is given by the equations below:

$$Q_\beta = X_\beta W_{Q_\beta}, K_\alpha = X_\alpha W_{K_\alpha}, V_\alpha = X_\alpha W_{V_\alpha} \tag{6.9}$$

$$Q_\beta \in \mathcal{R}^{B\times T_\beta \times d_k}, K_\alpha \in \mathcal{R}^{B\times T_\alpha \times d_k}, V_\alpha \in \mathcal{R}^{B\times T_\alpha \times d_v} \tag{6.10}$$

$$\begin{aligned} Y_\beta &= CM_{\alpha\rightarrow\beta}(X_\beta, X_\alpha) \\ &= softmax\left(\frac{Q_\beta K_\alpha^T}{\sqrt{d_k}}\right) V_\alpha \\ &= softmax\left(\frac{X_\beta W_{Q_\beta} W_{K_\alpha}^T X_\alpha^T}{\sqrt{d_k}}\right) X_\alpha W_{V_\alpha} \end{aligned} \tag{6.11}$$

The outputs of the crossmodal attention layers, i.e., Y_α and Y_β, are concatenated and passed through a global average pooling layer followed by a dense layer with two units. The proposed model is illustrated in Fig. 6.3.

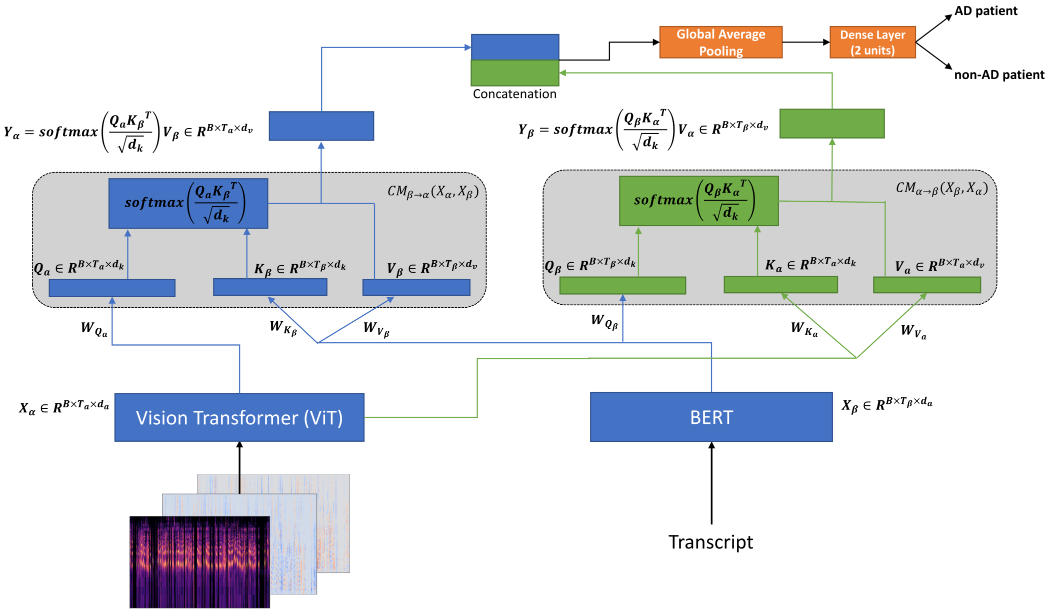

AD patient
Global Average Pooling
Dense Layer (2 units)
non-AD patient
Concatenation
$Y_\alpha = softmax\left(\frac{Q_a K_\beta^T}{\sqrt{d_k}}\right) V_\beta \in R^{B\times T_a \times d_v}$
$Y_\beta = softmax\left(\frac{Q_\beta K_\alpha^T}{\sqrt{d_k}}\right) V_\alpha \in R^{B\times T_\beta \times d_v}$
$softmax\left(\frac{Q_a K_\beta^T}{\sqrt{d_k}}\right)$
$CM_{\beta\rightarrow\alpha}(X_\alpha, X_\beta)$
$softmax\left(\frac{Q_\beta K_\alpha^T}{\sqrt{d_k}}\right)$
$CM_{\alpha\rightarrow\beta}(X_\beta, X_\alpha)$
$Q_a \in R^{B\times T_a \times d_k}$
$K_\beta \in R^{B\times T_\beta \times d_k}$
$V_\beta \in R^{B\times T_\beta \times d_v}$
$Q_\beta \in R^{B\times T_\beta \times d_k}$
$K_a \in R^{B\times T_a \times d_k}$
$V_a \in R^{B\times T_a \times d_v}$
W_{Q_a}
W_{K_β}
W_{V_β}
W_{Q_β}
W_{K_a}
W_{V_a}
$X_\alpha \in R^{B\times T_a \times d_a}$
Vision Transformer (ViT)
BERT
$X_\beta \in R^{B\times T_\beta \times d_a}$
Transcript

Figure 6.3: BERT + ViT + Crossmodal Attention

6.4.1 Experiments

All experiments are conducted on a single Tesla P100-PCIE-16GB GPU.

Comparison with state-of-the-art approaches

1. **Unimodal state-of-the-art approaches (only transcripts)**

   - BERT (Chapter 5): This method trains a BERT model using transcripts.

2. **Multimodal state-of-the-art approaches (speech and transcripts)**

- top-3 late fusion [258]: This method proposes a late fusion approach of the three best feature configurations, namely Temporal + char4grams, New + char4grams, and char4grams. The authors train a Random Forest Classifier.
- Audio + Text (Fusion) [92]: The authors introduce three models for detecting AD patients using only speech data and three models for detecting AD patients using only text data. Finally, they use a majority level approach, where the final prediction corresponds to the class getting the most votes from the six aforementioned models.
- SVM [251]: This method extracts lexicosyntactic, semantic, and acoustic features, performs feature selection using ANOVA, and finally trains a Support Vector Machine Classifier.
- Fusion Maj. (3-best) [63]: This method uses a majority vote of three approaches, namely Bag-of-Audio-Words, zero-frequency filtered (ZFF) signals, and BiLSTM-Attention network.
- LSTM with Gating (Acoustic + Lexical + Dis) [62]: This research work extracts a set of features from speech and transcripts, passes the respective sets of features through two branches of BiLSTMs, one branch for each modality. Next the authors introduce feed-forward highway layers with a gating mechanism.
- System 3: Phonemes and Audio [250]: This method transcribes the segment text into phoneme written pronunciation using CMUDict and combines this representation of features with features extracted via the audio.
- Fusion of system [67]: This method merges features extracted via speech and transcripts and trains a Support Vector Machine Classifier. Features of speech constitute the x-vectors. In terms of the language features, (i) a Global Maximum pooling, (ii) a bidirectional LSTM-RNNs provided with an attention module, and (iii) the second model augmented with part-of-speech (POS) embeddings are trained on the top of a pretrained BERT model.
- Bimodal Network (Ensembled Output) [263]: In this research work, the outputs of the top 5 bimodal networks with high validation results are ensembled and used as the final submission.
- GFI, NUW, Duration, Character 4-grams, Suffixes, POS tag, UD [95]: This method exploits the gunning fog index, number of unique words, duration of the audio file, character 4-grams, suffixes, pos-tags, and Universal dependency features in a tf-idf setting. Logistic Regression is trained with the corresponding feature sets.
- Acoustic & Transcript [90]: This method employs the scores from the whole training subset to train a final fusion GBR model that is used to perform the fusion of scores coming from the acoustic and transcript-based models for the challenge evaluation.

- Dual BERT [91]: This method employs a Speech BERT and a Text BERT and concatenates their representations.
- Model C [259]: This method extracts features from segmented audio and passes them through GRU layers. Regarding the transcripts, this method extracts pos-tags and passes both the transcripts and pos-tags through two separate CNN layers. Then the outputs of the CNN layers are passed through a BiLSTM layer coupled with an Attention Layer. The authors also extract a different set of features from both transcripts and audio files and pass them to a dense layer. The respective outputs are concatenated and passed to a dense layer, which gives the final output.
- Majority vote (NLP + Acoustic) [64]: This method obtains firstly the best-performing acoustic and language-based models. Next, it computes a weighted majority-vote ensemble meta-algorithm for classification. The authors choose the three best-performing acoustic models along with the best-performing language model, and compute a final prediction by taking a linear weighted combination of the individual model predictions.

Experimental Setup Firstly, we divide the train set provided by the Challenge into a train and a validation set (65-35%). Next, we train the proposed architectures five times with an Adam optimizer and a learning rate of 1e-5. We apply *ReduceLROnPlateau*, where we reduce the learning rate by a factor of 0.1, if the validation loss has stopped decreasing for three consecutive epochs. Also, we apply *EarlyStopping* and stop training if the validation loss has stopped decreasing for six consecutive epochs. We minimize the cross-entropy loss function. All models have been created using the PyTorch library [346]. We use the BERT base uncased version. We test the proposed models using the test set provided by the Challenge. We average the results obtained by the five repetitions.

Evaluation Metrics Accuracy, Precision, Recall, F1-Score, and Specificity have been used for evaluating the results of the introduced architectures. These metrics have been computed by regarding the dementia class as the positive one.

6.4.2 Results

The results of the proposed models mentioned in Section 6.4 are reported in Table 6.2. Also this table presents a comparison of our introduced models with both unimodal and multimodal state-of-the-art approaches.

Regarding our proposed transformer-based models with log-Mel spectrogram as input, one can observe that BERT+ViT+Crossmodal Attention constitutes our best performing model surpassing the other introduced models in F1-score and Accuracy, while it achieves equal Recall score with BERT+ViT+Gated Multimodal Unit. More specifically, BERT+ViT+Crossmodal Attention outperforms BERT+ViT in recall by a margin of 10.84%, in F1-score by 3.22%, and in accuracy by 2.08%, confirming that the crossmodal attention improves the performance of the multimodal models. Also, it outperforms BERT+ViT+Gated Multimodal Unit in F1-score by 2.77% and in Accuracy by 3.33%. In addition, BERT+ViT+Gated Multimodal Unit surpasses BERT+ViT in Recall and F1-score by 10.84% and 0.45% respectively. Although BERT+ViT surpasses the other proposed models in Specificity by 6.67-13.34%, it must be noted that F1-score is a more important metric than Specificity in health-related tasks, since high Specificity and low F1-score means that AD patients are misdiagnosed as non-AD ones.

As one can easily observe, our best performing model, namely BERT+ViT+Crossmodal Attention, surpasses the performance of the multimodal state-of-the-art models, except [258, 92], in Accuracy by 3.13-15.41%, while it outperforms the research works in Recall by 3.67-29.17% and in F1-score by 3.29-18.93%. At the same time, BERT+ViT+Crossmodal Attention obtains a higher accuracy score than BERT (Chapter 5) outperforming it by 0.83%. BERT+ViT+Crossmodal Attention outperforms BERT in F1-score by 1.96%. At the same time, the standard deviations of BERT+ViT+Crossmodal Attention in both F1-score and Accuracy are lower than the standard deviations of BERT (Chapter 5). This fact indicates the superiority of our introduced model and shows that it can capture effectively the interactions between the two modalities. Regarding BERT+ViT, we can observe that it surpasses the multimodal state-of-the-art models, except [258, 92], in Accuracy and F1-score by 1.05-13.33% and 0.07-15.71% respectively. Thus, the combination of transformer networks, i.e., BERT and ViT, outperforms or obtains comparable performance to the multimodal state-of-the-art approaches. Although BERT+ViT surpasses Fusion Maj. (3-best) [63] in F1-score by a small margin of 0.07%, it must be noted that our proposed model is more computationally and time effective, since the method in [63] trains three different models in order to enhance the classification performance. We observe also that BERT+ViT performs worse than BERT (Chapter 5). We speculate that this difference of 1.25% in Accuracy is attributable to the concatenation operation. In terms of BERT+ViT+Gated Multimodal Unit, it also outperforms the state-of-the-art approaches in F1-score and Accuracy except for [63, 258, 92]. Although BERT (Chapter 5) outperforms BERT+ViT+Gated Multimodal Unit in terms of F1-score and Accuracy, the results show that BERT+ViT+Gated Multimodal Unit can better capture the relevant information of the two modalities on the test set in comparison to the performances of the existing research initiatives proposing multimodal models.

Regarding our proposed transformer-based models with MFCCs as input, one can observe that BERT + ViT + Crossmodal Attention constitutes our best performing model attaining an Accuracy score of 87.92% and an F1-score of 87.99%. Specifically, it outperforms the introduced models in Accuracy by 2.50-3.76%, in F1-score by 1.92-3.65%, and in Recall by 3.33-10.00%. Similarly to the proposed transformer-based models with log-Mel spectrogram, we observe that the crossmodal attention yields better results than the concatenation operation and the gated multimodal unit. In addition, we observe that the BERT+ViT+Gated Multimodal Unit surpasses BERT+ViT in Accuracy by 1.26%.

However, BERT+ViT outperforms BERT+ViT+Gated Multimodal Unit in F1-score by 1.73%.

In comparison with the existing research initiatives, we observe that BERT + ViT + Crossmodal Attention improves the performance obtained by BERT (Chapter 5). Specifically, Accuracy is improved by 0.42%, F1-score sees an improvement of 1.26%, and Recall is improved by 7.50%. On the contrary, BERT+ViT and BERT+ViT+Gated Multimodal Unit obtain worse performance than BERT (Chapter 5). Compared with the multimodal state-of-the-art approaches, BERT+ViT+Crossmodal Attention surpasses the research works, except [258, 92], in Accuracy by 2.72-15.00%, in F1-score by 2.59-18.23%, and in Recall by 1.16-26.66%. BERT+ViT+Gated Multimodal Unit outperforms the research works, except [258, 92], in Accuracy by 0.22-12.50%. Finally, BERT+ViT surpasses the research works, except [258, 92, 63], in Accuracy by 1.16-11.24%, while it outperforms the research work [63] in F1-score by 0.67%.

6.5 Discussion

The identification of dementia from spontaneous speech constitutes a hot topic in recent years due to the fact that it is time and cost-efficient. Although several research works have been proposed towards diagnosing dementia from speech, there are still limitations. For example, most methods extract features from speech or transcripts and train traditional Machine Learning classifiers. Another significant limitation has to do with the way the different modalities, e.g., speech and transcripts, are combined in a single neural network. Specifically, research works train separately speech-based and text-based networks and then use majority voting approaches, thus increasing significantly the training time. Other research works add or concatenate the text and image representations, thus treating equally the two modalities and obtaining suboptimal performance. Furthermore, although transformers have achieved state-of-the-art results in many domains, their potential has not been fully exploited in the task of dementia detection using speech data. To the best of our knowledge, this is the first study employing the Vision Transformer for detecting dementia only from speech. This study aims also to fill gaps with regards to the usage of multimodal models by introducing the Gated Multimodal Unit and the crossmodal attention layers, which have not been applied before in the task of dementia identification from spontaneous speech. From the results obtained in this study, we found that:

- **Finding 1:** The Vision Transformer (receiving as input images consisting of log-Mel spectrogram, delta, and delta-delta) outperformed the other pretrained models, i.e., ResNet50, WideResNet-50-2, AlexNet, etc., in all the evaluation metrics except for Specificity. Similarly, the Vision Transformer (receiving as input images consisting of MFCCs, delta, and delta-delta) obtained higher scores by the other models in Accuracy, F1-score, and Precision. We believe that the Vision Transformer consti-

Table 6.2: Performance comparison among proposed models (using both speech and transcripts) and state-of-the-art approaches on the ADReSS Challenge test set. Reported values are mean ± standard deviation. Results are averaged across five runs.

	Evaluation metrics				
Architecture	**Precision**	**Recall**	**F1-score**	**Accuracy**	**Specificity**
Unimodal state-of-the-art approaches (only transcripts)					
BERT (Chapter 5)	87.19 ±3.25	81.66 ±5.00	86.73 ±4.53	87.50 ±4.37	93.33 ±5.65
Multimodal state-of-the-art approaches (speech and transcripts)					
top-3 late fusion [258]	-	-	-	93.75	-
Audio + Text (Fusion) [92]	-	87.50	-	89.58	91.67
SVM [251]	80.00	83.00	82.00	81.30	79.00
Fusion Maj. (3-best) [63]	-	-	85.40	85.20	-
LSTM with Gating (Acoustic + Lexical + Dis) [62]	81.82	75.00	78.26	79.17	83.33
System 3: Phonemes and Audio [250]	81.82	75.00	78.26	79.17	83.33
Fusion of system [67]	94.12	66.67	78.05	81.25	95.83
Bimodal Network (Ensembled Output) [263]	89.47	70.83	79.07	81.25	91.67
GFI,NUW,Duration,Character 4-grams,Suffixes, POS tag,UD [95]	-	-	-	77.08	-
Acoustic & Transcript [90]	70.00	88.00	78.00	75.00	83.00
Dual BERT [91]	83.04 ±3.97	83.33 ±5.89	82.92 ±1.86	82.92 ±1.56	82.50 ±5.53
Model C [259]	78.94	62.50	69.76	72.92	83.33
Majority vote (NLP + Acoustic) [64]	-	-	-	83.00	-
Proposed Transformer-based models (log-Mel Spectrogram)					
BERT+ViT	90.73 ±2.74	80.83 ±2.04	85.47 ±1.70	86.25 ±1.67	91.67 ±2.64
BERT+ViT+Gated Multimodal Unit	80.92 ±2.30	91.67 ±3.73	85.92 ±2.37	85.00 ±2.43	78.33 ±3.12
BERT+ViT+Crossmodal Attention	86.13 ±3.26	91.67 ±4.56	88.69 ±2.12	88.33 ±2.12	85.00 ±4.25
Proposed Transformer-based models (MFCCs)					
BERT+ViT	86.72 ±2.05	85.83 ±6.77	86.07 ±2.69	84.16 ±1.02	86.66 ±3.12
BERT+ViT+Gated Multimodal Unit	90.57 ±2.80	79.16 ±5.89	84.34 ±3.53	85.42 ±2.95	91.66 ±2.64
BERT+ViT+Crossmodal Attention	87.09 ±2.40	89.16 ±5.65	87.99 ±2.79	87.92 ±2.43	86.66 ±3.12

tutes our best performing model due to the transformer encoder and the multi-head self-attention. On the contrary, all the other pretrained models are based on convolutional neural networks.

- **Finding 2:** We compared the performance achieved between BERT and BERT+ViT and showed that BERT+ViT achieved slightly worse results. We speculated that this difference may be attributable to the usage of a simple concatenation of the text and image representations. A simple concatenation operation assigns equal importance to the different modalities. In addition, we compared the performance of BERT+ViT on the test set with 13 research works and showed that BERT+ViT outperformed most of the research works in F1-score and Accuracy. Thus, transformers achieve comparable performance to state-of-the-art approaches.

- **Finding 3:** Results on the ADReSS Challenge test set showed that BERT + ViT + Gated Multimodal Unit (with log-Mel spectrogram) yielded a higher F1-score than BERT + ViT (with log-Mel spectrogram), while BERT +ViT + Gated Multimodal Unit (with MFCCs) yielded a higher Accuracy score than BERT+ViT (with MFCCs). In addition, we compared the performance of BERT+ViT+Gated Multimodal Unit on the test set with 13 multimodal research works and showed that BERT+ViT+Gated Multimodal Unit achieved comparable performance.

- **Finding 4:** We presented a new method to detect AD patients consisting of BERT, ViT, and crossmodal attention layers and showed that crossmodal interactions outperform the competitive multimodal models. We compared our best performing model (BERT+ViT+Crossmodal Attention with log-Mel spectrogram as input) with 13 research works on the ADReSS Challenge test set and showed that our introduced model outperformed 11 of these strong baselines in Accuracy and F1-score by a large margin of 3.13-15.41% and 3.29-18.93% respectively. Moreover, the incorporation of the crossmodal attention enhanced the performance obtained by BERT by 0.83% in Accuracy and by 1.96% in F1-score. In terms of BERT+ViT+Crossmodal Attention (with MFCCs), we observed that it outperformed 11 of 13 strong baselines in Accuracy and F1-score by a large margin of 2.72-15.00% and 2.59-18.23% respectively, while it achieved better performance than BERT. Also, we observed that the variances of BERT + ViT + Crossmodal Attention by using either log-Mel Spectrogram or MFCCs are lower than BERT (Chapter 5).

  Also, we observed that BERT + ViT + Crossmodal Attention outperforms both BERT+ViT and BERT + ViT + Gated Multimodal Unit. Specifically, BERT + ViT + Crossmodal Attention performs better than BERT+ViT, since BERT+ViT fuses the features of different modalities through a concatenation operation. The concatenation operation ignores inherent correlations between different modalities. In addition, BERT + ViT + Crossmodal Attention outperforms BERT+ViT+Gated Multimodal Unit. This can be justified by the fact that the Gated Multimodal Unit is inspired by the flow control in recurrent architectures, such as GRU or LSTM. Specifically, the Gated Multimodal Unit controls only the information flow from each modality and does not capture interactions between text and image. On the contrary, the usage of the crossmodal attention layers captures the crossmodal interactions, enabling one modality for receiving information from another modality. More specifically, we pass textual information to speech and speech information to text. Therefore, we observe that controlling the flow of information from the two modalities is not sufficient. On the contrary, learning crossmodal interactions is more important.

  In addition, we observed that our best performing model, i.e., BERT + ViT +

Crossmodal Attention, outperforms most of the strong baselines. This fact justifies our initial hypothesis that early and late fusion strategies and the usage of concatenation or add operation introduced by other studies do not capture effectively the inter-modal interactions of different modalities, thus obtain in this way suboptimal performance.

One limitation of the current research work has to do with the limited number of samples in the ADReSS Challenge dataset, i.e., 78 AD and 78 non-AD patients. However, as mentioned in Section 3.3.5.2, one cannot overlook that this dataset is matched for gender and age, so as to mitigate bias in the prediction task. Concurrently, in contrast to other datasets, it has been carefully selected so as to mitigate common biases often overlooked in evaluations of AD detection methods, including repeated occurrences of speech from the same participant and variations in audio quality. Moreover, it is balanced, since it includes 78 AD and 78 non-AD patients. It is also used widely by a lot of research works dealing with the task of dementia identification from speech.

6.6 Summary

In this chapter, we proposed methods to differentiate AD from non-AD patients using either only speech or both speech and transcripts. Regarding the models using only speech, we exploited several pretrained models used extensively in the computer vision domain, with the Vision Transformer achieving the highest F1-score and accuracy accounting for 69.76% and 65.00% respectively. Next, we employed three neural network models in which we combined speech and transcripts. We exploited the Gated Multimodal Unit, in order to control the influence of each modality towards the final classification. In addition, we experimented with crossmodal interactions, where we used the crossmodal attention. Results showed that crossmodal attention can enhance the performance of competitive multimodal approaches and surpass state-of-the-art approaches. More specifically, models incorporating the crossmodal attention yielded accuracy equal to 88.83% on the ADReSS Challenge test set.

In Chapters 5 and 6, we concentrated our experiments on detecting AD patients, i.e., binary classification task. In Chapter 5, we divided the MMSE scores into four groups depending on the severity of dementia and performed a multitask learning framework, where the identification of the severity of dementia constituted the auxiliary task, i.e., multiclass classification task. However, the exact estimation of MMSE score is crucial. Therefore, in the next chapter, we will continue with proposing advanced fusion methods and will perform our experiments on both detecting the AD patients and predicting their MMSE scores.

Chapter 7

Multimodal Deep Learning Models for Detecting Dementia and Predicting Mini-Mental State Examination scores from Speech and Transcripts

7.1 Introduction

In the previous chapters, we introduced unimodal models exploiting either speech or transcripts for detecting AD patients. Multimodal models were also proposed. The main task was the classification of a single subject as AD patient or non-AD one. Therefore, in this chapter, we will proceed with experimenting with fusion methods and will extend our experiments on predicting the MMSE scores, i.e., regression task. Several research works have been proposed aiming to predict the Mini-Mental State Examination (MMSE) scores using the modalities of both speech and transcripts. However, the majority of them have introduced averaging approaches [169, 170, 61]. Specifically, they train several textual and acoustic models and they make the final prediction by simply averaging the predictions of the individual models.

In order to tackle the aforementioned limitations, in this chapter, we employ transformer-based networks, which can capture effectively the interaction between the different modalities and control the importance of each modality towards the final prediction. Compared with recent deep ensemble learning methods, which need to train models individually and then fuse the results of the classifiers, the proposed neural networks in this chapter can be trained in an end-to-end trainable manner. Similar to Chapter 6, we extract Log-Mel spectrograms, their delta, and delta-delta (acceleration values) and construct an image per audio file consisting of three channels. Next, we introduce a neural network consisting of BERT and Vision Transformer (ViT) for extracting textual and visual embeddings respectively, and add a co-attention mechanism over the respective embeddings, which can attend at the different modalities at the same time. In addition, we introduce an architecture, which integrates multimodal information into a BERT model via an Attention Gate called Multimodal Shifting Gate. To be more precise, we propose three variations of this architecture, where we inject *(a)* textual and visual, *(b)* textual and acoustic, and *(c)* textual, visual, and acoustic information into the BERT model. Finally, we propose an architecture, which can learn both the inter- and intra-modal interactions, i.e., image-image, text-text, text-image, and image-text, and show that it achieves state-of-the art results. Contrary to Chapter 6, in this chapter, we propose a self-attention layer which includes a gating mechanism. Compared with prior works, our methods provide important advantages, since they can learn more representative features regarding the different modalities and require also less time for training.

The contributions of this chapter can be summarized as follows:

- We conduct extensive experiments for detecting AD patients (AD classification task) and predicting the MMSE scores (MMSE regression task).
- We propose a multimodal model consisting of BERT, ViT, and a Co-Attention mechanism.
- We introduce an architecture, which incorporates a Multimodal Shifting Gate aiming to control the importance of text, acoustic, and visual representations. The conjunction of the textual, acoustic, and visual embeddings is fed to a BERT model.
- We propose an architecture aiming to model the inter- and intra-modal interactions of multimodal data.
- We achieve competitive results with state-of-the-art approaches on the ADReSS Challenge dataset both in the AD classification and MMSE regression task.
- Our best performing model achieves a new state-of-the-art result in the MMSE regression task.

7.2 Dataset

We use the ADReSS Challenge Dataset described in Section 3.3.5.2 for conducting our experiments.

7.3 Problem Statement

7.3.1 AD Classification Task

Let a labeled dataset consist of transcripts and their corresponding audio files belonging to AD patients and non-AD ones. Transcripts belonging to AD subjects are given the label

1, while transcripts belonging to the non-AD patients are given the label 0. The task is to identify, if a transcript along with its audio file belongs to a person suffering from dementia, or to a person belonging to the healthy control group (binary classification problem).

7.3.2 MMSE Regression Task

Let a dataset consist of transcripts and their corresponding audio files belonging to AD patients and non-AD ones. Each transcript along with the audio file has been assigned with a MMSE score ranging from 0 to 30, where a MMSE score of 25-30 is considered as normal, a MMSE score of 21-24 as mild, a MMSE score of 10-20 as moderate, and a MMSE score less than 10 as severe impairment [62]. Given the transcript and the audio file, the task is to predict the MMSE score (regression problem).

7.4 Predictive Models

In this section, we present the proposed predictive models for detecting dementia using speech and transcripts. We use the python library PyLangAcq [275] for having access to the manual transcripts, since the dataset has been created using the CHAT [348] coding system. Moreover, we employ the Python library librosa [334] for converting the audio files to Log-Mel spectrograms, their delta, and delta-delta (acceleration values). For all the experiments conducted, we use 224 Mel bands, hop length equal to 1024, and a Hanning window. Each image is resized to (224 × 224) pixels.

7.4.1 BERT + ViT + Co-Attention

We pass the transcripts through a BERT model [347, 26] and the corresponding images through a ViT model [349]. Then, we use a co-attention mechanism [79, 80] over the outputs of the aforementioned models, since it can help learn the attention weights of transcripts and image patches concurrently.

Formally, let $C \in \mathbb{R}^{d \times N}$ and $S \in \mathbb{R}^{d \times T}$ be the outputs of the BERT and ViT pretrained models respectively. Following the methodology proposed by [79], given the output of the BERT $\left(\mathbf{C} \in \mathbb{R}^{d \times N}\right)$ and the output of the ViT $\left(\mathbf{S} \in \mathbb{R}^{d \times T}\right)$, where d denotes the hidden size of the model, N and T the sequence length of the transcripts and image patches respectively, the affinity matrix $F \in \mathbb{R}^{N \times T}$ is calculated using the equation presented below:

$$F = \tanh\left(C^T W_l S\right) \tag{7.1}$$

where $W_l \in \mathbb{R}^{d \times d}$ is a matrix of learnable parameters. Next, this affinity matrix is considered as a feature and we learn to predict the transcript and image attention maps via the following,

$$H^s = \tanh\left(W_s S + \left(W_c C\right) F\right) \tag{7.2}$$

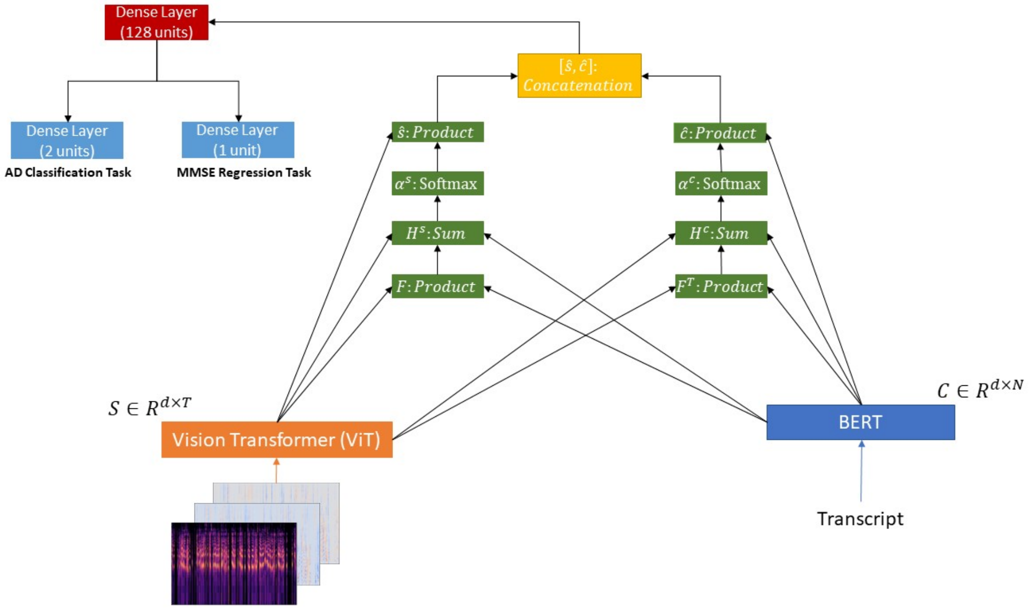


Figure 7.1: BERT + ViT + Co-Attention

$$H^c = \tanh\left(W_c C + (W_s S) F^T\right) \tag{7.3}$$

where $W_s, W_c \in \mathbb{R}^{k \times d}$ are matrices of learnable parameters. The attention probabilities for each word in the transcripts and each image patch are calculated through the softmax function as follows,

$$a^s = softmax\left(w_{hs}^T H^s\right) \tag{7.4}$$

$$a^c = softmax\left(w_{hc}^T H^c\right) \tag{7.5}$$

where $a_s \in \mathbb{R}^{1 \times T}$ and $a_c \in \mathbb{R}^{1 \times N}$. $w_{hs}, w_{hc} \in \mathbb{R}^{k \times 1}$ are the weight parameters. Based on the above attention weights, the attention vectors for text and image representations are obtained via the following equations:

$$\hat{s} = \sum_{i=1}^{T} a_i^s s^i, \quad \hat{c} = \sum_{j=1}^{N} a_j^c c^j \tag{7.6}$$

where $\hat{s} \in \mathbb{R}^{1 \times d}$ and $\hat{c} \in \mathbb{R}^{1 \times d}$.
Finally, these two vectors are concatenated.

Regarding the **AD detection problem** described in Section 7.3.1, the resulting vector $\left(p \in \mathbb{R}^{1 \times 2d}\right)$ is passed to a dense layer with 128 units and a ReLU activation function followed by a dense layer consisting of two units.

Regarding the **MMSE prediction problem** described in Section 7.3.2, the resulting vector $\left(p \in \mathbb{R}^{1 \times 2d}\right)$ is passed to a dense layer with 128 units and a ReLU activation function followed by a dense layer consisting of one unit with a ReLU activation function.

The proposed architecture is illustrated in Fig. 7.1.

7.4.2 Multimodal BERT

In this section, we exploit the method proposed in Chapter 4. First, we pass each transcript through a BERT model obtaining a text representation $X \in \mathbb{R}^{N \times d}$. Similarly, we pass each image through a ViT model and get the output of the ViT model ($z_0^L \in \mathbb{R}^{1 \times d}$). Then, we repeat the vector z_0^L N times, in order that the text and image representation matrices have the same size. Regarding the acoustic modality, we use the Python library openSMILE [318] for extracting the eGeMAPSv02 feature set per audio file. We obtain a vector of 88d per audio file, where we project the respective vector to a 256d vector and repeat it N times. Let $e^{(i)}, h_{\alpha}^{(i)}$, and $h_v^{(i)}$ denote word, acoustic, and image representation for the i-th word in a sequence. Next, we concatenate the representations (text-image and text-audio) using two attention gating mechanisms as described via the equations below:

$$w_v^{(i)} = \sigma\left(W_{hv}[h_v^{(i)}; e^{(i)}] + b_v\right) \tag{7.1}$$

$$w_{\alpha}^{(i)} = \sigma\left(W_{h\alpha}[h_{\alpha}^{(i)}; e^{(i)}] + b_{\alpha}\right) \tag{7.2}$$

where σ denotes the sigmoid activation function, $W_{hv}, W_{h\alpha}$ are two weight matrices, and $w_v^{(i)}, w_{\alpha}^{(i)}$ correspond to the visual and acoustic gates respectively. b_v and b_{α} are the scalar biases.

Next, we calculate a nonverbal shift vector $h_m^{(i)}$ by multiplying the visual embeddings with the visual gate and the acoustic embeddings with the acoustic gate.

$$h_m^{(i)} = w_v^{(i)} \cdot \left(W_v h_v^{(i)}\right) + w_{\alpha}^{(i)} \cdot \left(W_{\alpha} h_{\alpha}^{(i)}\right) + b_m^{(i)} \tag{7.3}$$

where W_a and W_v are weight matrices for acoustic and visual information respectively. $b_m^{(i)}$ is the bias vector.

Next, we apply the Multimodal Shifting component aiming to dynamically shift the word representations by integrating the nonverbal shift vector $h_m^{(i)}$ into the original word embedding.

$$e_m^{(i)} = e^{(i)} + \alpha h_m^{(i)} \tag{7.4}$$

$$\alpha = min\left(\frac{||e^{(i)}||_2}{||h_m^{(i)}||_2}\beta, 1\right) \tag{7.5}$$

, where β is a hyperparameter. Then, we apply a layer normalization [28] and dropout layer [350] to $e_m^{(i)}$. Finally, the combined embeddings are fed to a BERT model.

Regarding the **AD detection problem** described in Section 7.3.1, the CLS token constituting the output of the BERT model is passed through a dense layer with 128 units and a ReLU activation function followed by a dense layer with two units, which gives the final output.

Regarding the **MMSE prediction problem** described in Section 7.3.2, the CLS token constituting the output of the BERT model is passed through a dense layer with 128 units and a ReLU activation function followed by a dense layer with one unit and a ReLU activation function.

We experiment with injecting acoustic information **(Multimodal BERT - eGeMAPS)**, visual information **(Multimodal BERT - ViT)**, and both acoustic and visual information **(Multimodal BERT - eGeMAPS + ViT)**.

The architecture **(Multimodal BERT - eGeMAPS + ViT)** is illustrated in Fig. 7.2.

7.4.3 BERT + ViT + Gated Self-Attention

Similar to the aforementioned introduced models, we pass each transcript through a BERT model and each image through a ViT model. Let $X \in \mathbb{R}^{N \times d}$ and $Y \in \mathbb{R}^{T \times d}$ be the outputs of the BERT and ViT pretrained models respectively. In this section, our main aim is to model the intra-modal and inter-modal interactions at the same time (i.e., $X \rightarrow X$, $Y \rightarrow Y$, and $X \leftrightarrow Y$). Thus, we adopt the methodology introduced by [83].

After having obtained $X \in \mathbb{R}^{N \times d}$ and $Y \in \mathbb{R}^{T \times d}$, which correspond to the text and image representations respectively, we concatenate these two representations as follows:

$$Z = [X; Y] \tag{7.1}$$

Next, $Z \in \mathbb{R}^{m \times d}$, where $m = N + T$, is considered the query Q, key K, and value V, as follows:

$$Q = Z, K = Z, V = Z \tag{7.2}$$

Next, we adopt the gating model introduced by [83] as follows:

$$M = \sigma\left(FC^g\left(FC_q^g\left(Q\right) \odot FC_k^g\left(K\right)\right)\right) \tag{7.3}$$

where $FC_q^g, FC_k^g \in \mathbb{R}^{d \times d_g}$, $FC^g \in \mathbb{R}^{d_g \times 2}$ are three fully-connected layers, and d_g denotes the dimensionality of the projected space. $\odot$ denotes the element-wise product function and σ the sigmoid function. In addition, $M \in \mathbb{R}^{m \times 2}$ corresponds to the two masks $M_q \in \mathbb{R}^m$ and $M_k \in \mathbb{R}^m$ for the features Q and V respectively.

Next, the two masks M and K are tiled to $\tilde{M_q}, \tilde{M_k} \in \mathbb{R}^{m \times d}$ and then used for computing the attention map as following:

$$A^g = softmax\left(\frac{\left(Q \odot \tilde{M_q}\right)\left(K \odot \tilde{M_k}\right)^T}{\sqrt{d}}\right) \tag{7.4}$$

$$H = A^g V \tag{7.5}$$

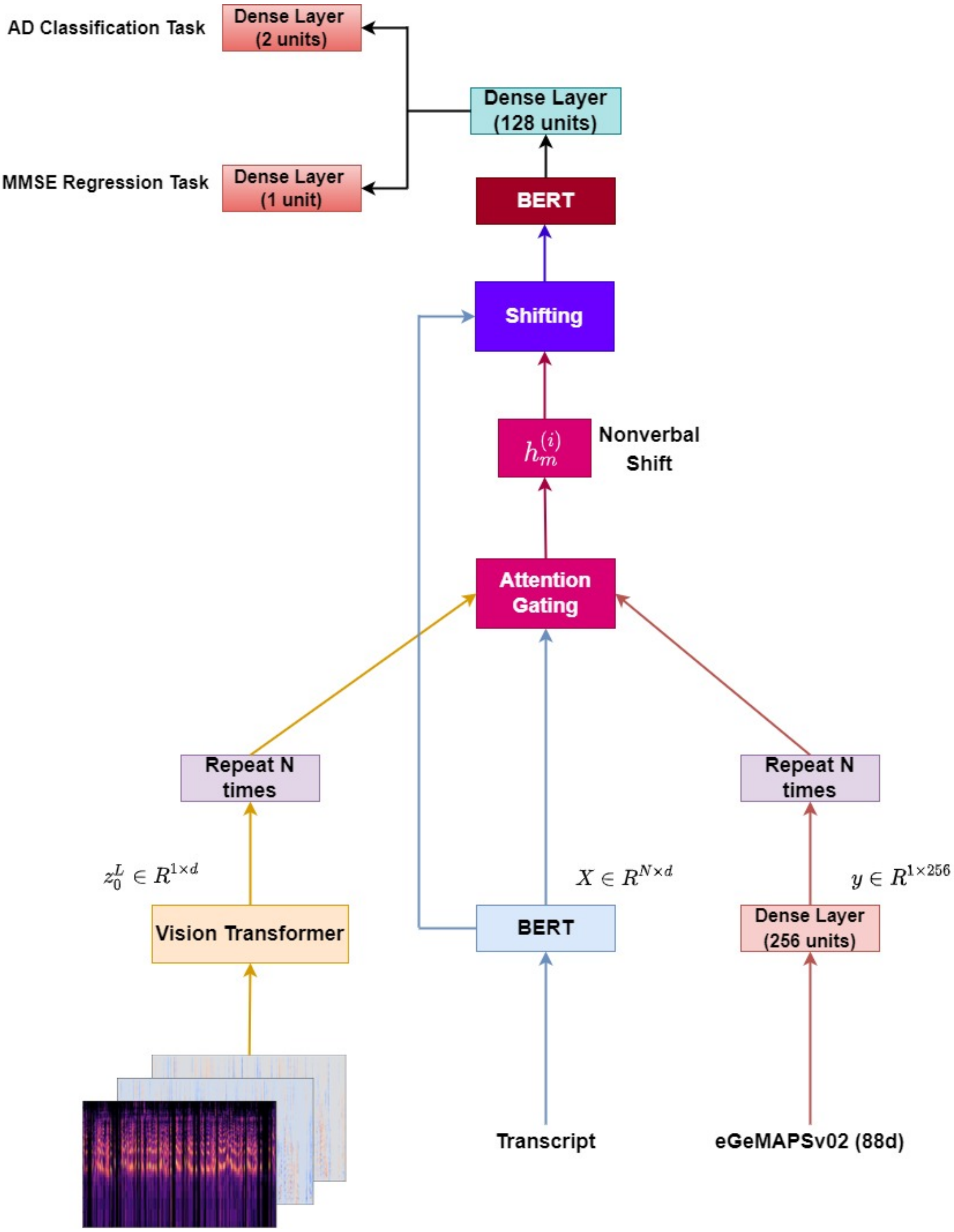


Figure 7.2: Multimodal BERT - eGeMAPS + ViT

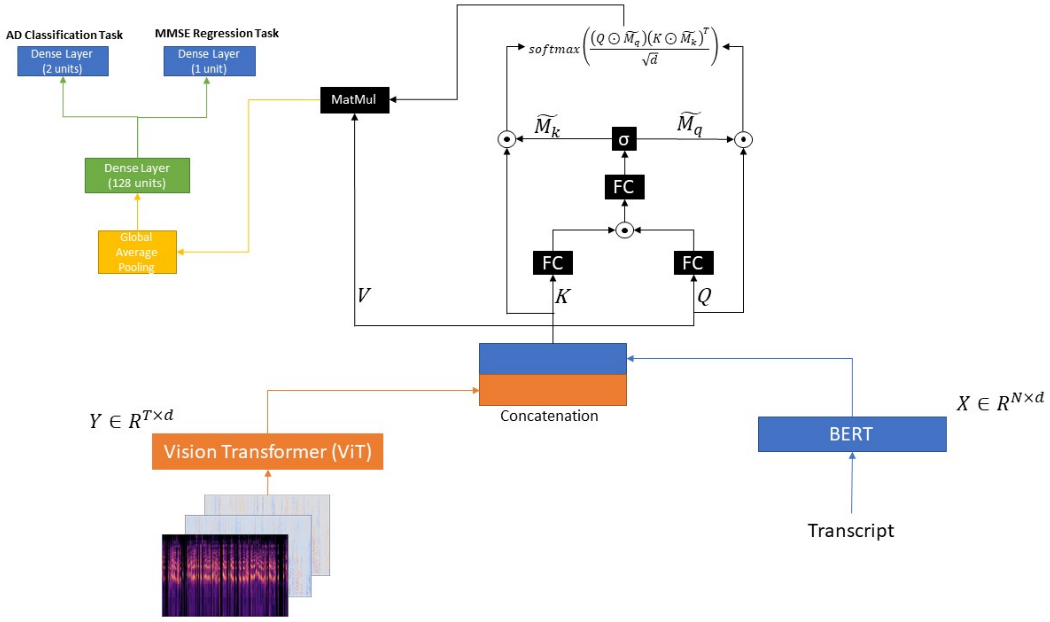


Figure 7.3: BERT + ViT + Gated Self-Attention

Then, the output H is passed through a global average pooling layer followed by a dense layer with 128 units and a ReLU activation function.

Regarding the **AD detection problem** described in Section 7.3.1, we use a dense layer with two units, which gives the final output.

Regarding the **MMSE prediction problem** described in Section 7.3.2, we use a dense layer with one unit and a ReLU activation function.

The proposed architecture is illustrated in Fig. 7.3.

7.5 Experiments

7.5.1 Comparison with state-of-the-art approaches

We compare our introduced models with research works proposing either unimodal or multimodal approaches. These research works have been selected due to the fact that they conduct their experiments on the ADReSS Challenge test set. These research works are reported in Tables 7.1, 7.2, and 7.3. More specifically, Table 7.1 refers to research works using multimodal approaches, Table 7.2 refers to research works proposing unimodal approaches using only text, and Table 7.3 refers to research works proposing unimodal approaches using only speech.

Table 7.1: Overview of the multimodal state-of-the-art approaches, which are later compared with our work.

Reference	Architecture	Features/Methodology	Task
[351]	Fusion Maj./W-avg (3-best)	Bag-of-Audio-Words, zero-frequency filtered (ZFF) signals, and BiLSTM-Attention network	AD/MMSE
[62]	LSTM with Gating (Acoustic + Lexical + Dis)	Acoustic, Linguistic Features, Bi-LSTM, gating mechanism	AD/MMSE
[250]	System 3: Phonemes and Audio	phoneme written pronunciation using CMUDict + acoustic features	AD
[67]	Fusion of System	fusion of x-vectors with linguistic features, train SVM	AD
[263]	Bimodal Network (Ensembled Output)	Ensemble (top-5 bimodal networks)	AD/MMSE
[95]	GFI, NUW, Duration, Character 4-grams, Suffixes, POS tag, UD	feature extraction, Logistic Regression Classifier	AD
[90]	Acoustic & Transcript	fusion of the acoustic (x-vectors) and transcript (BERT) model scores	AD
[90]	Acoustic+silence & Transcript	Average the scores from the different models, four silence features	MMSE
[91]	Dual BERT	concatenation of the representations obtained by BERT and Speech BERT	AD
[259]	Model C	Neural network consisting of CNN, BiLSTM, Attention, GRU, and Dense layers	AD
[64]	Majority vote (NLP + Acoustic)	final prediction by taking a linear weighted combination of the individual model predictions	AD
[64]	Random Forest (NLP) + gradient boosting (acoustic)	language/fluency/n-gram features, MFCC and delta coefficients, Dimensionality Reduction Techniques	MMSE

Continued on next page

Table 7.1 – continued from previous page

Reference	Architecture	Features/Methodology	Task
[61]	Audio + Text	majority level approach of six models, averaging-based fusion	AD/MMSE
[170]	Ensemble	Majority voting approach, average the predictions	AD/MMSE
[66]	Attempt 4	label fusion from the top-5 performing models from audio and text modalities (top-5 from each modality), average value of predictions of individual models	AD/MMSE
[254]	SELECTED-FEATURE	For selecting the features, a Random Forest regression model was trained. The authors retained only features having mean decrease impurity (MDI) values exceeding a predefined threshold	MMSE

Table 7.2: Overview of the unimodal state-of-the-art approaches using only text, which are later compared with our work.

Reference	Architecture	Features/Methodology	Task
[351]	bi-LSTM-Att	GloVe 100d as pretrained weights, maximum word number for each transcript is 200, Bi-LSTM with attention	AD/MMSE
[62]	LSTM (Lexical + Dis)	GloVe features of 100d, disfluency markers (self-repair), Bi-LSTM	AD/MMSE
[250]	System 2: Phonemes	The authors transcribed the segment text into phoneme written pronunciation using CMUDict. FastText was trained on the phoneme representation	AD
[67]	Sentence Embedding	sentence embeddings are computed by averaging the second to twelfth hidden layers of each word., train SVM	AD
[263]	Transformer-XL	The authors extracted textual features using Transformer-XL and trained a neural network consisting of CNN, Attention, Bi-LSTM, and Dense Layers.	AD/MMSE
[90]	Transcript	The authors train a BERT model.	AD/MMSE
[91]	Longformer	Training of Longformer	AD

Continued on next page

Table 7.2 – continued from previous page

Reference	Architecture	Features/Methodology	Task
[259]	Model A0	Neural network consisting of CNN, LSTM, and Dense layers	AD
[64]	Logistic Regression (NLP)	language and fluency features, n-gram features, Dimensionality Reduction Techniques	AD
[64]	Random Forest (NLP)	language and fluency features, n-gram features, Dimensionality Reduction Techniques	MMSE
[61]	Text (fusion)	fusion of top-3 performing models from the textual modality	AD/MMSE
[66]	Attempt 5	label fusion from the top-10 performing models from text modalities, average of MMSE score predictions from the top-10 performing models	AD/MMSE
[253]	BERT	Training of BERT model	AD
[254]	n-gram	All lexicosyntactic features, SVR training	MMSE
[65]	fastText, bi+trigram	The authors fit 21 models and the outputs are combined by a majority voting scheme for final classification. In the regression task, the outputs of these bootstrap models are averaged to arrive at the final MMSE score	AD/MMSE

Table 7.3: Overview of the unimodal state-of-the-art approaches using only speech, which are later compared with our work.

Reference	Architecture	Features/Methodology	Task
[351]	SiameseNet	a deep Siamese neural network consisting of convolutional layers. As an input, the model used either 8-second or 16-second segments.	AD
[351]	BoAW fusion (3-best)	MelFrequency Cepstral Coefficient (MFCC), log-Mel, and the COMPARE acoustic feature set	MMSE
[62]	LSTM (Acoustic)	higher-order statistics of COVAREP features. Bi-LSTM training	AD/MMSE

Continued on next page

Table 7.3 – continued from previous page

Reference	Architecture	Features/Methodology	Task
[250]	System 1: Audio	LDA posterior probabilities of ComParE2016 features	AD
[67]	x-vectors_SRE	The authors use both the SRE and the Voxceleb models for the x-vectors framework. train SVM	AD
[263]	VGGish	The authors used VGGish features and trained a neural network consisting of Attention Layer, CNN, Bi-LSTM, and Dense Layers.	AD/MMSE
[90]	Acoustic + Silence	silency features, x-vector PCA-transformed coefficients, Probabilistic Linear Discriminant Analysis (PLDA) for detection and Support Vector Regression (SVR) for MMSE prediction	AD/MMSE
[91]	YAMNet	The input of YAMNet is the Mel spectrogram from audio data with dimensions of (p, t, 1)	AD
[259]	Model B0 (emobase)	GRU taking in audio segment features and finally combining the features from the speech segments into a common vector	AD
[64]	Majority vote (Acoustic)	acoustic feature extraction across all speech segments, weighted majority vote classification on segments	AD
[64]	Gradient Boosting (Acoustic)	MFCC 1–16 features and their delta coefficients from 26 Mel-bands	MMSE
[61]	Audio (fusion)	majority level approach of three acoustic models, averaging-based fusion	AD/MMSE
[93]	DemCNN	convolutional neural network for speech classification using the raw waveform	AD
[352]	CNN - LSTM (MFCC)	21 models are fitted using the above 21 bootstrap samples and the outputs are combined by a majority voting scheme for final classification.	AD
[352]	pBLSTM-CNN (log-Mel)	bagging of 21 models by averaging the outputs.	MMSE

Continued on next page

Table 7.3 – continued from previous page

Reference	Architecture	Features/Methodology	Task
[254]	acoustic-all	Mel Frequency Cepstral Coefficients (MFCCs), mean value, variance, etc.	MMSE
[66]	Attempt 3	label fusion from the top-5 performing models from the audio modality, prediction from the BERT base uncased RangePool	AD/MMSE

7.5.2 Experimental Setup

7.5.2.1 Training and Evaluation - Implementation Details

In terms of the **MMSE regression task**, the ADReSS Challenge train set includes the MMSE scores for all the people except one. Thus, we remove this person from the train set in the **MMSE regression task**.

We follow a similar training strategy to the one adopted by [91]. Firstly, we divide the train set provided by the Challenge into a train and a validation set (65%-35%). Next, we train the proposed architectures five times with an Adam optimizer and a learning rate of 1e-5. Regarding the **AD detection problem** described in Section 7.3.1, we minimize the cross-entropy loss function, whereas with regards to the **MMSE prediction problem** described in Section 7.3.2, we minimize the RMSE. We apply *ReduceLROnPlateau*, where we reduce the learning rate by a factor of 0.1, if the validation loss has stopped decreasing for three consecutive epochs. Also, we apply *EarlyStopping* and stop training, if the validation loss has stopped decreasing for six consecutive epochs. We test the proposed models using the test set provided by the Challenge. We average the results obtained by the five repetitions. All models have been created using the PyTorch library [353]. We have used the Vision Transformer (with fixed-size patches of resolution 16 × 16) and the BERT base uncased version from the Transformers library [305]. The input to the BERT and ViT model is the output of the BERT tokenizer and ViT feature extractor respectively as defined by the Transformers library. All experiments are conducted on a single Tesla P100-PCIE-16GB GPU.

7.5.2.2 Hyperparameters

Regarding **BERT+ViT+Co-Attention**, we set k equal to 40. We use dropout after the output of the co-attention layer with a rate of 0.4, and a dropout layer after the dense layer consisting of 128 units with a rate of 0.2. Regarding **(Multimodal BERT - eGeMAPS)**, we set $\beta = 0.01$. In terms of **(Multimodal BERT - ViT)**, we set $\beta = 0.001$. Regarding **(Multimodal BERT - eGeMAPS + ViT)**, we set $\beta = 0.01$. With regards to the following models: **(Multimodal BERT - eGeMAPS)**, **(Multimodal BERT - ViT)**, and **(Multimodal BERT - eGeMAPS + ViT)**, we apply dropout with a rate of 0.4 at the output of (7.4) and freeze the weights of the first BERT model.

Also, we use a dropout layer after the output of the second BERT model with a rate of 0.2. With regards to **BERT+ViT+Gated Self-Attention**, we set $d_g = 64$. We use dropout after the global average pooling layer with a rate of 0.3. For all the experiments conducted, the hidden size of BERT and ViT denoted by d is equal to 768. Moreover, $N = 512$, since we pad each transcript to a maximum number of 512 tokens. T is equal to 197. Thus, m is equal to 709.

7.5.3 Evaluation Metrics

Regarding the **AD detection problem** described in Section 7.3.1, Accuracy, Precision, Recall, F1-Score, and Specificity have been used for evaluating the results of the introduced architectures. These metrics have been computed by regarding the dementia class as the positive one. We report the average and standard deviation of these metrics over five runs.

With regards to the **MMSE prediction problem** described in Section 7.3.2, the RMSE has been used for evaluating the results of the introduced architectures. We report the average and standard deviation of the RMSE scores across five runs. The RMSE is the metric used in the baseline paper provided by the ADReSS challenge.

7.6 Results

7.6.1 AD Classification Task

The results of the proposed models mentioned in Section 7.4 for the AD classification task are reported in Table 7.4. In addition, in this table we compare the results of our introduced models with research works proposing multimodal approaches, unimodal models using only text data, and unimodal approaches using only speech data.

Regarding our proposed models, one can observe from Table 7.4 that BERT + ViT + Gated Self-Attention outperforms all the introduced models in Accuracy and F1-score by a large margin of 2.50-11.25% and 3.13-9.59% respectively. This can be justified by the fact that the Gated Self-Attention aims to capture both the intra- and inter-modal interactions. Specifically, BERT+ViT+Gated Self-Attention outperforms BERT+ViT+Co-Attention in accuracy by 2.50%, in Recall by 7.5%, and in F1-score by 3.13%. Despite the fact that BERT+ViT+Co-Attention obtains a high specificity score accounting for 93.33% outperforming BERT+ViT+Gated Self-Attention by 2.5%, BERT+ViT+Co-Attention attains a low F1-score accounting for 86.81%. On the contrary, BERT+ViT+Gated Self-Attention yields an F1-score of 89.94% outperforming BERT+ViT+Co-Attention by 3.13%. This means that BERT+ViT+Gated Self-Attention can detect better the AD patients than BERT+ViT+Co-Attention, where AD patients are misdiagnosed as non-AD ones. In addition, although BERT+ViT+Gated Self-Attention obtains lower results in Precision and Recall by other introduced models, it surpasses them in F1-score, which constitutes the weighted average of recall and precision. Regarding the Multimodal BERT models, one can observe that Multimodal BERT-ViT outperforms Multimodal BERT-eGeMAPS in accuracy by 0.83%, in recall by 4.17%, and in F1-score by 1.44%. We speculate that Multimodal BERT-ViT performs better than Multimodal BERT-eGeMAPS due to the usage of the Vision Transformer. Thus, the visual modality obtained via ViT seems to perform slightly better than the acoustic modality. In addition, we observe that the injection of both the acoustic and visual information enhances the performance of the models having just one modality, be it either the acoustic modality or the visual one. More specifically, Multimodal BERT-eGeMAPS+ViT surpasses Multimodal BERT-eGeMAPS and Multimodal BERT-ViT in accuracy by 2.08% and 1.25% respectively. In comparison to the Multimodal BERT-eGeMAPS+ViT, BERT+ViT+Gated Self-Attention surpasses its performance in accuracy by 9.17%, in Precision by 14.30%, in F1-score by 7.66%, and in Specificity by 18.33%. Overall, BERT+ViT+Gated Self-Attention constitutes our best performing model, since it surpasses all the other introduced models in F1-score and Accuracy.

In comparison to the multimodal approaches, as one can easily observe from Table 7.4, BERT+ViT+Gated Self-Attention surpasses the state-of-the-art multimodal approaches in Recall by 1.17-26.67%, in F1-Score by 4.54-20.18%, and in Accuracy by 0.42-17.08%. These findings confirm our initial hypothesis that inter- and intra-modal interactions enhance the classification results obtained by approaches, which predict AD patients either by using majority voting on predictions of several individual models or adding/concatenating the text and image representations. In addition, although our best performing model outperforms *Audio+Text* [61] by a small margin of 0.42% in Accuracy and by a larger margin of 1.67% in Recall, it is worth mentioning that our proposed approach is more computational and time-efficient, since the method proposed by [61] employs six models and eventually uses a majority vote approach. In terms of BERT+ViT+Co-Attention, it outperforms all the research works, except *Audio+Text* [61], in Accuracy by 2.30-14.58%. Also, it surpasses all the research works, except *Fusion of System* [67] in Precision by 3.36-22.83%. Also, it surpasses all the research works in F1-score results by 1.41-17.05%. It outperforms four research works out of the eight ones, which report Recall results by 6.67-19.17%. Thus, the co-attention mechanism can yield better performance than the results obtained by the research initiatives, since it can attend to transcripts and images simultaneously. Finally, with regards to the proposed Multimodal BERT models, it seems that they are rather complex for our limited dataset. However, results suggest that Multimodal BERT - eGeMAPS+ViT surpasses six research works in Accuracy by 1.63-7.91%, five research works in F1-score by 3.21-12.52%, all the research works in the Recall score by 1.17-26.67%, and one research work in the Precision score by 6.57%.

In comparison to the unimodal approaches using only text data, as one can easily observe from Table 7.4, the approach proposed by [61] outperforms our best performing model in terms of accuracy, recall, and specificity by 1.67%, 2.50%, and 0.84% respectively. However, our best performing model outperforms all the other approaches in accuracy by 4.58-17.10%, in Recall by 5.84-35.01%, in Precision by 2.73-21.87%, in F1-score by 6.67-

26.53%, and in Specificity by 2.83-7.50%.

In comparison to the unimodal approaches using only speech data, as one can easily observe from Table 7.4, BERT+ViT+Gated Self-Attention outperforms the research initiatives in terms of Precision, Recall, F1-score, and Accuracy. More specifically, BERT + ViT + Gated Self-Attention surpasses the research works in Precision by 8.87-36.70%, in Recall by 5.84-51.17%, in F1-score by 19.14-38.94%, and in Accuracy by 8.75-35.83%. In addition, BERT+ViT+Co-Attention surpasses the research works in Precision by 10.83-38.66%, in F1-score by 16.01-35.81%, and in Accuracy by 6.25-33.33%. Additionaly, Multimodal BERT - eGeMAPS, Multimodal BERT - ViT, and Multimodal BERT - eGeMAPS + ViT outperform all the research initiatives except [61] in terms of the accuracy score by a margin of 5.83-24.58%, 6.66-25.41%, and 7.91-26.66% respectively.

It is obvious that the unimodal approaches exploiting only speech data achieve low evaluation results in comparison with unimodal approaches employing text data or multimodal models.

Table 7.4: AD Classification Task: Performance comparison among proposed models and state-of-the-art approaches on the ADReSS Challenge test set. Reported values are mean ± standard deviation. Results are averaged across five runs. Best results per evaluation metric are in bold.

Architecture	P.	R.	F1-score	Acc.	Spec.
State-of-the-art approaches (Multimodal)					
Fusion Maj (3-best) [351]	-	-	85.40	85.20	-
LSTM with Gating (Acoustic + Lexical + Dis) [62]	81.82	75.00	78.26	79.17	83.33
System 3: Phonemes and Audio [250]	81.82	75.00	78.26	79.17	83.33
Fusion of System [67]	**94.12**	66.67	78.05	81.25	**95.83**
Bimodal Network (Ensembled Output) [263]	89.47	70.83	79.07	81.25	91.67
GFI, NUW, Duration, Character 4-grams, Suffixes, POS tag, UD [95]	-	-	-	77.08	-
Acoustic & Transcript [90]	70.00	88.00	78.00	75.00	63.00
Dual BERT [91]	83.04±3.97	83.33±5.89	82.92±1.86	82.92±1.56	82.50 ± 5.53
Model C [259]	78.94	62.50	69.76	72.92	83.33
Majority vote (NLP + Acoustic) [64]	-	-	-	83.00	-
Audio + Text [61]	-	87.50	-	89.58	91.67
Ensemble [170]	83.00	83.00	83.00	83.00	-
Attempt 4 [66]	-	-	-	79.17	-
State-of-the-art approaches (only Text)					
bi-LSTM-Att [351]	-	-	81.20	81.30	-
LSTM (Lexical + Dis) [62]	76.19	66.67	71.11	72.92	79.10

System 2: Phonemes [250]	80.95	70.83	75.56	77.08	83.33
Sentence Embedding [67]	82.35	58.33	68.29	72.92	87.50
Transformer-XL [263]	80.00	83.33	81.63	81.25	79.17
Transcript [90]	69.00	83.00	75.00	72.92	63.00
Longformer [91]	88.14±2.09	74.17±5.53	80.44±3.55	82.08±2.83	90.00 ± 2.04
Model A0 [259]	76.47	54.16	63.41	68.75	83.33
Logistic Regression (NLP) [64]	-	-	-	85.00	-
Text (fusion) [61]	-	**91.67**	-	**91.67**	91.67
Attempt 5 [66]	-	-	-	85.42	-
BERT [253]	83.89	83.33	83.27	83.32	83.33
fastText, bi+trigram [65]	86.00	79.00	83.00	83.33	88.00
State-of-the-art approaches (only Speech)					
SiameseNet [351]	-	-	70.80	70.80	-
LSTM (Acoustic) [62]	-	-	-	66.60	-
System 1: Audio [250]	58.62	70.83	64.15	60.42	50.00
x-vectors_SRE [67]	54.17	54.17	54.17	54.17	54.17
VGGish [263]	78.95	62.50	69.77	72.92	83.33
Acoustic + Silence [90]	70.00	58.00	63.00	66.70	75.00
YAMNet [91]	64.40±3.93	73.40±8.82	68.60±4.84	66.20±4.79	59.20 ± 7.73
Model B0 (emobase) [259]	65.21	62.50	63.82	64.58	66.67
Majority vote (Acoustic) [64]	-	-	-	65.00	-
Audio (fusion) [61]	-	83.33	-	81.25	79.17
DemCNN [93]	62.50	62.50	62.50	62.50	62.50
CNN - LSTM (MFCC) [352]	82.00	38.00	51.00	64.58	92.00
Attempt 3 [66]	-	-	-	64.58	-
Proposed Transformer-based models					
BERT+ViT+Co-Attention	92.83±6.39	81.67±2.04	86.81±3.37	87.50±3.49	93.33±6.24
Multimodal BERT - eGeMAPS	74.51±1.01	87.50±6.45	80.35±2.77	78.75±2.04	70.00±3.12
Multimodal BERT - ViT	73.91±2.40	**91.67**±2.64	81.79±1.72	79.58±2.04	67.50±4.08
Multimodal BERT - eGeMAPS+ViT	76.57±3.74	89.17±5.65	82.28±3.49	80.83±3.58	72.50±5.65
BERT+ViT+Gated Self-Attention	90.87±3.50	89.17±2.04	**89.94**±1.36	90.00±1.56	90.83±4.08

7.6.2 MMSE Regression Task

The results of the proposed models mentioned in Section 7.4 for the MMSE regression task are reported in Table 7.5. In addition, in this table we compare the results of our introduced models with research works proposing multimodal approaches, unimodal models using only text data, and unimodal approaches using only speech data.

Regarding our proposed models, one can observe from Table 7.5 that BERT + ViT +

Gated Self-Attention obtains the lowest RMSE score accounting for 3.61 followed by BERT + ViT + Co-Attention, whose RMSE score is equal to 4.20. Regarding Multimodal BERT - eGeMAPS, Multimodal BERT - ViT, and Multimodal BERT - eGeMAPS + ViT, it is obvious that these neural networks are complex for the MMSE regression task achieving RMSE scores equal to 5.64, 5.50, and 5.62 respectively.

In comparison to the multimodal approaches, as one can easily observe from Table 7.5, BERT + ViT + Gated Self-Attention, which constitutes our best performing model, improves the RMSE score obtained by the multimodal state-of-the-art approaches by 0.15-2.40. Regarding BERT + ViT + Co-Attention, it improves the RMSE scores of all the existing research initiatives, except *Bimodal Network (Ensembled Output)* [263], by 0.14-1.41. In terms of the Multimodal BERT - eGeMAPS, Multimodal BERT - ViT, and Multimodal BERT - eGeMAPS + ViT, it seems that these architectures are rather complex for the MMSE regression task improving the RMSE score of only one research work [64].

In comparison with the unimodal approaches exploiting only text data, one can easily observe from Table 7.5 that BERT + ViT + Gated Self-Attention performs better than the existing research initiatives improving the current RMSE score by 0.13-2.25. In addition, BERT + ViT + Co-Attention achieves comparable performance to existing research works outperforming all the existing research works, except *Transformer-XL* [263] and *Text (fusion)* [61], by 0.10-1.66. Finally, Multimodal BERT - ViT obtains lower RMSE score than the one obtained by [90, 64].

In comparison with the unimodal approaches using only speech data, one can observe from Table 7.5 that BERT + ViT + Gated Self-Attention outperforms all the research initiatives by a large margin of 1.47-3.06. Similarly, BERT + ViT + Co-Attention obtains lower RMSE score than the scores achieved by all the research works. Specifically, the performance gain ranges from 0.48 to 2.47. Finally, Multimodal BERT - eGeMAPS, Multimodal BERT - ViT, and Multimodal BERT - eGeMAPS + ViT outperform all the state-of-the-art approaches, except *Attempt 3* [66] and *VGGish* [263], improving the RMSE score by 0.22-1.03, 0.36-1.17, and 0.24-1.05 respectively.

It is obvious that the research works exploiting only speech data obtain higher RMSE scores than the ones exploiting text data or the combination of text and speech data.

Table 7.5: MMSE Regression Task: Performance comparison among proposed models and state-of-the-art approaches on the ADReSS Challenge test set. Reported values are mean ± standard deviation. Results are averaged across five runs. Best results are in bold.

Architecture	**RMSE**
State-of-the-art approaches (Multimodal)	
Fusion Wavg (3-best) [351]	4.65
LSTM with Gating (Acoustic + Lexical + Dis) [62]	4.54
Bimodal Network (Ensembled Output) [263]	3.77

Acoustic+silence & Transcript [90]	5.32
Random Forest (NLP) + gradient boosting (acoustic) [64]	6.01
Audio + Text [61]	4.47
Ensemble [95]	5.06
Attempt 4 [66]	4.91
SELECTED FEATURE [254]	4.34
State-of-the-art approaches (only Text)	
bi-LSTM-Att [351]	4.66
LSTM (Lexical + Dis) [62]	4.88
Transformer-XL [263]	4.02
Transcript [90]	5.86
Random Forest (NLP) [64]	5.62
Text (fusion) [61]	3.74
Attempt 5 [66]	4.30
n-gram [254]	4.61
fastText, bi+trigram [65]	4.87
State-of-the-art approaches (only Speech)	
BoAW fusion (3-best) [351]	6.45
LSTM (Acoustic) [62]	5.93
VGGish [263]	5.08
Acoustic + Silence [90]	5.97
Gradient Boosting (Acoustic) [64]	6.67
Audio (fusion) [61]	5.86
pBLSTMCNN (log-Mel) [352]	5.90
acoustic-all [254]	6.42
Attempt 3 [66]	5.18
Proposed Transformer-based models	
BERT+ViT+Co-Attention	4.20 ±0.47
Multimodal BERT - eGeMAPS	5.64 ±0.11
Multimodal BERT - ViT	5.50 ±0.30
Multimodal BERT - eGeMAPS+ViT	5.62 ±0.12
BERT+ViT+Gated Self-Attention	**3.61** ±0.48

7.7 Discussion

The detection of dementia from spontaneous speech has emerged into a hot topic throughout the years due to the fact that it constitutes a time-effective procedure. Although dementia detection from speech is a hot topic and item of interest from several researchers around the world, there are still significant limitations that need to be addressed. The main limitation is pertinent to the way the different modalities, i.e., acoustic, visual, and textual, are combined in a single neural network. Research works having proposed multimodal methods tend to train separately acoustic, language, and visual models and then apply majority vote or average-based approaches for the AD classification and MMSE regression task respectively. In addition, they tend to add or concatenate the representations obtained by the different modalities, thus treating equally each modality. Therefore, in this study, we aim to tackle the aforementioned limitations and propose three novel architectures, which combine the different modalities effectively achieving competitive performance to existing research initiatives.

From the results obtained in this study for the AD classification task, we found that:

- **Finding 1:** The incorporation of a co-attention mechanism, which can learn the attention weights for words and image patches simultaneously, outperforms the multimodal research initiatives except one in terms of the Accuracy score.
- **Finding 2:** We propose a method to inject visual and acoustic modalities along with the textual one into a BERT model via a Multimodal Shifting Gate. We experiment with injecting only visual information, only acoustic information, and their combination. Findings state that the injection of both modalities performs better than the injection of single modalities.
- **Finding 3:** We introduce an approach aiming to model both the inter- and intra-modal interactions at the same time and show that this approach is the best performing one among the introduced approaches.

From the results obtained in this study for the MMSE regression task, we found that:

- **Finding 4:** The incorporation of the co-attention mechanism at the top of the pretrained models, i.e., BERT and ViT, obtains low RMSE improving all the state-of-the-art approaches except [263, 61].
- **Finding 5:** Multimodal BERT models do not perform well to the MMSE regression task. These architectures are rather complex for the limited dataset used in this study.
- **Finding 6:** BERT+ViT+Gated Self-Attention improves the RMSE score in the MMSE regression task by 0.13-3.06 obtaining a new state-of-the-art result. The ability of this architecture to perform well both in the AD classification task and in the MMSE regression task establishes the usefulness of this architecture for the dementia detection problem and indicates that both the inter- and intra-modal interactions are important.

Although the unimodal approach proposed by [61] outperforms our best performing model in the AD classification task, our best performing model obtains better results in the

MMSE regression task. In addition, our introduced model is more computationally and time-effective, since the approach by [61] extracts embeddings by employing transformer networks, applies feature aggregation techniques, trains traditional machine learning algorithms, and finally applies a majority voting approach of the top-3 performing models. Regarding the multimodal approach proposed by [61], it achieves lower evaluation results than the unimodal approach. We speculate that this degradation in performance is attributable to the fact that the majority-vote approach does not take the interactions between the different modalities into consideration.

7.8 Summary

In this chapter, we introduced three novel multimodal neural networks for detecting dementia (AD classification task) and predicting the MMSE scores (MMSE regression task) from spontaneous speech. First, we proposed a model consisting of BERT, ViT, and a co-attention mechanism at the top of the proposed architecture, which is capable of attending to both the words and the image patches simultaneously. Results indicated that the proposed model achieved an accuracy of 87.50% in the AD classification task outperforming all the research works proposing multimodal approaches except one. Regarding the MMSE regression task, our proposed architecture achieved an RMSE score equal to 4.20. Secondly, we introduced a deep learning architecture, where we injected information from the visual and acoustic modalities along with the textual one into a BERT model and used an attention gate mechanism to control the importance of each modality. Results for the AD classification task suggested that the injection of both the acoustic and visual modalities enhanced the performance of the models achieved when using only either the acoustic or the visual modality along with the textual one. Finally, we introduced a transformer-based network, where we concatenated the representations obtained via BERT and ViT and passed the representation through a self-attention mechanism incorporating a novel gating mechanism. Findings showed that this introduced model was the best performing one on the ADReSS Challenge test set reaching Accuracy and F1-score up to 90.00% and 89.94% respectively. In terms of the MMSE regression task, our best performing model obtained an RMSE score of 3.61 improving the state-of-the-art RMSE scores for the regression task of the ADReSS Challenge by 0.13-3.06.

In Chapters 5-7, we utilized manual transcripts for conducting our experiments. However, manual transcripts are not always available. Therefore, in the next chapter, we will continue with experimenting with multimodal fusion methods by utilizing both manual and automatic transcripts.

Chapter 8

Context-Aware Attention Layers coupled with Optimal Transport Domain Adaptation and Multimodal Fusion methods for recognizing dementia

8.1 Introduction

In the previous chapters, we introduced methods for fusing the different modalities utilizing manual transcripts. However, some limitations still exist. Specifically, manual transcripts are not available in clinical settings. Additionally, in Chapter 7, we proposed a method, which concatenates the representation vectors of the two modalities and exploits a self-attention layer incorporating a gated model. However, in terms of the textual modality recent studies have shown that Self-Attention layers treat the input sequence as a bag-of-word tokens and each token individually performs attention over the bag-of-word tokens. Consequently, the contextual information is not taken into account in the calculation of dependencies between elements. There have been proposed a number of studies enhancing the self-attention layers with contextual information [354, 355, 356, 357].

In addition, the reliability of a machine learning model's confidence in its predictions, denoted as calibration [11, 12], is critical for high risk applications, such as deciding whether to trust a medical diagnosis prediction [156, 157, 158]. However, no prior work has taken into account the calibration of the models, creating in this way overconfident models. According to [171], modern neural networks are not well-calibrated, while they are overconfident at the same time.

In order to tackle the aforementioned limitations, in this chapter, we introduce deep neural networks, which are trained in an end-to-end trainable manner and capture both the inter- and intra-modal interactions. Similar to the previous chapters, we convert the audio files into images consisting of three channels, namely log-Mel spectrograms, their delta, and delta-delta. Next, each transcript and image are passed through BERT [26] and DeiT [358] models respectively. In order to ensure that the sequence length of the vectors obtained by BERT and DeiT is the same, we exploit an Optimal Transport Kernel (OTK) Embedding. We pass the textual representation through an enhanced self-attention layer with contextual information. We exploit three main methods for the contextualization, including the global context, deep context, and deep-global context [359, 360]. Next, we pass the image representation through a self-attention mechanism with a novel gating model proposed by [83] to model the intra-modal interactions. Motivated by the study of [273], we use optimal transport based domain adaptation [361] methods for capturing the inter-modal interactions. Then, we propose two attention-based methods for fusing the self and cross-attention features. Finally, for preventing models becoming too overconfident, we use label smoothing. We use metrics for assessing both the performance and the calibration of our model. We verify the effectiveness of our approaches by conducting experiments on two publicly available datasets, namely ADReSS and ADReSSo Challenge datasets, and using both manual and automatically generated transcripts. We show that our introduced approaches obtain multiple advantages over the state-of-the-art approaches.

The contributions of this chapter can be summarized as follows:

- To the best of our knowledge, this is the first study utilizing DeiT, optimal transport kernel, and optimal transport domain adaptation methods in the task of dementia detection from spontaneous speech.
- This is the first study in the task of dementia detection from spontaneous speech exploiting label smoothing for preventing the models become too overconfident. We also evaluate our proposed models in terms of both the performance and the calibration.
- This is the first study in the task of dementia detection from speech data exploiting context-aware self-attention mechanisms and comparing two different approaches for fusing the self- and cross-attention features.
- We conduct a series of ablation experiments to demonstrate the effectiveness of the introduced approach. We evaluate our approaches on the ADReSS and ADReSSo Challenge datasets and show that they achieve competitive results to the existing research initiatives.

8.2 Data & Task

8.2.1 ADReSS Challenge Dataset

We use the ADReSS Challenge Dataset described in Section 3.3.5.2 for conducting our experiments.

8.2.2 ADReSSo Challenge Dataset

To further verify the effectiveness of our proposed approaches, we use the ADReSSo Challenge Dataset described in Section 3.3.5.3 for conducting our experiments. This dataset includes only audio files. No transcripts are provided. Therefore, one should convert the speech into text automatically via Automatic Speech Recognition (ASR) methods. Specifically, we use whisper[1] [362] and get the automatically generated transcripts per audio file.

8.2.3 Task

Let a labeled dataset consist of audio files and their corresponding transcripts. Each transcript along with its audio file belongs to an AD patient or non-AD patient. The task is to identify if a specific transcript along with its audio file corresponds to an AD patient or to a person belonging to the healthy control group (binary classification problem).

8.3 Predictive Models

8.3.1 Architecture

In this section, we describe our proposed deep learning architectures for detecting AD patients. The proposed architectures are illustrated in Fig. 8.4. Due to the fact that the manual transcripts have been annotated using the CHAT coding system [348], we use the PyLangAcq library [275] for having access to these transcripts. In addition, we use the Python library, called librosa [363, 312], and convert each audio file into a log-Mel spectrogram, its delta, and delta-delta. In this way, we create an image consisting of three channels. For all the experiments conducted, we use 224 Mel bands, hop length equal to 1024, and a Hanning window. Each image is resized to (224 × 224) pixels.

Firstly, we pass each transcript through a BERT [26] model and the corresponding image through a DeiT [358] model. Formally, let $X \in \mathbb{R}^{n \times D}$ and $Y \in \mathbb{R}^{T \times D}$ be the outputs of the BERT and DeiT pretrained models respectively. Next, we pass Y through an Optimal Transport Kernel introduced by [364], in order to ensure that the sequence length of Y is equal to the sequence length of X, i.e., $T = n$. Let $S \in \mathbb{R}^{T \times D}$, where $T = n$, denote the output representation of the Optimal Transport Kernel.

Context-Aware Self Attention for the textual modality: Fig. 8.2a illustrates the conventional self-attention mechanism, which individually calculates the attention weight of two items, i.e., ”the” and ”tomorrow”, ignoring the contextual information. In this study, we aim to enhance the self-attention layer by adding contextual information. Therefore, we exploit the context-based self-attention layer [359], which is illustrated in Fig. 8.1. We observe that this layer receives as input the input sequence denoted by X and the contextual information vector denoted by C.

[1] https://github.com/openai/whisper

We transform the input sequence X into a query, key, and value matrix, as described via the Equation 8.1:

$$Q = XW_q, K = XW_k, V = X \tag{8.1}$$

, where $W_q \in \mathbb{R}^{D \times D_q}, W_k \in \mathbb{R}^{D \times D_k}$ are learnable weight matrices.

As described in Equations 8.2 and 8.3, the context vector $C \in \mathbb{R}^{n \times D_c}$ is transformed to a contextual query matrix $Q_c \in \mathbb{R}^{n \times D_q}$ and a contextual key matrix $K_c \in \mathbb{R}^{n \times D_k}$:

$$Q_c = CW_q^c \tag{8.2}$$

, where $W_q^c \in \mathbb{R}^{D_c \times D_q}$ is a learnable weight matrix.

$$K_c = CW_k^c \tag{8.3}$$

, where $W_k^c \in \mathbb{R}^{D_c \times D_k}$ is a learnable weight matrix.

Next, we exploit gated sum, as illustrated in Fig. 8.1b, for quantifying the contribution of the input sequence X and the contextual vector C to the attention weight prediction. Finally, we get new query and key matrices denoted by $\overline{\mathrm{Q}} \in \mathbb{R}^{n \times D_q}$ and $\overline{\mathrm{K}} \in \mathbb{R}^{n \times D_k}$ respectively. We describe the equations governing the gated sum below:

$$g_q = \sigma\left(QW_g^Q + Q_cW_g^{Q_c}\right) \tag{8.4}$$

, where $W_g^Q, W_g^{Q_c} \in \mathbb{R}^{D_q \times 1}$ are learnable parameters.

$$g_k = \sigma\left(KW_g^K + K_cW_g^{K_c}\right) \tag{8.5}$$

where $W_g^K, W_g^{K_c} \in \mathbb{R}^{D_k \times 1}$ are learnable parameters.

q_q and g_k indicate the weight of the importance of the contextual information.

$$\overline{\mathrm{Q}} = (1 - g_q)Q + g_qQ_c \tag{8.6}$$

$$\overline{\mathrm{K}} = (1 - g_k)K + g_kK_c \tag{8.7}$$

Therefore, we obtain new query and key matrices. Finally, we calculate the self-attention via the equation mentioned below:

$$Attention(\overline{\mathrm{Q}}, \overline{\mathrm{K}}, V) = softmax\left(\frac{\overline{\mathrm{Q}} \cdot \overline{\mathrm{K}}^T}{\sqrt{D_k}}\right)V \tag{8.8}$$

Next, we describe three methods, namely Global Context, Deep Context, and Deep-Global Context, for calculating the contextual vector C. Specifically, we follow [359, 360] to represent the context vector (C), which is composed of internal representation.

- **Global Context:** Fig. 8.2b illustrates the global context strategy. More specifically, the global context indicates the mean operation over the input sequence for summarizing the input representation. Let $X = [x_1, x_2, ..., x_n] \in \mathbb{R}^{n \times D}$. We calculate the context representation C as defined in Eq. 8.9. Note that the output of Eq. 8.9 is a vector, i.e., $C \in \mathbb{R}^D$, instead of a matrix. To facilitate subsequent calculation operations, we use Eq. 8.10, where we obtain the contextual matrix $\mathbf{C} \in \mathbb{R}^{n \times D}$.

$$C = \overline{\mathrm{X}} = Avgpool(X) = \frac{1}{n}\sum_{1}^{n} x_i \tag{8.9}$$

$$\mathbf{C} = stack\,(C, C, ..., C) \tag{8.10}$$

- **Deep Context:** By deeply stacking self-attention layers, the model captures only high-level syntactic and semantic information neglecting the lower-level information. Therefore, as shown in Fig. 8.2c, the deep context strategy enables the layer to fuse different types of syntactic and semantic information captured by different layers.

  Formally, taking $X = [x_1, x_2, ..., x_n] \in \mathbb{R}^{n \times D}$ as the initial input sequence X^0, and the output of the L^{th} layer is $X^l = [x_1^l, x_2^l, ..., x_n^l] \in \mathbb{R}^{n \times D}$, the deep context matrix $C \in \mathbb{R}^{n \times D}$ can be represented as follows:

$$C = \hat{X} W_c^0 \tag{8.11}$$

$$\hat{X} = concat(X^0, X^1, ..., X^{l-1}) \in \mathbb{R}^{n \times lD} \tag{8.12}$$

  , where $W_c^0 \in \mathbb{R}^{lD \times D}$ is a learnable parameter matrix. *concat(.)* denotes join operation.

- **Deep-Global Context:** The deep-global context strategy combines the strategies of global context and deep context as described before. The deep-global context strategy is illustrated in Fig. 8.2d and is described via the equations below:

$$C = \overline{\mathrm{C}} W_{\overline{\mathrm{C}}}^0 \tag{8.13}$$

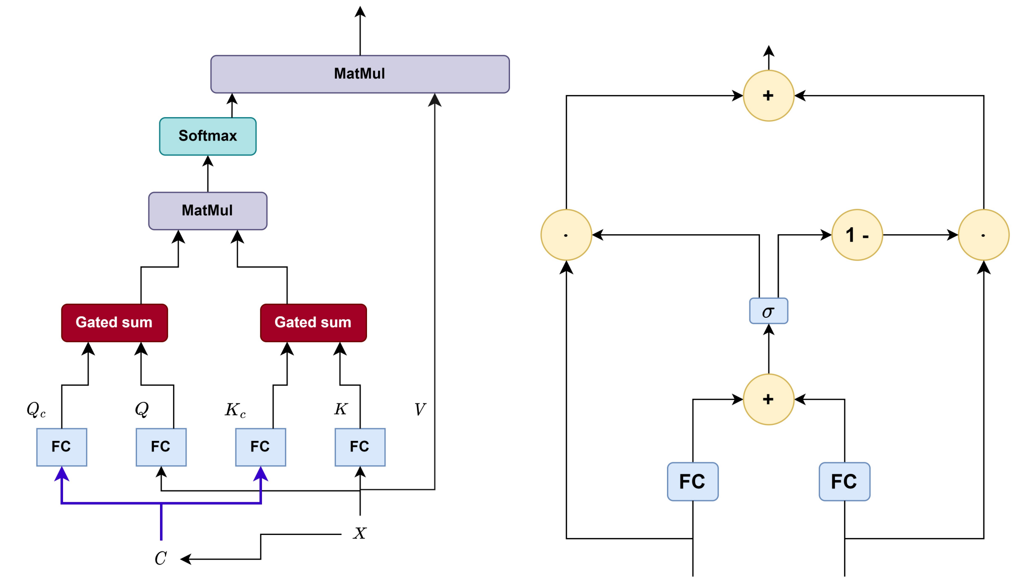
$$\overline{\mathrm{C}} = concat(C^0, C^1, ..., C^{l-1}) \tag{8.14}$$

  , where $C^j = Avgpool(X^j), C^j \in \mathbb{R}^D$. Therefore, $\overline{\mathrm{C}} \in \mathbb{R}^{lD}$. In addition, $W_{\overline{\mathrm{C}}}^0 \in \mathbb{R}^{lD \times D}$. Thus, we obtain C of Eq. 8.13, as $C \in \mathbb{R}^D$.

  As mentioned before, to facilitate subsequent calculation operations, matrix $\mathbf{C} \in \mathbb{R}^{n \times D}$ is obtained through the stack operation, as follows: $\mathbf{C} = stack(C, C, ..., C)$.

Let F be the output of the context-based self-attention mechanism corresponding to the textual modality denoted by X.

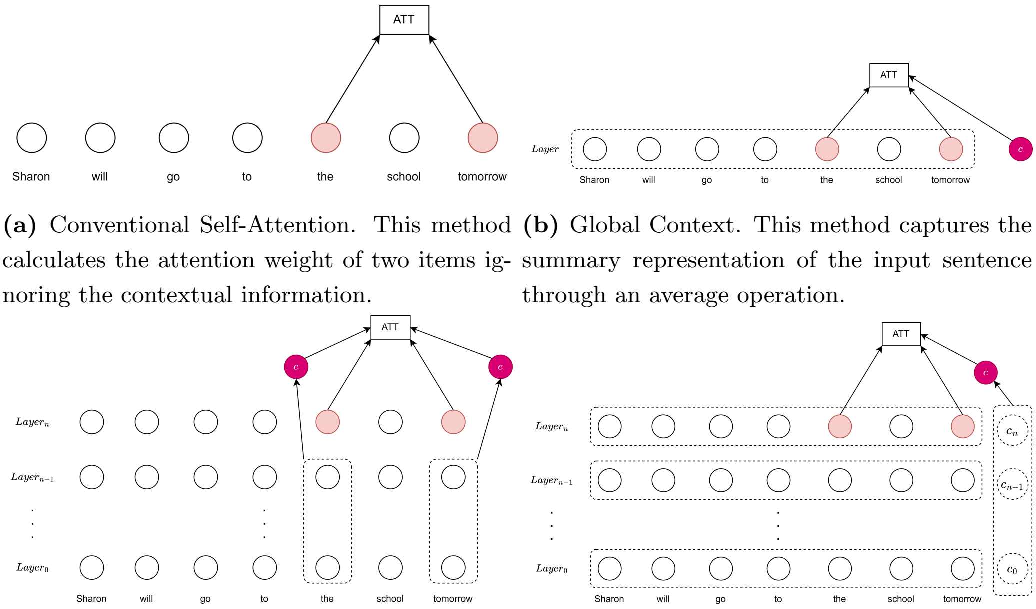
Gated Self-Attention for the image modality: Motivated by the work of [83], we pass S through a self-attention mechanism, which incorporates a novel gating model, for



(a) Context-based Self-Attention. This method is different from the conventional self-attention mechanism, since it exploits a contextual information vector C.

(b) Gated sum. This unit is used for quantifying the contribution of the original representation X and the context vector C to the attention weight prediction.

Figure 8.1: Context-based Self-Attention



(a) Conventional Self-Attention. This method calculates the attention weight of two items ignoring the contextual information.

(b) Global Context. This method captures the summary representation of the input sentence through an average operation.

(c) Deep Context. This method captures both the low- and high-level syntactic and semantic information.

(d) Deep-Global Context. This method combines the concepts of global and deep context.

Figure 8.2: Self-Attention based on different context-vectors

capturing the intra-modal interactions. This gated self-attention mechanism is illustrated in Fig. 8.3. The self-attention mechanism including the gating model is described via the equations below:

$$Q = S, K = S, V = S \tag{8.15}$$

$$M = \sigma\left(FC^{g}\left(FC_{q}^{g}(Q) \odot FC_{k}^{g}(K)\right)\right) \tag{8.16}$$

where $FC_q^g, FC_k^g \in \mathbb{R}^{D\times d_g}$, $FC^g \in \mathbb{R}^{d_g\times 2}$ are three fully-connected layers, and d_g denotes the dimensionality of the projected space and is equal to 64 units. $\odot$ denotes the element-wise product function and σ the sigmoid function. In addition, $M \in \mathbb{R}^{T\times 2}$ corresponds to the two masks $M_q \in \mathbb{R}^{T}$ and $M_k \in \mathbb{R}^{T}$ for the features Q and K respectively.

Next, the two masks M and K are tiled to $\tilde{M}_q, \tilde{M}_k \in \mathbb{R}^{T\times D}$ and then used for computing the attention map as following:

$$A^{g} = softmax\left(\frac{\left(Q \odot \tilde{M}_q\right)\left(K \odot \tilde{M}_k\right)^{T}}{\sqrt{D}}\right) \tag{8.17}$$

$$H = A^{g}V \tag{8.18}$$

Let H be the output of the self-attention mechanism corresponding to the visual modality denoted by S.

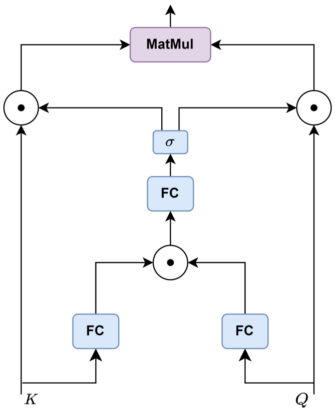


Figure 8.3: Gated Dot-product. This gating model is incorporated in the conventional self-attention mechanism for improving the quality of the learned attention. This method is based on low-rank bilinear pooling.

Optimal Transport: Next, we use optimal transport-based domain adaptation methods [361, 365, 366], i.e., Earth Mover's Distance (EMD) Transport, for transporting between each pair of modalities, which can be interpreted as domain adaptation across two modalities. Formally:

$$X' = OT(S \rightarrow X) \tag{8.19}$$

$$S' = OT(X \rightarrow S) \tag{8.20}$$

Concatenation: After that, we concatenate transported and self-attended features as follows:

$$C = [F, X'] \tag{8.21}$$

$$S = [H, S'] \tag{8.22}$$

Fusion: Next, we describe two methods for fusing C and S:

- **(i) Co-Attention Mechanism:** We exploit the fusion method proposed by [79] and implemented in Chapter 7. Specifically, given $\left(\mathbf{C} \in \mathbb{R}^{d' \times n}\right)$ and $\left(\mathbf{S} \in \mathbb{R}^{d' \times T}\right)$, where $d' = 2 \cdot D$, the affinity matrix $F \in \mathbb{R}^{n \times T}$ is calculated using the equation presented below:

  $$F = \tanh\left(C^T W_l S\right) \tag{8.23}$$

  where $W_l \in \mathbb{R}^{d' \times d'}$ is a matrix of learnable parameters. By treating the affinity matrix as a feature, we learn to predict the attention maps via the following,

  $$H^s = \tanh\left(W_s S + \left(W_c C\right) F\right) \tag{8.24}$$

  $$H^c = \tanh\left(W_c C + \left(W_s S\right) F^T\right) \tag{8.25}$$

  where $W_s, W_c \in \mathbb{R}^{k \times d'}$ are matrices of learnable parameters. We set k equal to 40. Then, we generate the attention weights through the softmax function as follows,

  $$a^s = softmax\left(w_{hs}^T H^s\right) \tag{8.26}$$

  $$a^c = softmax\left(w_{hc}^T H^c\right) \tag{8.27}$$

  where $a_s \in \mathbb{R}^{1 \times T}$ and $a_c \in \mathbb{R}^{1 \times n}$. $w_{hs}, w_{hc} \in \mathbb{R}^{k \times 1}$ are the weight parameters. Based on the above attention weights, the attention vectors are obtained via the following equations:

  $$\hat{s} = \sum_{i=1}^{T} a_i^s s^i, \quad \hat{c} = \sum_{j=1}^{n} a_j^c c^j \tag{8.28}$$

  where $\hat{s} \in \mathbb{R}^{1 \times d'}$ and $\hat{c} \in \mathbb{R}^{1 \times d'}$.

  Finally, these vectors are concatenated $p = [\hat{c}, \hat{s}]$. We apply a dropout layer with a rate of 0.5. Then, this vector is passed through a Dense Layer consisting of 128 units with a ReLU activation function. We apply also a dropout layer with a rate of 0.2. Finally, we use a dense layer consisting of two units, which gives the final output.

  The proposed architecture is illustrated in Fig. 8.4a.

- **(ii) Attention-based fusion:** Motivated by the work of [270], we design an attentional reduction model for C, as defined in Equation 8.21 (or S, as defined in Equation 8.22), for obtaining its attended feature $\tilde{c}$ (or $\tilde{s}$). To the best of our knowledge, this is the first study utilizing this fusion method in the task of dementia detection from spontaneous speech. Taking C as an example, we describe the attention reduction model used in this study via the equations presented below:

$$\alpha^c = softmax\left(MLP\left(C\right)\right) \tag{8.29}$$

, where α^c refers to the learned attention weights and MLP is given by the equation below:

$$MLP = FC(128) - ReLU - Dropout(0.1) - FC(1) \tag{8.30}$$

$$\tilde{c} = \sum_{i=1}^{n} \alpha_i^c c_i \tag{8.31}$$

, where we obtain the attended feature $\tilde{c}$ for C.

We obtain the attended feature $\tilde{s}$ using an independent attention reduction model in the same way. Having computed $\tilde{c}$ and $\tilde{s}$, we design the linear multimodal fusion function as follows:

$$z = LayerNorm\left(W_c^T \tilde{c} + W_s^T \tilde{s}\right) \tag{8.32}$$

, where $W_c, W_s \in \mathbb{R}^{d' \times d_z}$ are two linear projection matrices, d_z is the common dimensionality of the fused feature and is equal to 128, and LayerNorm [28] is used for stabilizing the training. Finally, we pass z to a dense layer consisting of two units, which gives the final prediction.

The proposed architecture is illustrated in Fig. 8.4b.

8.3.2 Model Calibration

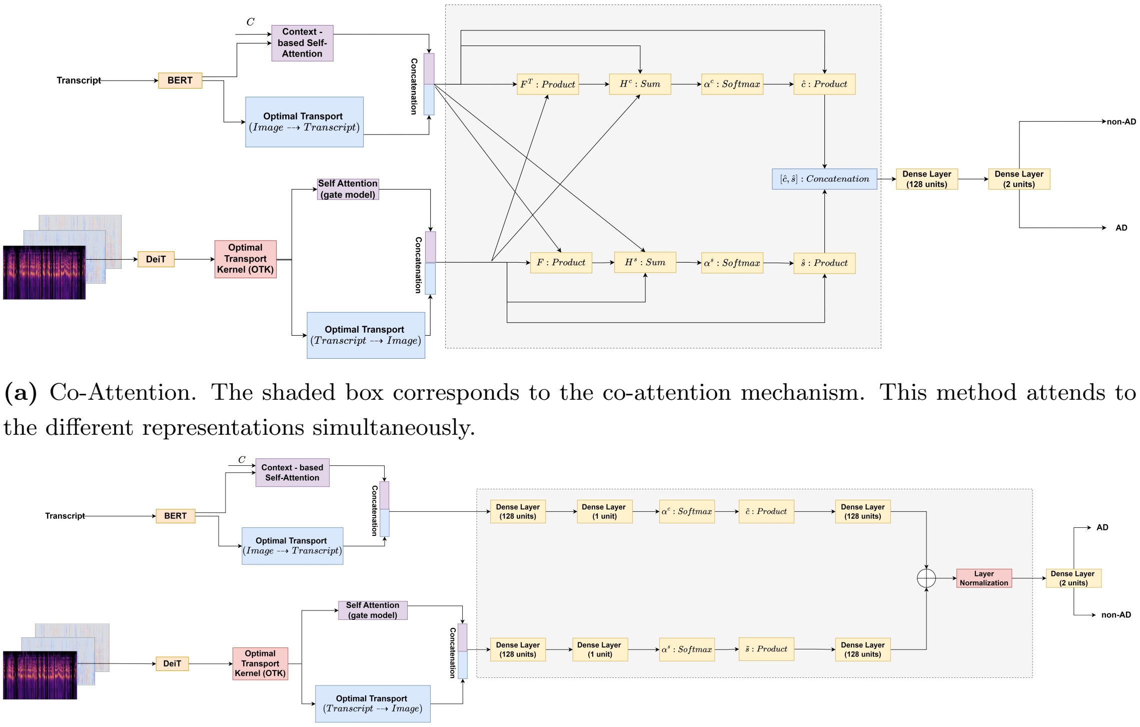
To prevent the model becoming too overconfident, we use label smoothing [30, 31], as described in Chapter 4. Specifically, label smoothing calibrates learned models so that the confidences of their predictions are more aligned with the accuracies of their predictions.

For a network trained with hard targets, the cross-entropy loss is minimized between the true targets y_k and the network's outputs p_k, as in $H(y,p) = \sum_{k=1}^{K} -y_k log(p_k)$, where y_k is "1" for the correct class and "0" for the other. For a network trained with label smoothing, we modify the true targets y_k to $y_k^{LS_u}$ as shown in Eq. 8.1:

$$y_k^{LS_u} = y_k \cdot (1 - \alpha) + \frac{\alpha}{K} \tag{8.1}$$

, where α is the smoothing parameter and K is the number of classes.

Finally, we minimize the cross-entropy between the modified targets $y_k^{LS_u}$ and the network's outputs p_k, as shown in Eq. 8.2:



(a) Co-Attention. The shaded box corresponds to the co-attention mechanism. This method attends to the different representations simultaneously.

(b) Attention-based Fusion. The shaded box shows this fusion method. This method exploits two independent attentional reduction models. Features are fused through an add operation, while a layer normalization is used for stabilizing training.

Figure 8.4: Illustration of our Proposed Architectures. For the textual modality, we use BERT, while for the image modality, we use DeiT and exploit an Optimal Transport Kernel. Next, we use optimal transport domain adaptation methods for transporting between each pair of modalities. Also, we pass the textual representation through context-based self-attention layers, while the image representation is passed through a gated self-attention layer. Finally, methods for fusing the self- and cross-attention features are presented, namely Co-Attention and Attention-based Fusion. Each shaded box shows the fusion method used, namely Co-Attention and Attention-based Fusion.

$$H(y,p) = \sum_{k=1}^{K} -y_k^{LS_u} \cdot \log\left(p_k\right) \tag{8.2}$$

8.4 Experiments

8.4.1 Baselines

Table 8.1: Baselines (ADReSS Challenge Dataset).

Reference/Architecture	Features/Methodology
Baselines - Unimodal state-of-the-art approaches (only transcripts)	
BERT (Chapter 5)	Fine-tune a BERT model

Baselines - Multimodal state-of-the-art approaches	
Fusion Maj. (3-best) [351]	Majority Vote of the BoAW-MFCC-C125, ZFF, and bi-LSTM-Att
System 3: Phonemes and Audio [250]	Acoustic features (emobase, eGeMAPS, ComParE2016) along with feature selection techniques, transcription of the segmented text into phoneme written pronunciation using CMUDict
Fusion of System [67]	merged the x-vectors features set with the combination of linguistic feature sets (GMax/LSTM-RNNs/LSTM-RNNs-Pos) and trained a SVM classifier
Bimodal Network (Ensembled Output) [263]	For the acoustic modality, the authors use VGGish, while for the textual modality, the authors exploit GloVe, Transformer-XL, POS and HC features. Finally, the authors combine the results of the models via an ensemble approach.
GFI, NUW, Duration, Character 4-grams, Suffixes, POS tag, UD [95]	feature extraction, early fusion approach, train a Logistic Regression Classifier
Acoustic & Transcript [90]	For transcripts, the authors exploited BERT, while for speech, the authors used x-vector PCA-transformed coefficients.
Dual BERT [91]	concatenation of the representations obtained by BERT and Speech BERT
Model C [259]	The authors extracted emobase, eGeMAPS, ComParE features. For the text modality, the used GloVe embeddings and pos-tags. Finally, they trained a Neural network consisting of CNN, BiLSTM, Attention, GRU, and Dense layers.
Majority vote (NLP + Acoustic) [64]	The authors extracted a set of acoustic and linguistic features. After training shallow machine learning classifiers, they chose the three best-performing acoustic models along with the best-performing language model, and computed a final prediction by taking a linear weighted combination of the individual model predictions.
Audio + Text [92]	majority level approach of six models
LSTM with Gating (Acoustic + Lexical + Dis) [62]	Acoustic, Linguistic Features, Bi-LSTM, feed-forward highway layers with gating units

Ensemble [170]	A majority vote was taken between the predictions of the three individual models. Specifically, the authors extracted three sets of features, namely disfluency, acoustic, and interventions, and trained three deep neural networks.
BERT+ViT (log-Mel spectrogram) (Chapter 6)	coversion of an audio file into an image of three channels, BERT for the text representation, Vision Transformer for the image representation, concatenation
BERT+ViT+Gated Multimodal Unit (log-Mel spectrogram) (Chapter 6)	Gated Multimodal Unit to control the information flow of the different modalities.
BERT+ViT+Crossmodal Attention (log-Mel spectrogram) (Chapter 6)	Similar to [76], the authors exploited a cross-attention mechanism.
BERT+ViT+Co-Attention (Chapter 7)	The authors used a co-attention mechanism to fuse the representation matrices of the two modalities.
Multimodal BERT - eGeMAPS (Chapter 7)	The authors injected acoustic information (eGeMAPS) into a BERT model.
Multimodal BERT - ViT (Chapter 7)	The authors injected image information (via ViT) into a BERT model.
Multimodal BERT - eGeMAPS+ViT (Chapter 7)	The authors injected both acoustic information (eGeMAPS) and image information (via ViT) into a BERT model.
BERT+ViT+Gated Self-Attention (Chapter 7)	The authors concatenated the outputs of BERT and ViT and passed the resulting matrix through a self-attention layer incorporating a gate model for capturing the inter- and intra-modal interactions.
Transcript+Image+Acoustic [367]	The authors used a Tensor Fusion Layer for fusing the different modalities.
Introduced Approaches without Label Smoothing	
	Our proposed approaches described in Section 8.3 without label smoothing.

Table 8.2: Baselines (ADReSSo Challenge Dataset).

Reference/Architecture	Features/Methodology
Baselines - Unimodal state-of-the-art approaches (only transcripts)	
BERT	We exploit a BERT model, get the [CLS] token, and pass it through two dense layers consisting of 128 and 2 units respectively.

Model C: LR[Comp] + LR[DisFl] + Ernie + Bert (stacking) [368]	The authors employed model stacking to combine two logistic regression models (LR) using complexity and (dis)fluency features respectively, and the two pretrained language models, i.e. BERT and ERNIE.
Model 5: [262]	The authors concatenate the last three states of the BERT sequence classifier with the confidence score input. The confidence score input is generated by the ASR system.
Label Fusion selected models [369]	The authors extracted a set of handcrafted features, namely syntactic, readability, and lexical diversity, and a set of deep textual embeddings, including BERT and so on. Finally, the authors trained Logistic Regression and SVM classifiers.
Mp1 [370]	The authors add sentence-level pauses to ASR transcripts and exploit a BERT model.
Baselines - Multimodal state-of-the-art approaches	
LSTM w/ Gating (Words + Acoustic + Disf + Pse + WP) [264]	extraction of acoustic and language features, feed-forward highway layers with gating units
Global Fusion [256]	fusion of BERT (ASR) and acoustic models, namely x-vectors, x-vectors with 250ms frame-length, and encoder-decoder ASR embeddings (SB Enc/Dec).
Top-10 Avg. [169]	Average fusion of predicted class probabilities of the 10 best performing models
Attempt 1: [371]	The authors used acoustic features, linguistic features, and embedding features. For each type of feature, they exploited a deep neural network consisting of multihead attention layers, convolutional layers, and dilated convolutional layers. They used an attention layer for fusing the outputs of the different branches.
Introduced Approaches without Label Smoothing	
	Our proposed approaches described in Section 8.3 without label smoothing.

We compare our introduced approaches with the following research works reported in Tables 8.1 and 8.2, since these research works have conducted their experiments on the ADReSS and ADReSSo test set. Specifically, Table 8.1 describes the baselines used in terms of the ADReSS Challenge dataset, while Table 8.2 reports the baselines used regarding the ADReSSo Challenge dataset. Regarding Table 8.1, we are using existing published results for all the baselines except for *Introduced Approaches without Label Smoothing.* In terms of Table 8.2, we are using existing published results for all the baselines except for: (i) *BERT*, and (ii) *Introduced Approaches without Label Smoothing.*

8.4.2 Experimental Setup

We divide the ADReSS Challenge train set into a train and a validation set (65%-35%). We use a batch size of 4. We train the introduced architectures five times and report the results on the ADReSS Challenge test set via mean ± standard deviation. Similarly, we divide the ADReSSo Challenge train set into a train and a validation set (65%-35%). We train the introduced architectures five times and report the results on the ADReSSo Challenge test set via mean ± standard deviation. We use *EarlyStopping*, where we stop training if the validation loss has stopped decreasing for eight consecutive epochs. Also, we apply *StepLR* with a step_size of 4 and a gamma of 0.1. We set α of Eq. 8.1 equal to 0.001. We set $D = D_c = 768$. We set $D_k = D_q = 64$. Regarding the global context strategy, we use one layer of the contextual self-attention mechanism. In terms of the deep-context strategy, we use three layers of the contextual self-attention mechanism. With regards to the deep-global context strategy, we use two layers of the contextual self-attention mechanism. We use the BERT base uncased version and the DeiT[2] model from the Transformers library [305]. For the optimal transport methods, we use the Python library Optimal Transport [372]. All the models have been created using the PyTorch library [346]. All experiments are conducted on a single Tesla P100-PCIE-16GB GPU.

8.4.3 Evaluation Metrics

8.4.3.1 Performance Metrics

Accuracy (Acc.), Precision (Prec.), Recall (Rec.), F1-Score, and Specificity (Spec.) have been used for evaluating the results of the introduced architectures. These metrics have been computed by regarding the dementia class as the positive one. We report the average and standard deviation of these metrics over five runs.

8.4.3.2 Calibration Metrics

We evaluate the calibration of our model using the metrics proposed by [308, 309, 171]. Specifically, we use the metrics mentioned below:

- **Expected Calibration Error (ECE).** The calibration error is the difference between the fraction of predictions in the bin that are correct (accuracy) and the mean of the probabilities in the bin (confidence). First, we divide the predictions into M equally spaced bins (size $1/M$).

$$acc(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} 1(\hat{y}_i = y_i) \tag{8.1}$$

[2] facebook/deit-base-distilled-patch16-224

$$conf(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i \tag{8.2}$$

, where y_i and $\hat{y}_i$ are the true and predicted labels for the sample i and $\hat{p}_i$ is the confidence (predicted probability value) for sample i.

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} \left| acc\left(B_m\right) - conf\left(B_m\right) \right| \tag{8.3}$$

, where N is the total number of data points and B_m is the group of samples whose predicted probability values falls into the interval $I_m = \left(\frac{m-1}{M}, \frac{m}{M}\right]$.

Perfectly calibrated models have an ECE of 0.

- **Adaptive Calibration Error (ACE).** Adaptive Calibration Error uses an adaptive scheme which spaces the bin intervals so that each contains an equal number of predictions.

$$ACE = \frac{1}{KR} \sum_{k=1}^{K} \sum_{r=1}^{R} |acc(r,k) - conf(r,k)| \tag{8.4}$$

, where $acc(r,k)$ and $conf(r,k)$ are the accuracy and confidence of adaptive calibration range r for class label k, respectively; and N is the total number of data points. Calibration range r defined by the $[N/R]$th index of the sorted and thresholded predictions.

8.5 Results

The results of our introduced models are reported in Tables 8.3 and 8.4. Specifically, Table 8.3 reports the results on the ADReSS Challenge dataset, while Table 8.4 reports the results on the ADReSSo Challenge dataset. Also, these tables present a comparison of our introduced approaches with existing research initiatives, which have proposed either unimodal or multimodal approaches. In order to compare models, we use the Almost Stochastic Order (ASO) test [47, 48] of statistical significance implemented by [49]. We use $confidence\_level = 0.95$ and $num\_comparisons = 50$. Generally, the ASO test determines whether a stochastic order [50] exists between two models or algorithms, i.e., A and B. This method computes a score (ϵ_{min}) which represents how far the first is from being significantly better in respect to the second. When $\epsilon_{min} = 0$, then one can claim that A is truly stochastically dominant over B. When $\epsilon_{min} < 0.5$, one can claim that A is almost stochastically dominant over B. For $\epsilon_{min} = 0.5$, no order can be determined. (†) means that Attention-based Fusion (Deep Context) with label smoothing is stochastically dominant over the respective models. Similarly, in terms of the ADReSSo Challenge dataset, (†) means that Co-Attention (Deep Context) with label smoothing is stochastically dominant over the respective models. ($\star$) denotes almost stochastic dominance of the Attention-based Fusion (Deep Context) with label smoothing over the respective approaches. Similarly, in terms of the ADReSSo Challenge dataset, ($\star$) means that Co-Attention (Deep Context) with label smoothing is almost stochastically dominant over the respective models. Note that we cannot compare our approaches with all the existing research initiatives, since we do not have access to the multiple runs or the other approaches have not used multiple runs. In terms of the ECE and ACE calibration metrics, we use ASO for comparing our best performing model, namely Attention-based Fusion (Deep Context) or Co-Attention (Deep Context) with label smoothing, with the respective model without label smoothing.

Table 8.3: Performance comparison among proposed models and state-of-the-art approaches on the ADReSS Challenge test set. Reported values are mean ± standard deviation. Results are averaged across five runs. (†) means that Attention-based Fusion (Deep Context) with label smoothing is stochastically dominant over the respective models. (⋆) denotes almost stochastic dominance of the Attention-based Fusion (Deep Context) with label smoothing over the respective approaches.

Architecture	**P. (%)**	**R. (%)**	**F1 (%)**	**Acc. (%)**	**Spec. (%)**	**ECE**	**ACE**
Baselines - Unimodal state-of-the-art approaches (only transcripts)							
BERT (Chapter 5)	87.19 ±3.25	81.66 ±5.00	86.73† ±4.53	87.50† ±4.37	93.33 ±5.65		
Baselines - Multimodal state-of-the-art approaches							
Fusion Maj. (3-best) [351]	-	-	85.40	85.20	-		
System 3: Phonemes and Audio [250]	81.82	75.00	78.26	79.17	83.33		
Fusion of system [67]	94.12	66.67	78.05	81.25	95.83		
Bimodal Network (Ensembled Output) [263]	89.47	70.83	79.07	81.25	91.67		
GFI, NUW, Duration, Character 4-grams, Suffixes, POS tag, UD [95]	-	-	-	77.08	-		
Acoustic & Transcript [90]	70.00	88.00	78.00	75.00	63.00		
Dual BERT [91]	83.04 ±3.97	83.33 ±5.89	82.92 ±1.86	82.92 ±1.56	82.50 ±5.53		
Model C [259]	78.94	62.50	69.76	72.92	83.33		
Majority vote (NLP+Acoustic) [64]	-	-	-	83.00	-		
Audio + Text [92]	-	87.50	-	89.58	91.67		
LSTM with Gating (Acoustic + Lexical + Dis) [62]	81.82	75.00	78.26	79.17	83.33		
Ensemble [170]	83.00	83.00	83.00	83.00	-		
BERT+ViT (Chapter 6) (log-Mel spectrogram)	90.73 ±2.74	80.83 ±2.04	85.47† ±1.70	86.25† ±1.67	91.67 ±2.64		
BERT+ViT+Gated Multimodal Unit (Chapter 6) (log-Mel spectrogram)	80.92 ±2.30	91.67 ±3.73	85.92† ±2.37	85.00† ±2.43	78.33 ±3.12		
BERT+ViT+Crossmodal Attention (Chapter 6) (log-Mel spectrogram)	86.13 ±3.26	91.67 ±4.56	88.69⋆ ±2.12	88.33⋆ ±2.12	85.00 ±4.25		
BERT+ViT+Co-Attention (Chapter 7)	92.83 ±6.39	81.67 ±2.04	86.81⋆ ±3.37	87.50⋆ ±3.49	93.33 ±6.24		
Multimodal BERT - eGeMAPS (Chapter 7)	74.51 ±1.01	87.50 ±6.45	80.35† ±2.77	78.75† ±2.04	70.00 ±3.12		
Multimodal BERT - ViT (Chapter 7)	73.91 ±2.40	91.67 ±2.64	81.79† ±1.72	79.58† ±2.04	67.50 ±4.08		
Multimodal BERT - eGeMAPS+ViT (Chapter 7)	76.57 ±3.74	89.17 ±5.65	82.28† ±3.49	80.83† ±3.58	72.50 ±5.65		
BERT+ViT+Gated Self-Attention (Chapter 7)	90.87 ±3.50	89.17 ±2.04	89.94⋆ ±1.36	90.00⋆ ±1.56	90.83 ±4.08		
Transcript+Image+Acoustic [367]	90.88 ±3.60	80.83 ±2.04	85.48† ±0.76	86.25† ±1.02	91.66 ±3.73		
Baselines - Introduced models (without label smoothing)							
Co-Attention (Global Context)	89.62 ±1.75	85.83 ±3.33	87.63† ±1.80	87.92† ±1.56	90.00 ±2.04	0.1208 ±0.2296	0.1660 ±0.0335
Co-Attention (Deep Context)	88.25 ±1.56	87.50 ±2.64	87.85⋆ ±1.66	87.92† ±1.56	88.33 ±1.66	0.1384 ±0.0109	0.1532 ±0.0110
Co-Attention (Deep-Global Context)	90.26 ±1.70	85.00 ±4.25	87.51⋆ ±2.69	87.92⋆ ±2.43	90.83 ±1.66	0.1355 ±0.0183	0.1648 ±0.0119
Attention-based Fusion (Global Context)	89.55 ±7.31	85.83 ±6.24	87.32⋆ ±4.35	87.50⋆ ±4.37	89.16 ±8.58	0.1256 ±0.0291	0.1279 ±0.0277
Attention-based Fusion (Deep Context)	91.06 ±5.04	89.16 ±3.33	89.95⋆ ±1.91	90.00⋆ ±2.04	90.83 ±5.53	0.0975⋆ ±0.0188	0.1046⋆ ±0.0173
Attention-based Fusion (Deep-Global Context)	90.45 ±2.93	85.83 ±2.04	88.04⋆ ±1.65	88.33⋆ ±1.66	90.83 ±3.12	0.1173 ±0.0134	0.1065 ±0.0153
Introduced models (with label smoothing)							
Co-Attention (Global Context)	88.65 ±4.63	88.33 ±1.66	88.39⋆ ±1.76	88.33⋆ ±2.12	88.33 ±5.53	0.1075 ±0.0198	0.1710 ±0.0281
Co-Attention (Deep Context)	93.57 ±2.08	84.16 ±4.86	88.53⋆ ±2.79	89.16⋆ ±2.43	94.16 ±2.04	0.1082 ±0.0184	0.1316 ±0.0296
Co-Attention (Deep-Global Context)	87.88 ±3.73	87.50 ±6.97	87.39⋆ ±2.45	87.50† ±1.86	87.50 ±4.56	0.1176 ±0.0167	0.1568 ±0.0306
Attention-based Fusion (Global Context)	90.51 ±3.40	85.00 ±4.25	87.53† ±1.75	87.92† ±1.56	90.83 ±4.08	0.1094 ±0.0086	0.1168 ±0.0099
Attention-based Fusion (Deep Context)	93.08 ±2.03	89.17 ±2.04	91.06 ±1.60	91.25 ±1.56	93.33 ±2.04	0.0859 ±0.0130	0.0830 ±0.0158
Attention-based Fusion (Deep-Global Context)	89.87 ±5.52	83.33 ±4.56	86.20† ±0.90	86.66† ±1.02	90.00 ±5.65	0.1397 ±0.0102	0.1508 ±0.0123

Table 8.4: Performance comparison among proposed models and state-of-the-art approaches on the ADReSSo Challenge test set. Reported values are mean ± standard deviation. Results are averaged across five runs. (†) means that Co-Attention (Deep Context) with label smoothing is stochastically dominant over the respective models. (⋆) denotes almost stochastic dominance of the Co-Attention (Deep Context) with label smoothing over the respective approaches.

Architecture	**P. (%)**	**R. (%)**	**F1 (%)**	**Acc. (%)**	**Spec. (%)**	**ECE**	**ACE**
Baselines - Unimodal state-of-the-art approaches (only transcripts)							
BERT	83.35 ±0.86	74.29 ±2.55	78.53† ±1.43	80.00† ±1.05	85.55 ±1.11	- -	- -
Model C: [368]	85.00	80.00	82.00	83.00	86.00	-	-
Model 5: [262]	81.58	88.57	84.93	84.51	80.56	-	-
Label Fusion selected models [369]	-	-	-	84.51	-	-	-
Mp1 [370]	87.10	77.14	81.82	83.10	88.89	-	-
Baselines - Multimodal state-of-the-art approaches							
LSTM w/ Gating (Words +Acoustic+Disf+Pse+WP) [264]	-	-	-	84.00	-	-	-
Global Fusion [256]	92.00	74.00	83.00	84.51	94.00	-	-
Top-10 Avg. [169]	-	-	88.89	81.69	80.00	-	-
Attempt 1: [371]	75.00	91.67	82.50	80.28	68.57	-	-
Baselines - Introduced models (without label smoothing)							
Co-Attention (Global Context)	83.77 ±4.59	81.13 ±9.13	81.85⋆ ±3.01	82.54⋆ ±1.69	83.88 ±6.66	0.1536 ±0.0311	0.2017 ±0.0214
Co-Attention (Deep Context)	82.22 ±1.79	84.00 ±4.28	83.01⋆ ±1.63	83.10⋆ ±1.26	82.22 ±2.83	0.1349⋆ ±0.0135	0.1845 ±0.0169
Co-Attention (Deep-Global Context)	83.03 ±2.07	80.57 ±2.79	81.73⋆ ±1.23	82.25† ±1.13	83.88 ±2.72	0.1414 ±0.0091	0.1948 ±0.0265
Attention-based Fusion (Global Context)	83.44 ±1.16	74.86 ±2.14	78.90† ±1.51	80.28† ±1.26	85.56 ±1.11	0.1633 ±0.0207	0.1825 ±0.0140
Attention-based Fusion (Deep Context)	81.52 ±3.47	81.14 ±5.59	81.08† ±1.58	81.41† ±1.05	81.66 ±5.44	0.1442 ±0.0284	0.1737 ±0.0089
Attention-based Fusion (Deep-Global Context)	79.58 ±2.69	85.71 ±4.78	82.38⋆ ±1.59	81.97† ±1.38	78.33 ±4.78	0.1671 ±0.0201	0.1820 ±0.0193
Introduced models (with label smoothing)							
Co-Attention (Global Context)	84.77 ±2.39	81.71 ±3.43	83.12⋆ ±0.95	83.66⋆ ±0.69	85.55 ±3.24	0.1282 ±0.0053	0.1630 ±0.0179
Co-Attention (Deep Context)	84.43 ±1.59	86.29 ±4.19	85.27 ±1.78	85.35 ±1.44	84.43 ±2.19	0.1178 ±0.0209	0.1800 ±0.0213
Co-Attention (Deep-Global Context)	82.45 ±0.99	82.86 ±4.78	82.55⋆ ±2.03	82.82⋆ ±1.38	82.77 ±2.08	0.1443 ±0.0046	0.1749 ±0.0082
Attention-based Fusion (Global Context)	80.44 ±1.65	81.71 ±4.98	80.95† ±2.04	81.13† ±1.44	80.55 ±3.04	0.1540 ±0.0195	0.1920 ±0.0215
Attention-based Fusion (Deep Context)	85.10 ±0.53	81.71 ±3.43	83.35⋆ ±2.04	83.94⋆ ±1.69	86.11 ±0.04	0.1336 ±0.0190	0.1660 ±0.0144
Attention-based Fusion (Deep-Global Context)	81.45 ±1.32	85.14 ±4.92	83.18⋆ ±2.42	83.10⋆ ±1.99	81.11 ±2.08	0.1690 ±0.0245	0.1938 ±0.0112

8.5.1 ADReSS Challenge Dataset

Regarding our proposed models, one can observe that Attention-based Fusion (Deep Context) constitutes our best performing model outperforming all the other introduced models in all the evaluation metrics except Precision and Specificity. Specifically, Attention-based Fusion (Deep Context) outperforms the other introduced models with label smoothing in Accuracy by 2.09-4.59%, in Recall by 0.84-5.84%, and in F1-score by 2.53-4.86%. Despite the fact that Attention-based Fusion (Deep Context) obtains a lower Precision score by other introduced models, it surpasses them in F1-score, which constitutes the weighted average of Precision and Recall. Although it achieves lower specificity scores by Co-Attention (Deep Context), it must be noted that in health related studies, F1-score is more important than Specificity, since high F1-score means that the model can detect better the AD patients, while high Specificity and low F1-score means that AD patients are misdiagnosed as non-AD ones. In addition, Co-Attention (Deep Context) constitutes our second best performing model attaining an Accuracy of 89.16%. It achieves the highest precision and specificity scores accounting for 93.57% and 94.16% respectively, while it achieves an F1-score of 88.53%. It outperforms all the introduced models, except Attention-based Fusion (Deep Context), in Accuracy by 0.83-2.50% and in F1-score by 0.14-2.33%. It outperforms all the models in Precision and Specificity by 0.49-5.69% and 0.83-6.66% respectively.

Next, we compare our introduced approaches with label smoothing with the ones without applying label smoothing. As one can easily observe, label smoothing leads to both performance improvement and better calibration of the proposed approaches. Specifically, we observe that Attention-based Fusion (Deep Context) with label smoothing obtains a higher Accuracy score than the one obtained by the respective model without label smoothing by 1.25%, Attention-based Fusion (Global Context) with label smoothing surpasses Attention-based Fusion (Global Context) without label smoothing in Accuracy by 0.42%, etc. In terms of the calibration metrics, namely ECE and ACE, one can observe that label smoothing leads to better calibrated models. For instance, Attention-based Fusion (Deep Context) with label smoothing obtains an ECE of 0.0859 and an ACE of 0.0830, which are significantly better than the ones obtained by Attention-based Fusion (Deep Context) without label smoothing by 0.0116 and 0.0216 respectively.

In comparison with the unimodal and multimodal state-of-the-art approaches, one can observe that our best performing model, namely Attention-based Fusion (Deep Context) with label smoothing, outperforms the research works in Accuracy by 1.25-18.33% and in F1-score by 1.12-21.30%. These differences in performance are attributable to the fact that our best performing model captures both the inter- and intra-modal interactions through the self-attention mechanisms and optimal transport domain adaptation methods, enhances the self-attention mechanism with contextual information, and applies label smoothing in contrast to the research initiatives. In addition, Co-Attention (Deep Context) outperforms the research works, except [177, 92], in Accuracy by 0.83-16.24%.

8.5.2 ADReSSo Challenge Dataset

As one can easily observe in Table 8.4, Co-Attention (Deep Context) with label smoothing constitutes our best performing model attaining an Accuracy of 85.35%, a Recall of 86.29%, and a F1-score of 85.27%. It surpasses the other introduced models (with label smoothing) in Accuracy by 1.41-4.22%, in Recall by 1.15-4.58%, and in F1-score by 1.92-4.32%. In addition, we observe that Co-Attention (Deep Context) with label smoothing achieves better performance than the one obtained by the respective model without label smoothing. Specifically, the Accuracy is improved by 2.25%, the Recall is improved by 2.29%, the F1-score presents a surge of 2.26%, the Precision is increased by 2.21%, and the Specificity is improved by 2.21%. In terms of the calibration metrics, we observe that the ECE is improved by 0.0171 (ASO test indicates almost stochastic dominance).

Comparing our introduced models with label smoothing with the ones without label smoothing, we observe that in most cases label smoothing contributes to both the performance improvement and better calibration. For instance, Co-Attention (Global Context) with label smoothing improves Accuracy by 1.12% compared with the respective model without label smoothing, while ECE and ACE are also improved by 0.0254 and 0.0387 respectively. Similarly, Attention-based Fusion (Deep Context) with label smoothing outperforms the respective model without label smoothing in F1-score and Accuracy by 2.27% and 2.53% respectively, while the ECE and ACE also present a decline of 0.0106 and 0.0077 respectively.

In comparison with the unimodal and multimodal baselines, we observe that our best performing model, namely Co-Attention (Deep Context) with label smoothing, outperforms these baselines in Accuracy by 0.84-5.35%. Also, it outperforms all the research works, except for [169], in F1-score by 0.34-6.74%. We observe that our best performing model attains a better performance than BERT (ASO test indicates stochastic dominance), verifying our initial hypothesis that both modalities, i.e., transcripts and audio files, contribute to a better performance. In addition, we observe that our second best performing model, namely Attention-based Fusion (Deep Context) outperforms some research works, except for [262, 369, 264, 256], in Accuracy by 0.84-3.94%.

8.6 Ablation Study

In this section, we run a series of ablation experiments using the ADReSS Challenge dataset to explore the effectiveness of the introduced architecture described in Section 8.3. We report the results of the ablation study in Tables 8.5 and 8.6.

First, we explore the effectiveness of the context-based self-attention. To do this, we remove the contextual information and exploit the conventional self-attention mechanism introduced by [46]. We observe that the Accuracy score drops from 91.25% to 87.08%, while the F1-score presents a decline of 4.60%. Also, we observe that the removal of contextual information yields to higher standard deviations of the performance metrics.

Table 8.5: Ablation Study. Reported values are mean ± standard deviation. Results are averaged across five runs.

Architecture	**Prec. (%)**	**Rec. (%)**	**F1-score (%)**	**Acc. (%)**	**Spec. (%)**
without contextual vector in self-attention	91.34 ±7.35	83.33 ±9.50	86.46 ±4.64	87.08 ±4.04	90.83 ±10.00
self-attention without gate model	92.99 ±4.28	84.16 ±3.12	88.22 ±0.88	88.75 ±1.02	93.33 ±4.25
without optimal transport and OTK	87.60 ±2.02	87.50 ±3.73	87.47 ±1.52	87.50 ±1.32	87.50 ±2.64
repeat vector instead of OTK	86.08 ±3.37	91.66 ±2.64	88.73 ±1.97	88.33 ±2.12	85.00 ±4.25
Concatenation - Without fusion	87.23 ±4.99	88.33 ±3.12	87.65 ±2.64	87.50 ±2.95	86.66 ±6.12
Proposed Framework	**93.08** **±2.03**	**89.17** **±2.04**	**91.06** **±1.60**	**91.25** **±1.56**	**93.33** **±2.04**

Table 8.6: Ablation Study. Reported values are mean ± standard deviation. Results are averaged across five runs.

Layers	**Prec. (%)**	**Rec. (%)**	**F1-score (%)**	**Acc. (%)**	**Spec. (%)**
1	90.37 ±3.33	83.33 ±5.27	86.52 ±1.94	87.08 ±1.56	90.83 ±4.08
2	88.09 ±1.96	91.66 ±3.73	89.77 ±1.45	89.58 ±1.32	87.50 ±2.64
3 **(Our best performing model)**	**93.08** **±2.03**	**89.17** **±2.04**	**91.06** **±1.60**	**91.25** **±1.56**	**93.33** **±2.04**
4	92.05 ±3.70	76.66 ±5.65	83.55 ±4.28	85.00 ±3.58	93.33 ±3.33
5	88.67 ±4.20	83.33 ±3.73	85.84 ±2.86	86.25 ±2.83	89.16 ±4.25

Next, we investigate the efficacy of the gate model, which is incorporated into the self-attention mechanism. To do this, we remove the gate model and exploit the conventional self-attention mechanism. We observe that Accuracy and F1-score present a decline of 2.50% and 2.84% respectively.

Moreover, we explore the effectiveness of the optimal transport domain adaptation method and the Optimal Transport Kernel. To do this, we remove these components from the introduced architecture. We observe that the Accuracy score is equal to 87.50%, which is lower by 3.75% than the one obtained by our best performing model. Also, this approach yields an F1-score accounting for 87.47%, which is lower by 3.59% than the one achieved by Attention-based Fusion (Deep Context).

Next, we explore the effectiveness of the Optimal Transport Kernel. To do this, we remove this component, exploit the average operation over the sequence length, and finally repeat the vector n times, so as to ensure that both the textual and image modalities have the same sequence length. As one can observe, this method presents a decline in Accuracy score by 2.92%, while the F1-score is also reduced by 2.33%.

In addition, we explore the effectiveness of the fusion method. To prove this, we remove the fusion method, apply the average operation over C (Eq. 8.21) and S (Eq. 8.22) and concatenate these two representation vectors. We observe that the concatenation of features yields an Accuracy and F1-score of 87.50% and 87.65% respectively. This difference in performance can be justified by the fact that the concatenation operation does not capture the inherent correlations between the modalities.

Finally, we vary the layers of the context-based self-attention mechanism. The results of this ablation study are reported in Table 8.6. As the number of layers increases from 1 to 3, we observe that both the Accuracy and F1-score also increase. This justifies our initial hypothesis that stacking attention layers and fusing the outputs of different layers into one context vector, yields to better evaluation results, since the model captures both high-level and low-level syntactic and semantic information. However, we observe that the performance of our approach starts to present a decline by stacking four or five layers of context-based self-attention by applying the deep-context strategy. We assume that this decline in performance is attributable to the limited dataset used and consequently to the problem of overfitting.

8.7 Discussion

From the results obtained in this study, we found that:

- **Finding 1:** We proposed a context-based self-attention mechanism and exploited three approaches of adding contextual information to self-attention layers. Results on the ADReSS and ADReSSo Challenge datasets showed that the fusion of the outputs (low-level syntactic and semantic information) of different layers as a deep context vector yielded the highest evaluation results.

- **Finding 2:** We compared our proposed approaches with and without label smoothing. Findings suggested that label smoothing contributes to both the performance improvement and improvements in terms of the calibration metrics.

- **Finding 3:** We exploited two methods for fusing the self and cross-attention features. Findings of the experiments conducted on the ADReSS Challenge dataset suggested that the usage of two independent attentional reduction models, the add operation, and the layer normalization achieved better performance than the usage of a co-attention mechanism. On the other hand, results on the ADReSSo Challenge dataset showed that the co-attention mechanism as a fusion method achieved the best evaluation results.

- **Finding 4:** Findings from a series of ablation studies showed the effectiveness of the introduced architecture.

- **Finding 5:** Our proposed models yielded competitive performances to the existing state-of-the-art approaches. We also used the Almost Stochastic Order test to test for statistical significance. This test does not make any assumptions about the distributions of the scores.

- **Finding 6:** We observed that in most cases the performance of the multi-modal models (baselines) was inferior to the transcript only BERT baseline. We hypothesize that this difference in performance is attributable to the fact that the multimodal approaches propose early and late fusion strategies or add / concatenate the representation vectors of different modalities during training. In this way, the intermodal interactions cannot be captured effectively. This difference in performance justifies our initial motivation that more effective fusion methods must be explored for capturing the inter-modal interactions.

Our approaches entail some limitations, which are described below:

- Hyperparameter Tuning: In this study, we did not perform hyperparameter tuning due to the limited access to GPU resources. Hyperparameter tuning yields to an increase of the classification performance.
- Explainability: In this study, we did not apply explainability techniques, namely LIME, Integrated Gradients, and so on, for explaining the predictions of our introduced approaches.
- Self-Supervised Learning: Contrary to self-supervised learning methods, our approach relies heavily on plenty of training data.

8.8 Summary

In this chapter, we introduced some new approaches to detect AD patients from speech and transcripts, which capture the inter- and intra-modal interactions, enhance the conventional self-attention mechanism with contextual information, and deal with the problem of creating overconfident models by applying label smoothing. Our proposed architectures consist of BERT, DeiT, self-attention mechanism incorporating a gating model, context-based self-attention, optimal transport domain adaptation methods, and one new method for fusing the self and cross-attention features in the task of dementia detection from speech data. Furthermore, we designed extensive ablation experiments to explore the effectiveness of the components of the proposed architecture. Extensive experiments conducted on the ADReSS and ADReSSo Challenge datasets demonstrate the efficacy of the proposed architectures reaching Accuracy up to 91.25% and 83.94% respectively. Also, findings suggested that label smoothing contributes to both the performance improvement and calibration of our model.

In Chapters 5-8, we designed fixed deep neural networks for identifying AD patients. Moving forward to the next chapter, we aim to incorporate the power of Neural Architecture Search (NAS) methods into our deep neural network for finding automatically the best performing architecture for our specific task.

Chapter 9

Neural Architecture Search with Multimodal Fusion Methods for Recognizing Dementia

9.1 Introduction

In the previous chapters, we designed fixed deep neural networks. In this chapter, we will introduce a method for generating automatically a deep neural network.

Several research works have been introduced, which employ Convolutional Neural Networks (CNNs) for classifying subjects into AD patients and non-AD ones. Specifically, some of them use as input to CNNs embeddings of transcript data, i.e., GloVE, word2vec, etc. [65]. Other approaches use as input the raw audio signal [63, 93], while others transform the speech signal to log-Mel spectrograms and Mel-frequency Cepstral Coefficients (MFCCs) [63, 91, 94].However, constructing high-performance deep learning models requires extensive engineering and domain knowledge. Neural Architecture Search (NAS) has emerged as class of approaches that automate the generation of state-of-the-art neural network architectures, thus limiting the human effort [373, 374, 375]. A powerful NAS method, namely DARTS [84], has achieved great discovered high-performance convolutional architectures for image classification problems. DARTS uses a continuous relaxation of the architecture representation and then applies gradient descent to discover the best architecture. In this chapter, we present the first study that incorporates DARTS into a neural network for diagnosing dementia from spontaneous speech.

In this chapter, we propose a multimodal neural network, where we pass each transcript through a BERT model [26] and obtain a text representation. Next, we convert audio files into images consisting of log-Mel spectrograms, delta, and delta-delta. We pass each image through the DARTS model. Finally, we exploit a variety of fusion methods for modelling the inter-modal interactions, including Tucker decomposition, a method based on the block-superdiagonal tensor decomposition, etc. To the best of our knowledge, this is the first study to propose such a framework, which combines a NAS approach, a language model, and fusion methods in an end-to-end neural network.

The contributions of this chapter can be summarized as follows:

- We employ a neural architecture search approach, namely DARTS, to automatically generate the best CNN architecture.
- We introduce several fusion methods for combining the representations of the CNN and the BERT model effectively.
- We perform extensive ablation studies to study the impact of the depth of the CNN architecture.
- We perform a series of experiments and show that our introduced architecture yields comparable performance to state-of-the-art approaches.

9.2 Task and Data

Given a labelled dataset consisting of AD and non-AD patients, the task is to identify if an audio file along with its transcript belongs to an AD patient or to a non-AD one.

We use the ADReSS Challenge dataset described in Section 3.3.5.2 for conducting our experiments.

9.3 Predictive Models

In this section, we describe the functionality of the modules, which constitute our introduced architecture. First, we introduce the basic notation and describe the data preprocessing steps. Next, we present the neural architecture search algorithm, namely DARTS [84] that automatically finds the best CNN architecture to process the input speech. Then, we describe the module that process the text modality, using the BERT language model. Finally, we present the multimodal fusion methods, that combine the two modalities and make the final prediction. The whole architecture is end-to-end trainable and is illustrated in Figure 9.1.

Preliminaries. Each input sample consists of a speech signal, a text description of the speech, and the label that indicates if the subject is an AD patient or a non-AD one. We use *librosa* [312] and convert the audio files into images consisting of three channels, namely log-Mel spectrogram, delta, and delta-delta. We use 224 Mel bands, hop length accounting for 1024, and a Hanning window. Each image is resized to 224 × 224 pixels. We denote each image i as $\boldsymbol{X_{I_i}} \in R^{224 \times 224 \times 3}$.

We exploit the python library called *PyLangAcq* [275] for reading the manual transcripts. We use the *BertTokenizer* and pad each transcript to a maximum length of 512 tokens, while transcripts with number of tokens greater than 512 are truncated. *BertTokenizer* returns the attention mask and the input_ids per transcript. We denote each attention mask and input_ids of a transcript i as $\boldsymbol{X_{\alpha_i}} \in R^{512}$ and $\boldsymbol{X_{T_i}} \in R^{512}$ respectively. We further denote the binary label of each sample i as $y_i \in \{0, 1\}$. Therefore each sample i is represented as the tuple $(\boldsymbol{X_{I_i}}, \boldsymbol{X_{\alpha_i}}, \boldsymbol{X_{T_i}}, y_i)$. Our goal is to learn a function $f(\boldsymbol{X_{I_i}}, \boldsymbol{X_{\alpha_i}}, \boldsymbol{X_{T_i}})$, that takes as input the speech and text sample, and predicts the label of the subject.

Speech-Neural Architecture Search. CNNs have achieved great performance in image classification tasks, but require extensive architecture engineering. Therefore, in our work we aim to automatically learn the optimal CNN architecture using the DARTS model [84]. Following previous works [84, 376], since the final CNN architecture can have many layers, to reduce the computation complexity of the model, we search for a computational cell and then we stack this cell many times to construct the CNN architecture.

Each cell can be represented as a directed acyclic graph (DAG), with 7 nodes. Every node x_i denotes a feature map, and every edge (i, j) transforms x_i based on the operation of the edge $o_{(i,j)}$. Since the cell is a DAG, there exists a topological ordering of the nodes. Therefore, we can compute the feature map of each node, based on all the predecessors nodes, using the following equations:

$$x_j = \sum_{i<j} o_{(i,j)}(x_i) \tag{9.1}$$

The goal of the NAS algorithm then is to learn the operations on the edges. In our settings, we search operations in the following set O: $\{$3 × 3 and 5 × 5 separable convolutions, 3 × 3 and 5 × 5 dilated separable convolutions, 3 × 3 max pooling, 3 × 3 average pooling, identity, and zero, which indicates no connection$\}$.

However, gradient-based optimization is not directly applicable in a discrete search space. Therefore, we apply a continuous relaxation in the search space, by learning a set of weights a for each edge operation. The discrete choice of each operation is transformed to a softmax over all operations:

$$\hat{o}_{(i,j)} = \sum_{o \in O} \frac{exp(a_{o_{(i,j)}})}{\sum_{\hat{o} \in O} exp(a_{\hat{o}_{(i,j)}})} o(x) \tag{9.2}$$

To obtain the final CNN architecture, we replace each operation $\hat{o}_{(i,j)}$ with the operation with the largest weight $o_{i,j} = \text{argmax}_{o \in O}\, a_{o_{(i,j)}}$.

Text-Language Models. Bidirectional Encoder Representations from Transformers (BERT) is a multi-layer bidirectional Transformer encoder. It is trained on masked language modeling, where some percentage of the input tokens are masked at random aiming to predict those masked tokens based on the context only. We pass to the BERT model the attention mask and the input_ids denoted by $\boldsymbol{X_{\alpha_i}} \in R^{512}$ and $\boldsymbol{X_{T_i}} \in R^{512}$ respectively. We extract the classification token denoted by [CLS], where its dimensionality is equal to 768. Finally, we project its dimensionality to $d = 64$.

Multimodal Methods. Let $z^t \in \mathcal{R}^{64}$ denote the representation vector of the textual modality. Let $z^v \in \mathcal{R}^{64}$ denote the representation vector of the acoustic modality, extracted by the output of the CNN. We fuse the two modalities, i.e., textual and acoustic, denoted by the vectors z^t and z^v by employing the following fusion methods:

- Tucker decomposition [85]: a bilinear interaction where the tensor is expressed as a Tucker decomposition.
- Multimodal Factorized Bilinear pooling (MFB) [86]: This approach enjoys the dual benefits of compact output features of Multimodal Lowrank Bilinear (MLB) pooling and robust expressive capacity of Multimodal Compact Bilinear (MCB) pooling.
- Multimodal Factorized High-order pooling (MFH) [86]: The MFH approach is developed by cascading multiple MFB blocks.
- BLOCK [87]: Block Superdiagonal Fusion framework for multimodal representation based on the block-term tensor decomposition [314]. It combines the strengths of the Candecomp/PARAFAC (CP) [315] and Tucker decompositions.
- Concatenation: We concatenate z^t and z^v, as $p = [z^t, z^v]$, where $p \in \mathcal{R}^{128}$. We pass p through a dense layer consisting of 16 units with a ReLU activation function.

Finally, we obtain the fused vector, denoted by $z^f \in \mathcal{R}^{16}$, and we pass it through a dense layer with two units, which makes the final prediction. We optimize the model using gradient descent by minimizing the cross-entropy loss.

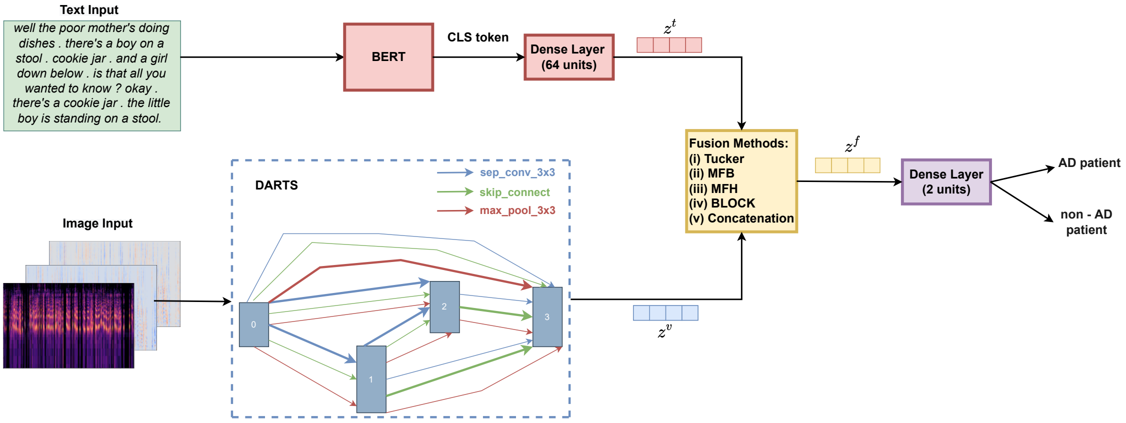


Figure 9.1: Illustration of our introduced architecture. For the text modality, we use a BERT language model to obtain the textual representation. In terms of the acoustic modality, we use the DARTS algorithm for obtaining the optimal CNN architecture and the acoustic representation. We fuse the two representations with fusion methods and pass the fused vector to a dense layer, which makes the prediction.

9.4 Experiments

9.4.1 Comparison with state-of-the-art approaches

We compare our approach with **(i)** unimodal approaches employing only the textual modality, i.e., BERT (Chapter 5), **(ii)** unimodal approaches employing only the acoustic modality, i.e., DARTS, AT-LSTM (x-vector) [88], ECAPA-TDNN [89], SiameseNet [63], x-vectors_SRE[67], Acoustic+Silence [90], YAMNet [91], Majority vote (Acoustic) [64], Audio (Fusion) [92], DemCNN [93], CNN-LSTM (MFCC) [94], and **(iii)** Multimodal approaches employing both the textual and acoustic modality, i.e., Audio + Text (Fusion)[92], Fusion Maj. (3-best) [63], Fusion of system [67], GFI, NUW, Duration, Character 4-grams, Suffixes, POS tag, UD [95], Acoustic & Transcript [90], Dual BERT (Concat/Joint, BERT large) [91], Majority vote (NLP + Acoustic) [64], Attention-based Fusion (Deep Context) of Chapter 8.

9.4.2 Experimental Setup

We minimize the cross-entropy loss function. We use a batch size of 8. We train the models on the ADReSS Challenge train set and report their performance on the test set. We divide the train set into a train and a validation set. We train the model for 50 epochs. We choose the epoch with the smallest validation loss and evaluate the performance of the model on the test set. We repeat the experiments five times and report the mean and standard deviation. We use Weights & Biases [377] for tuning the hyperparameters. Specifically, we perform a random search to optimize the following hyperparameters: number of CNN layers, learning rate for CNN, learning for alpha parameters of DARTS, learning rate for BERT, weight decay, fusion hidden dimension. We use the BERT base uncased version provided via the Transformers library [305]. All models are created using the PyTorch library and trained in a single NVIDIA RTX A6000 48GB GPU.

9.4.3 Evaluation Metrics

Accuracy, Precision, Recall, F1-Score, and Specificity have been used for evaluating the results of the introduced architectures. These metrics have been computed by regarding the dementia class as the positive one. We report the average and standard deviation of these metrics over five runs.

9.5 Results

The results are reported in Table 9.1.

Regarding our proposed multimodal models, we observe that DARTS + BERT + BLOCK is our best performing model reaching Accuracy and F1-score up to 92.08% and 91.94% respectively. It surpasses the introduced multimodal models in Recall by 0.83-6.66%, in F1-score by 2.14-5.21%, and in Accuracy by 2.50-5.00%. DARTS + BERT + MFB constitutes our second best performing model achieving an Accuracy of 89.58% and an F1-score of 89.80%. It outperforms the introduced models, except for DARTS + BERT + BLOCK, in F1-score by 1.15-3.07% and in Accuracy by 0.84-2.50%. In addition, DARTS + BERT + MFH and DARTS + BERT + Concatenation yield almost equal Accuracy results, with DARTS + BERT + MFH surpassing DARTS + BERT +

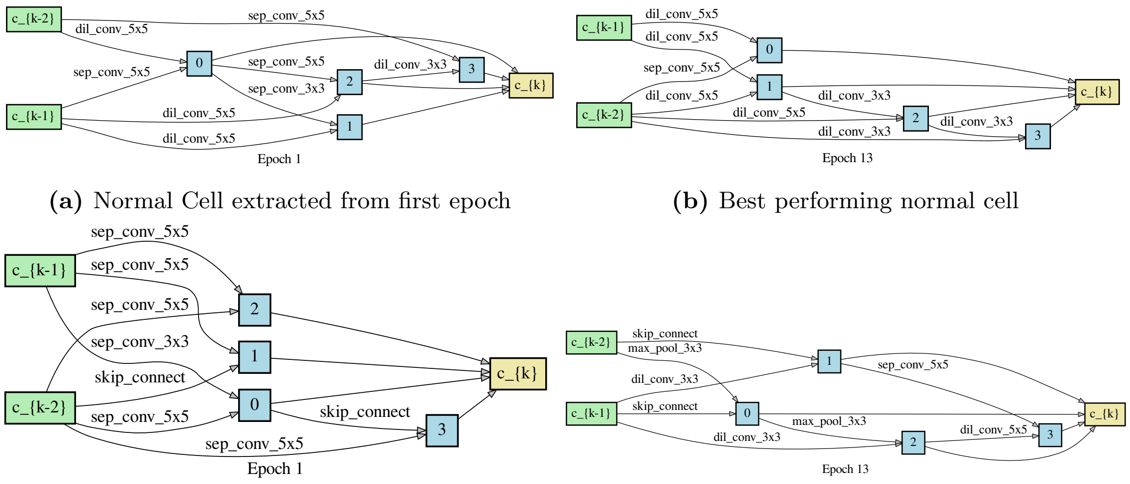
Table 9.1: Performance comparison among proposed models and state-of-the-art approaches on the ADReSS Challenge test set. Reported values are mean ± standard deviation. Results are averaged across five runs. Best results per evaluation metric are in bold.

	Evaluation metrics				
Architecture	**Precision**	**Recall**	**F1-score**	**Accuracy**	**Specificity**
Unimodal state-of-the-art approaches (only transcripts)					
BERT (Chapter 5)	87.19 ±3.25	81.66 ±5.00	86.73 ±4.53	87.50 ±4.37	93.33 ±5.65
Unimodal state-of-the-art approaches (only Speech)					
DARTS	70.04 ±3.84	89.99 ±2.04	76.09 ±0.87	72.92 ±2.28	62.3 ±7.05
AT-LSTM (x-vector) [88]	66.00	69.00	67.00	67.00	-
ECAPA-TDNN [89]	-	-	-	66.70	-
SiameseNet [63]	-	-	70.80	70.80	-
x-vectors_SRE [67]	54.17	54.17	54.17	54.17	54.17
Acoustic+Silence [90]	70.00	58.00	63.00	66.70	75.00
YAMNet [91]	64.40±3.93	73.40±8.82	68.60±4.84	66.20±4.79	59.20±7.73
Majority vote (Acoustic) [64]	-	-	-	65.00	-
Audio (Fusion) [92]	-	83.33	-	81.25	79.17
DemCNN [93]	62.50	62.50	62.50	62.50	62.50
CNN-LSTM (MFCC) [94]	82.00	38.00	51.00	64.58	92.00
Multimodal state-of-the-art approaches (speech and transcripts)					
Audio + Text (Fusion) [92]	-	87.50	-	89.58	91.67
Fusion Maj. (3-best) [63]	-	-	85.40	85.20	-
Fusion of system [67]	94.12	66.67	78.05	81.25	**95.83**
GFI,NUW,Duration,Character 4-grams, Suffixes,POS tag,UD [95]	-	-	-	77.08	-
Acoustic & Transcript [90]	70.00	88.00	78.00	75.00	83.00
Dual BERT [91]	83.04 ±3.97	83.33 ±5.89	82.92 ±1.86	82.92 ±1.56	82.50 ±5.53
Majority vote (NLP + Acoustic) [64]	-	-	-	83.00	-
Attention-based Fusion (Deep Context) (Chapter 8)	93.08 ±2.03	89.17 ±2.04	91.06 ±1.60	91.25 ±1.56	93.33 ±2.04
Our Proposed Architecture					
DARTS+BERT+Tucker Decomposition	89.16 ± 3.96	85.00 ± 6.24	86.73 ± 1.57	87.08 ± 0.83	89.16 ± 5.00
DARTS+BERT+MFB	91.29 ±0.34	88.29 ±3.13	89.80 ±1.76	89.58 ±1.86	91.66 ±1.26
DARTS+BERT+MFH	**94.46** ± 3.38	86.66 ± 3.11	88.31 ± 0.71	88.74 ± 1.02	94.16 ± 3.34
DARTS+BERT+BLOCK	94.09 ±2.61	**91.66** ±6.97	**91.94** ±1.98	**92.08** ±1.56	94.16 ±3.33
DARTS+BERT+Concatenation	86.68 ±3.35	90.83 ±1.66	88.65 ±1.36	88.33 ±1.66	85.83 ±4.25

Concatenation in Accuracy by 0.41%. On the contrary, DARTS + BERT + Concatenation outperforms DARTS + BERT + MFH in F1-score by a small margin of 0.34%. We speculate that DARTS + BERT + MFB performs better than DARTS + BERT + MFH, since the MFH approach is developed by cascading multiple MFB blocks, thus is more complex for our limited dataset. In addition, we observe that the fusion method of Tucker decomposition yields the worst results reaching Accuracy and F1-score up to 87.08% and 86.73% respectively.

Compared with unimodal approaches (employing only text), we observe that our introduced approaches, except for DARTS + BERT + Tucker Decomposition, outperform BERT (Chapter 5). Specifically, DARTS + BERT + BLOCK improves the performance obtained by BERT (Chapter 5) in Precision by 6.90%, in Recall by 10.00%, in F1-score by 5.21%, in Accuracy by 4.58%, and in Specificity by 0.83%. At the same time, we observe that the standard deviations over five runs are lower than BERT in all the evaluation metrics, except Recall.

Compared with unimodal approaches (employing only speech), we observe that DARTS + BERT + BLOCK surpasses these approaches in Precision by 12.09-39.92%, in Recall by 1.67-53.66%, in F1-score by 15.85-40.94%, in Accuracy by 10.83-37.91%, and in Specificity by 2.16-39.99%. We also compare our best performing model with DARTS and show



(a) Normal Cell extracted from first epoch

(b) Best performing normal cell

(c) Reduce cell extracted from first epoch

(d) Best performing reduce cell

Figure 9.2: We visualize the initial normal and reduce cells and the best performing cells obtained from DARTS. These cells are stacked to create the convolutional neural network architecture.

that our best performing model outperforms DARTS in Precision by 24.05%, in Recall by 1.67%, in F1-score by 15.85%, in Accuracy by 19.16%, and in Specificity by 31.86%. Next, we compare our approach, i.e., DARTS, with the existing research initiatives employing only speech. We observe that DARTS outperforms all the research works, except for Audio (Fusion) [92], in Accuracy by 2.12-18.75%. DARTS attains a Recall score accounting for 89.99% and outperforms the state-of-the-art approaches, including Audio (Fusion), in Recall by 6.66-51.99%. DARTS outperforms also the existing research initiatives in terms of F1-score by 5.29-25.09%.

In comparison with multimodal state-of-the-art approaches, we observe that our best performing model outperforms the existing research initiatives in Recall by 3.66-24.99%, in F1-score by 0.88-13.94%, and in Accuracy by 0.83-17.08%. Although Fusion of system [67] obtains a better Specificity score by our best performing model, our best performing model surpasses this approach in Recall, F1-score, and Accuracy. It is worth noting that Recall is a more important metric than Specificity, since high Specificity and low Recall means that AD patients are misdiagnosed as non-AD ones. We observe that DARTS + BERT + BLOCK outperforms the approach proposed in Chapter 8, namely Attention-based Fusion (Deep Context), in terms of all the evaluation metrics, verifying our initial hypothesis that automatically learning an optimal CNN architecture during training yields improvements in the performance.

We further visualize the initialized and the best performing CNN architecture obtained by DARTS, in Figure 9.2. We observe that the best performing cell has different operations and different structure than the initial one, showing how the neural architecture search algorithm converges to an optimal cell, by altering the operations and the connections in the convolutional architecture.

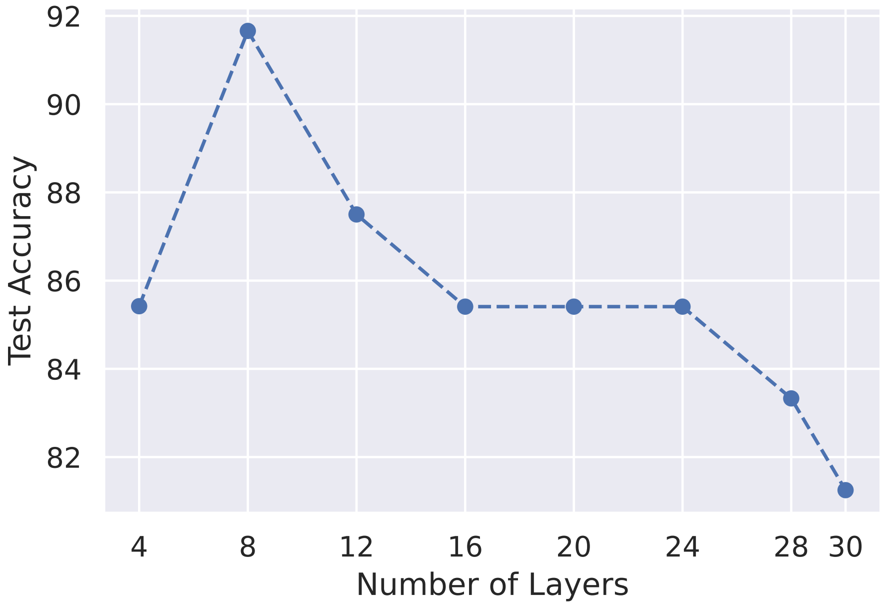


Figure 9.3: Test accuracy of our proposed model with respect to the number of CNN layers generated from DARTS.

9.6 Ablation Study

We perform a series of ablation experiments, where we vary the layers of the CNN architecture, obtained by DARTS. Specifically, we set the number of CNN layers to 4, 8, 12, 16, 20, 24, 28, and 30. We report the accuracy obtained via these experiments in Fig. 9.3. We observe that the best accuracy accounting for 91.66% is obtained by using 8 layers. As the number of layers increases, the accuracy decreases. Specifically, the worst accuracy score is equal to 83.33% and is obtained, when we use 30 CNN layers. We speculate that architectures with many layers are so complex for the dataset, and therefore the model overfits.

9.7 Summary

We presented the first study, which exploits Neural Architecture Search methods and fusion methods based on Tucker Decomposition, Factorized Bilinear Pooling, and block-term tensor decomposition, in the task of dementia detection. Specifically, we proposed an end-to-end trainable multimodal model, which combines an automatically discovered CNN architecture obtained from the NAS algorithm as well as a language model for processing the text information. We integrate the two modalities using a variety of fusion methods. Our approach exhibited comparable performance with the state-of-the-art baselines.

Chapter 10

Multimodal Detection of Epilepsy with Deep Neural Networks

10.1 Introduction

There have been a number of studies proposing methods for detecting epileptic seizures. The majority of these studies first extract both time-domain and frequency domain features from the electroencephalogram (EEG) signals. For instance, the authors apply the Discrete Wavelet Transform (DWT) [97, 98] for decomposing the EEG signals into sub-bands and then extract features from each sub-band. After having extracted a large number of features, the authors usually exploit feature selection or dimensionality reduction techniques for finding the best subset of features or reducing the dimension of the feature vector respectively. The last step of the proposed methods includes the train of traditional machine learning classifiers, i.e., Logistic Regression (LR), Support Vector Machines (SVMs), Random Forests (RF), Decision Trees, etc. However, these methods are time-consuming, since they demand some level of domain expertise for extracting the best representative features. Only a few number of studies [99, 100, 101, 102] have exploited deep neural networks, i.e., CNNs, LSTMs, or BiLSTMs in the task of epilepsy detection and prediction. However, most of these methods still rely on handcrafted features [100, 101, 99]. Another limitation is the fact that existing research works split the EEG signals into segments and propose majority-voting approaches [102]. Thus, they have to train multiple models separately increasing substantially the computation time. At the same time, most of the CNN models are not able to model effectively the temporal dependencies among the EEG data. Although LSTMs and BiLSTMs can capture the temporal dependencies in EEG data, they usually have high model complexity.

In order to tackle these limitations, we propose two new methods to distinguish healthy, interictal, and ictal cases. Firstly, we introduce a unimodal approach, where we apply the short-time fourier transform (STFT) to the EEG signal and we construct an image consisting of three channels, namely the db-scaled (after having computed the absolute values) STFT spectrogram, its delta, and delta-delta. Next, we employ several pretrained models of the domain of computer vision, including AlexNet, VGG16, EfficientNet, etc. and compare their performances. Secondly, we introduce a deep neural network, which distinguishes healthy, interictal, and ictal cases in an end-to-end trainable manner without requiring the exhaustive and tedious procedure of feature extraction. We aim through this neural network to automate the process of the feature extraction by exploiting the capabilities of deep learning. Specifically, each EEG signal is passed through two branches of convolutional neural networks (CNNs) with different filter sizes, in order to capture both the temporal and the frequency information. Next, we apply the short-time fourier transform (STFT) to the EEG signal and we construct an image consisting of three channels, namely the db-scaled (after having computed the absolute values) STFT spectrogram, its delta, and delta-delta. Each image is passed through a pretrained EfficientNet-B7 model. Finally, the EEG representation vector and the image vector are passed through a Gated Multimodal Unit for suppressing the irrelevant information. We perform extensive experiments (and ablation studies) on a publicly available dataset, namely the EEG database of the University of Bonn, and experimental results demonstrate that our introduced model can achieve valuable advantages over existing research initiatives.

The contributions of this chapter can be summarized as follows:

- We propose a unimodal approach for detecting healthy, interictal, and ictal cases. Specifically, we apply the STFT algorithm to the single-channel EEG signals. We construct an image consisting of the db-scaled (after having computed the absolute values) spectrogram, the delta, and delta-delta. Each image is passed through pretrained models used extensively in the computer vision domain, such as AlexNet, VGG16, EfficientNet, etc. We compare the performance of the pretrained models. To the best of our knowledge, there is no prior work creating images in this way towards the epileptic seizures detection task.

- We propose a multimodal neural network, which employs (i) two branches of CNNs with different kernel sizes for processing EEG signals, (ii) an EfficientNet-B7 model for obtaining a visual representation vector from an image consisting of the db-scaled (after having computed the absolute values) spectrogram, the delta, and delta-delta, and (iii) a gated multimodal unit, which controls the importance of each modality towards the final prediction. To the best of our knowledge, this is the first study proposing a multimodal deep neural network with these components.

- We conduct our experiments on a publicly available dataset and consider five cases for classification.

- We run a series of ablation experiments to explore the effectiveness of the components of our introduced deep learning architecture.

- Our introduced model obtains comparable performance to the state-of-the-art approaches.

Table 10.1: Description of cases considered for classification

Case	Classes	Description
I	AB, CD, E	healthy, interictal, ictal
II	A, E	healthy, seizure
III	AB, CD	healthy, interictal
IV	AB, CDE	healthy, epileptic
V	A, C, E	healthy, interictal, ictal

10.2 Dataset

We use the publicly available EEG dataset of University of Bonn for conducting our experiments [103].

In this paper, we have considered five different cases for conducting our experiments, all of which are presented below and reported in Table 10.1.

- Case I (AB - CD - E)
- Case II (A - E)
- Case III (AB - CD)
- Case IV (AB - CDE)
- Case V (A - C - E)

10.3 Predictive Unimodal Models

In this section, we present our unimodal approaches using only image data. First, we apply the short-time fourier transform (STFT) with a Hanning window to the raw EEG signals. After calculating the absolute values of the STFT spectrogram (STFT's magnitude), we compute the db-scaled spectrogram, the delta, and delta-delta. Thus, we construct an image consisting of three channels, i.e., db-scaled spectrogram, delta, and delta-delta. We scale each image to $[0, 1]$. Each image is resized to (224×224) pixels. The pixel values of all images are normalized.

Next, we pass each image through the following pre-trained models: ResNet50 & ResNet18 [336], WideResNet-50-2 [337], AlexNet [44], SqueeezeNet 1.1 [338], DenseNet-201 [339], ResNeXt-50 $(32 \times 4d)$ [342], VGG16 [343], and EfficientNet B7[1] [344].

We modify the output layer of the aforementioned models. Specifically, for cases II, III, and IV, the output layer consists of two units. For cases I and V, the output layer consists of three units.

[1] We experimented also with EfficientNet-B0 to B6, but EfficientNet-B7 was the best performing model.

10.3.1 Experiments

10.3.1.1 Experimental Setup

We use a 10-fold stratified cross-validation procedure to train and test the proposed models. In each iteration of this procedure, we split the train set into a train and a validation set. All models have been trained with an Adam optimizer and a learning rate of 1e-5. We apply *ReduceLROnPlateau*, where we reduce the learning rate by a factor of 0.1, if the validation loss has stopped decreasing for three consecutive epochs. Also, we apply *EarlyStopping* and stop training, if the validation loss has stopped decreasing for eight consecutive epochs. We minimize the cross-entropy loss function. All models have been created using the PyTorch library [306]. All experiments are conducted on a single Tesla P100-PCIE-16GB GPU with a running time of approximately one hour.

10.3.1.2 Evaluation Metrics

Accuracy, Precision, Recall, and F1-Score have been used for evaluating the results of the introduced models.

Regarding the binary classification task (Cases II, III, and IV), these metrics have been computed by regarding the seizure/interictal/epilepsy class as the positive one.

Regarding the multiclass classification task (Cases I and V), we report the precision, recall, and F1-score for each class separately.

For all the cases, results are presented via mean ± standard deviation (over 10 folds).

10.3.2 Results

The results of the proposed models mentioned before are reported in Tables 10.2-10.6. More specifically, in Table 10.2 we report the results for case I (AB - CD - E), in Table 10.3 we report the results for case II (A - E), in Table 10.4 we report the results for case III (AB - CD), in Table 10.5 we report the results for case IV (AB - CDE), and in Table 10.6 we report the results for case V (A - C - E).

Regarding case I (AB - CD - E), as one can easily observe from Table 10.2, EfficientNet-B7 constitutes our best performing model achieving an accuracy score equal to 95.20% surpassing the other models by 1.21-11.40%. In terms of F1-score, which constitutes the weighted average of precision and recall, our best performing model achieves the highest score for all the classes except CD in comparison to the other models. F1-scores accounting for 96.57% and 90.71% are obtained for the AB (healthy volunteers) and E (ictal state) classes respectively. EfficientNet-B7 improves the F1-score by 2.00-12.80% and 0.10-17.76% for AB and E class respectively. The highest F1-score accounting for 96.28% for the CD class is obtained by the VGG16 model. AlexNet is the second best performing model attaining an accuracy score equal to 93.99%. ResNet50 achieves the worst performance among the introduced unimodal models with accuracy accounting for 83.80%.

In terms of case II (A - E), one can easily observe from Table 10.3 that EfficientNet-B7 outperforms the other pretrained models in terms of Accuracy, Recall, and F1-score by 2.00-8.00%, 4.00-14.00%, and 2.43-9.71% respectively. This fact renders EfficientNet-B7 the best performing model for case II (A - E). The highest precision score is obtained by WideResNet-50-2 and is equal to 99.00%. Similarly to case I (AB - CD - E), ResNet50 obtains the worst classification results with accuracy accounting for 87.50%. The accuracy achieved by the pretrained models, except EfficientNet-B7 and ResNet50, ranges from 90.00% to 93.50%, with AlexNet being the second best performing model in terms of accuracy and F1-score.

With regards to case III (AB - CD), looking at Table 10.4, one can observe that EfficientNet-B7 attains the highest accuracy and F1-score accounting for 97.50% and 97.45% respectively. Specifically, EfficientNet-B7 surpasses the other models in terms of Accuracy and F1-score by a margin of 0.25-4.50% and 0.22-4.97% respectively. It is worth noting that although EfficientNet-B7 achieves lower precision and recall scores by other models, it surpasses them in F1-score, which constitutes the weighted average of precision and recall. In addition, VGG16 and DenseNet201 obtain equal accuracy scores accounting for 97.25%, with VGG16 surpassing DenseNet201 in terms of F1-score by a small margin of 0.08%. ResNet18 achieves the lowest classification results with accuracy reaching up to 93.00%.

In terms of case IV (AB - CDE), one can observe in Table 10.5 that EfficientNet-B7 constitutes the best performing model surpassing the other models in accuracy by 0.40-10.00%, in precision by 1.62-10.45%, and in F1-score by 0.26-8.39%. In addition, AlexNet is the second best performing model obtaining an accuracy score accounting for 96.00% and the highest recall score equal to 97.00%. DenseNet201 achieves the worst performance in this case reaching accuracy up to 86.40%.

Finally, looking at Table 10.6 for the case V (A - C - E), one can observe that EfficientNet-B7 yields the highest accuracy accounting for 93.00% surpassing the other models' performance by 3.00-13.00%. With regards to the F1-score, EfficientNet-B7 obtains the highest scores for all the classes, i.e., A, C, and E. Specifically, scores accounting for 91.99%, 93.79%, and 93.02% are obtained for the classes A, C, and E respectively. The second best accuracy score is obtained by VGG16 and is equal to 90.00%. The other models obtain an accuracy score, which ranges from 80.00% to 89.33% with DenseNet201 and ResNet50 obtaining the lower accuracy results equal to 80.00% and 80.33% respectively.

Overall, EfficientNet-B7 constitutes the best performing model in terms of the accuracy for all the cases considered for classification.

10.4 Proposed Multimodal Model

In this section, we describe our introduced architecture for detecting epilepsy using EEG signals and STFT spectrograms. The proposed architecture is illustrated in Fig. 10.1.

Table 10.2: Performance comparison among unimodal proposed models via cross-validation (AB - CD - E). Reported values are mean ± standard deviation. Best results per evaluation metric are in bold.

Model	**Evaluation metrics**									
	Precision			**Recall**			**F1-score**			**Accuracy**
	AB	CD	E	AB	CD	E	AB	CD	E	
ResNet50	85.94	86.94	80.11	82.99	91.49	70.00	83.77	88.55	72.95	83.80
	±7.28	±9.41	±14.15	±11.45	±7.09	±14.14	±6.29	±4.39	±9.63	±4.77
ResNet18	85.42	95.87	87.86	92.50	90.49	77.00	88.20	92.97	79.59	88.60
	±10.75	±3.85	±11.02	±6.80	±6.49	±20.52	±5.86	±4.24	±13.94	±5.14
WideResNet	83.73	**97.33**	81.36	92.50	90.00	74.00	87.73	93.45	76.37	87.80
50-2	±5.05	±2.67	±13.59	±4.03	±4.99	±14.97	±2.82	±3.27	±11.05	±3.28
AlexNet	94.27	95.02	**93.77**	94.99	95.50	89.00	94.57	94.98	90.61	93.99
	±4.54	±4.78	±8.66	±2.24	±6.49	±11.36	±2.59	±2.84	±7.62	±3.22
SqueezeNet	86.22	92.18	81.18	89.49	92.99	72.00	87.58	92.55	75.58	87.40
1.1	±8.15	±3.62	±9.31	±7.23	±3.32	±15.36	±6.07	±2.88	±10.90	±4.48
DenseNet	89.02	89.68	77.91	82.50	89.49	82.00	84.73	88.98	78.80	85.20
201	±12.07	±7.39	±13.05	±9.29	±8.79	±13.27	±6.65	±3.84	±8.95	±5.46
ResNeXt-50	88.93	91.36	84.56	86.00	93.49	84.00	86.93	92.18	83.96	88.59
($32 \times 4d$)	±6.11	±6.40	±7.20	±9.17	±5.50	±13.56	±4.53	±3.89	±10.19	±4.00
VGG16	90.49	96.30	90.58	94.50	96.50	80.00	92.25	**96.28**	84.20	92.40
	±6.40	±5.43	±7.97	±5.22	±5.02	±13.42	±3.98	±4.17	±8.63	±3.88
EfficientNet	**96.42**	96.11	93.03	**97.00**	**96.00**	**90.00**	**96.57**	95.97	**90.71**	**95.20**
B7	±5.09	±4.19	±9.08	±3.32	±5.39	±11.83	±2.40	±3.98	±7.73	±3.25

Table 10.3: Performance comparison among unimodal proposed models via cross-validation (A - E). Reported values are mean ± standard deviation. Best results per evaluation metric are in bold.

	Evaluation metrics			
Architecture	**Precision**	**Recall**	**F1-score**	**Accuracy**
ResNet50	93.92 ±8.46	81.00 ±17.00	85.79 ±11.27	87.50 ±9.01
ResNet18	96.35 ±5.64	83.00 ±11.00	88.92 ±8.26	90.00 ±7.07
WideResNet 50-2	**99.00** ±2.99	87.00 ±9.00	92.32 ±5.25	93.00 ±4.58
AlexNet	96.98 ±4.64	90.00 ±8.94	93.07 ±5.61	93.50 ±5.02
SqueezeNet 1.1	92.89 ±6.89	88.00 ±11.66	89.97 ±7.57	90.50 ±6.87
DenseNet 201	97.75 ±4.53	86.00 ±9.17	91.27 ±6.37	92.00 ±5.57
ResNeXt-50 $(32 \times 4d)$	97.89 ±4.23	85.00 ±8.06	90.74 ±5.07	91.50 ±4.50
VGG16	94.82 ±5.26	91.00 ±9.43	92.67 ±6.59	93.00 ±6.00
EfficientNet B7	96.42 ±5.77	**95.00** ±4.99	**95.50** ±3.38	**95.50** ±3.50

- **EEG signal:** As illustrated in Fig. 10.1, we implement two branches of CNNs with different kernel sizes to process the raw EEG signals. The choice of these two branches of CNNs with small and large filter sizes is inspired by [104, 105], where the authors state that the small filter is able to capture temporal information, while the larger filter is capable of capturing frequency information.

  Each branch consists of three convolutional layers and two max-pooling layers, where each convolutional layer includes a batch normalization layer [106] and a ReLU activation function. As one can observe from Fig. 10.1, the first convolutional block of each branch shows the filter size, the number of filters, and the stride size. The next two convolutional blocks of each branch show the filter size and the number of filters. The stride size is equal to 1. Each max-pool block shows the pooling size and the stride size. For reducing overfitting, we apply dropout with a rate of 0.5 after the first max-pool block of each branch and after the concatenation of both branches. Finally, we flatten the matrix to a 1d vector.

  Let the output of this part of the architecture be f^t.

Table 10.4: Performance comparison among unimodal proposed models via cross-validation (AB - CD). Reported values are mean ± standard deviation. Best results per evaluation metric are in bold.

	Evaluation metrics			
Architecture	**Precision**	**Recall**	**F1-score**	**Accuracy**
ResNet50	93.89 ±6.85	95.00 ±7.07	94.08 ±4.31	94.00 ±4.50
ResNet18	97.18 ±3.69	89.00 ±9.17	92.48 ±4.09	93.00 ±3.32
WideResNet *50-2*	98.47 ±2.33	92.50 ±4.61	95.31 ±2.57	95.50 ±2.45
AlexNet	95.25 ±2.88	**98.50** ±3.20	96.80 ±2.26	96.75 ±2.25
SqueezeNet *1.1*	95.99 ±2.93	94.00 ±2.99	94.95 ±2.32	95.00 ±2.24
DenseNet *201*	**99.02** ±1.95	95.50 ±4.72	97.15 ±2.51	97.25 ±2.36
ResNeXt-50 ($32 \times 4d$)	93.98 ±4.81	93.50 ±5.02	93.49 ±1.22	93.50 ±1.22
VGG16	97.55 ±2.46	97.00 ±3.32	97.23 ±2.12	97.25 ±2.08
EfficientNet *B7*	98.55 ±2.22	96.50 ±3.91	**97.45** ±3.91	**97.50** ±1.94

- **Image representation:** We apply the short-time fourier transform (STFT) to the raw EEG signals. After calculating the absolute values of the STFT spectrogram (STFT's magnitude), we compute the db-scaled spectrogram, the delta, and delta-delta. Thus, we construct an image consisting of three channels, i.e., db-scaled spectrogram, delta, and delta-delta. We scale each image to $[0, 1]$. Each image is resized to (224×224) pixels. The pixel values of all images are normalized.

  As shown in Fig. 10.1, each image is fed to a pretrained EfficientNet-B7 model followed by a dropout layer with a rate of 0.5. We choose the EfficientNet-B7, since it is the best performing model as shown in Section 10.3.2. We also remove the last layer of EfficientNet used for classification. Thus, the pretrained EfficientNet-B7 model acts as a feature extractor.

  Let the output of this part of the architecture be f^v.

- **Gated Multimodal Unit:** We apply the Gated Multimodal Unit introduced by [75] and implemented in Chapter 6, in order to assign more importance to the relevant modality suppressing the irrelevant information. Given f^t and f^v as computed

Table 10.5: Performance comparison among unimodal proposed models via cross-validation (AB - CDE). Reported values are mean ± standard deviation. Best results per evaluation metric are in bold.

	Evaluation metrics			
Architecture	**Precision**	**Recall**	**F1-score**	**Accuracy**
ResNet50	93.71 ±3.31	92.00 ±3.71	92.77 ±2.37	91.40 ±2.84
ResNet18	88.20 ±5.77	87.67 ±7.75	89.81 ±5.03	88.20 ±5.55
WideResNet 50-2	94.85 ±4.87	89.00 ±2.99	91.79 ±3.39	90.40 ±4.08
AlexNet	96.42 ±2.54	**97.00** ±3.14	96.67 ±2.13	96.00 ±2.53
SqueezeNet 1.1	92.32 ±4.56	88.67 ±4.76	90.32 ±3.10	88.60 ±3.69
DenseNet 201	90.44 ±6.27	87.33 ±5.54	88.54 ±2.69	86.40 ±3.56
ResNeXt-50 ($32 \times 4d$)	91.42 ±6.25	90.33 ±7.81	90.40 ±3.12	88.60 ±3.35
VGG16	97.03 ±2.24	94.67 ±4.27	95.75 ±1.99	95.00 ±2.24
EfficientNet B7	**98.65** ±1.65	95.33 ±3.06	**96.93** ±1.69	**96.40** ±1.96

above, we calculate the multimodal representation h, as follows:

$$h^t = \tanh\left(W^t f^t + b^t\right) \tag{10.1}$$

$$h^v = \tanh\left(W^v f^v + b^v\right) \tag{10.2}$$

$$z = \sigma(W^z \left[f^v; f^t\right] + b^z) \tag{10.3}$$

$$h = z * h^v + (1 - z) * h^t \tag{10.4}$$

$$\Theta = \{W^t, W^v, W^z\} \tag{10.5}$$

where Θ denote the parameters to be learned, and $[.;.]$ the concatenation operation. We project the f^t, f^v, and the concatenated vector $[f^v; f^t]$ to obtain the same dimensionality ($d_{proj} = 256$).

- **Output Layer:** The multimodal representation h is passed to a dropout layer with a rate of 0.5 followed by a dense layer, which gives the final output. The number of units in the dense layer depends on each case considered for classification and can be either two (binary classification) or three units (multiclass classification).

Table 10.6: Performance comparison among unimodal proposed models via cross-validation (A - C - E). Reported values are mean ± standard deviation. Best results per evaluation metric are in bold.

Model	**Evaluation metrics**									
	Precision			**Recall**			**F1-score**			**Accuracy**
	A	C	E	A	C	E	A	C	E	
ResNet50	84.31	75.29	86.98	61.00	**92.00**	88.00	70.31	82.19	86.67	80.33
	±6.81	±11.64	±9.68	±13.00	±7.48	±11.66	±10.70	±7.58	±7.19	±6.23
ResNet18	76.53	83.86	89.47	76.00	87.99	84.00	75.43	85.34	86.07	82.66
	±10.48	±10.02	±8.21	±19.08	±11.66	±9.17	±13.61	±8.41	±5.28	±6.63
WideResNet	81.42	87.49	90.55	83.00	85.00	82.99	79.42	84.37	85.63	83.66
50-2	±11.94	±12.10	±9.66	±19.52	±14.32	±13.45	±12.45	±7.38	±8.16	±6.57
AlexNet	84.46	89.77	**95.05**	89.00	87.99	90.99	86.46	88.71	92.83	89.33
	±10.69	±10.45	±6.44	±10.44	±12.49	±6.99	±9.55	±11.04	±5.68	±3.22
SqueezeNet	78.23	86.16	86.77	84.00	87.99	75.00	80.28	86.56	79.74	82.33
1.1	±8.31	±12.15	±9.76	±15.62	±8.72	±10.25	±9.91	±8.53	±7.23	±7.31
DenseNet	79.92	76.18	89.82	67.00	89.00	84.00	71.79	80.98	86.23	80.00
201	±10.41	±11.35	±7.93	±11.87	±11.36	±12.81	±6.93	±7.16	±9.06	±5.37
ResNeXt-50	80.78	76.39	95.58	73.00	89.00	85.00	75.47	81.77	89.44	82.33
(32 × 4*d*)	±13.41	±6.62	±7.49	±13.45	±12.21	±8.06	±9.98	±7.61	±4.11	±5.59
VGG16	84.72	96.08	92.46	94.00	90.99	85.00	88.52	93.18	88.33	90.00
	±10.94	±6.17	±8.57	±7.99	±9.43	±11.18	±6.92	±6.46	±9.09	±6.99
EfficientNet	**89.68**	**96.52**	94.55	**95.00**	**92.00**	**92.00**	**91.99**	**93.79**	**93.02**	**93.00**
B7	±8.12	±6.98	±9.27	±12.04	±8.72	±7.48	±9.29	±5.69	±7.17	±6.74

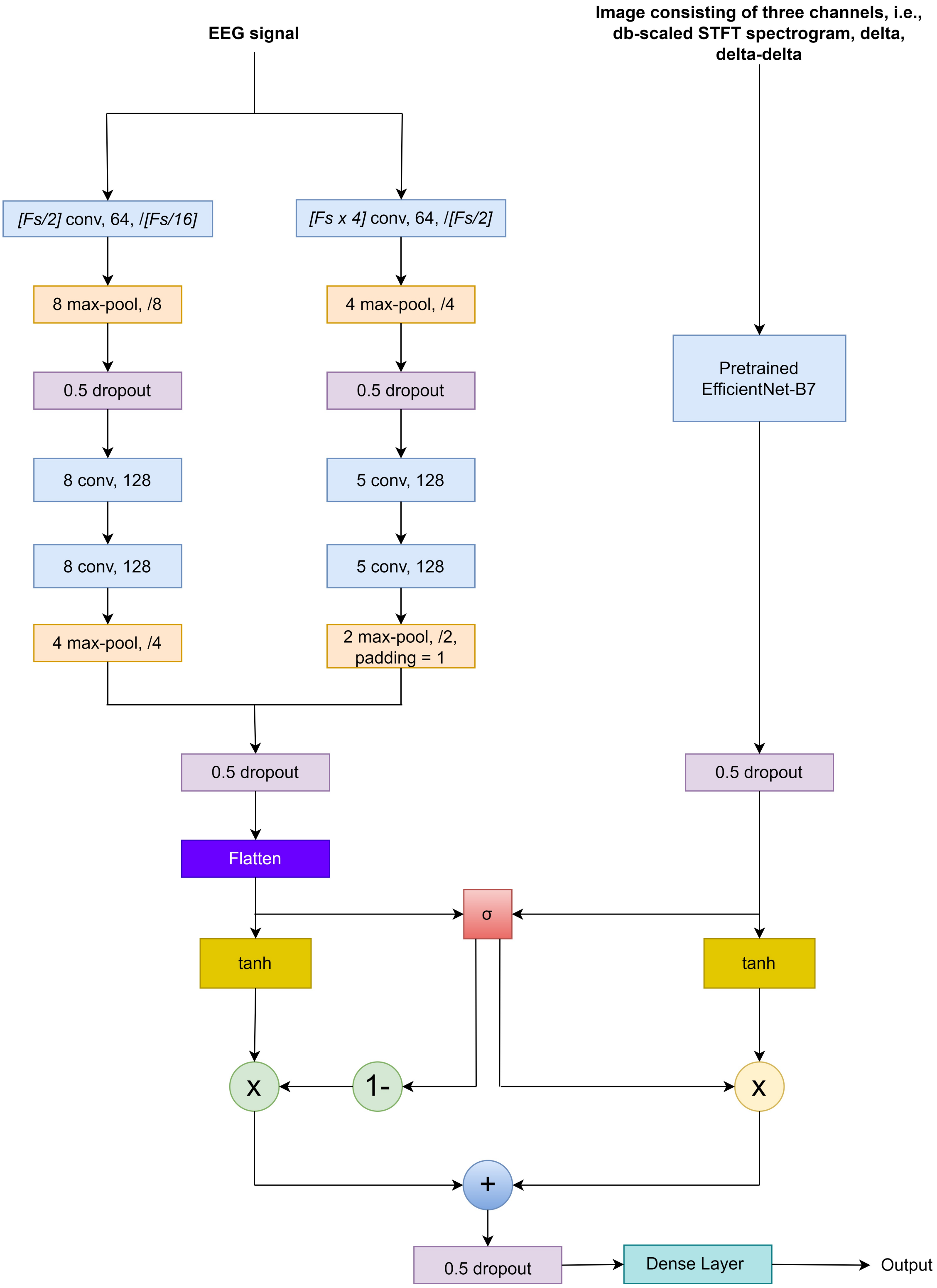

EEG signal
Image consisting of three channels, i.e., db-scaled STFT spectrogram, delta, delta-delta
[Fs/2] conv, 64, /[Fs/16]
[Fs x 4] conv, 64, /[Fs/2]
8 max-pool, /8
4 max-pool, /4
0.5 dropout
0.5 dropout
8 conv, 128
5 conv, 128
8 conv, 128
5 conv, 128
4 max-pool, /4
2 max-pool, /2, padding = 1
Pretrained EfficientNet-B7
0.5 dropout
0.5 dropout
Flatten
σ
tanh
tanh
x
1-
x
+
0.5 dropout
Dense Layer
Output

Figure 10.1: Proposed Architecture

10.5 Experiments

10.5.1 Comparison with state-of-the-art approaches

- **Case I (AB - CD - E)**
  - Novel RF [107]: This method applies a short-time Fourier transform (STFT) and extracts the mean energy, standard deviation, and high amplitude gamma frequency of signals. The dimensionality of the respective feature set is reduced via the Principal Component Analysis (PCA) and then is fed to a Random Forest Classifier. A grid search optimization technique has also been exploited.
  - EMD, higher order moments, ANN [108]: This method extracts the variance, kurtosis, and skewness from the intrinsic mode function (IMF) obtained by the empirical mode decomposition (EMD) of the EEG signal. These features are fed to an artificial neural network (ANN) for the classification of the EEG signals.
  - BiLSTM [99]: This work extracts the instantaneous frequency and the spectral entropy of the EEG signals and trains a Bi-LSTM neural network.
  - DWT + Kmeans + Multilayer perceptron neural network (MLPNN) [98]: This method applies DWT for decomposing the EEG signal into a set of sub-bands. K-means clustering for the wavelet coefficients in each sub-band is then used. Finally, the probability of belonging of wavelet coefficients to a cluster for each sub-band is fed to a MLPNN.
  - CNN [109]: This method proposes a deep neural network consisting of three convolutional blocks followed by three fully connected layers to classify EEG signals.
  - Random Forest [110]: This method applies the STFT to the EEG signals and extracts the alpha band. Then, the mean, variance, skewness, and kurtosis of the alpha band are used as features for training traditional machine learning classifiers with the Random Forest achieving the highest classification results.
  - Matrix Determinant and MLP [111]: This research work proposes the arrangement of the EEG time series in square matrix form of order 13, 16, 23, and 32 and then introduces the matrix determinant as a significant feature.
  - EMD and SVM [112]: This method applies empirical mode decomposition of the EEG signals for getting the intrinsic mode function (IMF). The authors select the first three IMFs for further preprocessing. Next, they calculate the temporal and spectral characteristics of the IMFs creating in this way a feature set, which is fed to a Support Vector Machine Classifier.
  - dual-tree complex wavelet transform domain [113]: This method decomposes the EEG signal into sub-bands using the dual-tree complex wavelet transform

  (DT-CWT). Then, the parameters of a normal inverse Gaussian (NIG) probability density function (pdf) are estimated from the various sub-bands and are used as features for training a Support Vector Machine Classifier.

  - statistical dual-tree complex wavelet transform domain [114]: This paper applies a dual-tree complex wavelet transform to decompose the EEG signal into sub-bands. Variances calculated from the EEG signals and the sub-bands are used as features and fed to ANN and SVM.
  - ANN, hierarchical multi-class SVM with new kernel [115]: This work first decomposes the EEG signal into six sub-bands using the wavelet transform. Next, the authors extract six features from each sub-band, namely the approximate entropy, largest Lyapounox exponent, minimum, maximum, mean, and standard deviation. Finally, they introduce a hierarchical multiclass SVM with extreme learning machine (ELM) as kernel for the classification.
  - Random Forest, wavelets [116]: This method applies a five level decomposition of the EEG signal using the Discrete Wavelet Transform. After extracting five features per sub-band, the authors train a Random Forest Classifier.
  - CNN [117]: This method introduces a 13-layer deep convolutional neural network to classify EEG signals into normal, pre-ictal, and seizure classes.
  - OPF [118]: This method applies the DWT to the EEG signals, extracts statistical features from each sub-band, and applies feature selection techniques, including Relief, InfoGain, and correlation-based feature subset selection. Finally, the authors employ the optimum path forest (OPF) classifier.
  - Symlets wavelets, statistical mean energy std and PCA, GBM-GSO, RF, SVM [119]: This method adopts fourth-order Symlet wavelets for decomposing the EEG data into five frequencies sub-bands. Next, statistical features are computed and used as classification features.

- **Case II (A - E)**
  - Relative Wavelet Energy [378]: This method applies DWT for decomposing the EEG signal into sub-bands, extracts the relative wavelet energy, and trains an ANN.
  - Permutation entropy, SVM [379]: This method applies wavelet decomposition, extracts the permutation entropy, and trains a Probabilistic Neural Network.
  - stacked sparse autoencoders [380]: This method trains a stacked sparse autoencoder with a softmax classifier.
  - Cross-correlation aided SVM classifier [381]: This method extracts cross-correlation and trains an SVM classifier.
  - Permutation entropy - SVM classifier [382]: This method extracts Permutation entropy and trains an SVM classifier.

  - ME [97]: This method decomposes the EEG signal into sub-bands via DWT and trains a Mixture of Experts (ME) model consisting of a gating network and several expert networks, where a double-loop Expectation-Maximization (EM) algorithm has been introduced.
  - multiwavelet transform based approximate entropy and ANN [383]: This method uses approximate entropy features derived via multiwavelet transform for training an artificial neural network.

- **Case III (AB - CD)**
  - ATFFWT and FD, LS-SVM [384]: This method employs analytic time-frequency flexible wavelet transform (ATFFWT) for decomposing the EEG signals into sub-bands. Next, the authors calculate the fractal dimension for each sub-band, and use the fractal dimensions as features for training a least-square support vector machine classifier.
  - Novel RF [107]
  - Random Forest [110]
  - novel signal modeling [385]: This method introduces a new 3-level multirate filterbank structure based on DCT, and a new statistical modeling of brain rhythms. Finally, the authors use the hurst exponent values and ARMA parameters as features for training an SVM classifier.
  - Symlets wavelets, statistical mean energy std and PCA, GBM-GSO, RF, SVM [119]
  - MFDFA features + SVM [386]: This approach exploits multifractal detrended fluctuation analysis, extracts a set of 14 features, and trains an SVM classifier.

- **Case IV (AB - CDE)**
  - Random Forest [110]
  - APN [387]: This research work decomposes the EEG signal into sub-bands by applying a discrete wavelet transform, exploits the minimize entropy principle approach, and finally constructs an associative Petri net model.
  - Alpha band (Blackman window) [388]: This method applies STFT to the EEG signals exploiting the Blackman window, extracts the alpha band from the t-f plane, extracts statistical features from the alpha band of the tf-plane, and trains a Random Forest classifier.

- **Case V (A - C - E)**
  - LSTM [101]: This method applies a multi-rate DCT filter which divides each EEG signal into five sub-bands of different bandwidths. Next, Hurst and ARMA features are extracted for each sub-band, which generate a total of 20 features for an EEG signal. Finally, the authors train an LSTM architecture.

- Matrix Determinant and MLP [111]
- DWT and neural network [389]: This method utilizes DWT to decompose the EEG signal into sub-bands and then extracts statistical features per sub-band, namely maximum, minimum, mean, and standard deviation. These features are fed to a NN for the classification.
- DWT and ensemble classifier [390]: This method utilizes DWT to decompose the EEG signal into sub-bands and then extracts statistical features per sub-band. Finally, it proposes an ensemble classifier combining four classification algorithms, namely ANN, Bayes, k-NN, and SVM.
- CNN [391]: This method employs a convolutional neural network.

10.5.2 Experimental Setup

We use a 10-fold stratified cross-validation procedure to train and test the proposed model. In each iteration of this procedure, we split the train set into a train and a validation set. The proposed model has been trained with an Adam optimizer and a learning rate of 1e-4. We apply *ReduceLROnPlateau*, where we reduce the learning rate by a factor of 0.1, if the validation loss has stopped decreasing for three consecutive epochs. Also, we apply *EarlyStopping* and stop training, if the validation loss has stopped decreasing for eight consecutive epochs. We minimize the cross-entropy loss function. We use the PyTorch library [306]. All experiments are conducted on a single Tesla P100-PCIE-16GB GPU with a running time of approximately two hours.

10.5.3 Evaluation Metrics

Accuracy, Precision, Recall, and F1-Score have been used for evaluating the results of the introduced architecture.

Regarding the binary classification task (Cases II, III, and IV), these metrics have been computed by regarding the seizure/interictal/epilepsy class as the positive one.

Regarding the multiclass classification task (Cases I and V), we report the precision, recall, and F1-score for each class separately. Also, we report the macro metrics.

For all the cases, results are presented via mean ± standard deviation (over 10 folds).

10.6 Results

The results for all the cases considered for classification of our introduced multimodal model described in Section 10.4 are reported in Tables 10.7-10.12. Tables 10.7 and 10.11 present the results of the proposed model for cases I and V respectively and report the precision, recall, and F1-score for each class separately. In Table 10.12, we report the macro results (precision, recall, and F1-score) for cases I and V. Tables 10.8, 10.9, and 10.10 present the results of the introduced approach for cases II, III, and IV respectively.

In addition, Tables 10.13-10.17 present a comparison of the results between our proposed model and state-of-the-art approaches in terms of the accuracy score. Specifically, Table 10.13 presents the comparison for case I. In Table 10.14, we compare the results of our proposed model on case II with existing research initiatives. Similarly, Tables 10.15 and 10.16 show the comparison of the proposed multimodal model for cases III and IV respectively with state-of-the-art approaches. Finally, Table 10.17 compares the results of our approach with research works on case V.

For case I (AB - CD - E), as observed in Table 10.7, our model achieves an accuracy score accounting for 97.00%. F1-scores equal to 97.52%, 96.77%, and 96.41% are obtained for the AB (healthy), CD (interictal), and E (ictal) classes respectively. Our model obtains also a macro F1-score accounting for 96.90% as shown in Table 10.12. In terms of case II, as one can observe from Table 10.8, our model attains an Accuracy and F1-score accounting for 96.50% and 96.31% respectively. This case is pertinent to epilepsy diagnosis based on the presence of seizure activity only. In case III, 98.75% accuracy and 98.77% F1-score are obtained as shown in Table 10.9, indicating that the proposed multimodal model can discriminate healthy and interictal cases very well. With regards to case IV, our model attains 97.20% accuracy and 97.65% F1-score as seen in Table 10.10. In case V, looking at Table 10.11, our model attains an accuracy score equal to 95.33%. F1-scores accounting for 94.05%, 94.78%, and 97.31% are obtained for the classes A, C, and E respectively. Also, one can observe from Table 10.12 that a macro F1-score accounting for 95.38% is achieved by our model.

One can observe from Table 10.13 that our model outperforms 15 research initiatives in accuracy by 0.50-17.00% for case I. For case II, as one can observe from Table 10.14, our model surpasses state-of-the-art approaches by a margin of 0.50-3.00% in accuracy. Table 10.15 provides the results for case III and shows that our introduced model improves the accuracy score by 1.05-12.75%. Results reported in Table 10.16 for case IV indicate that the proposed model surpasses the state-of-the-art approaches by 3.40-8.80% in terms of accuracy. Finally, with regards to case V, one can observe from Table 10.17 that the introduced architecture presents a surge in accuracy outperforming existing research works by a margin of 0.33-5.33%. Overall, our introduced approach yields a better accuracy in all the cases compared to the other methods proposed.

10.7 Ablation Study

In this section, we run a series of ablation experiments to explore the effectiveness and robustness of the introduced architecture described in Section 10.4. Regarding the cases II, III, and IV, the results of the ablation studies are reported in Table 10.19. Regarding case I and case V, the results of the ablation studies are reported in Tables 10.18 and 10.20 respectively.

First, we explore the effectiveness of the gated multimodal unit. Specifically, we remove the gated multimodal unit and concatenate the representations h^t and h^v. The resulting

Table 10.7: Performance of the proposed multimodal model via cross-validation (AB - CD - E). Reported values are mean ± standard deviation.

Model	Evaluation metrics									
	Precision			Recall			F1-score			Accuracy
	AB	CD	E	AB	CD	E	AB	CD	E	
Case I (AB - CD - E)										
Proposed Model	97.14	97.16	97.18	97.99	96.49	96.00	97.52	96.77	96.41	97.00
	±3.10	±3.75	±4.31	±2.45	±2.29	±6.63	±1.91	±1.88	±4.09	±1.84

Table 10.8: Performance of the proposed multimodal model via cross-validation (A - E). Reported values are mean ± standard deviation.

	Evaluation metrics			
Architecture	**Precision**	**Recall**	**F1-score**	**Accuracy**
Case II (A - E)				
Proposed	99.00	94.00	96.31	96.50
Model	±2.99	±6.63	±4.13	±3.91

Table 10.9: Performance of the proposed multimodal model via cross-validation (AB - CD). Reported values are mean ± standard deviation.

	Evaluation metrics			
Architecture	**Precision**	**Recall**	**F1-score**	**Accuracy**
Case III (AB - CD)				
Proposed	98.57	99.00	98.77	98.75
Model	±3.05	±2.00	±2.26	±2.30

Table 10.10: Performance of the proposed multimodal model via cross-validation (AB - CDE). Reported values are mean ± standard deviation.

	Evaluation metrics			
Architecture	**Precision**	**Recall**	**F1-score**	**Accuracy**
Case IV (AB - CDE)				
Proposed	98.03	97.33	97.65	97.20
Model	±2.13	±2.91	±1.88	±2.22

Table 10.11: Performance of the proposed multimodal model via cross-validation (A - C - E). Reported values are mean ± standard deviation.

Model	Evaluation metrics									
	Precision			**Recall**			**F1-score**			**Accuracy**
	A	C	E	A	C	E	A	C	E	
Case V (A - C - E)										
Proposed Model	89.99	98.00	1.00	99.00	92.00	95.00	94.05	94.78	97.31	95.33
	±8.30	±3.99	±0.00	±2.99	±5.99	±6.71	±4.87	±4.03	±3.66	±3.71

Table 10.12: Macro Precision, Recall, and F1-score for Cases I (AB - CD - E) and V (A - C - E) obtained by the proposed multimodal model. Reported values are mean ± standard deviation.

Model	Evaluation metrics		
	M. Precision	**M. Recall**	**M. F1-score**
Case I (AB - CD - E)			
Proposed	97.16	96.83	96.90
Model	±1.73	±2.52	±2.13
Case V (A - C - E)			
Proposed	95.99	95.33	95.38
Model	±3.04	±3.71	±3.65

Table 10.13: Performance comparison among proposed multimodal model and state-of-the-art approaches (AB - CD - E). Reported values are mean ± standard deviation. Best results are in bold.

	Evaluation metric
Architecture	**Accuracy**
State-of-the-art approaches	
Novel RF [107]	96.70
EMD, higher order moments, ANN [108]	80.00
BiLSTM [99]	88.00
DWT + Kmeans + MLPNN [98]	95.60
CNN [109]	96.97
Random Forest [110]	87.00
Matrix Determinant and MLP [111]	96.50
EMD and SVM [112]	93.00
dual-tree complex wavelet transform domain [113]	96.28
statistical dual-tree complex wavelet transform domain [114]	83.50
ANN, hierarchical multi-class SVM with new kernel [115]	95.00
Random Forest, wavelets [116]	95.84
CNN [117]	88.67
OPF [118]	89.20
Symlets wavelets, statistical mean energy std and PCA, GBM-GSO, RF, SVM [119]	96.50
Proposed Architecture	
	97.00 ±1.84

concatenated vector is passed to a dropout layer with a rate of 0.5 followed by a dense layer (with either two or three units), which gives the final prediction. In terms of the Case I (AB - CD - E), one can observe from Tables 10.18 and 10.7 that the removal of the gated multimodal unit leads to a decrease of the accuracy score by 0.80%. In terms of the Case II (A - E), one can observe from Tables 10.19 and 10.8 that the removal of the gated multimodal unit leads to a decrease of the accuracy and F1-score by 1.50% and 1.82% respectively. With regards to the Case III (AB - CD), one can observe from Tables 10.19 and 10.9 a decrease in F1-score and Accuracy by 1.38% and 1.25% respectively. With regards to the Case IV (AB - CDE), one can observe from Tables 10.19 and 10.10 a decrease in F1-score and Accuracy by 1.69% and 1.80% respectively. With regards to the

Table 10.14: Performance comparison among proposed multimodal model and state-of-the-art approaches (A - E). Reported values are mean ± standard deviation. Best results are in bold.

	Evaluation metric
Architecture	**Accuracy**
State-of-the-art approaches	
Relative Wavelet Energy [378]	95.20
Permutation entropy, SVM [379]	93.50
stacked sparse autoencoders [380]	95.50
Cross-correlation aided SVM classifier [381]	95.96
Permutation entropy - SVM classifier [382]	93.55
ME [97]	94.50
multiwavelet transform based approximate entropy and ANN [383]	96.00
Proposed Architecture	
	96.50 ±3.91

Table 10.15: Performance comparison among proposed multimodal model and state-of-the-art approaches (AB - CD). Reported values are mean ± standard deviation. Best results are in bold.

	Evaluation metric
Architecture	**Accuracy**
State-of-the-art approaches	
ATFFWT and FD, LS-SVM [384]	92.50
Novel RF [107]	93.20
Random Forest [110]	86.00
novel signal modeling [385]	97.70
Symlets wavelets, statistical mean energy std and PCA, GBM-GSO, RF, SVM [119]	93.20
MFDFA features + SVM [386]	96.25
Proposed Architecture	
	98.75 ±2.30

Case V (A - C - E), one can observe from Tables 10.20 and 10.11 a decrease in Accuracy by 4.00%.

Next, we conduct the ablation studies to explore the effects of the part of the architecture corresponding to the image modality. To facilitate this, we remove both the

Table 10.16: Performance comparison among proposed multimodal model and state-of-the-art approaches (AB - CDE). Reported values are mean ± standard deviation. Best results are in bold.

	Evaluation metric
Architecture	**Accuracy**
State-of-the-art approaches	
Random Forest [110]	88.40
APN [387]	93.80
Alpha band (Blackman window) [388]	92.00
Proposed Architecture	
	97.20 ±2.22

Table 10.17: Performance comparison among proposed multimodal model and state-of-the-art approaches (A - C - E). Reported values are mean ± standard deviation. Best results are in bold.

	Evaluation metric
Architecture	**Accuracy**
State-of-the-art approaches	
LSTM [101]	94.81
Matrix Determinant and MLP [111]	94.75
DWT and neural network [389]	95.00
DWT and ensemble classifier [390]	90.00
CNN [391]	90.10
Proposed Architecture	
	95.33 ±3.71

image representation part and the gated multimodal unit and experiment with detecting epileptic seizures by using only the two branches of the CNN layers. In terms of the Case I (AB - CD - E), one can observe from Tables 10.18 and 10.7 a decrease of the accuracy score by 1.80%. In terms of the Case II (A - E), one can observe from Tables 10.19 and 10.8 a decrease of the accuracy score and F1-score by 1.01% and 1.29% respectively. With regards to the Case III (AB - CD), one can observe from Tables 10.19 and 10.9 a decrease in F1-score and Accuracy by 2.76% and 2.50% respectively. With regards to the case IV (AB - CDE), one can observe from Tables 10.19 and 10.10 a decrease in F1-score and Accuracy by 3.48% and 3.60% respectively. With regards to the case V (A - C - E), one can observe from Tables 10.20 and 10.11 a decrease in Accuracy by 2.00%.

Then, we investigate the efficacy of the branch of the CNN architecture with the small filter. To do this, we remove this branch and the EEG signal is passed only through the branch having the larger kernel size. In terms of the Case I (AB - CD - E), one can observe from Tables 10.18 and 10.7 a decrease of the accuracy score by 2.20%. In terms of the Case II (A - E), one can observe from Tables 10.19 and 10.8 a decrease of the accuracy and F1-score by 1.50% and 1.94% respectively. With regards to the case III (AB - CD), one can observe from Tables 10.19 and 10.9 a decrease in F1-score and Accuracy by 1.85% and 1.75% respectively. With regards to the Case IV (AB - CDE), one can observe from Tables 10.19 and 10.10 a decrease in F1-score and Accuracy by 0.42% and 0.40% respectively. With regards to the case V (A - C - E), one can observe from Tables 10.20 and 10.11 a decrease in Accuracy by 3.00%.

Finally, we explore the effectiveness of the branch of the CNN architecture with the large filter. To do this, we remove this branch and the EEG signal is passed only through the branch having the small kernel size. In terms of the Case I (AB - CD - E), one can observe from Tables 10.18 and 10.7 a decrease of the accuracy score by 2.00%. In terms of the case II (A - E), one can observe from Tables 10.19 and 10.8 a decrease of the accuracy and F1-score by 1.01% and 1.22% respectively. With regards to the Case III (AB - CD), one can observe from Tables 10.19 and 10.9 a decrease in F1-score and accuracy by 2.41% and 2.26% respectively. With regards to the Case IV (AB - CDE), one can observe from Tables 10.19 and 10.10 a decrease in F1-score and Accuracy by 2.08% and 2.40% respectively. With regards to the Case V (A - C - E), one can observe from Tables 10.20 and 10.11 a decrease in Accuracy by 1.33%.

10.8 Discussion

The early diagnosis of epilepsy is very important, since people receiving treatment are able to live seizure-free for their entire life. Although several research works have been proposed for detecting and predicting epilepsy, there are still significant limitations that need to be addressed. The main limitation is pertinent to the exhaustive and tedious procedure of feature extraction. Specifically, most research works extract features from EEG signals both from time and frequency domain and train shallow machine learning classifiers. Due to the fact that the feature extraction demands a lot of expertise, there is a probability that someone will not extract the most representative set of features for each dataset. Motivated by this limitation, in this paper we aim to automate the process of feature extraction by utilizing two branches of CNNs with different kernel sizes. Concurrently, we employ pretrained models on the computer vision domain to extract vector representations from db-scaled (after having computed the absolute values) spectrograms, their delta, and delta-delta. Finally, we propose a gated multimodal unit receiving as input the two different modalities and trying to suppress the irrelevant information. From the results obtained in this study, we found that:

- **Finding 1:** We applied STFT to the raw EEG signals and constructed an image consisting of three channels, i.e., db-scaled (after having computed the absolute val-

Table 10.18: Ablation Study (AB - CD - E). Reported values are mean ± standard deviation.

Model	Evaluation metrics									
	Precision			Recall			F1-score			Accuracy
	AB	CD	E	AB	CD	E	AB	CD	E	
Case I (AB - CD - E)										
Concatenation	95.86	98.09	94.45	99.00	94.50	94.00	97.34	96.15	93.98	96.20
	±4.17	±3.11	±6.09	±2.00	±4.72	±6.63	±2.23	±2.43	±4.34	±2.27
Predicting from	97.44	92.90	96.51	93.00	95.50	99.00	95.08	94.06	97.61	95.20
EEG signals	±3.40	±5.16	±5.67	±5.57	±6.10	±2.99	±3.75	±4.62	±3.14	±3.59
Predicting without	93.33	96.68	97.09	97.50	94.50	90.00	95.11	95.41	92.86	94.80
the left branch of CNNs	±8.31	±4.24	±4.45	±4.03	±5.68	±11.83	±5.05	±3.31	±7.19	±4.12
Predicting without	94.56	97.05	93.65	97.99	94.50	90.00	96.14	95.63	91.31	95.00
the right branch of CNNs	±4.93	±2.41	±8.88	±2.45	±5.22	±10.00	±2.32	±2.44	±6.95	±2.41

Table 10.19: Ablation Study. Cases II, III & IV. Reported values are mean ± standard deviation.

	Evaluation metrics			
Architecture	**Precision**	**Recall**	**F1-score**	**Accuracy**
Case II (A - E)				
Concatenation	1.00 ±0.00	90.00 ±8.94	94.49 ±5.20	95.00 ±4.47
Predicting from EEG signals	1.00 ±0.00	90.99 ±9.43	95.02 ±5.45	95.49 ±4.72
Predicting without the left branch of CNNs	1.00 ±0.00	90.00 ±10.95	94.37 ±6.45	95.00 ±5.48
Predicting without the right branch of CNNs	1.00 ±0.00	90.99 ±8.31	95.09 ±4.61	95.49 ±4.15
Case III (AB - CD)				
Concatenation	99.52 ±1.43	95.50 ±4.72	97.39 ±2.40	97.50 ±2.24
Predicting from EEG signals	97.74 ±4.06	95.00 ±8.94	96.01 ±5.04	96.25 ±4.37
Predicting without the left branch of CNNs	98.47 ±2.33	95.50 ±4.15	96.92 ±2.76	97.00 ±2.69
Predicting without the right branch of CNNs	97.99 ±2.46	95.00 ±5.92	96.36 ±3.46	96.49 ±3.20
Case IV (AB - CDE)				
Concatenation	99.38 ±1.88	93.00 ±5.67	95.96 ±2.90	95.40 ±3.10
Predicting from EEG signals	99.01 ±2.12	90.33 ±9.24	94.17 ±5.20	93.60 ±5.35
Predicting without the left branch of CNNs	99.66 ±0.99	95.00 ±3.73	97.23 ±1.97	96.80 ±2.23
Predicting without the right branch of CNNs	98.01 ±2.98	93.33 ±2.58	95.57 ±1.84	94.80 ±2.23

ues) spectrogram, delta, and delta-delta. Several pretrained models were exploited, including ResNet18, ResNet50, AlexNet, VGG16, DenseNet201, EfficientNet, etc. Results showed that EfficientNet-B7 was the best performing model for the five cases considered for classification.

- **Finding 2:** We introduced a multimodal deep neural network. Results indicated that the proposed approach achieved comparable performance to the existing research initiatives without the exhaustive procedure of feature extraction.

Table 10.20: Ablation Study (A - C - E). Reported values are mean ± standard deviation.

	Evaluation metrics									
Model	**Precision**			**Recall**			**F1-score**			**Accuracy**
	A	C	E	A	C	E	A	C	E	
Case V (A - C - E)										
Concatenation	81.64	98.18	1.00	99.00	94.00	81.00	89.13	95.75	88.60	91.33
	±9.45	±5.45	±0.00	±2.99	±9.17	±15.78	±5.42	±6.25	±10.38	±4.52
Predicting from	97.97	85.44	1.00	84.00	98.00	98.00	89.95	90.96	98.95	93.33
EEG signals	±4.07	±8.68	±0.00	±11.14	±3.99	±3.99	±6.60	±4.65	±2.11	±3.65
Predicting without	85.16	95.18	1.00	98.00	95.00	84.00	90.89	94.94	91.10	92.33
the left branch of CNNs	±7.54	±6.58	±0.00	±3.99	±8.06	±8.00	±4.30	±6.42	±4.69	±3.96
Predicting without	89.55	95.61	99.09	97.00	92.00	93.00	92.83	93.38	95.72	94.00
the right branch of CNNs	±6.34	±7.06	±2.73	±6.40	±7.48	±7.81	±3.90	±4.49	±4.12	±2.91

- **Finding 3:** We ran a series of ablation experiments and explored the effectiveness of each component included in the proposed architecture. Results suggested that the removal of some components of the proposed model led to a decrease in the evaluation results.

10.9 Summary

In this paper, we introduced both unimodal and multimodal approaches for classifying healthy, interictal, and ictal cases. Regarding the unimodal approaches, we applied the STFT to the EEG data and created an image for each EEG signal consisting of db-scaled (after having computed the absolute values) spectrogram, delta, and delta-delta. We passed each image through pretrained models and showed that EfficientNet-B7 outperformed all the models for all the cases considered for classification achieving accuracy scores ranging from 93.00% to 97.50%. Next, we introduced a multimodal deep neural network. First, each EEG signal was passed through two branches of CNNs with different kernel sizes, i.e., large and small, aiming to automate the process of feature extraction and capture both the temporal (i.e., when certain of EEG patterns appear) and frequency information (i.e., frequency components). Similarly to the unimodal approach, we created an image and passed it through the EfficientNet-B7 pretrained model. Finally, a gated multimodal unit was incorporated in the top of the architecture for controlling the importance of each modality towards the final classification. Extensive experiments conducted on the dataset provided by the University of Bonn indicated that the introduced architecture obtained comparable performance to the existing research initiatives with an accuracy ranging from 95.33% to 98.75% for the five different cases considered for the classification.

Chapter 11

Conclusions and Future Work

11.1 Conclusions

We investigated the latest machine learning methods for *(i)* identifying depression through posts in social media and spontaneous speech, *(ii)* detecting AD patients and predicting their MMSE scores from spontaneous speech, and *(iii)* identifying epileptic patients through single-channel EEG signals. This thesis attempted to find answers to a number of research questions which were listed in Chapter 2.

- **Do transformer-based networks, i.e., BERT, ALBERT, etc. achieve better performance than traditional techniques, i.e., LSTMs, CNNs, etc.?** In terms of this research question, we exploited and fine-tuned transformer-based networks, including BERT, BioBERT, BioClinicalBERT, ConvBERT, RoBERTa, ALBERT, and XLNet (Chapter 5.4.1.1).

- **Can we provide explanations, which will show how our models reach their decisions? Especially in health-related tasks, it is very important for a clinician to be informed why the deep neural network classified a person as an AD patient or a non-AD one. At the same time, according to the European Union General Data Protection Regulation (GDPR) [172] each person has the right to the explanation. Also, can we propose interpretable models, which will achieve comparable performance to existing research initiatives?** Considering this research question, we introduced an interpretable deep neural network, which incorporates a co-attention mechanism for detecting AD patients (Chapter 5.4.1.2). Also, we exploited LIME to explain the predictions made by our best performing model and showed which pos-tags are used by AD patients mainly (Chapter 5.7.5).

- **Can we propose multi-task learning models, consisting of primary and auxiliary tasks, to explore if the axiliary tasks help the primary one in improving its performance?** With respect to this research question, we presented two approaches. Specifically, we presented a method which investigates if the estimation of age, gender, and education level helps the depression identification task (Chapter 4.3). We also introduced two deep neural networks, which detect AD patients and predict the MMSE scores at the same time (Chapter 5.4.2).

- **How can we combine the representation vectors of the different modalities (multimodal approaches) effectively?** Regarding this research question, we introduced several methods for *(i)* combining effectively the modalities of speech and transcripts without losing information, and *(ii)* in terms of the task of epilepsy (Chapter 4, 6-9).

- **Instead of creating fixed deep neural networks, can we create automatically architectures which will perform best for our specific task?** In terms of this research question, we incorporated a NAS approach, called DARTS, into a deep neural network, which is capable of generating a CNN architecture automatically. This CNN architecture receives as input an image of log-Mel spectrogram (of the input speech signal), its delta, and delta-delta, and extracts a visual representation. This research question is answered in Chapter 9.

- **How can we improve self-attention networks through capturing the richness of context?** We exploit several strategies for contextualization, including global context, deep context, and deep-global context. This research question is addressed in Chapter 8.

- **How can we prevent deep learning models from becoming too overconfident?** In terms of this research question, we used label smoothing and evaluated our proposed deep learning models in terms of both the performance and the calibration metrics. This research question is answered in Chapters 4,8.

In the following paragraphs, we present our detailed conclusions per chapter.

In Chapter 4, we presented two methods for detecting depression by utilizing social media posts and spontaneous speech. Firstly, we introduced a method for identifying depression in social media text by injecting linguistic information into transformer-based models. Also, it is the first study exploiting label smoothing, in order to ensure that our model is calibrated. We evaluated our proposed methods on two publicly available datasets, which include two depression detection datasets (binary classification and multiclass classification - severity of depression). Findings suggested that transformer-based networks combined with linguistic information lead to performance improvement in comparison with transformer-based networks. Also, applying label smoothing yielded both to the performance improvement and better calibration of the proposed models. Specifically, in terms of the Depression_Mixed dataset, we found that the injection of top2vec features into BERT and MentalBERT models along with label smoothing obtained the highest F1-score and Accuracy. With regards to the Depression_Severity dataset, findings showed that the injection of NRC features into the BERT model and the integration of features derived by LDA topics, namely GOSS features, into the MentalBERT model yielded the highest weighted F1-scores. We also conducted a linguistic analysis and showed that depressive posts are full of sadness, anxiety, and negative tone. Secondly, we presented the first study utilizing a cross-attention scaling layer and multimodal fusion methods in a single neural network for detecting depression from spontaneous speech in the Italian language through speech and automatic transcripts. This is also the first study experimenting with a multitask learning setting to investigate if the prediction of gender, age, and education level as auxiliary tasks aid the depression detection task (primary task) in increasing its performance. Results showed that our introduced approach improves competitive baselines in Accuracy by 1.21-21.99% and in F1-score by 1.32-22.23%. Results also showed that the introduced single-task learning model outperforms the multitask learning ones. Finally, we performed an ablation study, where we removed several parts of the proposed architecture and observe differences in performance. Findings showed degradation in performance in terms of Accuracy by 1.29-3.19%.

In Chapter 5, we introduced both single-task and multi-task learning models. Regarding single-task learning models, we employed several transformer-based networks and compared their performances. Results showed that BERT achieved the highest classification performance with accuracy accounting for 87.50%. Concurrently, we introduced siamese networks coupled with a co-attention mechanism which can detect AD patients with an accuracy up to 83.75%. In terms of the multi-task learning setting, it consisted of two tasks, the primary and the auxiliary one. The primary task was the identification of dementia (binary classification), whereas the auxiliary task was the categorization of the severity of dementia into one of the four categories -healthy, mild/moderate/severe dementia- (multiclass classification). Specifically, we proposed two multi-task learning models. Results showed that our model achieves competitive results in the MTL framework reaching accuracy up to 86.25% on the detection of AD patients. Next, we performed an in-depth linguistic analysis, in order to understand better the differences in language between AD and non-AD patients. Finally, we employed LIME, in order to shed light on how our best performing model works. Findings suggest that AD patients tend to use personal pronouns, interjection, adverbs, verbs in the past tense, and the token ”and” at the beginning of utterances in a high frequency. On the contrary, healthy people use verbs in present participle or gerund, nouns as well as determiners.

In Chapter 6, we proposed methods to differentiate AD from non-AD patients using either only speech or both speech and transcripts. Regarding the models using only speech, we exploited several pretrained models used extensively in the computer vision domain, with the Vision Transformer achieving the highest F1-score and accuracy accounting for 69.76% and 65.00% respectively. Next, we employed three neural network models in which we combined speech and transcripts. We exploited the Gated Multimodal Unit, in order to control the influence of each modality towards the final classification. In addition, we experimented with crossmodal interactions, where we used the crossmodal attention. Results showed that crossmodal attention can enhance the performance of competitive multimodal approaches and surpass state-of-the-art approaches. More specifically, models incorporating the crossmodal attention yielded accuracy equal to 88.83% on the ADReSS Challenge test set.

In Chapter 7, we introduced three novel multimodal neural networks for detecting dementia (AD classification task) and predicting the MMSE scores (MMSE regression task) from spontaneous speech. First, we proposed a model consisting of BERT, ViT, and a co-attention mechanism at the top of the proposed architecture, which is capable of attending to both the words and the image patches simultaneously. Results indicated that the proposed model achieved an accuracy of 87.50% in the AD classification task outperforming all the research works proposing multimodal approaches except one. Regarding the MMSE regression task, our proposed architecture achieved an RMSE score equal to 4.20. Secondly, we introduced a deep learning architecture, where we injected information from the visual and acoustic modalities along with the textual one into a BERT model and used an attention gate mechanism to control the importance of each modality. Results for the AD classification task suggested that the injection of both the acoustic and visual modalities enhanced the performance of the models achieved when using only either the acoustic or the visual modality along with the textual one. Finally, we introduced a transformer-based network, where we concatenated the representations obtained via BERT and ViT and passed the representation through a self-attention mechanism incorporating a novel gating mechanism. Findings showed that this introduced model was the best performing one on the ADReSS Challenge test set reaching Accuracy and F1-score up to 90.00% and 89.94% respectively. In terms of the MMSE regression task, our best performing model obtained an RMSE score of 3.61 improving the state-of-the-art RMSE scores for the regression task of the ADReSS Challenge by 0.13-3.06.

In Chapter 8, we introduced some new approaches to detect AD patients from speech and transcripts, which capture the inter- and intra-modal interactions, enhance the conventional self-attention mechanism with contextual information, and deal with the problem of creating overconfident models by applying label smoothing. Our proposed architectures consist of BERT, DeiT, self-attention mechanism incorporating a gating model, context-based self-attention, optimal transport domain adaptation methods, and one new method for fusing the self and cross-attention features in the task of dementia detection from speech data. Furthermore, we designed extensive ablation experiments to explore the effectiveness of the components of the proposed architecture. Extensive experiments conducted on the ADReSS and ADReSSo Challenge datasets demonstrate the efficacy of the proposed architectures reaching Accuracy up to 91.25% and 83.94% respectively. Also, findings suggested that the label smoothing contributes to both the performance improvement and calibration of our model.

In Chapter 9, we presented the first study, which exploits Neural Architecture Search methods and fusion methods based on Tucker Decomposition, Factorized Bilinear Pooling, and block-term tensor decomposition, in the task of dementia detection. Specifically, we proposed an end-to-end trainable multimodal model, which combines an automatically discovered CNN architecture obtained from the NAS algorithm as well as a language model for processing the text information. We integrated the two modalities using a variety of fusion methods. Our approach exhibited comparable performance with the state-of-the-art baselines.

In Chapter 10, we introduced both unimodal and multimodal approaches for classifying healthy, interictal, and ictal cases. Regarding the unimodal approaches, we applied the STFT to the EEG data and created an image for each EEG signal consisting of db-scaled (after having computed the absolute values) spectrogram, delta, and delta-delta. We passed each image through pretrained models and showed that EfficientNet-B7 outperformed all the models for all the cases considered for classification achieving accuracy scores ranging from 93.00% to 97.50%. Next, we introduced a multimodal deep neural network. First, each EEG signal was passed through two branches of CNNs with different kernel sizes, i.e., large and small, aiming to automate the process of feature extraction and capture both the temporal (i.e., when certain of EEG patterns appear) and frequency information (i.e., frequency components). Similarly to the unimodal approach, we created an image and passed it through the EfficientNet-B7 pretrained model. Finally, a gated multimodal unit was incorporated in the top of the architecture for controlling the importance of each modality towards the final classification. Extensive experiments conducted on the dataset provided by the University of Bonn indicated that the introduced architecture obtained comparable performance to the existing research initiatives with an accuracy ranging from 95.33% to 98.75% for the five different cases considered for the classification.

11.2 Limitations

The studies in this thesis include the following list of limitations:

- **Lack of Explainability Methods in terms of the Multimodal Approaches.** The multimodal approaches are not accompanied with explainable AI algorithms. Therefore, the user is not capable of understanding the reasons of correct and incorrect predictions.

- **Lack of Longitudinal Tracking.** The datasets used for the detection of depression and Alzheimer's dementia do not allow for investigating how these brain disorders progress over time, since each participant is recorded only once.

- **Hyperparameter Tuning.** The studies in this thesis do not include a hyperparameter tuning procedure due to limited access to GPU resources. It is known that hyperparameter tuning leads to a performance improvement.

- **Need for Labelled Data.** The studies in this thesis require access to labelled datasets. However, collecting labelled datasets in the healthcare domain is often a difficult task due to privacy reasons. On the contrary, self-supervised learning approaches have been developed, which address the need of labels' scarcity.

11.3 Future Work

- **Interpretable multimodal deep learning models.** The clinician must be informed why the ML algorithm reached a specific decision. For this reason, we aim to apply post-hoc explainability techniques for rendering the proposed multimodal approaches explainable. Specifically, GRAD-CAM and Integrated Gradients are two explainability techniques which can be applied for explaining the results of any ML algorithm.

- **Labels' Scarcity.** Collecting large labelled datasets for training AI/ML algorithms is crucial. For this reason, we plan to apply self-supervised learning approaches in the future to address the need of large labelled datasets.

- **Detection of MCI.** In the future, we aim to apply our introduced approaches in the VAS dataset proposed in [120, 121]. This dataset, includes AD patients, non-AD ones, and Mild Cognitive Impairment (MCI) subjects. Detecting MCI subjects is challenging and has been proved to be crucial. Specifically, the progression of the disease can be delayed substantially by detecting timely subjects in the MCI condition.

- **Privacy Issues - Federated Learning.** Processing healthcare data entails privacy issues. To be more precise, the majority of existing approaches rely on centralized settings, where data are gathered on a central server. On the contrary, federated learning [392] addresses this issue by distributing the training process to end-user devices.

- **Data Augmentation.** Generative Adversarial Networks (GANs) can also be exploited for creating signals, i.e., speech signals, EEG, and more. Specifically, the deep neural networks can be trained with artificially generated data, while their performance can be tested on real data.

- **Apply our methods in other brain disorders.** Our introduced approaches can be applied to other diseases as well. For instance, research has showed that Parkinson's disease affects speech, thus Parkinson's disease might be detected through speech and transcripts.

- **Use of multi-channel EEG data.** In the future, we aim to use multichannel EEG signals [122, 123].

- **Multilingual Approaches.** We plan to apply our introduced approaches in a multilingual framework. Specifically, we aim to train our models in one language and evaluate the performance of the models in another language. For instance, one could exploit the MADReSS Challenge dataset [124]. One can train models based on English speech data and assess the models' performance on spoken Greek data.

- **Knowledge Distillation.** For addressing the need of creating large models, which entail computational issues, we aim to exploit Knowledge Distillation approaches [125, 126]. In this way, a large neural network is compressed into a smaller and simpler one without sacrificing its performance.

- **Adapters.** In this thesis, we fine-tuned some pretrained models based on transformers. For example, see Chapter 5. However, some information is lost during fine-tuning, since only task-specific data are used for updating the models' parameters. This phenomenon is known as catastrophic forgetting [127]. Therefore, in the future, we plan to use adapters [128, 129].

- **Longitudinal Applications.** Since depression and Alzheimer's dementia evolves over time, it is important to be diagnosed early. Longitudinal disease tracking is of great importance nowadays. For instance, one of the tasks in terms of the ADReSSo Challenge is the cognitive decline (disease progression) inference task, where one can create a model to predict changes in cognitive status over time.

Bibliography

[1] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, “Detection of depression-related posts in reddit social media forum,” *IEEE Access*, vol. 7, pp. 44 883–44 893, 2019.

[2] S. Chandra Guntuku, A. Buffone, K. Jaidka, J. C. Eichstaedt, and L. H. Ungar, “Understanding and measuring psychological stress using social media,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 13, no. 01, pp. 214–225, Jul. 2019. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/3223

[3] J. S. L. Figuerêdo, A. L. L. Maia, and R. T. Calumby, “Early depression detection in social media based on deep learning and underlying emotions,” *Online Social Networks and Media*, vol. 31, p. 100225, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2468696422000283

[4] L. Ansari, S. Ji, Q. Chen, and E. Cambria, “Ensemble hybrid learning methods for automated depression detection,” *IEEE Transactions on Computational Social Systems*, pp. 1–9, 2022.

[5] N. Poerner, U. Waltinger, and H. Schütze, “E-BERT: Efficient-yet-effective entity embeddings for BERT,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 803–818. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.71

[6] N. Kassner and H. Schütze, “Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7811–7818. [Online]. Available: https://aclanthology.org/2020.acl-main.698

[7] N. Peinelt, M. Rei, and M. Liakata, “GiBERT: Enhancing BERT with linguistic information using a lightweight gated injection method,” in *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2322–2336. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.200

[8] R. Wang, D. Tang, N. Duan, Z. Wei, X. Huang, J. Ji, G. Cao, D. Jiang, and M. Zhou, "K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021.* Online: Association for Computational Linguistics, Aug. 2021, pp. 1405–1418. [Online]. Available: https://aclanthology.org/2021.findings-acl.121

[9] J. Lu, D. Batra, D. Parikh, and S. Lee, *ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks.* Red Hook, NY, USA: Curran Associates Inc., 2019.

[10] M. E. Peters, M. Neumann, R. Logan, R. Schwartz, V. Joshi, S. Singh, and N. A. Smith, "Knowledge enhanced contextual word representations," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).* Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 43–54. [Online]. Available: https://aclanthology.org/D19-1005

[11] A. P. Dawid, "The well-calibrated bayesian," *Journal of the American Statistical Association*, vol. 77, no. 379, pp. 605–610, 1982. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1982.10477856

[12] A. H. Murphy and E. S. Epstein, "Verification of probabilistic predictions: A brief review," *Journal of Applied Meteorology and Climatology*, vol. 6, no. 5, pp. 748 – 755, 1967. [Online]. Available: https://journals.ametsoc.org/view/journals/apme/6/5/1520-0450_1967_006_0748_voppab_2_0_co_2.xml

[13] I. Pirina and Ç. Çöltekin, "Identifying depression on Reddit: The effect of training data," in *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task.* Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 9–12. [Online]. Available: https://aclanthology.org/W18-5903

[14] N. Ramirez-Esparza, C. Chung, E. Kacewic, and J. Pennebaker, "The psychology of word use in depression forums in english and in spanish: Testing two text analytic approaches," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 2, no. 1, pp. 102–108, Sep. 2021. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/18623

[15] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, and M. Mitchell, "CLPsych 2015 shared task: Depression and PTSD on Twitter," in *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality.* Denver, Colorado: Association for Computational Linguistics, Jun. 5 2015, pp. 31–39. [Online]. Available: https://aclanthology.org/W15-1204

[16] A. Yates, A. Cohan, and N. Goharian, "Depression and self-harm risk assessment in online forums," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 2968–2978. [Online]. Available: https://aclanthology.org/D17-1322

[17] U. Naseem, A. G. Dunn, J. Kim, and M. Khushi, "Early identification of depression severity levels on reddit using ordinal classification," in *Proceedings of the ACM Web Conference 2022*, ser. WWW '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 2563–2572. [Online]. Available: https://doi.org/10.1145/3485447.3512128

[18] W. Rahman, M. K. Hasan, S. Lee, A. Bagher Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Online: Association for Computational Linguistics, Jul. 2020, pp. 2359–2369. [Online]. Available: https://aclanthology.org/2020.acl-main.214

[19] M. Jin and N. Aletras, "Complaint identification in social media with transformer networks," in *Proceedings of the 28th International Conference on Computational Linguistics.* Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 1765–1771. [Online]. Available: https://aclanthology.org/2020.coling-main.157

[20] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word–emotion association lexicon," *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8640.2012.00460.x

[21] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[22] R. L. Boyd, A. Ashokkumar, S. Seraj, and J. W. Pennebaker, "The development and psychometric properties of liwc-22," *Austin, TX: University of Texas at Austin*, 2022.

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. null, p. 993–1022, mar 2003.

[24] L. Liu, Y. Lu, Y. Luo, R. Zhang, L. Itti, and J. Lu, "Detecting "smart" spammers on social network: A topic model approach," in *Proceedings of the NAACL Student Research Workshop.* San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 45–50. [Online]. Available: https://aclanthology.org/N16-2007

[25] D. Angelov, "Top2vec: Distributed representations of topics," 2020.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[27] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "MentalBERT: Publicly available pretrained language models for mental healthcare," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 7184–7190. [Online]. Available: https://aclanthology.org/2022.lrec-1.778

[28] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html

[30] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[31] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/f1748d6b0fd9d439f71450117eba2725-Paper.pdf

[32] N. Seneviratne and C. Espy-Wilson, "Multimodal Depression Severity Score Prediction Using Articulatory Coordination Features and Hierarchical Attention Based Text Embeddings," in *Proc. Interspeech 2022*, 2022, pp. 3353–3357.

[33] M. R. Morales and R. Levitan, "Speech vs. text: A comparative analysis of features for depression detection systems," in *2016 IEEE SLT*, 2016, pp. 136–143.

[34] S. Guohou, Z. Lina, and Z. Dongsong, "What reveals about depression level? the role of multimodal features at the level of interview questions," *Information & Management*, vol. 57, no. 7, p. 103349, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378720620302871

[35] F. Tao, X. Ge, W. Ma, A. Esposito, and A. Vinciarelli, "Multi-local attention for speech-based depression detection," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[36] Z. Zhang, W. Lin, M. Liu, and M. Mahmoud, "Multimodal deep learning framework for mental disorder recognition," in *IEEE FG 2020*, pp. 344–350.

[37] M. Niu, K. Chen, Q. Chen, and L. Yang, "Hcag: A hierarchical context-aware graph attention model for depression detection," in *ICASSP 2021*, 2021, pp. 4235–4239.

[38] G. Lam, H. Dongyan, and W. Lin, "Context-aware deep learning for multi-modal depression detection," in *ICASSP*, 2019.

[39] J. Ye, Y. Yu, Q. Wang, W. Li, H. Liang, Y. Zheng, and G. Fu, "Multi-modal depression detection based on emotional audio and evaluation text," *Journal of Affective Disorders*, vol. 295, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165032721008958

[40] E. Villatoro-Tello, S. P. Dubagunta, J. Fritsch, G. R. de-la Rosa, P. Motlicek, and M. Magimai-Doss, "Late Fusion of the Available Lexicon and Raw Waveform-Based Acoustic Modeling for Depression and Dementia Recognition," in *Proc. Interspeech 2021*, 2021, pp. 1927–1931.

[41] F. Ceccarelli and M. Mahmoud, "Multimodal temporal machine learning for bipolar disorder and depression recognition," *Pattern Anal. Appl.*, vol. 25, no. 3, p. 493–504, aug 2022. [Online]. Available: https://doi.org/10.1007/s10044-021-01001-y

[42] F. Tao, A. Esposito, and A. Vinciarelli, "The Androids Corpus: A New Publicly Available Benchmark for Speech Based Depression Detection," in *Proc. INTERSPEECH 2023*, 2023, pp. 4149–4153.

[43] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*. JMLR.org, 2023.

[44] A. Krizhevsky, "One weird trick for parallelizing convolutional neural networks," *arXiv preprint arXiv:1404.5997*, 2014.

[45] T. Sachan, N. Pinnaparaju, M. Gupta, and V. Varma, "Scate: shared cross attention transformer encoders for multimodal fake news detection," in *ASONAM '21*. New York, NY, USA: ACM, 2022. [Online]. Available: https://doi.org/10.1145/3487351.3490965

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[47] E. Del Barrio, J. A. Cuesta-Albertos, and C. Matrán, "An optimal transportation approach for assessing almost stochastic order," in *The Mathematics of the Uncertain.* Springer, 2018, pp. 33–44.

[48] R. Dror, S. Shlomov, and R. Reichart, "Deep dominance - how to properly compare deep neural models," in *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, A. Korhonen, D. R. Traum, and L. Màrquez, Eds. Association for Computational Linguistics, 2019, pp. 2773–2785. [Online]. Available: https://doi.org/10.18653/v1/p19-1266

[49] D. Ulmer, C. Hardmeier, and J. Frellsen, "deep-significance-easy and meaningful statistical significance testing in the age of neural networks," *arXiv preprint arXiv:2204.06815*, 2022.

[50] N. Reimers and I. Gurevych, "Why comparing single performance scores does not allow to draw conclusions about machine learning approaches," *arXiv preprint arXiv:1803.09578*, 2018.

[51] World Health Organization, "*Dementia*," Available online at: https://www.who.int/news-room/fact-sheets/detail/dementia, 2021.

[52] A. A. (2021)., "*What Is Dementia? Alzheimer's Disease and Dementia.*" Available online at: https://www.https://www.alz.org/alzheimers-dementia/what-is-dementia, accessed: 2021-07-30.

[53] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge," in *Proceedings of INTERSPEECH 2020*, Shanghai, China, 2020. [Online]. Available: https://arxiv.org/abs/2004.06833

[54] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The Natural History of Alzheimer's Disease: Description of Study Cohort and Accuracy of Diagnosis," *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 06 1994. [Online]. Available: https://doi.org/10.1001/archneur.1994.00540180063015

[55] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[56] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, and M. McDermott, "Publicly available clinical BERT embeddings," in *Proceedings of the 2nd Clinical Natural Language Processing Workshop.* Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 72–78. [Online]. Available: https://aclanthology.org/W19-1909

[57] Z.-H. Jiang, W. Yu, D. Zhou, Y. Chen, J. Feng, and S. Yan, “Convbert: Improving bert with span-based dynamic convolution,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 837–12 848. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/96da2f590cd7246bbde0051047b0d6f7-Paper.pdf

[58] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” 2019.

[59] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “Albert: A lite bert for self-supervised learning of language representations,” *arXiv preprint arXiv:1909.11942*, 2019.

[60] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, “Xlnet: Generalized autoregressive pretraining for language understanding,” in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf

[61] Z. S. Syed, M. S. S. Syed, M. Lech, and E. Pirogova, “Automated recognition of alzheimer’s dementia using bag-of-deep-features and model ensembling,” *IEEE Access*, vol. 9, pp. 88 377–88 390, 2021.

[62] M. Rohanian, J. Hough, and M. Purver, “Multi-Modal Fusion with Gating Using Audio, Lexical and Disfluency Features for Alzheimer’s Dementia Recognition from Spontaneous Speech,” in *Proc. Interspeech 2020*, 2020, pp. 2187–2191.

[63] N. Cummins, Y. Pan, Z. Ren, J. Fritsch, V. S. Nallanthighal, H. Christensen, D. Blackburn, B. W. Schuller, M. Magimai-Doss, H. Strik, and A. Härmä, “A Comparison of Acoustic and Linguistics Methodologies for Alzheimer’s Dementia Recognition,” in *Proc. Interspeech 2020*, 2020, pp. 2182–2186.

[64] Z. Shah, J. Sawalha, M. Tasnim, S.-a. Qi, E. Stroulia, and R. Greiner, “Learning language and acoustic models for identifying alzheimer’s dementia from speech,” *Frontiers in Computer Science*, vol. 3, p. 4, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/fcomp.2021.624659

[65] A. Meghanani, C. S. Anoop, and A. G. Ramakrishnan, “Recognition of alzheimer’s dementia from the transcriptions of spontaneous speech using fasttext and cnn models,” *Frontiers in Computer Science*, vol. 3, p. 7, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/fcomp.2021.624558

[66] M. S. S. Syed, Z. S. Syed, M. Lech, and E. Pirogova, “Automated Screening for Alzheimer’s Dementia Through Spontaneous Speech,” in *Proc. Interspeech 2020*, 2020, pp. 2222–2226.

[67] A. Pompili, T. Rolland, and A. Abad, “The INESC-ID Multi-Modal System for the ADReSS 2020 Challenge,” in *Proc. Interspeech 2020*, 2020, pp. 2202–2206.

[68] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman *et al.*, “Personality, gender, and age in the language of social media: The open-vocabulary approach,” *PloS one*, vol. 8, no. 9, p. e73791, 2013.

[69] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: a practical and powerful approach to multiple testing,” *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[70] A. Almor, D. Kempler, M. C. MacDonald, E. S. Andersen, and L. K. Tyler, “Why do alzheimer patients have difficulty with pronouns? working memory, semantics, and reference in comprehension and production in alzheimer’s disease,” *Brain and language*, vol. 67, no. 3, pp. 202–227, 1999.

[71] C. M. Watson, “An analysis of trouble and repair in the natural conversations of people with dementia of the alzheimer’s type,” *Aphasiology*, vol. 13, no. 3, pp. 195–218, 1999.

[72] L. J. Garcia and Y. Joanette, “Analysis of conversational topic shifts: A multiple case study,” *Brain and language*, vol. 58, no. 1, pp. 92–114, 1997.

[73] R. Bastiaanse, “Why reference to the past is difficult for agrammatic speakers,” *Clinical Linguistics & Phonetics*, vol. 27, no. 4, pp. 244–263, 2013, pMID: 23339396. [Online]. Available: https://doi.org/10.3109/02699206.2012.751626

[74] M. T. Ribeiro, S. Singh, and C. Guestrin, “” why should i trust you?” explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[75] J. Arevalo, T. Solorio, M. Montes-y Gomez, and F. A. González, “Gated multimodal networks,” *Neural Computing and Applications*, pp. 1–20, 2020.

[76] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, “Multimodal transformer for unaligned multimodal language sequences,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 6558–6569. [Online]. Available: https://aclanthology.org/P19-1656

[77] D. Sánchez Villegas and N. Aletras, "Point-of-interest type prediction using text and images," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.* Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 7785–7797. [Online]. Available: https://aclanthology.org/2021.emnlp-main.614

[78] B. Sharma, M. Madhavi, and H. Li, "Leveraging acoustic and linguistic embeddings from pretrained speech and language models for intent classification," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7498–7502.

[79] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 289–297.

[80] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "Defend: Explainable fake news detection," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 395–405. [Online]. Available: https://doi.org/10.1145/3292500.3330935

[81] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 7216–7223, Jul. 2019. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/4706

[82] M. Jin and N. Aletras, "Modeling the severity of complaints in social media," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Online: Association for Computational Linguistics, Jun. 2021, pp. 2264–2274. [Online]. Available: https://aclanthology.org/2021.naacl-main.180

[83] Z. Yu, Y. Cui, J. Yu, D. Tao, and Q. Tian, "Multimodal unified attention networks for vision-and-language interactions," *arXiv preprint arXiv:1908.04107*, 2019.

[84] H. Liu, K. Simonyan, and Y. Yang, "DARTS: Differentiable architecture search," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=S1eYHoC5FX

[85] H. Ben-younes, R. Cadene, M. Cord, and N. Thome, "Mutan: Multimodal tucker fusion for visual question answering," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2631–2639.

[86] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 12, pp. 5947–5959, 2018.

[87] H. Ben-younes, R. Cadene, N. Thome, and M. Cord, "Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 8102–8109, Jul. 2019. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/4818

[88] A. Ablimit, C. Botelho, A. Abad, T. Schultz, and I. Trancoso, "Exploring dementia detection from speech: Cross corpus analysis," in *ICASSP*, 2022.

[89] J. Li, J. Yu, Z. Ye, S. Wong, M. Mak, B. Mak, X. Liu, and H. Meng, "A comparative study of acoustic and linguistic features classification for alzheimer's disease detection," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6423–6427.

[90] R. Pappagari, J. Cho, L. Moro-Velázquez, and N. Dehak, "Using State of the Art Speaker Recognition and Natural Language Processing Technologies to Detect Alzheimer's Disease and Assess its Severity," in *Proc. Interspeech 2020*, 2020, pp. 2177–2181.

[91] Y. Zhu, X. Liang, J. A. Batsis, and R. M. Roth, "Exploring deep transfer learning techniques for alzheimer's dementia detection," *Frontiers in Computer Science*, vol. 3, p. 22, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/fcomp.2021.624683

[92] Z. S. Syed, M. S. S. Syed, M. Lech, and E. Pirogova, "Automated recognition of alzheimer's dementia using bag-of-deep-features and model ensembling," *IEEE Access*, vol. 9, pp. 88 377–88 390, 2021.

[93] K. Chlasta and K. Wołk, "Towards computer-based automated screening of dementia through spontaneous speech," *Frontiers in Psychology*, vol. 11, p. 4091, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2020.623237

[94] A. Meghanani, A. C. S., and A. G. Ramakrishnan, "An exploration of log-mel spectrogram and mfcc features for alzheimer's dementia recognition from spontaneous speech," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 670–677.

[95] M. Martinc and S. Pollak, "Tackling the ADReSS Challenge: A Multimodal Approach to the Automated Recognition of Alzheimer's Dementia," in *Proc. Interspeech 2020*, 2020, pp. 2157–2161.

[96] World Health Organization, "*Epilepsy*," Available online at: https://www.who.int/news-room/fact-sheets/detail/epilepsy, 2019.

[97] A. Subasi, "Eeg signal classification using wavelet feature extraction and a mixture of expert model," *Expert Systems with Applications*, vol. 32, no. 4, pp. 1084–1093, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417406000844

[98] U. Orhan, M. Hekim, and M. Ozer, "Eeg signals classification using the k-means clustering and a multilayer perceptron neural network model," *Expert Systems with Applications*, vol. 38, no. 10, pp. 13 475–13 481, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417411006762

[99] E. Tuncer and E. Doğru Bolat, "Classification of epileptic seizures from electroencephalogram (eeg) data using bidirectional short-term memory (bi-lstm) network architecture," *Biomedical Signal Processing and Control*, vol. 73, p. 103462, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809421010594

[100] K. M. Tsiouris, V. C. Pezoulas, M. Zervakis, S. Konitsiotis, D. D. Koutsouris, and D. I. Fotiadis, "A long short-term memory deep learning network for the prediction of epileptic seizures using eeg signals," *Computers in biology and medicine*, vol. 99, pp. 24–37, 2018.

[101] M. U. Abbasi, A. Rashad, A. Basalamah, and M. Tariq, "Detection of epilepsy seizures in neo-natal eeg using lstm architecture," *IEEE Access*, vol. 7, pp. 179 074–179 085, 2019.

[102] I. Ullah, M. Hussain, E. ul Haq Qazi, and H. Aboalsamh, "An automated system for epilepsy detection using eeg brain signals based on deep learning approach," *Expert Systems with Applications*, vol. 107, pp. 61–71, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417418302513

[103] R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, and C. E. Elger, "Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state," *Physical Review E*, vol. 64, no. 6, p. 061907, 2001.

[104] A. Supratak, H. Dong, C. Wu, and Y. Guo, "Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 11, pp. 1998–2008, 2017.

[105] E. Eldele, Z. Chen, C. Liu, M. Wu, C.-K. Kwoh, X. Li, and C. Guan, "An attention-based deep learning approach for sleep stage classification with single-channel eeg," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 809–818, 2021.

[106] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 448–456. [Online]. Available: https://proceedings.mlr.press/v37/ioffe15.html

[107] X. Wang, G. Gong, N. Li, and S. Qiu, "Detection analysis of epileptic eeg using a novel random forest model combined with grid search optimization," *Frontiers in Human Neuroscience*, vol. 13, p. 52, 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnhum.2019.00052

[108] S. M. S. Alam and M. I. H. Bhuiyan, "Detection of seizure and epilepsy using higher order statistics in the emd domain," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 2, pp. 312–318, 2013.

[109] W. Zhao, W. Zhao, W. Wang, X. Jiang, X. Zhang, Y. Peng, B. Zhang, and G. Zhang, "A novel deep neural network for robust detection of seizures using eeg signals," *Computational and mathematical methods in medicine*, vol. 2020, 2020.

[110] M. Sameer and B. Gupta, "Detection of epileptical seizures based on alpha band statistical features," *Wireless Personal Communications*, vol. 115, no. 2, pp. 909–925, 2020.

[111] S. Raghu, N. Sriraam, A. S. Hegde, and P. L. Kubben, "A novel approach for classification of epileptic seizures using matrix determinant," *Expert Systems with Applications*, vol. 127, pp. 323–341, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417419301836

[112] F. Riaz, A. Hassan, S. Rehman, I. K. Niazi, and K. Dremstrup, "Emd-based temporal and spectral features for the classification of eeg signals using supervised learning," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 1, pp. 28–35, 2016.

[113] A. B. Das, M. I. H. Bhuiyan, and S. S. Alam, "Classification of eeg signals using normal inverse gaussian parameters in the dual-tree complex wavelet transform domain for seizure detection," *Signal, Image and Video Processing*, vol. 10, no. 2, pp. 259–266, 2016.

[114] A. B. Das, M. I. H. Bhuiyan, and S. M. S. Alam, "A statistical method for automatic detection of seizure and epilepsy in the dual tree complex wavelet transform domain," in *2014 International Conference on Informatics, Electronics & Vision (ICIEV)*, 2014, pp. 1–6.

[115] A. M. Murugavel and S. Ramakrishnan, "Hierarchical multi-class svm with elm kernel for epileptic eeg signal classification," *Medical & biological engineering & computing*, vol. 54, no. 1, pp. 149–161, 2016.

[116] K. D. Tzimourta, A. T. Tzallas, N. Giannakeas, L. G. Astrakas, D. G. Tsalikakis, P. Angelidis, and M. G. Tsipouras, "A robust methodology for classification of epileptic seizures in eeg signals," *Health and Technology*, vol. 9, no. 2, pp. 135–142, 2019.

[117] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, and H. Adeli, "Deep convolutional neural network for the automated detection and diagnosis of seizure using eeg signals," *Computers in Biology and Medicine*, vol. 100, pp. 270–278, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482517303153

[118] T. M. Nunes, A. L. Coelho, C. A. Lima, J. P. Papa, and V. H. C. de Albuquerque, "Eeg signal classification for epilepsy diagnosis via optimum path forest – a systematic assessment," *Neurocomputing*, vol. 136, pp. 103–123, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092523121400174X

[119] X. Wang, G. Gong, and N. Li, "Automated recognition of epileptic eeg states using a combination of symlet wavelet processing, gradient boosting machine, and grid search optimizer," *Sensors*, vol. 19, no. 2, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/2/219

[120] X. Liang, J. A. Batsis, Y. Zhu, T. M. Driesse, R. M. Roth, D. Kotz, and B. MacWhinney, "Evaluating voice-assistant commands for dementia detection," *Computer Speech & Language*, vol. 72, p. 101297, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S088523082100098X

[121] E. Kurtz, Y. Zhu, T. Driesse, B. Tran, J. A. Batsis, R. M. Roth, and X. Liang, "Early detection of cognitive decline using voice assistant commands," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[122] D. Zhang, L. Yao, K. Chen, S. Wang, X. Chang, and Y. Liu, "Making sense of spatio-temporal preserving representations for eeg-based human intention recognition," *IEEE Transactions on Cybernetics*, vol. 50, no. 7, pp. 3033–3044, 2020.

[123] X. Wei, L. Zhou, Z. Chen, L. Zhang, and Y. Zhou, "Automatic seizure detection using three-dimensional cnn based on multi-channel eeg," *BMC medical informatics and decision making*, vol. 18, pp. 71–80, 2018.

[124] S. Luz, F. Haider, D. Fromm, I. Lazarou, I. Kompatsiaris, and B. MacWhinney, "Multilingual alzheimer's dementia recognition through spontaneous speech: A signal processing grand challenge," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–2.

[125] Z. Wei, H. Pan, L. Qiao, X. Niu, P. Dong, and D. Li, "Cross-modal knowledge distillation in multi-modal fake news detection," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 4733–4737.

[126] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.

[127] M. McCloskey and N. J. Cohen, “Catastrophic interference in connectionist networks: The sequential learning problem,” ser. Psychology of Learning and Motivation, G. H. Bower, Ed. Academic Press, 1989, vol. 24, pp. 109–165. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0079742108605368

[128] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, “Parameter-efficient transfer learning for NLP,” in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 2790–2799. [Online]. Available: https://proceedings.mlr.press/v97/houlsby19a.html

[129] X. He, C. Li, P. Zhang, J. Yang, and X. E. Wang, “Parameter-efficient model adaptation for vision transformers,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, pp. 817–825, Jun. 2023. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/25160

[130] S. Koops, G. S. Brederoo, N. J. de Boer, G. F. Nadema, E. A. Voppel, and E. I. Sommer, “Speech as a biomarker for depression,” *CNS & Neurological Disorders - Drug Targets*, vol. 22, no. 2, pp. 152–160, 2023. [Online]. Available: http://www.eurekaselect.com/article/119378

[131] J. D. Bernard, J. L. Baddeley, B. F. Rodriguez, and P. A. Burke, “Depression, language, and affect: An examination of the influence of baseline depression and affect induction on language,” *Journal of Language and Social Psychology*, vol. 35, no. 3, pp. 317–326, 2016. [Online]. Available: https://doi.org/10.1177/0261927X15589186

[132] T. Inoue, T. Tanaka, S. Nakagawa, Y. Nakato, R. Kameyama, S. Boku, H. Toda, T. Kurita, and T. Koyama, “Utility and limitations of phq-9 in a clinic specializing in psychiatric care,” *BMC psychiatry*, vol. 12, pp. 1–6, 2012.

[133] Alzheimer’s Society, “*Dementia and language*,” Available online at: https://www.alzheimers.org.uk/about-dementia/symptoms-and-diagnosis/symptoms/dementia-and-language, 2021.

[134] D. F. Tang-Wai and N. L. Graham, “Assessment of language function in dementia,” *Geriatrics and Aging*, vol. 11, no. 2, pp. 103–110, 2008.

[135] S. H. Ferris and M. Farlow, “Language impairment in alzheimer’s disease and benefits of acetylcholinesterase inhibitors,” *Clinical interventions in aging*, vol. 8, p. 1007, 2013.

[136] A. Society, "*The progression and stages of dementia*," Available online at: https://www.alzheimers.org.uk/about-dementia/symptoms-and-diagnosis/how-dementia-progresses/progression-stages-dementia, accessed: 2022-07-30.

[137] ——, "*Early-stage signs and symptoms of dementia*," Available online at: https://www.alzheimers.org.uk/about-dementia/symptoms-and-diagnosis/how-dementia-progresses/early-stages-dementia#content-start, accessed: 2022-07-30.

[138] ——, "*The middle stage of dementia*," Available online at: https://www.alzheimers.org.uk/about-dementia/symptoms-and-diagnosis/how-dementia-progresses/middle-stage-dementia#content-start, accessed: 2022-07-30.

[139] D. Potkins, P. Myint, C. Bannister, G. Tadros, R. Chithramohan, A. Swann, J. O'Brien, J. Fossey, E. George, C. Ballard *et al.*, "Language impairment in dementia: impact on symptoms and care needs in residential homes," *International Journal of Geriatric Psychiatry*, vol. 18, no. 11, pp. 1002–1006, 2003.

[140] A. Society, "*The later stage of dementia* ," Available online at: https://www.alzheimers.org.uk/about-dementia/symptoms-and-diagnosis/how-dementia-progresses/later-stages-dementia#content-start, accessed: 2022-07-30.

[141] M. F. Folstein, S. E. Folstein, and P. R. McHugh, ""mini-mental state": a practical method for grading the cognitive state of patients for the clinician," *Journal of psychiatric research*, vol. 12, no. 3, pp. 189–198, 1975.

[142] T. N. Tombaugh and N. J. McIntyre, "The mini-mental state examination: a comprehensive review," *Journal of the American Geriatrics Society*, vol. 40, no. 9, pp. 922–935, 1992.

[143] Z. S. Nasreddine, N. A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, J. L. Cummings, and H. Chertkow, "The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment," *Journal of the American Geriatrics Society*, vol. 53, no. 4, pp. 695–699, 2005.

[144] R. F. Coen, D. A. Robertson, R. A. Kenny, and B. L. King-Kallimanis, "Strengths and limitations of the moca for assessing cognitive functioning: Findings from a large representative sample of irish older adults," *Journal of geriatric psychiatry and neurology*, vol. 29, no. 1, pp. 18–24, 2016.

[145] P. Mathuranath, P. Nestor, G. Berrios, W. Rakowicz, and J. Hodges, "A brief cognitive test battery to differentiate alzheimer's disease and frontotemporal dementia," *Neurology*, vol. 55, no. 11, pp. 1613–1620, 2000.

[146] E. Mioshi, K. Dawson, J. Mitchell, R. Arnold, and J. R. Hodges, "The addenbrooke's cognitive examination revised (ace-r): a brief cognitive test battery for dementia screening," *International Journal of Geriatric Psychiatry: A journal of the psychiatry of late life and allied sciences*, vol. 21, no. 11, pp. 1078–1085, 2006.

[147] S. Hsieh, S. Schubert, C. Hoon, E. Mioshi, and J. R. Hodges, "Validation of the addenbrooke's cognitive examination iii in frontotemporal dementia and alzheimer's disease," *Dementia and geriatric cognitive disorders*, vol. 36, no. 3-4, pp. 242–250, 2013.

[148] H. Goodglass, E. Kaplan, and S. Weintraub, "Boston naming test," 1983.

[149] C. Roth, *Boston Naming Test.* New York, NY: Springer New York, 2011, pp. 430–433. [Online]. Available: https://doi.org/10.1007/978-0-387-79948-3_869

[150] D. Wechsler, "Wechsler adult intelligence scale–," *Archives of Clinical Neuropsychology*, 1955.

[151] ——, "Wechsler memory scale." 1945.

[152] W. G. Rosen, R. C. Mohs, and K. L. Davis, "A new rating scale for alzheimer's disease." *The American journal of psychiatry*, 1984.

[153] J. K. Kueper, M. Speechley, and M. Montero-Odasso, "The alzheimer's disease assessment scale–cognitive subscale (adas-cog): modifications and responsiveness in pre-dementia populations. a narrative review," *Journal of Alzheimer's Disease*, vol. 63, no. 2, pp. 423–444, 2018.

[154] H. Brodaty, D. Pond, N. M. Kemp, G. Luscombe, L. Harding, K. Berman, and F. A. Huppert, "The gpcog: a new screening test for dementia designed for general practice," *Journal of the American Geriatrics Society*, vol. 50, no. 3, pp. 530–534, 2002.

[155] A. D. Diagnosis, "*GPCOG Screening Tool for Dementia*," Available online at: https://www.verywellhealth.com/what-is-the-gpcog-the-general-practitioner-assessment-of-cognition-98641, accessed: 2022-07-30.

[156] C. S. Crowson, E. J. Atkinson, and T. M. Therneau, "Assessing calibration of prognostic risk scores," *Statistical Methods in Medical Research*, vol. 25, no. 4, pp. 1692–1706, 2016, pMID: 23907781. [Online]. Available: https://doi.org/10.1177/0962280213497434

[157] X. Jiang, M. Osl, J. Kim, and L. Ohno-Machado, "Calibrating predictive model estimates to support personalized medicine," *Journal of the American Medical Informatics Association*, vol. 19, no. 2, pp. 263–274, 10 2011. [Online]. Available: https://doi.org/10.1136/amiajnl-2011-000291

[158] M. Raghu, K. Blumer, R. Sayres, Z. Obermeyer, B. Kleinberg, S. Mullainathan, and J. Kleinberg, "Direct uncertainty prediction for medical second opinions," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 5281–5290. [Online]. Available: https://proceedings.mlr.press/v97/raghu19a.html

[159] R. Sawhney, A. Neerkaje, and M. Gaur, "A risk-averse mechanism for suicidality assessment on social media," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).* Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 628–635. [Online]. Available: https://aclanthology.org/2022.acl-short.70

[160] L. Fiorillo, P. Favaro, and F. D. Faraci, "Deepsleepnet-lite: A simplified automatic sleep stage scoring model with uncertainty estimates," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 2076–2085, 2021.

[161] J. Weiner and T. Schultz, "Selecting features for automatic screening for dementia based on speech," in *International Conference on Speech and Computer.* Springer, 2018, pp. 747–756.

[162] L. Calzà, G. Gagliardi, R. Rossini Favretti, and F. Tamburini, "Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia," *Computer Speech & Language*, vol. 65, p. 101113, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230820300462

[163] K. C. Fraser, K. Lundholm Fors, M. Eckerström, F. Öhman, and D. Kokkinakis, "Predicting mci status from multimodal language data using cascaded classifiers," *Frontiers in Aging Neuroscience*, vol. 11, p. 205, 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnagi.2019.00205

[164] S. Nasreen, M. Rohanian, J. Hough, and M. Purver, "Alzheimer's dementia recognition from spontaneous speech using disfluency and interactional features," *Frontiers in Computer Science*, vol. 3, p. 49, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/fcomp.2021.640669

[165] A. Khodabakhsh, S. Kuşxuoğlu, and C. Demiroğlu, "Natural language features for detection of alzheimer's disease in conversational speech," in *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI).* IEEE, 2014, pp. 581–584.

[166] J. Chen, J. Zhu, and J. Ye, "An attention-based hybrid network for automatic detection of alzheimer's disease from narrative speech." in *INTERSPEECH*, 2019, pp. 4085–4089.

[167] F. Di Palo and N. Parde, “Enriching neural models with targeted features for dementia detection,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop.* Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 302–308. [Online]. Available: https://aclanthology.org/P19-2042

[168] S. Karlekar, T. Niu, and M. Bansal, “Detecting linguistic characteristics of Alzheimer’s dementia by interpreting neural models,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers).* New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 701–707. [Online]. Available: https://aclanthology.org/N18-2110

[169] J. Chen, J. Ye, F. Tang, and J. Zhou, “Automatic Detection of Alzheimer’s Disease Using Spontaneous Speech Only,” in *Proc. Interspeech 2021*, 2021, pp. 3830–3834.

[170] U. Sarawgi, W. Zulfikar, N. Soliman, and P. Maes, “Multimodal Inductive Transfer Learning for Detection of Alzheimer’s Dementia and its Severity,” in *Proc. Interspeech 2020*, 2020, pp. 2212–2216.

[171] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 1321–1330. [Online]. Available: https://proceedings.mlr.press/v70/guo17a.html

[172] B. Goodman and S. Flaxman, “European union regulations on algorithmic decision-making and a “right to explanation”,” *AI magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[173] L. Ilias, S. Mouzakitis, and D. Askounis, “Calibration of transformer-based models for identifying stress and depression in social media,” *IEEE Transactions on Computational Social Systems*, vol. 11, no. 2, pp. 1979–1990, 2024.

[174] L. Ilias and D. Askounis, “A cross-attention layer coupled with multimodal fusion methods for recognizing depression from spontaneous speech,” in *Interspeech 2024*, 2024, pp. 912–916.

[175] ——, “Explainable identification of dementia from transcripts using transformer networks,” *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 4153–4164, 2022.

[176] L. Ilias, D. Askounis, and J. Psarras, “Detecting dementia from speech and transcripts using transformers,” *Computer Speech & Language*, vol. 79, p. 101485, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230823000049

[177] L. Ilias and D. Askounis, "Multimodal deep learning models for detecting dementia from speech and transcripts," *Frontiers in Aging Neuroscience*, vol. 14, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnagi.2022.830943

[178] ——, "Context-aware attention layers coupled with optimal transport domain adaptation and multimodal fusion methods for recognizing dementia from spontaneous speech," *Knowledge-Based Systems*, vol. 277, p. 110834, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705123005841

[179] M. Chatzianastasis, L. Ilias, D. Askounis, and M. Vazirgiannis, "Neural architecture search with multimodal fusion methods for diagnosing dementia," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[180] L. Ilias, D. Askounis, and J. Psarras, "Multimodal detection of epilepsy with deep neural networks," *Expert Systems with Applications*, vol. 213, p. 119010, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417422020280

[181] J. Liu and M. Shi, "A hybrid feature selection and ensemble approach to identify depressed users in online social media," *Frontiers in Psychology*, vol. 12, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.802821

[182] T. Nguyen, D. Phung, B. Dao, S. Venkatesh, and M. Berk, "Affective and content analysis of online depression communities," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 217–226, 2014.

[183] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki, "Recognizing depression from twitter activity," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ser. CHI '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 3187–3196. [Online]. Available: https://doi.org/10.1145/2702123.2702280

[184] M. A. Wani, M. A. ELAffendi, K. A. Shakil, A. S. Imran, and A. A. A. El-Latif, "Depression screening in humans with ai and deep learning techniques," *IEEE Transactions on Computational Social Systems*, pp. 1–0, 2022.

[185] J. Kim, J. Lee, E. Park, and J. Han, "A deep learning model for detecting mental illness from user content on social media," *Scientific reports*, vol. 10, no. 1, pp. 1–6, 2020.

[186] S. Ghosh and T. Anwar, "Depression intensity estimation via social media: A deep learning approach," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 6, pp. 1465–1474, 2021.

[187] H. Kour and M. K. Gupta, “An hybrid deep learning approach for depression prediction from user tweets using feature-rich cnn and bi-directional lstm,” *Multimedia Tools and Applications*, vol. 81, no. 17, pp. 23 649–23 685, 2022.

[188] H. Zogan, I. Razzak, S. Jameel, and G. Xu, “Hierarchical convolutional attention network for depression detection on social media and its impact during pandemic,” *IEEE Journal of Biomedical and Health Informatics*, pp. 1–9, 2023.

[189] L. Ren, H. Lin, B. Xu, S. Zhang, L. Yang, and S. Sun, “Depression detection on reddit with an emotion-based attention network: Algorithm development and validation,” *JMIR Med Inform*, vol. 9, no. 7, p. e28754, Jul 2021. [Online]. Available: https://medinform.jmir.org/2021/7/e28754

[190] M. Trotzek, S. Koitka, and C. M. Friedrich, “Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 3, pp. 588–601, 2020.

[191] V. Borba de Souza, J. Campos Nobre, and K. Becker, “Dac stacking: A deep learning ensemble to classify anxiety, depression, and their comorbidity from reddit texts,” *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 7, pp. 3303–3311, 2022.

[192] H. Zogan, I. Razzak, X. Wang, S. Jameel, and G. Xu, “Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media,” *World Wide Web*, vol. 25, no. 1, pp. 281–304, 2022.

[193] A.-S. Uban, B. Chulvi, and P. Rosso, “An emotion and cognitive based analysis of mental health disorders from social media data,” *Future Generation Computer Systems*, vol. 124, pp. 480–494, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X21001825

[194] H. Song, J. You, J.-W. Chung, and J. C. Park, “Feature attention network: Interpretable depression detection from social media,” in *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation.* Hong Kong: Association for Computational Linguistics, 1–3 Dec. 2018. [Online]. Available: https://aclanthology.org/Y18-1070

[195] S. Boinepelli, T. Raha, H. Abburi, P. Parikh, N. Chhaya, and V. Varma, “Leveraging mental health forums for user-level depression detection on social media,” in *Proceedings of the Thirteenth Language Resources and Evaluation Conference.* Marseille, France: European Language Resources Association, Jun. 2022, pp. 5418–5427. [Online]. Available: https://aclanthology.org/2022.lrec-1.580

[196] K. Anantharaman, A. S, R. Sivanaiah, S. Madhavan, and S. M. Rajendram, “SSN_MLRG1@LT-EDI-ACL2022: Multi-class classification using BERT models

for detecting depression signs from social media text," in *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion.* Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 296–300. [Online]. Available: https://aclanthology.org/2022.ltedi-1.44

[197] F. Nilsson and G. Kovács, "FilipN@LT-EDI-ACL2022-detecting signs of depression from social media: Examining the use of summarization methods as data augmentation for text classification," in *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion.* Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 283–286. [Online]. Available: https://aclanthology.org/2022.ltedi-1.41

[198] H. Zogan, I. Razzak, S. Jameel, and G. Xu, "Depressionnet: Learning multi-modalities with user post summarization for depression detection on social media," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 133–142. [Online]. Available: https://doi.org/10.1145/3404835.3462938

[199] S. Ghosh, A. Ekbal, and P. Bhattacharyya, "What does your bio say? inferring twitter users' depression status from multimodal profile information using deep learning," *IEEE Transactions on Computational Social Systems*, vol. 9, no. 5, pp. 1484–1494, 2022.

[200] Z. Li, Z. An, W. Cheng, J. Zhou, F. Zheng, and B. Hu, "Mha: a multimodal hierarchical attention model for depression detection in social media," *Health Information Science and Systems*, vol. 11, no. 1, p. 6, 2023.

[201] J. C. Cheng and A. L. Chen, "Multimodal time-aware attention networks for depression detection," *Journal of Intelligent Information Systems*, vol. 59, no. 2, pp. 319–339, 2022.

[202] G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.-S. Chua, and W. Zhu, "Depression detection via harvesting social media: A multimodal dictionary learning solution," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 3838–3844. [Online]. Available: https://doi.org/10.24963/ijcai.2017/536

[203] T. Gui, L. Zhu, Q. Zhang, M. Peng, X. Zhou, K. Ding, and Z. Chen, "Cooperative multimodal approach to depression detection in twitter," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 110–117, Jul. 2019. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/3775

[204] D. Zhou, J. Yuan, and J. Si, "Health issue identification in social media based on multi-task hierarchical neural networks with topic attention," *Artificial*

Intelligence in Medicine, vol. 118, p. 102119, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365721001123

[205] Y. Wang, Z. Wang, C. Li, Y. Zhang, and H. Wang, "Online social network individual depression detection using a multitask heterogenous modality fusion approach," *Information Sciences*, vol. 609, pp. 727–749, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S002002552200799X

[206] L. Ilias and D. Askounis, "Multitask learning for recognizing stress and depression in social media," *Online Social Networks and Media*, vol. 37-38, p. 100270, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2468696423000290

[207] M. Rodrigues Makiuchi, T. Warnita, K. Uto, and K. Shinoda, "Multimodal fusion of bert-cnn and gated cnn representations for depression detection," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, ser. AVEC '19. New York, NY, USA: ACM, 2019, p. 55–63. [Online]. Available: https://doi.org/10.1145/3347320.3357694

[208] T. Al Hanai, M. Ghassemi, and J. Glass, "Detecting Depression with Audio/Text Sequence Modeling of Interviews," in *Proc. Interspeech 2018*, 2018, pp. 1716–1720.

[209] E. Toto, M. Tlachac, and E. A. Rundensteiner, "Audibert: A deep transfer learning multimodal classification framework for depression screening," in *CIKM '21*. ACM, 2021, p. 4145–4154. [Online]. Available: https://doi.org/10.1145/3459637.3481895

[210] M. Muzammel, H. Salam, and A. Othmani, "End-to-end multimodal clinical depression recognition using deep neural networks: A comparative analysis," *Computer Methods and Programs in Biomedicine*, vol. 211, p. 106433, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260721005071

[211] P.-C. Wei, K. Peng, A. Roitberg, K. Yang, J. Zhang, and R. Stiefelhagen, "Multi-modal depression estimation based on sub-attentional fusion," in *Computer Vision – ECCV 2022 Workshops*, L. Karlinsky, T. Michaeli, and K. Nishino, Eds. Cham: Springer Nature Switzerland, 2023, pp. 623–639.

[212] M. Rohanian, J. Hough, and M. Purver, "Detecting Depression with Word-Level Multimodal Fusion," in *Proc. Interspeech 2019*, 2019, pp. 1443–1447.

[213] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, and L.-P. Morency, "The distress analysis interview corpus of human and computer interviews," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk,

and S. Piperidis, Eds. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 3123–3128. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/508_Paper.pdf

[214] Y. Shen, H. Yang, and L. Lin, "Automatic depression detection: an emotional audio-textual corpus and a gru/bilstm-based model," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6247–6251.

[215] M. Tasnim, M. Ehghaghi, B. Diep, and J. Novikova, "DEPAC: a corpus for depression and anxiety detection from speech," in *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, A. Zirikly, D. Atzil-Slonim, M. Liakata, S. Bedrick, B. Desmet, M. Ireland, A. Lee, S. MacAvaney, M. Purver, R. Resnik, and A. Yates, Eds. Seattle, USA: Association for Computational Linguistics, Jul. 2022, pp. 1–16. [Online]. Available: https://aclanthology.org/2022.clpsych-1.1

[216] H. Cai, Z. Yuan, Y. Gao, S. Sun, N. Li, F. Tian, H. Xiao, J. Li, Z. Yang, X. Li *et al.*, "A multi-modal open dataset for mental-disorder analysis," *Scientific Data*, vol. 9, no. 1, p. 178, 2022.

[217] L. Gauder, L. Pepino, L. Ferrer, and P. Riera, "Alzheimer Disease Recognition Using Speech-Based Embeddings From Pre-Trained Models," in *Proc. Interspeech 2021*, 2021, pp. 3795–3799.

[218] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[219] J. Shor, A. Jansen, R. Maor, O. Lang, O. Tuval, F. de Chaumont Quitry, M. Tagliasacchi, I. Shavitt, D. Emanuel, and Y. Haviv, "Towards Learning a Universal Non-Semantic Representation of Speech," in *Proc. Interspeech 2020*, 2020, pp. 140–144.

[220] X. Li, S. Dalmia, J. Li, M. Lee, P. Littell, J. Yao, A. Anastasopoulos, D. R. Mortensen, G. Neubig, A. W. Black, and F. Metze, "Universal phone recognition with a multilingual allophone system," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8249–8253.

[221] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[222] A. Balagopalan and J. Novikova, “Comparing Acoustic-Based Approaches for Alzheimer’s Disease Detection,” in *Proc. Interspeech 2021*, 2021, pp. 3800–3804.

[223] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, “Cnn architectures for large-scale audio classification,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 131–135.

[224] R. B. Ammar and Y. B. Ayed, “Evaluation of acoustic features for early diagnosis of alzheimer disease,” in *International Conference on Intelligent Systems Design and Applications*. Springer, 2019, pp. 172–181.

[225] S. Al-Hameed, M. Benaissa, and H. Christensen, “Detecting and predicting alzheimer’s disease severity in longitudinal acoustic data,” in *Proceedings of the International Conference on Bioinformatics Research and Applications 2017*, ser. ICBRA 2017. New York, NY, USA: Association for Computing Machinery, 2017, p. 57–61. [Online]. Available: https://doi.org/10.1145/3175587.3175589

[226] F. Haider, S. de la Fuente, and S. Luz, “An assessment of paralinguistic acoustic features for detection of alzheimer’s dementia in spontaneous speech,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 272–281, 2020.

[227] F. Eyben, F. Weninger, F. Gross, and B. Schuller, “Recent developments in opensmile, the munich open-source multimedia feature extractor,” in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM ’13. New York, NY, USA: Association for Computing Machinery, 2013, p. 835–838. [Online]. Available: https://doi.org/10.1145/2502081.2502224

[228] J. Chen, Y. Wang, and D. Wang, “A feature study for classification-based speech separation at low signal-to-noise ratios,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1993–2002, 2014.

[229] F. Bertini, D. Allevi, G. Lutero, L. Calzà, and D. Montesi, “An automatic alzheimer’s disease classifier based on spontaneous spoken english,” *Computer Speech & Language*, vol. 72, p. 101298, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230821000991

[230] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, “audeep: Unsupervised learning of representations from audio with deep recurrent neural networks,” *Journal of Machine Learning Research*, vol. 18, no. 173, pp. 1–5, 2018. [Online]. Available: http://jmlr.org/papers/v18/17-406.html

[231] J. Laguarta and B. Subirana, “Longitudinal speech biomarkers for automated alzheimer’s detection,” *Frontiers in Computer Science*, vol. 3, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcomp.2021.624694

[232] B. MacWhinney and J. Wagner, "Transcribing, searching and data sharing: The clan software and the talkbank data repository," *Gesprachsforschung: Online-Zeitschrift zur verbalen Interaktion*, vol. 11, p. 154, 2010.

[233] Y. Pan, B. Mirheidari, Z. Tu, R. O'Malley, T. Walker, A. Venneri, M. Reuber, D. Blackburn, and H. Christensen, "Acoustic Feature Extraction with Interpretable Deep Neural Network for Neurodegenerative Related Disorder Classification," in *Proc. Interspeech 2020*, 2020, pp. 4806–4810.

[234] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 1021–1028.

[235] Y. F. Khan, B. Kaushik, M. K. I. Rahmani, and M. E. Ahmed, "Stacked deep dense neural network model to predict alzheimer's dementia using audio transcript data," *IEEE Access*, vol. 10, pp. 32 750–32 765, 2022.

[236] S. Wankerl, E. Nöth, and S. Evert, "An N-Gram Based Approach to the Automatic Diagnosis of Alzheimer's Disease from Spoken Language," in *Proc. Interspeech 2017*, 2017, pp. 3162–3166.

[237] J. Fritsch, S. Wankerl, and E. Nöth, "Automatic diagnosis of alzheimer's disease using neural network language models," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5841–5845.

[238] T. Cohen and S. Pakhomov, "A tale of two perplexities: Sensitivity of neural language models to lexical retrieval deficits in dementia of the Alzheimer's type," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 1946–1957. [Online]. Available: https://aclanthology.org/2020.acl-main.176

[239] A. H. Alkenani, Y. Li, Y. Xu, and Q. Zhang, "Predicting prodromal dementia using linguistic patterns and deficits," *IEEE Access*, vol. 8, pp. 193 856–193 873, 2020.

[240] ——, "Predicting alzheimer's disease from spoken and written language using fusion-based stacked generalization," *Journal of Biomedical Informatics*, vol. 118, p. 103803, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046421001325

[241] S. Adhikari, S. Thapa, U. Naseem, P. Singh, H. Huo, G. Bharathy, and M. Prasad, "Exploiting linguistic information from nepali transcripts for early detection of alzheimer's disease using natural language processing and machine learning techniques," *International Journal of Human-Computer Studies*, vol. 160, p. 102761, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1071581921001798

[242] S. Al-Hameed, M. Benaissa, and H. Christensen, "Detecting and predicting alzheimer's disease severity in longitudinal acoustic data," in *Proceedings of the International Conference on Bioinformatics Research and Applications 2017*, 2017, pp. 57–61.

[243] F. Haider, S. De La Fuente, and S. Luz, "An assessment of paralinguistic acoustic features for detection of alzheimer's dementia in spontaneous speech," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 272–281, 2019.

[244] Y. Pan, B. Mirheidari, M. Reuber, A. Venneri, D. Blackburn, and H. Christensen, "Automatic hierarchical attention neural network for detecting ad." in *Interspeech*, 2019, pp. 4105–4109.

[245] W. Kong, H. Jang, G. Carenini, and T. Field, "A neural model for predicting dementia from language," in *Machine Learning for Healthcare Conference.* PMLR, 2019, pp. 270–286.

[246] Y. Pan, V. S. Nallanthighal, D. Blackburn, H. Christensen, and A. Härmä, "Multitask estimation of age and cognitive decline from speech," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2021, pp. 7258–7262.

[247] R. Haulcy and J. Glass, "Classifying alzheimer's disease using audio and text-based representations of speech," *Frontiers in Psychology*, vol. 11, p. 3833, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2020.624137

[248] A. Khodabakhsh, S. Kuşxuoğlu, and C. Demiroğlu, "Natural language features for detection of alzheimer's disease in conversational speech," in *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2014, pp. 581–584.

[249] L. Yang, W. Wei, S. Li, J. Li, and T. Shinozaki, "Augmented Adversarial Self-Supervised Learning for Early-Stage Alzheimer's Speech Detection," in *Proc. Interspeech 2022*, 2022, pp. 541–545.

[250] E. Edwards, C. Dognin, B. Bollepalli, and M. Singh, "Multiscale System for Alzheimer's Dementia Recognition Through Spontaneous Speech," in *Proc. Interspeech 2020*, 2020, pp. 2197–2201.

[251] A. Balagopalan, B. Eyre, F. Rudzicz, and J. Novikova, "To BERT or not to BERT: Comparing Speech and Language-Based Approaches for Alzheimer's Disease Detection," in *Proc. Interspeech 2020*, 2020, pp. 2167–2171.

[252] P. A. Pérez-Toro, J. C. Vásquez-Correa, T. Arias-Vergara, P. Klumpp, M. Sierra-Castrillón, M. E. Roldán-López, D. Aguillón, L. Hincapié-Henao, C. A. Tóbon-Quintero, T. Bocklet, M. Schuster, J. R. Orozco-Arroyave, and E. Nöth, "Acoustic and linguistic analyses to assess early-onset and genetic alzheimer's disease," in

ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 8338–8342.

[253] A. Balagopalan, B. Eyre, J. Robin, F. Rudzicz, and J. Novikova, "Comparing pre-trained and feature-based models for prediction of alzheimer's disease based on speech," *Frontiers in Aging Neuroscience*, vol. 13, p. 189, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnagi.2021.635945

[254] S. Farzana and N. Parde, "Exploring MMSE Score Prediction Using Verbal and Non-Verbal Cues," in *Proc. Interspeech 2020*, 2020, pp. 2207–2211.

[255] A. Mittal, S. Sahoo, A. Datar, J. Kadiwala, H. Shalu, and J. Mathew, "Multi-modal detection of alzheimer's disease from speech and text," 2021.

[256] R. Pappagari, J. Cho, S. Joshi, L. Moro-Velázquez, P. Żelasko, J. Villalba, and N. Dehak, "Automatic Detection and Assessment of Alzheimer Disease Using Speech and Language Technologies in Low-Resource Scenarios," in *Proc. Interspeech 2021*, 2021, pp. 3825–3829.

[257] E. L. Campbell, L. Docio-Fernandez, J. Jiménez-Raboso, and C. Gacia-Mateo, "Alzheimer's Dementia Detection from Audio and Language Modalities in Spontaneous Speech," in *Proc. IberSPEECH 2021*, 2021, pp. 270–274.

[258] M. Martinc, F. Haider, S. Pollak, and S. Luz, "Temporal integration of text transcripts and acoustic features for alzheimer's diagnosis based on spontaneous speech," *Frontiers in Aging Neuroscience*, vol. 13, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnagi.2021.642647

[259] P. Mahajan and V. Baths, "Acoustic and language based deep learning approaches for alzheimer's dementia detection from spontaneous speech," *Frontiers in Aging Neuroscience*, vol. 13, p. 20, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnagi.2021.623607

[260] K. Kumar, C. Kim, and R. M. Stern, "Delta-spectral cepstral coefficients for robust speech recognition," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4784–4787.

[261] S. Furui, "Speaker-independent isolated word recognition based on emphasized spectral dynamics," in *ICASSP '86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 11, 1986, pp. 1991–1994.

[262] Y. Pan, B. Mirheidari, J. M. Harris, J. C. Thompson, M. Jones, J. S. Snowden, D. Blackburn, and H. Christensen, "Using the Outputs of Different Automatic Speech Recognition Paradigms for Acoustic- and BERT-Based Alzheimer's Dementia Detection Through Spontaneous Speech," in *Proc. Interspeech 2021*, 2021, pp. 3810–3814.

[263] J. Koo, J. H. Lee, J. Pyo, Y. Jo, and K. Lee, "Exploiting Multi-Modal Features from Pre-Trained Networks for Alzheimer's Dementia Recognition," in *Proc. Interspeech 2020*, 2020, pp. 2217–2221.

[264] M. Rohanian, J. Hough, and M. Purver, "Alzheimer's Dementia Recognition Using Acoustic, Lexical, Disfluency and Speech Pause Features Robust to Noisy Inputs," in *Proc. Interspeech 2021*, 2021, pp. 3820–3824.

[265] D. Sánchez Villegas, S. Mokaram, and N. Aletras, "Analyzing online political advertisements," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021.* Online: Association for Computational Linguistics, Aug. 2021, pp. 3669–3680. [Online]. Available: https://aclanthology.org/2021.findings-acl.321

[266] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning.* PMLR, 2019, pp. 6105–6114.

[267] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1800–1807.

[268] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Hybrid attention based multimodal network for spoken language classification," in *Proceedings of the 27th International Conference on Computational Linguistics.* Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 2379–2390. [Online]. Available: https://aclanthology.org/C18-1201

[269] H. Pan, Z. Lin, P. Fu, Y. Qi, and W. Wang, "Modeling intra and inter-modality incongruity for multi-modal sarcasm detection," in *Findings of the Association for Computational Linguistics: EMNLP 2020.* Online: Association for Computational Linguistics, Nov. 2020, pp. 1383–1392. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.124

[270] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[271] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 1103–1114. [Online]. Available: https://aclanthology.org/D17-1115

[272] Y. Cai, H. Cai, and X. Wan, "Multi-modal sarcasm detection in twitter with hierarchical fusion model," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2506–2515.

[273] S. Pramanick, A. Roy, and V. M. Patel, "Multimodal learning using optimal transport for sarcasm and humor detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2022, pp. 3930–3940.

[274] B. MacWhinney, *The CHILDES project: Tools for analyzing talk, Volume II: The database.* Psychology Press, 2014.

[275] J. L. Lee, R. Burkholder, G. B. Flinn, and E. R. Coppess, "Working with chat transcripts in python," Department of Computer Science, University of Chicago, Tech. Rep. TR-2016-02, 2016.

[276] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Detecting Cognitive Decline Using Speech Only: The ADReSSo Challenge," in *Proc. Interspeech 2021*, 2021, pp. 3780–3784.

[277] A. L. Benton, "Differential behavioral effects in frontal lobe disease," *Neuropsychologia*, vol. 6, no. 1, pp. 53–60, 1968. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0028393268900389

[278] R. A. Li, I. Hajjar, F. Goldstein, and J. D. Choi, "Analysis of hierarchical multi-content text classification model on B-SHARP dataset for early detection of Alzheimer's disease," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing.* Suzhou, China: Association for Computational Linguistics, Dec. 2020, pp. 358–365. [Online]. Available: https://aclanthology.org/2020.aacl-main.38

[279] D. Gkoumas, B. Wang, A. Tsakalidis, M. Wolters, A. Zubiaga, M. Purver, and M. Liakata, "A longitudinal multi-modal dataset for dementia monitoring and diagnosis," *arXiv preprint arXiv:2109.01537*, 2021.

[280] C. Pope and B. H. Davis, "Finding a balance: The carolinas conversation collection," vol. 7, no. 1, pp. 143–161, 2011. [Online]. Available: https://doi.org/10.1515/cllt.2011.007

[281] B. Mirheidari, D. Blackburn, R. O'Malley, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Computational cognitive assessment: Investigating the use of an intelligent virtual agent for the detection of early signs of dementia," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2732–2736.

[282] B. Mirheidari, D. Blackburn, K. Harkness, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "An avatar-based system for identifying individuals likely to develop dementia," in *Proc. Interspeech 2017*, 2017, pp. 3147–3151. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-690

[283] S. Raghu, N. Sriraam, S. V. Rao, A. S. Hegde, and P. L. Kubben, "Automated detection of epileptic seizures using successive decomposition index and support vector machine classifier in long-term eeg," *Neural Computing and Applications*, vol. 32, no. 13, pp. 8965–8984, 2020.

[284] M. Zhou, C. Tian, R. Cao, B. Wang, Y. Niu, T. Hu, H. Guo, and J. Xiang, "Epileptic seizure detection based on eeg signals and cnn," *Frontiers in Neuroinformatics*, vol. 12, 2018. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fninf.2018.00095

[285] A. Subasi, J. Kevric, and M. A. Canbaz, "Epileptic seizure detection using hybrid machine learning methods," *Neural Computing and Applications*, vol. 31, no. 1, pp. 317–325, 2019.

[286] H. A. Glory, C. Vigneswaran, S. S. Jagtap, R. Shruthi, G. Hariharan, and V. S. Sriram, "Ahw-bgoa-dnn: a novel deep learning model for epileptic seizure detection," *Neural Computing and Applications*, vol. 33, no. 11, pp. 6065–6093, 2021.

[287] M. Thakur, U. Snekhalatha, M. N. Shafi, S. R. Gupta, S. R. Roy, and S. Vineetha, "Epileptic seizure detection using deep bidirectional long short-term memory network," in *Sentimental Analysis and Deep Learning.* Springer, 2022, pp. 893–906.

[288] D. Ahmedt-Aristizabal, C. Fookes, K. Nguyen, and S. Sridharan, "Deep classification of epileptic signals," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 332–335.

[289] T. D.K., P. B.G., and F. Xiong, "Epileptic seizure detection and prediction using stacked bidirectional long short term memory," *Pattern Recognition Letters*, vol. 128, pp. 529–535, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865519303125

[290] Ö. Yıldırım, U. B. Baloglu, and U. R. Acharya, "A deep convolutional neural network model for automated identification of abnormal eeg signals," *Neural Computing and Applications*, vol. 32, no. 20, pp. 15 857–15 868, 2020.

[291] I. Aliyu and C. G. Lim, "Selection of optimal wavelet features for epileptic eeg signal classification with lstm," *Neural Computing and Applications*, pp. 1–21, 2021.

[292] M. S. Hossain, S. U. Amin, M. Alsulaiman, and G. Muhammad, "Applying deep learning for epilepsy seizure detection and brain mapping visualization," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 15, no. 1s, feb 2019. [Online]. Available: https://doi.org/10.1145/3241056

[293] A. Abdelhameed and M. Bayoumi, "A deep learning approach for automatic seizure detection in children with epilepsy," *Frontiers in Computational Neuroscience*,

vol. 15, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fncom.2021.650050

[294] R. Akut, "Wavelet based deep learning approach for epilepsy detection," *Health information science and systems*, vol. 7, no. 1, pp. 1–9, 2019.

[295] T. Uyttenhove, A. Maes, T. V. Steenkiste, D. Deschrijver, and T. Dhaene, "Interpretable epilepsy detection in routine, interictal eeg data using deep learning," in *Proceedings of the Machine Learning for Health NeurIPS Workshop*, ser. Proceedings of Machine Learning Research, E. Alsentzer, M. B. A. McDermott, F. Falck, S. K. Sarkar, S. Roy, and S. L. Hyland, Eds., vol. 136. PMLR, 11 Dec 2020, pp. 355–366. [Online]. Available: https://proceedings.mlr.press/v136/uyttenhove20a.html

[296] S. Jonas, A. O. Rossetti, M. Oddo, S. Jenni, P. Favaro, and F. Zubler, "Eeg-based outcome prediction after cardiac arrest with convolutional neural networks: Performance and visualization of discriminative features," *Human brain mapping*, vol. 40, no. 16, pp. 4606–4617, 2019.

[297] S. Muhammad Usman, S. Khalid, and S. Bashir, "A deep learning based ensemble learning method for epileptic seizure prediction," *Computers in Biology and Medicine*, vol. 136, p. 104710, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482521005047

[298] K. Akyol, "Stacking ensemble based deep neural networks modeling for effective epileptic seizure detection," *Expert Systems with Applications*, vol. 148, p. 113239, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417420300658

[299] X. Yao, X. Li, Q. Ye, Y. Huang, Q. Cheng, and G.-Q. Zhang, "A robust deep learning approach for automatic classification of seizures against non-seizures," *Biomedical Signal Processing and Control*, vol. 64, p. 102215, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809420303475

[300] I. Obeid and J. Picone, "The temple university hospital eeg data corpus," *Frontiers in Neuroscience*, vol. 10, 2016. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2016.00196

[301] A. H. Shoeb, "Application of machine learning to epileptic seizure onset detection and treatment," Ph.D. dissertation, Massachusetts Institute of Technology, 2009.

[302] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[303] P. Detti, G. Vatti, and G. Zabalo Manrique de Lara, "Eeg synchronization analysis for seizure prediction: A study on data of noninvasive recordings," *Processes*, vol. 8, no. 7, 2020. [Online]. Available: https://www.mdpi.com/2227-9717/8/7/846

[304] N. Stevenson, K. Tapani, L. Lauronen, and S. Vanhatalo, "A dataset of neonatal EEG recordings with seizures annotations," Jun. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1280684

[305] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[306] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf

[307] P. Mishra, M. Del Tredici, H. Yannakoudakis, and E. Shutova, "Author profiling for abuse detection," in *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1088–1098. [Online]. Available: https://aclanthology.org/C18-1093

[308] M. P. Naeini, G. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using bayesian binning," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[309] J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, and D. Tran, "Measuring calibration in deep learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[310] K. Yang, T. Zhang, and S. Ananiadou, "A mental state knowledge–aware and contrastive network for early stress and depression detection on social media," *Information Processing & Management*, vol. 59, no. 4, p. 102961, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457322000796

[311] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17. JMLR.org, 2017, p. 3319–3328.

[312] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015, pp. 18–25.

[313] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=r1rhWnZkg

[314] L. De Lathauwer, "Decompositions of a higher-order tensor in block terms—part ii: Definitions and uniqueness," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 3, pp. 1033–1066, 2008. [Online]. Available: https://doi.org/10.1137/070690729

[315] J. D. Carroll and J.-J. Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970.

[316] L. Zhao, G. Han, Y. Zhao, Y. Jin, T. Ge, W. Yang, R. Cui, S. Xu, and B. Li, "Gender differences in depression: Evidence from genetics," *Frontiers in Genetics*, vol. 11, 2020. [Online]. Available: https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2020.562316

[317] B. Patria, "The longitudinal effects of education on depression: Finding from the indonesian national survey," *Frontiers in Public Health*, vol. 10, 2022. [Online]. Available: https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2022.1017995

[318] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1459–1462. [Online]. Available: https://doi.org/10.1145/1873951.1874246

[319] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper/2016/file/9dcb88e0137649590b755372b040afad-Paper.pdf

[320] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, “defend: Explainable fake news detection,” in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 395–405.

[321] Y.-J. Lu and C.-T. Li, “GCAN: Graph-aware co-attention networks for explainable fake news detection on social media,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Online: Association for Computational Linguistics, Jul. 2020, pp. 505–514. [Online]. Available: https://aclanthology.org/2020.acl-main.48

[322] R. Caruana, “Multitask learning,” *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[323] S. Rajamanickam, P. Mishra, H. Yannakoudakis, and E. Shutova, “Joint modelling of emotion and abusive language detection,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Online: Association for Computational Linguistics, Jul. 2020, pp. 4270–4279. [Online]. Available: https://aclanthology.org/2020.acl-main.394

[324] B. Portelli, E. Lenzi, E. Chersoni, G. Serra, and E. Santus, “BERT prescriptions to avoid unwanted headaches: A comparison of transformer architectures for adverse drug event detection,” in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume.* Online: Association for Computational Linguistics, Apr. 2021, pp. 1740–1747. [Online]. Available: https://aclanthology.org/2021.eacl-main.149

[325] E. A. Ríssola, M. Aliannejadi, and F. Crestani, “Beyond modelling: understanding mental disorders in online social media,” in *European Conference on Information Retrieval.* Springer, 2020, pp. 296–310.

[326] T. Koike, K. Qian, Q. Kong, M. D. Plumbley, B. W. Schuller, and Y. Yamamoto, “Audio for audio is better? an investigation on transfer learning models for heart sound classification,” in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020, pp. 74–77.

[327] H. Meng, T. Yan, F. Yuan, and H. Wei, “Speech emotion recognition from 3d log-mel spectrograms with deep learning network,” *IEEE Access*, vol. 7, pp. 125 868–125 881, 2019.

[328] Z. Zhao, Q. Li, N. Cummins, B. Liu, H. Wang, J. Tao, and B. W. Schuller, “Hybrid Network Feature Extraction for Depression Assessment from Speech,” in *Proc. Interspeech 2020*, 2020, pp. 4956–4960.

[329] A. Ebrahimi-Ghahnavieh, S. Luo, and R. Chiong, “Transfer learning for alzheimer’s disease detection on mri images,” in *2019 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, 2019, pp. 133–138.

[330] S. Raghu, N. Sriraam, Y. Temel, S. V. Rao, and P. L. Kubben, "Eeg based multi-class seizure type classification using convolutional neural network and transfer learning," *Neural Networks*, vol. 124, pp. 202–212, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608020300198

[331] A. D. Roy and M. M. Islam, "Detection of epileptic seizures from wavelet scalogram of eeg signal using transfer learning with alexnet convolutional neural network," in *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, 2020, pp. 1–5.

[332] P. Giannopoulos, I. Perikos, and I. Hatzilygeroudis, "Deep learning approaches for facial emotion recognition: A case study on fer-2013," in *Advances in hybridization of intelligent methods.* Springer, 2018, pp. 1–16.

[333] R. M. Ghoniem, "Deep genetic algorithm-based voice pathology diagnostic system," in *International Conference on Applications of Natural Language to Information Systems.* Springer, 2019, pp. 220–233.

[334] B. McFee, A. Metsai, M. McVicar, S. Balke, C. Thomé, C. Raffel, F. Zalkow, A. Malek, Dana, K. Lee, O. Nieto, D. Ellis, J. Mason, E. Battenberg, S. Seyfarth, R. Yamamoto, viktorandreevichmorozov, K. Choi, J. Moore, R. Bittner, S. Hidaka, Z. Wei, nullmightybofo, D. Hereñú, F.-R. Stöter, P. Friesch, A. Weiss, M. Vollrath, T. Kim, and Thassilo, "librosa/librosa: 0.8.1rc2," May 2021. [Online]. Available: https://doi.org/10.5281/zenodo.4792298

[335] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[336] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[337] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[338] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[339] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.

[340] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[341] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le, “Mnasnet: Platform-aware neural architecture search for mobile,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2815–2823.

[342] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5987–5995.

[343] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[344] M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 6105–6114. [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html

[345] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[346] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[347] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[348] B. MacWhinney, “The CHILDES Project: Tools for Analyzing Talk (third edition): Volume I: Transcription format and programs, Volume II: The database,” *Computational Linguistics*, vol. 26, no. 4, pp. 657–657, 12 2000. [Online]. Available: https://doi.org/10.1162/coli.2000.26.4.657

[349] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[350] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[351] N. Cummins, Y. Pan, Z. Ren, J. Fritsch, V. S. Nallanthighal, H. Christensen, D. Blackburn, B. W. Schuller, M. Magimai-Doss, H. Strik *et al.*, "A comparison of acoustic and linguistics methodologies for alzheimer's dementia recognition," in *Interspeech 2020.* ISCA-International Speech Communication Association, 2020, pp. 2182–2186.

[352] A. Meghanani, C. Anoop, and A. Ramakrishnan, "An exploration of log-mel spectrogram and mfcc features for alzheimer's dementia recognition from spontaneous speech," in *2021 IEEE Spoken Language Technology Workshop (SLT).* IEEE, 2021, pp. 670–677.

[353] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, *PyTorch: An Imperative Style, High-Performance Deep Learning Library.* Red Hook, NY, USA: Curran Associates Inc., 2019.

[354] Z. Tu, Y. Liu, Z. Lu, X. Liu, and H. Li, "Context gates for neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 87–99, 2017. [Online]. Available: https://aclanthology.org/Q17-1007

[355] B. Zhang, D. Xiong, J. Su, and H. Duan, "A context-aware recurrent encoder for neural machine translation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2424–2432, 2017.

[356] L. Wang, Z. Tu, A. Way, and Q. Liu, "Exploiting cross-sentence context for neural machine translation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 2826–2831. [Online]. Available: https://aclanthology.org/D17-1301

[357] E. Voita, P. Serdyukov, R. Sennrich, and I. Titov, "Context-aware neural machine translation learns anaphora resolution," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 1264–1274. [Online]. Available: https://aclanthology.org/P18-1117

[358] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers &amp; distillation through attention," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 10 347–10 357. [Online]. Available: https://proceedings.mlr.press/v139/touvron21a.html

[359] C. Chen, D. Han, and C.-C. Chang, "Caan: Context-aware attention network for visual question answering," *Pattern Recognition*, vol. 132, p. 108980, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320322004605

[360] B. Yang, J. Li, D. F. Wong, L. S. Chao, X. Wang, and Z. Tu, "Context-aware self-attention networks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 387–394, Jul. 2019. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/3809

[361] C. Villani, "Optimal transport, old and new. notes for the 2005 saint-flour summer school," *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer*, 2008.

[362] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[363] B. McFee, A. Metsai, M. McVicar, S. Balke, C. Thomé, C. Raffel, F. Zalkow, A. Malek, Dana, K. Lee, O. Nieto, D. Ellis, J. Mason, E. Battenberg, S. Seyfarth, R. Yamamoto, viktorandreevichmorozov, K. Choi, J. Moore, R. Bittner, S. Hidaka, Z. Wei, nullmightybofo, A. Weiss, D. Hereñú, F.-R. Stöter, L. Nickel, P. Friesch, M. Vollrath, and T. Kim, "librosa/librosa: 0.9.2," Jun. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6759664

[364] G. Mialon, D. Chen, A. d'Aspremont, and J. Mairal, "A trainable optimal transport embedding for feature aggregation and its relationship to attention," in *International Conference on Learning Representations*, 2021.

[365] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy, "Optimal transport for domain adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 9, pp. 1853–1865, 2017.

[366] S. Ferradans, N. Papadakis, G. Peyré, and J.-F. Aujol, "Regularized discrete optimal transport," *SIAM Journal on Imaging Sciences*, vol. 7, no. 3, pp. 1853–1882, 2014.

[367] L. Ilias, D. Askounis, and J. Psarras, "A multimodal approach for dementia detection from spontaneous speech with tensor fusion layer," in *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2022, pp. 1–5.

[368] Y. Qiao, X. Yin, D. Wiechmann, and E. Kerz, "Alzheimer's Disease Detection from Spontaneous Speech Through Combining Linguistic Complexity and (Dis)Fluency Features with Pretrained Language Models," in *Proc. Interspeech 2021*, 2021, pp. 3805–3809.

[369] Z. S. Syed, M. S. S. Syed, M. Lech, and E. Pirogova, "Tackling the ADRESSO Challenge 2021: The MUET-RMIT System for Alzheimer's Dementia Recognition from Spontaneous Speech," in *Proc. Interspeech 2021*, 2021, pp. 3815–3819.

[370] Y. Zhu, A. Obyat, X. Liang, J. A. Batsis, and R. M. Roth, "WavBERT: Exploiting Semantic and Non-Semantic Speech Using Wav2vec and BERT for Dementia Detection," in *Proc. Interspeech 2021*, 2021, pp. 3790–3794.

[371] N. Wang, Y. Cao, S. Hao, Z. Shao, and K. Subbalakshmi, "Modular Multi-Modal Attention Network for Alzheimer's Disease Detection Using Patient Audio and Language Data," in *Proc. Interspeech 2021*, 2021, pp. 3835–3839.

[372] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer, "Pot: Python optimal transport," *Journal of Machine Learning Research*, vol. 22, no. 78, pp. 1–8, 2021. [Online]. Available: http://jmlr.org/papers/v22/20-451.html

[373] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," *Journal of Machine Learning Research*, vol. 20, no. 55, pp. 1–21, 2019. [Online]. Available: http://jmlr.org/papers/v20/18-598.html

[374] C. Liu, B. Zoph, M. Neumann, J. Shlens, W. Hua, L.-J. Li, L. Fei-Fei, A. Yuille, J. Huang, and K. Murphy, "Progressive neural architecture search," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[375] M. Chatzianastasis, G. Dasoulas, G. Siolas, and M. Vazirgiannis, "Graph-based neural architecture search with operation embeddings," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021, pp. 393–402.

[376] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8697–8710.

[377] L. Biewald, "Experiment tracking with weights and biases," 2020, software available from wandb.com. [Online]. Available: https://www.wandb.com/

[378] L. Guo, D. Rivero, J. A. Seoane, and A. Pazos, "Classification of eeg signals using relative wavelet energy and artificial neural networks," in *Proceedings of the First*

ACM/SIGEVO Summit on Genetic and Evolutionary Computation, ser. GEC '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 177–184. [Online]. Available: https://doi.org/10.1145/1543834.1543860

[379] T. Gandhi, B. K. Panigrahi, and S. Anand, "A comparative study of wavelet families for eeg signal classification," *Neurocomputing*, vol. 74, no. 17, pp. 3051–3057, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231211003158

[380] Q. Lin, S.-q. Ye, X.-m. Huang, S.-y. Li, M.-z. Zhang, Y. Xue, and W.-S. Chen, "Classification of epileptic eeg signals with stacked sparse autoencoder based on deep learning," in *International Conference on Intelligent Computing.* Springer, 2016, pp. 802–810.

[381] S. Chandaka, A. Chatterjee, and S. Munshi, "Cross-correlation aided support vector machine classifier for classification of eeg signals," *Expert Systems with Applications*, vol. 36, no. 2, Part 1, pp. 1329–1336, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417407005696

[382] N. Nicolaou and J. Georgiou, "Detection of epileptic electroencephalogram based on permutation entropy and support vector machines," *Expert Systems with Applications*, vol. 39, no. 1, pp. 202–209, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417411009705

[383] L. Guo, D. Rivero, and A. Pazos, "Epileptic seizure detection using multiwavelet transform based approximate entropy and artificial neural networks," *Journal of Neuroscience Methods*, vol. 193, no. 1, pp. 156–163, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165027010004905

[384] M. Sharma, R. B. Pachori, and U. Rajendra Acharya, "A new approach to characterize epileptic seizures using analytic time-frequency flexible wavelet transform and fractal dimension," *Pattern Recognition Letters*, vol. 94, pp. 172–179, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865517300995

[385] A. Gupta, P. Singh, and M. Karlekar, "A novel signal modeling approach for classification of seizure and seizure-free eeg signals," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 5, pp. 925–935, 2018.

[386] R. Bose, S. Pratiher, and S. Chatterjee, "Detection of epileptic seizure employing a novel set of features extracted from multifractal spectrum of electroencephalogram signals," *IET Signal Processing*, vol. 13, no. 2, pp. 157–164, 2019. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-spr.2018.5258

[387] H.-S. Chiang, M.-Y. Chen, and Y.-J. Huang, “Wavelet-based eeg processing for epilepsy detection using fuzzy entropy and associative petri net,” *IEEE Access*, vol. 7, pp. 103 255–103 262, 2019.

[388] M. Sameer and B. Gupta, “Time–frequency statistical features of delta band for detection of epileptic seizures,” *Wireless Personal Communications*, vol. 122, no. 1, pp. 489–499, 2022.

[389] K. Abualsaud, M. Mahmuddin, R. Hussein, and A. Mohamed, “Performance evaluation for compression-accuracy trade-off using compressive sensing for eeg-based epileptic seizure detection in wireless tele-monitoring,” in *2013 9th International Wireless Communications and Mobile Computing Conference (IWCMC)*, 2013, pp. 231–236.

[390] K. Abualsaud, M. Mahmuddin, M. Saleh, and A. Mohamed, “Ensemble classifier for epileptic seizure detection for imperfect eeg data,” *The Scientific World Journal*, vol. 2015, 2015.

[391] J. Liu and B. Woodson, “Deep learning classification for epilepsy detection using a single channel electroencephalography (eeg),” in *Proceedings of the 2019 3rd International Conference on Deep Learning Technologies*, ser. ICDLT 2019. New York, NY, USA: Association for Computing Machinery, 2019, p. 23–26. [Online]. Available: https://doi.org/10.1145/3342999.3343008

[392] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, “Communication-Efficient Learning of Deep Networks from Decentralized Data,” in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Singh and J. Zhu, Eds., vol. 54. PMLR, 20–22 Apr 2017, pp. 1273–1282. [Online]. Available: https://proceedings.mlr.press/v54/mcmahan17a.html

Appendices

Appendix A

List of Publications

1. **Publications in Journals**

   - **L. Ilias** and D. Askounis, “Explainable Identification of Dementia From Transcripts Using Transformer Networks,” in *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 4153-4164, Aug. 2022, doi: https://doi.org/10.1109/JBHI.2022.3172479
   - **L. Ilias**, D. Askounis, and J. Psarras,“Detecting dementia from speech and transcripts using transformers,” *Computer Speech & Language*, vol. 79, p. 101485, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230823000049
   - **L. Ilias** and D. Askounis, “Multimodal deep learning models for detecting dementia from speech and transcripts,” *Frontiers in Aging Neuroscience*, vol. 14, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnagi.2022.830943
   - **L. Ilias** and D. Askounis, “Context-aware attention layers coupled with optimal transport domain adaptation and multimodal fusion methods for recognizing dementia from spontaneous speech,” *Knowledge-Based Systems*, vol. 277, p. 110834, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705123005841
   - **L. Ilias**, S. Mouzakitis and D. Askounis, “Calibration of Transformer-Based Models for Identifying Stress and Depression in Social Media,” in *IEEE Transactions on Computational Social Systems*, vol. 11, no. 2, pp. 1979-1990, April 2024, doi: https://doi.org/10.1109/TCSS.2023.3283009.
   - **L. Ilias** and D. Askounis, “Multitask learning for recognizing stress and depression in social media,” *Online Social Networks and Media*, vol. 37-38, p. 100270, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2468696423000290
   - **L. Ilias**, D. Askounis, and J. Psarras, “Multimodal detection of epilepsy with deep neural networks,” *Expert Systems with Applications*, vol. 213, p. 119010,

2023. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0957417422020280`

- **L. Ilias** and I. Roussaki, “Detecting malicious activity in twitter using deep learning techniques,” *Applied Soft Computing*, vol. 107, p. 107360, 2021. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S1568494621002830`
- **L. Ilias**, I. M. Kazelidis and D. Askounis, “Multimodal Detection of Bots on X (Twitter) using Transformers,” in *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 7320-7334, 2024, doi: `https://doi.org/10.1109/TIFS.2024.3435138`.
- **L. Ilias**, E. Sarmas, V. Marinakis, D. Askounis, and H. Doukas, “Unsupervised domain adaptation methods for photovoltaic power forecasting,” *Applied Soft Computing*, vol. 149, p. 110979, 2023. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S1568494623009973`
- **L. Ilias**, G. Tsapelas, P. Kapsalis, V. Michalakopoulos, G. Kormpakis, S. Mouzakitis, and D. Askounis, “Leveraging extreme scale analytics, AI and digital twins for maritime digitalization: the VesselAI architecture,” *Frontiers in Big Data*, vol. 6, 2023. [Online]. Available: `https://www.frontiersin.org/articles/10.3389/fdata.2023.1220348`
- **L. Ilias**, P. Kapsalis, S. Mouzakitis and D. Askounis, “A Multitask Learning Framework for Predicting Ship Fuel Oil Consumption,” in *IEEE Access*, vol. 11, pp. 132576-132589, 2023, doi: `https://doi.org/10.1109/ACCESS.2023.3335905`.
- **L. Ilias**, G. Doukas, M. Kontoulis, K. Alexakis, A. Michalitsi-Psarrou, C. Ntanos, and D. Askounis, “Overview of methods and available tools used in complex brain disorders,” *Open Research Europe*, vol. 3, no. 152, 2023. [Online]. Available: `https://doi.org/10.12688/openreseurope.16244.1`
- M. Kerasiotis, **L. Ilias**, and D. Askounis, “Depression Detection in Social Media Posts Using Transformer-Based Models and Auxiliary Features,” *Social Network Analysis and Mining*, vol. 14, 196 (2024). `https://doi.org/10.1007/s13278-024-01360-4`
- K. Psychogyios, **L. Ilias**, C. Ntanos and D. Askounis, “Missing Value Imputation Methods for Electronic Health Records,” in *IEEE Access*, vol. 11, pp. 21562-21574, 2023, doi: `https://doi.org/10.1109/ACCESS.2023.3251919`.
- G. Tzoumanekas, M. Chatzianastasis, **L. Ilias**, G. Kiokes, J. Psarras, and D. Askounis, “A Graph Neural Architecture Search Approach for Identifying Bots in Social Media,” *Frontiers in Artificial Intelligence*, vol. 7, 2024, doi: `https://doi.org/10.3389/frai.2024.1509179`

2. **Publications in Conferences**

- **L. Ilias**, D. Askounis and J. Psarras, "A Multimodal Approach for Detecting Dementia from Spontaneous Speech with Tensor Fusion Layer," *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, Ioannina, Greece, 2022, pp. 1-5, doi: `https://doi.org/10.1109/BHI56158.2022.9926818`.
- K. Psychogyios, **L. Ilias** and D. Askounis, "Comparison of missing data imputation methods using the Framingham Heart study dataset," *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, Ioannina, Greece, 2022, pp. 1-5, doi: `https://doi.org/10.1109/BHI56158.2022.9926882`.
- M. Chatzianastasis*, **L. Ilias***, D. Askounis and M. Vazirgiannis, "Neural Architecture Search with Multimodal Fusion Methods for Diagnosing Dementia," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1-5, doi: `https://doi.org/10.1109/ICASSP49357.2023.10096579`.
  ** The first two authors contributed equally.*
- **L. Ilias** and D. Askounis, "A Cross-Attention Layer coupled with Multimodal Fusion Methods for Recognizing Depression from Spontaneous Speech," *Proc. Interspeech 2024*, 2024, pp. 912-916, doi: `https://doi.org/10.21437/Interspeech.2024-188`
- V. Michalakopoulos, **L. Ilias**, P. Kapsalis, S. Mouzakitis and D. Askounis, "Comparison of Machine Learning Algorithms For Predicting CO_2 Emissions in the maritime domain," *2023 14th International Conference on Information, Intelligence, Systems & Applications (IISA)*, Volos, Greece, 2023, pp. 1-4, doi: `https://doi.org/10.1109/IISA59645.2023.10345936`.
- I. Stavropoulos, M. Kontoulis, **L. Ilias**, G. Doukas, E. Nianiou, M. Pina, M. Dolzan, F. Benninger, S. Gimenez Badia, A. Glik, I. Goldberg, E. Karageorgiou, C. Ntanos, O. Keret, A. Valentin, "Mes-CoBraD: an open-source research platform for integrated epilepsy-EEG data collection and analysis," *Epilepsia*, Volume 64, Issue S2, Available at: `https://doi.org/10.1111/epi.17787`.
- L. Nerantzis, I. Vourcachis, C. Santorinaios, **L. Ilias**, C. Ntanos, I. Benekos, "Search and Rescue: Emerging technologies for the Early location of Entrapped victims under Collapsed Structures and Advanced Wearables for risk assessment and First Responder's Safety in SAR operations," *SafeThessaloniki 2022 – 9th International Conference on Civil Protection & New Technologies*, Thessaloniki, Greece, 2022, pp. 313-316, Available at: `https://safegreece.org/safethessaloniki2022/images/docs/safethessaloniki_proceedings.pdf`.

3. **Preprints**

- **L. Ilias**, F. Soldner, and B. Kleinberg, "Explainable verbal deception detection

using transformers," *arXiv preprint arXiv:2210.03080*, 2022.
https://arxiv.org/pdf/2210.03080.pdf

- **L. Ilias**, G. Doukas, V. Lamprou, C. Ntanos, and D. Askounis, "Convolutional Neural Networks and Mixture of Experts for Intrusion Detection in 5G Networks and beyond," *arXiv preprint arXiv:2412.03483*, 2024.
https://arxiv.org/pdf/2412.03483

Appendix B

List of Abbreviations

ACE	Adaptive Calibration Error
ACE	Addenbrooke's Cognitive Examination
AD	Alzheimer's Disease
ADAS	Alzheimer's Disease Assessment Scale
ADR	Active Data Representation
AI	Artificial Intelligence
ANEW	affective norms for english words
ASO	Almost Stochastic Order
ASR	Automatic Speech Recognition
ATFFWT	analytic time-frequency flexible wavelet transform
BART	Bidirectional Auto-regressive Transformers
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	bidirectional long short-term memory
BNT	Boston Naming Test
CNN	Convolutional Neural Network
CP	Candecomp/PARAFAC
CV	Computer Vision
DARTS	Differentiable Architecture Search
DeiT	Data-Efficient Image Transformer
DWT	Discrete Wavelet Transform
ECE	Expected Calibration Error
EEG	Electroencephalogram
EMD	Earth's Mover Distance
FMD	functional memory disorder
GA	Genetic Algorithm
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
GOSS	Global Outlier Standard Score
GMU	Gated Multimodal Unit

GPCOG	General Practitioner assessment of Cognition
HAN	Hierarchical Attention Network
HC	Healthy Control
INV	Interviewer
IVA	Intelligent Virtual Agent
k-NN	k nearest neighbours
KL divergence	Kullback-Leibler divergence
LDA	Latent Dirichlet Allocation
LIWC	Linguistic Inquiry and Word Count
LM	Language Model
LOOCV	leave-one-out cross-validation
LR	Logistic Regression
M-BERT	Multimodal BERT
M-MentalBERT	Multimodal MentalBERT
MCB	Multimodal Compact Bilinear
MCI	Mild Cognitive Impairement
MFB	Multimodal Factorized Bilinear
MFCC	Mel Frequency Cepstral Coefficients
MFH	Multimodal Factorized High-order
ML	Machine Learning
MLB	Multimodal Lowrank Bilinear
MLPNN	Multilayer perceptron neural network
MMSE	Mini-Mental State Examination
MRI	Magnetic Resonance Imaging
MTL	Multitask Learning
MoCA	Montreal Cognitive Assessment
NAS	Neural Architecture Search
ND	Neurodegenerative Disorders
NLP	Natural Language Processing
OTK	Optimal Transport Kernel
OVBM	Open Voice Brain Model
PAR	Participant
PCA	Principal Component Analysis
PET	Positron Emission Tomography
PHQ-9	Patient Health Questionnaire-9
PLSR	partial least squares regressor
PoAD	Possible Alzheimer's Disease
PRI	Perceptual Reasoning Index
PSI	Processing Speed Index
PSO	Particle Swarm Optimization
ReLU	Rectified Linear Unit

RF	Random Forest
RMSE	Root Mean Squared Error
SDI	successive decomposition index
SGD	Stochastic Gradient Descent
STFT	Short-time Fourier Transform
STL	Single-task learning
SVM	Support Vector Machine
SVR	Support Vector Regression
VCI	Verbal Comprehension Index
ViT	Vision Transformer
WAIS	Wechsler Adult Intelligence Scale
WHO	World Health Organization
WMI	Working Memory Index
WMS	Wechsler Memory Scale