ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Speech-based Depression Estimation

Διπλωματική Εργασία

της

Πυλαρινού Άρτεμις

Εξωτερικός Επιβλέπων: Θεόδωρος Γιαννακόπουλος
Β΄ Ερευνητής ΕΚΕΦΕ Δημόκριτος

Επιβλέπων Ε.Μ.Π.: Γιώργος Στάμου
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2025

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Speech-based Depression Estimation

Διπλωματική Εργασία

της

Πυλαρινού Άρτεμις

Εξωτερικός Επιβλέπων: Θεόδωρος Γιαννακόπουλος
Β΄ Ερευνητής ΕΚΕΦΕ Δημόκριτος

Επιβλέπων Ε.Μ.Π.: Γιώργος Στάμου
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή επιτροπή την 3η Ιουλίου 2025.

.....	.....	.....
Γεώργιος Στάμου Καθηγητής Ε.Μ.Π.	Αθανάσιος Βουλόδημος Επίκουρος Καθηγητής Ε.Μ.Π.	Α.-Γ. Σταφυλοπάτης Ομότιμος Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2025

.....

Πυλαρινού Άρτεμις
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η παρούσα διπλωματική εστιάζει στη χρήση της μηχανικής μάθησης για την ανάπτυξη αντικειμενικών μεθόδων εκτίμησης της κατάθλιψης, αντιμετωπίζοντας τους περιορισμούς στις τρέχουσες διαγνωστικές πρακτικές. Η έρευνα εισάγει μια καινοτόμο διαδικασία για την εξαγωγή χαρακτηριστικών ήχου και embeddings κειμένου από το σύνολο δεδομένων DAIC-WOZ. Συγκεκριμένα, η βιβλιοθήκη PyAudioAnalysis χρησιμοποιήθηκε για την εξαγωγή χαρακτηριστικών ήχου και τα GloVe embeddings για τα χαρακτηριστικά κειμένου. Παράλληλα εφαρμόστηκε μια μέθοδος εξαγωγής βάσει ρόλων, με σκοπό την ανεξάρτητη επεξεργασία των χαρακτηριστικών του συμμετέχοντα και του συνεντευκτή, ώστε να αναδειχθεί η σημασία κάθε ρόλου στην εκτίμηση της κατάθλιψης και η επίδραση της δυναμικής της αλληλεπίδρασης στην ακρίβεια της πρόβλεψης. Σε αυτή τη μελέτη εφαρμόζονται τεχνικές μηχανικής μάθησης όπως τα Support Vector Machines (SVM) και τα μοντέλα XGBoost. Ο πρωταρχικός στόχος είναι η αναγνώριση του πιο αποτελεσματικού συνδυασμού χαρακτηριστικών και αλγορίθμων που μπορούν να ενισχύσουν την ακρίβεια και την αξιοπιστία των μοντέλων πρόβλεψης της κατάθλιψης. Τα βασικά ευρήματα δείχνουν ότι τα χαρακτηριστικά βασισμένα στο κείμενο, ιδιαίτερα τα embeddings GloVe, υπερέχουν των παραδοσιακών χαρακτηριστικών ήχου, επιτυγχάνοντας βαθμολογία AUC 0.74 για τα μοντέλα βασισμένα στο κείμενο έναντι 0.66 για τα μοντέλα βασισμένα στον ήχο. Η μελέτη διερευνά επίσης τεχνικές εξισορρόπησης, επισημαίνοντας ότι ενώ η μέθοδος SMOTE βελτίωσε την απόδοση των μοντέλων, η επιλογή των χαρακτηριστικών παραμένει κρίσιμη.

Λέξεις-κλειδιά: Κατάθλιψη, Ανάλυση Ομιλίας, Ανάλυση Κειμένου, Μηχανική Μάθηση, Αυτόματη Εκτίμηση Κατάθλιψης

Abstract

This thesis focuses on using machine learning to develop objective methods for estimating depression, thus addressing the limitations in current diagnostic practices. The research introduces a novel pipeline for extracting audio features and text embeddings from the DAIC-WOZ dataset. Specifically the PyAudioAnalysis library was utilized for audio feature extraction and GloVe embeddings for text features. A role-based extraction method was implemented to independently process features for the participant and the interviewer, providing insights into the significance of each role in depression estimation and the influence of interaction dynamics on predictive accuracy. In this study machine learning techniques are applied such as Support Vector Machines (SVM) and XGBoost models, to improve depression detection. The primary goal is to identify the most effective combination of features and algorithms that can enhance the accuracy and reliability of depression prediction models. Key findings indicate that text-based features, particularly GloVe embeddings, outperform traditional audio features, achieving an AUC score of 0.74 for text-based models compared to 0.66 for audio-based models. The study also explores balancing techniques, noting that while SMOTE improved model performance, the choice of features remains critical.

Keywords: Depression, Speech Analysis, Text Analysis, Machine Learning, Automatic Depression Estimation

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά όλους τους ανθρώπους που με στήριξαν και πίστεψαν σε εμένα. Ιδιαίτερες ευχαριστίες οφείλω στον κ. Γεώργιο Στάμου, για την καθοδήγηση και πολύτιμη υποστήριξή του.

Θα ήθελα ακόμη να εκφράσω την ευγνωμοσύνη μου στους φίλους και στην οικογένεια μου για την στήριξη τους όλα αυτά τα χρόνια.

Contents

List of Figures

List of Tables

Εκτενής Ελληνική Περίληψη

1 Εισαγωγή

Η κατάθλιψη αποτελεί κύρια αιτία αναπηρίας παγκοσμίως και αντιπροσωπεύει μια σημαντική πρόκληση για τα συστήματα δημόσιας υγείας. Η κατάθλιψη (επίσης μείζων κατάθλιψη, μείζων καταθλιπτική διαταραχή ή κλινική κατάθλιψη) είναι η πιο διαδεδομένη διαταραχή διάθεσης, που χαρακτηρίζεται από επίμονα χαμηλή διάθεση, μειωμένο ενδιαφέρον και μια σειρά επιπρόσθετων συμπτωμάτων που διαταράσσουν την καθημερινή λειτουργικότητα [43, 47].

Η κατάθλιψη συνδέεται με σημαντική νοσηρότητα και θνησιμότητα και συνδέεται στενά με υψηλά ποσοστά αυτοκτονίας. Επιπλέον, η κατάθλιψη συχνά συνυπάρχει με άλλες χρόνιες ασθένειες, συνδυασμός που συχνά οδηγεί σε πιο σοβαρά προβλήματα υγείας απ' ό,τι θα προκαλούσε κάθε πάθηση μεμονωμένα [53].

Τα παραπάνω και το υψηλό ποσοστό αυτοκτονιών, υπογραμμίζουν τη σημασία της προτεραιοποίησης της διάγνωσης και της θεραπείας της κατάθλιψης. Παρά την σημασία της διάγνωσης, πολλοί άνθρωποι με κατάθλιψη παραμένουν δίχως διάγνωση αυτής λόγω του στίγματος, της έλλειψης πρόσβασης σε υπηρεσίες ψυχικής υγείας και των περιορισμών των τρεχουσών διαγνωστικών μεθόδων [53, 68, 90].

Η τρέχουσα διάγνωση της κατάθλιψης βασίζεται σε κλινική εξέταση, με τα κριτήρια του DSM-V να αποτελούν τη βασική προσέγγιση [81]. Σύμφωνα με το DSM-V, πέντε ή περισσότερα από τα ακόλουθα συμπτώματα πρέπει να υπάρχουν κατά την ίδια περίοδο δύο εβδομάδων. Τουλάχιστον ένα από τα συμπτώματα πρέπει να είναι είτε καταθλιπτική διάθεση είτε απώλεια ενδιαφέροντος ή ευχαρίστησης, ώστε να ταξινομηθεί ως κατάθλιψη. Συνολικά, υπάρχουν εννέα εξίσου σημαντικά συμπτώματα που αξιολογούν τη διάθεση του ασθενούς, την κόπωση, την απώλεια ενδιαφέροντος και συγκέντρωσης, καθώς και αλλαγές στον ύπνο, την ανησυχία και το βάρος [24, 81].

Αυτή η προσέγγιση βασίζεται στην ικανότητα του ασθενούς να αναφέρει τα συμπτώματά του και να απαντά στις ερωτήσεις του ιατρού. Ωστόσο, αυτές οι αναφορές είναι συχνά υποκειμενικές και μπορεί να επηρεαστούν από διάφορους παράγοντες [66]. Επιπλέον, υποκειμενικοί παράγοντες όπως οι εκφράσεις των ασθενών, μπορούν να περιπλέξουν τη διάγνωση της κατάθλιψης, αυξάνοντας την πιθανότητα λανθασμένης διάγνωσης [66]. Ένα ακόμη μειονέκτημα της τρέχουσας μεθόδου είναι ότι οι κλινικοί γιατροί μπορεί να παραβλέψουν την κατάθλιψη, εκτός εάν ο ασθενής παρουσιάζει σαφή σημάδια θλίψης [81].

Αυτές οι προκλήσεις υπογραμμίζουν την επείγουσα ανάγκη για αντικειμενικές, βασισμένες σε δεδομένα μεθόδους που θα υποστηρίξουν και θα ενισχύσουν τη διάγνωση της

κατάθλιψης. Οι εξελίξεις στις ψηφιακές τεχνολογίες υγείας και στην τεχνητή νοημοσύνη προσφέρουν ελπιδοφόρες προοπτικές για τη συμπλήρωση των παραδοσιακών κλινικών αξιολογήσεων.

1.1 Κίνητρο για τη Συγγραφή της Διπλωματικής Εργασίας

Δεδομένων των προβλημάτων στις τρέχουσες διαγνωστικές πρακτικές, η μηχανική μάθηση αποτελεί μέσο για την ανάπτυξη πιο αντικειμενικών και έγκυρων εργαλείων εκτίμησης της κατάθλιψης.

Στην παρούσα διπλωματική εργασία χρησιμοποιείται το σύνολο δεδομένων DAIC-WOZ για την εξαγωγή χαρακτηριστικών, και δημιουργούνται πολλαπλά σύνολα δεδομένων για την αξιολόγηση της επίδοσης μοντέλων μηχανικής μάθησης στη διάγνωση της κατάθλιψης.

Όσον αφορά την εξαγωγή χαρακτηριστικών, η μελέτη αυτή επικεντρώνεται στην εξαγωγή ηχητικών χαρακτηριστικών χρησιμοποιώντας τη βιβλιοθήκη pyAudioAnalysis, ηχητικών embeddings και γλωσσικών embeddings. Παράλληλα εφαρμόζεται εξαγωγή βάσει ρόλου, όπου οι παραπάνω τύποι χαρακτηριστικών εξάγονται ξεχωριστά τόσο για τον συμμετέχοντα όσο και για τον συνεντευκτή. Αυτή η προσέγγιση αποσκοπεί στην αξιολόγηση της σημασίας κάθε ρόλου στη διαδικασία πρόβλεψης της κατάθλιψης. Ο στόχος δεν είναι μόνο η εξαγωγή αυτών των χαρακτηριστικών, αλλά και η αξιολόγηση του κατά πόσο ο συνδυασμός τους μπορεί να ενισχύσει την ακρίβεια της εκτίμησης της κατάθλιψης.

1.2 Συνεισφορά της Διπλωματικής Εργασίας

Αυτή η διπλωματική εργασία προσφέρει σημαντικές συνεισφορές στον τομέα της αυτόματης εκτίμησης της κατάθλιψης. Οι κύριες συνεισφορές είναι οι εξής:

- Ανάπτυξη Pipeline Εξαγωγής Χαρακτηριστικών.
- Ανάλυση Χαρακτηριστικών βάσει Ρόλου.
- Δημιουργία Διαφορετικών Εκδόσεων Συνόλου Δεδομένων από το DAIC-WOZ.
- Προόδος στην Αυτόματη Εκτίμηση Κατάθλιψης

1.3 Δομή του Ελληνικού Κειμένου της Διπλωματικής Εργασίας

Η παρούσα διπλωματική εργασία οργανώνεται σε πέντε κεφάλαια. Πιο συγκεκριμένα: Στο Κεφάλαιο 1 παρουσιάζεται μια σύντομη εισαγωγή, με στόχο να διευκρινιστεί το ερευνητικό πρόβλημα και τα κίνητρα της μελέτης, ενώ παράλληλα περιγράφεται η συνολική δομή της διπλωματικής εργασίας. Στο Κεφάλαιο 2 γίνεται επισκόπηση της κατάθλιψης καθώς και βασικών μεθόδων της τεχνητής νοημοσύνης που χρησιμοποιούνται στο πλαίσιο αυτής της εργασίας. Στο Κεφάλαιο 3 αναλύονται τα εργαλεία και οι μέθοδοι που χρησιμοποιήθηκαν στην εργασία. Στο Κεφάλαιο 4 περιγράφεται η πειραματική διάταξη και τα αποτελέσματα των πειραμάτων. Στο Κεφάλαιο 5 συνοψίζονται τα συμπεράσματα των πειραμάτων και γίνεται λόγος για μελλοντικές προεκτάσεις της εργασίας.

2 Θεωρητικό Υπόβαθρο και Σχετικές Εργασίες

2.1 Συμπτώματα Κατάθλιψης και Βιοδείκτες

Η κατάθλιψη, μια πολύπλοκη διαταραχή ψυχικής υγείας, επηρεάζει σημαντικά τη συναισθηματική κατάσταση, τις γνωστικές λειτουργίες, τη σωματική ευεξία ενός ατόμου και μπορεί επίσης να εκδηλωθεί στην ομιλία του [34]. Παρατηρούνται αλλοιωμένα πρότυπα ομιλίας σε άτομα με ψυχιατρικές διαταραχές, σημειώνοντας χαρακτηριστικά όπως η μονότονη ομιλία στην κατάθλιψη [54]. Η κατανόηση του τρόπου με τον οποίο η κατάθλιψη αλλοιώνει την ακουστική της φωνής και ο εντοπισμός μετρήσιμων χαρακτηριστικών μπορούν να παρέχουν πολύτιμες πληροφορίες για τη διάγνωση, την αξιολόγηση της σοβαρότητας και την παρακολούθηση της θεραπείας [34, 54].

2.1.1 Συμπτώματα Κατάθλιψης και η Επίδρασή τους στην Ομιλία

Το Διαγνωστικό και Στατιστικό Εγχειρίδιο Ψυχικών Διαταραχών (DSM-5) προσδιορίζει την ψυχοκινητική δυσλειτουργία ως βασικό χαρακτηριστικό της κατάθλιψης. Αυτή η δυσλειτουργία συχνά εκδηλώνεται μέσω μειωμένης έντασης ομιλίας, διακύμανσης, ποικιλίας περιεχομένου και μπορεί ακόμη να συνδέεται με αφωνία [54]. Επιπλέον, η σχέση μεταξύ νευροδιαβιβαστών και φωνητικών χαρακτηριστικών είναι τόσο πολύπλοκη όσο και σημαντική. Οι αλλαγές στη μυϊκή τάση μπορούν να επηρεάσουν τη δυναμική της φωνητικής οδού, περιορίζοντας έτσι τις αρθρωτικές κινήσεις και συμβάλλοντας περαιτέρω στις ανωμαλίες ομιλίας που παρατηρούνται στην κατάθλιψη [54].

Επιπρόσθετα, οι ανισορροπίες στη σεροτονίνη, τη ντοπαμίνη και τη νορεπινεφρίνη διαταράσσουν τη ρύθμιση της διάθεσης και τις γνωστικές λειτουργίες, οι οποίες με τη σειρά τους επηρεάζουν την προσωδία της φωνής και τη ρευστότητα της ομιλίας. Η νευροφλεγμονή και η δυσρύθμιση του άξονα υποθαλάμου-υπόφυσης-επινεφριδίων (HPA) επηρεάζουν το αυτόνομο νευρικό σύστημα, οδηγώντας σε αλλαγές στην τάση των φωνητικών χορδών και στα αναπνευστικά πρότυπα. Αυτές οι φυσιολογικές αλλοιώσεις συχνά οδηγούν σε πιο επίπεδο επιτονισμό, βραδύτερο ρυθμό ομιλίας και αυξημένα σφάλματα άρθρωσης μεταξύ των ατόμων με κατάθλιψη. Ψυχοκοινωνικοί παράγοντες, όπως η κοινωνική απόσυρση και η μειωμένη κινητοποίηση, επιτείνουν περαιτέρω αυτές τις επιδράσεις, καθιστώντας την ομιλία μια πλούσια πηγή βιοδεικτών για την ανίχνευση της κατάθλιψης [51, 77].

2.1.2 Ακουστικά Χαρακτηριστικά που Επηρεάζονται από την Κατάθλιψη

Τα συμπτώματα της κατάθλιψης, μπορούν να μετρηθούν συστηματικά μέσω ακουστικών χαρακτηριστικών όπως η μεταβλητότητα του τόνου, ο ρυθμός ομιλίας, η διάρκεια παύσης και δείκτες ποιότητας φωνής όπως το jitter [52].

Για τον προσδιορισμό των κατάλληλων βιοδεικτών, είναι σημαντικό να εξεταστεί η διαδικασία της ομιλίας. Ο εγκέφαλος οργανώνει προσωδιακές πληροφορίες, παράγει νευρομυϊκές οδηγίες που ελέγχουν τις δραστηριότητες των μυών και των ιστών που σχετίζονται με την κίνηση φώνησης. Στη συνέχεια, η ροή του αέρα από τους πνεύμονες είτε προκαλεί τη δόνηση των φωνητικών χορδών (όταν η γλωττίδα είναι κλειστή) είτε περνά ομαλά μέσα από τη φωνητική χορδή (όταν η γλωττίδα είναι ανοιχτή). Ο στοματοφαρυγγικός μυς σχηματίζει το κύριο κανάλι φώνησης, το οποίο ισοδυναμεί με ένα φίλτρο που μπορεί να

ενισχύσει ή να εξασθενίσει τον ήχο μιας συγκεκριμένης συχνότητας.

2.2 Συστήματα Ανάλυσης Ήχου

Τα συστήματα ανάλυσης ήχου είναι υπολογιστικά εργαλεία που εξάγουν σημαντικά πρότυπα από ηχητικά σήματα συνδυάζοντας επεξεργασία σήματος και μηχανική μάθηση. Τα συστήματα ανάλυσης ήχου χρησιμοποιούν τεχνικές εξαγωγής χαρακτηριστικών για να μετατρέψουν τα ακατέργαστα ηχητικά σήματα σε αναπαραστάσεις υψηλότερου επιπέδου, επιτρέποντας εφαρμογές όπως η ανίχνευση συναισθημάτων, η αναγνώριση ομιλητή και η ταξινόμηση γεγονότων [84, 75].

Βασικά συστατικά των συστημάτων ανάλυσης ήχου συνήθως περιλαμβάνουν την απόκτηση σήματος, την εξαγωγή χαρακτηριστικών και την ενσωμάτωση μηχανικής μάθησης. Η διαδικασία εξαγωγής χαρακτηριστικών περιλαμβάνει την ανάλυση των χρονικών, φασματικών και cepstral τομέων για να συλλάβει μια ολοκληρωμένη αναπαράσταση του ηχητικού σήματος. Αυτά τα εξαγόμενα χαρακτηριστικά χρησιμοποιούνται στη συνέχεια από αλγόριθμους μηχανικής μάθησης για να εντοπίσουν πρότυπα και να κάνουν προβλέψεις ή ταξινομήσεις βάσει του ηχητικού περιεχομένου [84].

2.2.1 Ηχητικά Χαρακτηριστικά για την Κατάθλιψη

Όπως αναφέρθηκε παραπάνω, υπάρχουν αρκετά ακουστικά χαρακτηριστικά που λειτουργούν ως βιοδείκτες κατάθλιψης. Αυτά τα χαρακτηριστικά μπορούν να συσχετιστούν άμεσα με συγκεκριμένα ηχητικά χαρακτηριστικά. Αυτό σημαίνει ότι χρησιμοποιώντας ηχητικά χαρακτηριστικά, μπορούμε να καθιερώσουμε μια σαφή συσχέτιση με την κατάθλιψη και έτσι να την προβλέψουμε χρησιμοποιώντας συστήματα ανάλυσης ήχου [54, 58].

2.2.2 Επεξεργασία Ήχου

2.2.2.1 Βραχυπρόθεσμη Επεξεργασία Ήχου

Η βραχυπρόθεσμη επεξεργασία είναι μια τεχνική στην οποία το ηχητικό σήμα διαιρείται σε μικρά επικαλυπτόμενα ή μη επικαλυπτόμενα πλαίσια (ή παράθυρα), που συνήθως διαρκούν 20–100 χιλιοστά του δευτερολέπτου. Ο διαχωρισμός σε παράθυρα είναι σημαντικός επειδή τα ηχητικά σήματα δεν είναι στατικά στο χρόνο, αντιθέτως αυτή η τμηματοποίηση υποθέτει ότι το σήμα παραμένει στάσιμο εντός κάθε πλαισίου, που σημαίνει ότι οι στατιστικές του ιδιότητες δεν αλλάζουν σημαντικά κατά τη διάρκεια αυτής της σύντομης περιόδου [85].

2.2.2.2 Μεσοπρόθεσμη Επεξεργασία Ήχου

Στη μεσοπρόθεσμη επεξεργασία, το ηχητικό σήμα αρχικά διαιρείται σε μεγαλύτερα τμήματα, που αναφέρονται ως μεσοπρόθεσμα παράθυρα, τα οποία συνήθως κυμαίνονται σε διάρκεια από 1 έως 10 δευτερόλεπτα. Κάθε μεσοπρόθεσμο τμήμα υποβάλλεται σε βραχυπρόθεσμη επεξεργασία για την εξαγωγή χαρακτηριστικών. Αυτά τα μεσοπρόθεσμα παράθυρα χαρακτηρίζονται από ομοιογένεια στη συμπεριφορά τους, καθιστώντας κατάλληλο τον υπολογισμό στατιστικών χαρακτηριστικών σε βάση τμήματος προς τμήμα.

2.2.3 Ηχητικά Χαρακτηριστικά

Ένα ηχητικό σήμα είναι ένας τύπος σήματος που μεταφέρει πληροφορίες εντός του εύρους των ηχητικών συχνοτήτων. Η αναπαράσταση ήχου περιλαμβάνει την εξαγωγή βασικών ιδιοτήτων ή χαρακτηριστικών ενός ηχητικού σήματος που αντικατοπτρίζουν την ακουστική του σύνθεση—τόσο στο πεδίο του χρόνου όσο και στο πεδίο της συχνότητας—καθώς και τη συμπεριφορά του με την πάροδο του χρόνου. Αυτή η διαδικασία συνήθως συνδυάζεται με την επιλογή χαρακτηριστικών, η οποία προσδιορίζει τα καταλληλότερα χαρακτηριστικά για την προβλεπόμενη εφαρμογή του ηχητικού σήματος. Ο κύριος στόχος είναι να εξαχθούν χαρακτηριστικά από ηχητικά δεδομένα (όπως η ομιλία) που μπορούν να παρέχουν πολύτιμες πληροφορίες για την εκπαίδευση μοντέλων.

Το ηχητικό σήμα διαιρείται σε βραχυπρόθεσμα παράθυρα, και συγκεκριμένα χαρακτηριστικά υπολογίζονται για κάθε παράθυρο. Από αυτά, υπολογίζονται στατιστικές τιμές σε μεσοπρόθεσμα παράθυρα για να συνοψιστούν οι ιδιότητες του σήματος. Υπάρχουν πολυάριθμες μετρικές που μπορούν να χρησιμοποιηθούν ως χαρακτηριστικά στην ανάλυση ήχου, και αυτή η ενότητα περιγράφει σύντομα ορισμένα από τα χαρακτηριστικά που χρησιμοποιούνται στο σχεδιασμό συστημάτων.

2.2.3.1 Χαρακτηριστικά pyAudioAnalysis

Το pyAudioAnalysis αποτελεί μία βιβλιοθήκη Python σχεδιασμένη για εργασίες ανάλυσης ήχου, όπως εξαγωγή χαρακτηριστικών, τμηματοποίηση, ταξινόμηση και οπτικοποίηση. Υλοποιεί τόσο βραχυπρόθεσμες όσο και μεσοπρόθεσμες μεθοδολογίες επεξεργασίας. Η βιβλιοθήκη υποστηρίζει διάφορες εργασίες ανάλυσης ήχου, συμπεριλαμβανομένης της εξαγωγής χαρακτηριστικών από το πεδίο του χρόνου και της συχνότητας, της ταξινόμησης, της παλινδρόμησης και της τμηματοποίησης [83].

Είναι επίσης σημαντικό να σημειωθεί ότι υπάρχει πληθώρα βιβλιοθηκών εξαγωγής χαρακτηριστικών ήχου διαθέσιμες, καθεμία με τα δικά της πλεονεκτήματα και περιορισμούς. Το pyAudioAnalysis, το οποίο χρησιμοποιείται στη συγκεκριμένη εργασία, προσφέρει ισορροπία μεταξύ ευκολίας χρήσης και ενός ισχυρού συνόλου χαρακτηριστικών σχεδιασμένων για ανάλυση ομιλίας και μουσικής, καθιστώντας το κατάλληλο για εργασίες εκτίμησης κατάθλιψης. Ωστόσο, κάθε βιβλιοθήκη παρουσιάζει προκλήσεις ως προς την υπολογιστική αποδοτικότητα, την κλιμακωσιμότητα και τη συμβατότητα με δεδομένα πραγματικού κόσμου που περιέχουν θόρυβο.

2.2.3.2 Προεκπαιδευμένα Embeddings Ήχου

Η διαδικασία ανάλυσης ήχου, όπως αναφέρθηκε προηγουμένως, περιλαμβάνει ένα ευρύ φάσμα εργασιών όπως αναγνώριση ομιλίας, ταυτοποίηση ομιλητή και αναγνώριση συναισθημάτων. Κεντρικής σημασίας για όλες αυτές τις εργασίες είναι η ανάγκη για αποτελεσματικές αναπαραστάσεις ήχου. Παραδοσιακά, χρησιμοποιούνται handcrafted χαρακτηριστικά ήχου, όπως αυτά που περιγράφηκαν στην προηγούμενη ενότητα.

Αν και τα handcrafted χαρακτηριστικά έχουν επιτύχει αξιοσημείωτα αποτελέσματα, συνοδεύονται από σημαντικούς περιορισμούς. Αρχικά, απαιτούν εκτεταμένη χειροκίνητη εργασία και λεπτομερή ρύθμιση για προσαρμογή σε διαφορετικές εργασίες ήχου. Επιπλέον, τα χαρακτηριστικά αυτά αποτυπώνουν κυρίως λεπτομέρειες χαμηλού επιπέδου και συχνά

αποτυγχάνουν να αναπαραστήσουν πληροφορίες υψηλότερου επιπέδου, όπως συναισθήματα ή το περιβάλλον. Επιπλέον, τα handcrafted χαρακτηριστικά μπορεί να είναι ευαίσθητα σε θόρυβο και συνθήκες ηχογράφησης, γεγονός που μειώνει την ανθεκτικότητά τους.

Για την αντιμετώπιση αυτών των αδυναμιών, οι ερευνητές χρησιμοποιούν ολοένα και περισσότερο προεκπαιδευμένα embeddings ήχου. Αυτά τα embeddings είναι διανυσματικές αναπαραστάσεις που παράγονται από βαθιά νευρωνικά δίκτυα εκπαιδευμένα σε μεγάλες και ποικίλες συλλογές δεδομένων ήχου, συχνά με εποπτευόμενες ή αυτοεποπτευόμενες μεθόδους μάθησης. Σε αντίθεση με τα handcrafted χαρακτηριστικά, τα embeddings μαθαίνονται αυτόματα, επιτρέποντας στο μοντέλο να ανακαλύψει βέλτιστες αναπαραστάσεις για τη σύλληψη σύνθετων ακουστικών και σημασιολογικών μοτίβων [49].

wav2vec 2.0: Τα προεκπαιδευμένα embeddings ήχου έχουν μεγάλη αλλαγή στην επεξεργασία ακουστικών σημάτων. Ανάμεσα στα διάφορα embeddings, το wav2vec 2.0 ξεχωρίζει ως ένα πρωτοποριακό μοντέλο που μαθαίνει αναπαραστάσεις ομιλίας απευθείας από ακατέργαστα κυματομορφικά σήματα μέσω αυτο-εποπτευόμενης μάθησης. Σε αντίθεση με τα παραδοσιακά handcrafted χαρακτηριστικά ή τα παλαιότερα embeddings νευρωνικών δικτύων, το wav2vec 2.0 συλλαμβάνει αποτελεσματικά τόσο χαμηλού επιπέδου ακουστικά μοτίβα όσο και υψηλότερου επιπέδου γλωσσικές δομές, οδηγώντας σε καλύτερα αποτελέσματα σε ποικίλες εργασίες που σχετίζονται με την ομιλία [33].

Η βασική ιδέα πίσω από το wav2vec 2.0 είναι η κωδικοποίηση ακατέργαστου ήχου ομιλίας μέσω ενός πολυεπίπεδου συνελικτικού νευρωνικού δικτύου, που παράγει λανθάνουσες αναπαραστάσεις ομιλίας. Τμήματα αυτών των λανθανόντων χαρακτηριστικών αποκρύπτονται, παρόμοια με το masked language modeling στην επεξεργασία φυσικής γλώσσας. Οι αποκρυμμένες λανθάνουσες αναπαραστάσεις περνούν από ένα δίκτυο Transformer, το οποίο κατασκευάζει συμφραζόμενα embeddings, συλλαμβάνοντας εξαρτήσεις σε όλη τη χρονική ακολουθία. Το μοντέλο εκπαιδεύεται με αντιθετική απώλεια, όπου πρέπει να αναγνωρίσει τη σωστή λανθάνουσα αναπαράσταση μεταξύ πολλών παραπλανητικών, ενισχύοντας έτσι την εκμάθηση ουσιαστικών και διακριτικών χαρακτηριστικών [33].

2.3 Συστήματα Ανάλυσης Κειμένου

Παρόμοια με τα συστήματα ανάλυσης ήχου, υπάρχουν δύο κύριες εργασίες στα συστήματα ανάλυσης κειμένου. Η πρώτη είναι η προεπεξεργασία κειμένου και η δημιουργία embeddings κειμένου.

2.3.1 Προεπεξεργασία Κειμένου

Η προεπεξεργασία κειμένου αποτελεί ένα κρίσιμο βήμα, στο οποίο μετατρέπεται το ακατέργαστο, μη δομημένο κείμενο σε μια δομημένη μορφή κατάλληλη για ανάλυση και μοντελοποίηση. Αυτή η διαδικασία περιλαμβάνει διάφορες τεχνικές για την αφαίρεση θορύβου και ασυνεπειών, καθιστώντας τα δεδομένα πιο ομοιόμορφα και διαχειρίσιμα για τα μοντέλα επεξεργασίας φυσικής γλώσσας (NLP).

2.3.2 Embeddings Κειμένου

Μετά την προεπεξεργασία, το επόμενο βήμα είναι η εξαγωγή embeddings, που μπορεί να γίνει είτε μέσω embeddings λέξεων είτε μέσω embeddings προτάσεων.

- **Embeddings Λέξεων**: Τεχνικές όπως οι Word2Vec και GloVe δημιουργούν embeddings σε επίπεδο λέξης με βάση τα συμφραζόμενα. Το Word2Vec χρησιμοποιεί δειγματοληψία παραθύρων κειμένου για τη δημιουργία embeddings για μεμονωμένες λέξεις, ενώ το GloVe χρησιμοποιεί παγκόσμια παραγοντοποίηση μητρώου.
- **Embeddings Προτάσεων**: Μοντέλα όπως τα BERT, SBERT και τα μοντέλα text-embedding της OpenAI παράγουν embeddings για ολόκληρες προτάσεις ή έγγραφα, αποτυπώνοντας αποτελεσματικά τα συμφραζόμενα.

2.3.3 GloVe Embeddings

2.3.3.1 Εισαγωγή

Το GloVe (Global Vectors for Word Representation) είναι μια ευρέως χρησιμοποιούμενη μέθοδος εκμάθησης embeddings λέξεων (πυκνών διανυσματικών αναπαραστάσεων λέξεων) αξιοποιώντας τα παγκόσμια στατιστικά συν-εμφάνισης λέξεων από ένα σώμα κειμένου. Σε αντίθεση με παλαιότερες μεθόδους που εστιάζουν είτε στην παγκόσμια παραγοντοποίηση μητρώου (όπως το LSA) είτε σε τοπικές προβλέψεις παραθύρων συμφραζομένων (όπως το word2vec), το GloVe συνδυάζει αποτελεσματικά τα πλεονεκτήματα και των δύο προσεγγίσεων ώστε να παράγει embeddings που αποτυπώνουν τόσο τα παγκόσμια στατιστικά όσο και τη γραμμική υποδομή της σημασίας των λέξεων.

2.3.3.2 Θεωρητικό Υπόβαθρο

Η βασική ιδέα πίσω από το GloVe είναι ότι η σημασία μιας λέξης μπορεί να αποτυπωθεί εξετάζοντας τη συχνότητα με την οποία συν-εμφανίζεται με άλλες λέξεις σε ένα μεγάλο σώμα κειμένου. Συγκεκριμένα, το GloVe μοντελοποιεί τα πηλίκα πιθανοτήτων συν-εμφάνισης. Έστω δύο λέξεις-στόχοι, i και j, και μία λέξη-συμφραζόμενο k. Το πηλίκο των πιθανοτήτων ότι το k εμφανίζεται στο συμφραζόμενο του i έναντι του j (P_{ik}/P_{jk}) μπορεί να αναδείξει πτυχές της σημασίας που διαφοροποιούν το i από το j.

Για παράδειγμα, το πηλίκο της πιθανότητας ότι η λέξη «στερεό» εμφανίζεται με τη λέξη «πάγος» έναντι της «ατμός» είναι πολύ μεγαλύτερο από το ένα, υποδεικνύοντας ισχυρή συσχέτιση με τον «πάγο». Αντίθετα, το πηλίκο για τη λέξη «αέριο» είναι πολύ μικρότερο της μοναδας, υποδεικνύοντας ισχυρή συσχέτιση με τον «ατμό». Πηλίκα κοντά στο ένα (π.χ. για το «νερό») δείχνουν λέξεις εξίσου σχετικές και με τις δύο.

2.3.3.3 Διατύπωση του Μοντέλου

Έστω $\mathbf{X}$ η μήτρα συν-εμφάνισης λέξεων, όπου X_{ij} είναι ο αριθμός φορών που η λέξη j εμφανίζεται στο συμφραζόμενο της λέξης i. Η πιθανότητα ότι η λέξη k εμφανίζεται στο συμφραζόμενο της λέξης i είναι:

$$P_{ik} = \frac{X_{ik}}{X_i}$$

όπου:

$$X_i = \sum_k X_{ik}$$

είναι το συνολικό πλήθος εμφανίσεων συμφραζόμενων για τη λέξη i.

Το GloVe επιδιώκει να βρει διανύσματα λέξεων $\mathbf{w}_i$ και διανύσματα λέξεων-συμφραζομένων $\tilde{\mathbf{w}}_k$ ώστε το εσωτερικό τους γινόμενο, συν τους όρους μετατόπισης, να προσεγγίζει τον λογάριθμο του πλήθους συν-εμφάνισης:

$$\mathbf{w}_i^\top \tilde{\mathbf{w}}_k + b_i + \tilde{b}_k \approx \log(X_{ik})$$

Ο παραπάνω τύπος αποσκοπεί στο να μπορεί να γίνει ανταλλαγή ρόλων λέξης και συμφραζομένου, και στην κωδικοποίηση των γραμμικών σχέσεων που συναντώνται σε αναλογίες λέξεων.

2.3.3.4 Διαδικασία Εκπαίδευσης

Το μοντέλο GloVe εκπαιδεύεται με ελαχιστοποίηση της συνάρτησης

$$J = \sum_{i,j=1}^{V} f(X_{ij}) \left(\mathbf{w}_i^\top \tilde{\mathbf{w}}_j + b_i + \tilde{b}_j - \log(X_{ij}) \right)^2$$

ως προς όλα τα διανύσματα λέξεων και συμφραζομένων και τους όρους μετατόπισης, χρησιμοποιώντας στοχαστική καθοδική κλίση ή παρόμοιες μεθόδους βελτιστοποίησης. Η εκπαίδευση επαναλαμβάνεται πάνω στις μη μηδενικές τιμές της μήτρας συν-εμφανίσεων, ενημερώνοντας τα διανύσματα και τους όρους ώστε να προσαρμόζονται καλύτερα στους παρατηρούμενους λογαρίθμους των συν-εμφανίσεων.

Μετά την εκπαίδευση, το τελικό embedding κάθε λέξης μπορεί να ληφθεί ως το άθροισμα (ή ο μέσος όρος) των διανυσμάτων λέξης και συμφραζομένου.

2.3.3.5 Ιδιότητες και Πλεονεκτήματα

- **Γραμμική Υποδομή:** Τα GloVe embeddings αποτυπώνουν γραμμικές σχέσεις, το οποίο τα καθιστά κατάλληλα για αναλογικές εργασίες (π.χ. «βασιλιάς» − «άντρας» + «γυναίκα» ≈ «βασίλισσα»).
- **Αποδοτική Χρήση Στατιστικών:** Εστιάζοντας στις μη μηδενικές συν-εμφανίσεις και σταθμίζοντάς τες κατάλληλα, το GloVe αξιοποιεί αποδοτικά τα στατιστικά δεδομένα μεγάλων σωμάτων κειμένου.
- **Κλιμάκωση:** Το μοντέλο μπορεί να εκπαιδευτεί σε πολύ μεγάλα σώματα κειμένου, παράγοντας embeddings υψηλής ποιότητας για εκτεταμένα λεξιλόγια.

Τα GloVe embeddings αποτελούν έναν ισχυρό και αποδοτικό τρόπο εκμάθησης αναπαραστάσεων λέξεων. Κωδικοποιούν τόσο σημασιολογικές όσο και συντακτικές κανονικότητες. Με την άμεση μοντελοποίηση των παγκόσμιων στατιστικών συν-εμφάνισης, το GloVe παράγει διανυσματικούς χώρους με ουσιαστική υποδομή, ξεπερνώντας πολλές προηγούμενες μεθόδους σε εργασίες όπως η ομοιότητα και οι αναλογίες λέξεων. Τα embeddings που προκύπτουν αποτελούν πλέον βασικό εργαλείο σε σύγχρονες εφαρμογές NLP [48].

2.3.4 SBERT Embeddings

2.3.4.1 Εισαγωγή στα BERT embeddings

Το BERT (Bidirectional Encoder Representations from Transformers) είναι ένα πρωτοποριακό μοντέλο που εισήγαγε η Google το 2018 και άλλαξε τα δεδομένα στην επεξεργασία φυσικής γλώσσας, προσφέροντας βαθιά, embeddings λέξεων βασισμένα στα συμφραζόμενα. Σε αντίθεση με προηγούμενα μοντέλα που διαβάζουν το κείμενο είτε από αριστερά προς τα δεξιά είτε από δεξιά προς τα αριστερά, το BERT διαβάζει το κείμενο αμφίδρομα, λαμβάνοντας υπόψη τα συμφραζόμενα πριν και μετά τη λέξη ταυτόχρονα. Αυτή η αμφίδρομη προσέγγιση επιτρέπει στο BERT να αποτυπώνει πιο λεπτές σημασίες και σχέσεις στη γλώσσα. Το BERT χρησιμοποιεί αρχιτεκτονική μετασχηματιστή (transformer) που εφαρμόζει μηχανισμούς αυτοπροσοχής (self-attention) για να σταθμίζει τη σημασία κάθε λέξης σε μια πρόταση σε σχέση με τις υπόλοιπες. Παράγει αναπαραστάσεις λέξεων που αλλάζουν ανάλογα με τα συμφραζόμενα, σε αντίθεση με στατικές embeddings όπως το Word2Vec ή το GloVe. Το BERT προεκπαιδεύεται σε μεγάλα σώματα κειμένου με μη επιβλεπόμενες εργασίες όπως η μάσκα λέξεων (masked language modeling) και η πρόβλεψη επόμενης πρότασης (next sentence prediction), και στη συνέχεια εξειδικεύεται σε συγκεκριμένες εργασίες NLP όπως η απάντηση σε ερωτήσεις, η ανάλυση συναισθήματος και η αναγνώριση οντοτήτων.

Ωστόσο, ενώ το BERT διαπρέπει στην κατανόηση συμφραζομένων σε επίπεδο λέξης, δεν έχει σχεδιαστεί αρχικά για να παράγει embeddings σταθερού μεγέθους για προτάσεις που μπορούν να συγκριθούν απευθείας. Για εργασίες όπως η σημασιολογική ομοιότητα ή η ομαδοποίηση, απαιτείται μία μοναδική διανυσματική αναπαράσταση ανά πρόταση. Το πρόβλημα προκύπτει επειδή για να συγκριθούν δύο προτάσεις με το BERT, πρέπει να εισαχθούν μαζί ως ζεύγος, κάτι που οδηγεί σε υψηλό υπολογιστικό κόστος.

2.3.4.2 Εισαγωγή στα SBERT embeddings

Το Sentence-BERT (SBERT) αντιμετωπίζει το πρόβλημα των BERT embeddings, τροποποιώντας την αρχιτεκτονική του BERT ώστε να δημιουργεί μια διπλή (ή δίδυμη) δομή δικτύου. Έτσι κάθε πρόταση κωδικοποιείται ανεξάρτητα σε ένα διάνυσμα σταθερού μεγέθους. Αυτό σημαίνει ότι κάθε πρόταση περνά ξεχωριστά από το δίκτυο, παράγοντας ένα διάνυσμα σταθερού μήκους που αναπαριστά το νόημα της πρότασης. Αυτές οι embeddings προτάσεων μπορούν στη συνέχεια να συγκριθούν αποδοτικά χρησιμοποιώντας απλά μέτρα ομοιότητας, όπως η ομοιότητα συνημιτόνου.

Η αρχιτεκτονική φαίνεται στην παρακάτω εικόνα:

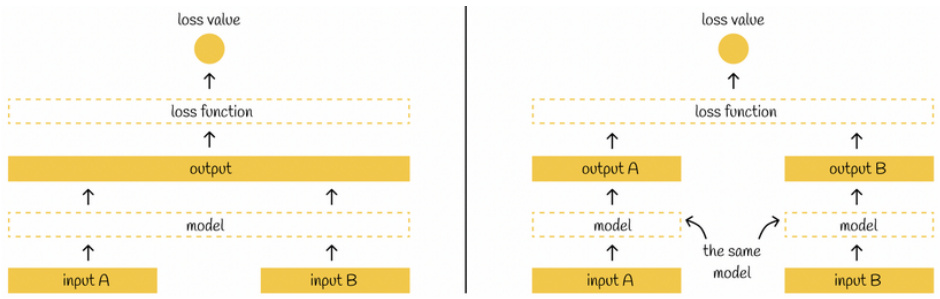

Figure 1: Το BERT μοντέλο, αριστερό, επεξεργάζει και τις δύο εισόδους ταυτόχρονα. Αντίθετα, το μοντέλο διπλού-κωδικοποιητή (SBERT), δεξιά, χειρίζεται τις εισόδους ανεξάρτητα και παράλληλα, έτσι κάθε έξοδος παράγεται ανεξάρτητα [88].

2.4 Μοντέλα και Έννοιες Μηχανικής Μάθησης

2.4.1 Μοντέλα Μηχανικής Μάθησης

2.4.1.1 Support Vector Machines (SVM)

Τα SVM είναι επιβλεπόμενα μοντέλα ταξινόμησης που στοχεύουν στο να βρουν το βέλτιστο υπερεπίπεδο (hyperplane), το οποίο θα διαχωρίζει τις κλάσεις έτσι ώστε να έχουν το μέγιστο περιθώριο (margin). Είναι ιδιαίτερα αποτελεσματικά σε προβλήματα με υψηλή διάσταση και μπορούν να χρησιμοποιούν πυρήνες (kernels) για να αντιμετωπίσουν μη γραμμικά προβλήματα.

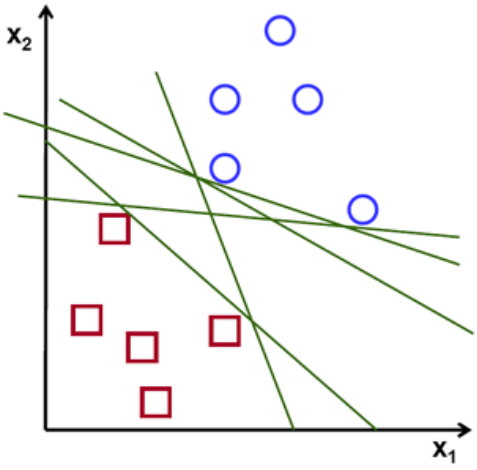

Figure 2: Αποδεκτό hyperplanes

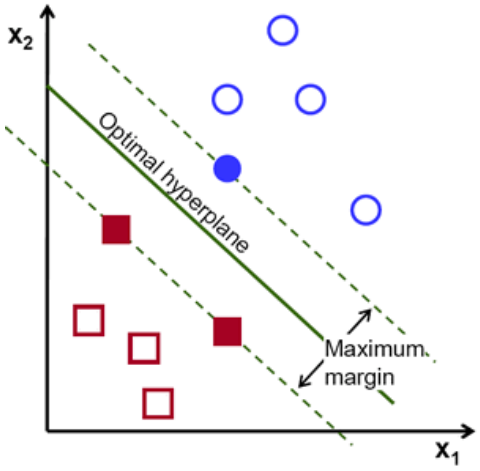

Figure 3: Βέλτιστο hyperplane

2.4.1.2 Gradient Boosting

Το gradient boosting είναι μια τεχνική μηχανικής μάθησης που κατασκευάζει ένα ισχυρό προγνωστικό μοντέλο συνδυάζοντας πολλαπλά αδύναμα μοντέλα, συνήθως δέντρα αποφάσεων, σε ένα ισχυρότερο σύνολο (ensemble). Ως αδύναμα μοντέλα ορίζονται συνήθως μοντέλα που αποδίδουν ελαφρώς καλύτερα από την τυχαία πρόβλεψη [65]. Αυτή η μέθοδος εστιάζει στη διόρθωση των σφαλμάτων των προηγούμενων μοντέλων μέσω της βελτιστοποίησης μιας συγκεκριμένης συνάρτησης απώλειας χρησιμοποιώντας gradient descent [25, 60].

Κύρια στοιχεία του gradient boosting περιλαμβάνουν:

- Αδύναμοι Μαθητές (Weak Learners): Τα δέντρα αποφάσεων χρησιμοποιούνται συχνά λόγω της απλότητάς τους και της ικανότητάς τους να μοντελοποιούν αποτελεσματικά μη γραμμικές σχέσεις [65].
- Προσθετικό Μοντέλο (Additive Model): Οι προβλέψεις από όλους τους αδύναμους μαθητές αθροίζονται για να σχηματίσουν το τελικό αποτέλεσμα, με κάθε νέο μαθητή να εκπαιδεύεται πάνω στα υπολείμματα των προηγούμενων βημάτων [60].
- Συνάρτηση Απώλειας (Loss Function): Η επιλογή της συνάρτησης απώλειας εξαρτάται από τον τύπο του προβλήματος και πρέπει να είναι διαφορίσιμη για να διευκολύνει τη βελτιστοποίηση [22].

2.4.1.3 XGBoost

Το XGBoost είναι μια υλοποίηση του αλγορίθμου gradient boosting που χρησιμοποιεί δέντρα απόφασης ως weak learners. Το XGBoost έχει γίνει γνωστό για την αποδοτικότητά του στην επεξεργασία μεγάλων συνόλων δεδομένων. Ενσωματώνει αρκετά βασικά χαρακτηριστικά που το διακρίνουν από άλλους αλγορίθμους gradient boosting, όπως η ικανότητά του να χειρίζεται αποδοτικά αραιά δεδομένα χρησιμοποιώντας τον αλγόριθμο weighted quantile sketch.Συνολικά, ο συνδυασμός ταχύτητας, κλιμάκωσης και αξιόπιστης απόδοσης έχει καταστήσει το XGBoost δημοφιλή επιλογή σε διαγωνισμούς μηχανικής μάθησης και εφαρμογές στον πραγματικό κόσμο.

2.4.2 Cross-Validation

Το cross-validation αποτελεί μια θεμελιώδη τεχνική στη μηχανική μάθηση για την αξιολόγηση της απόδοσης των μοντέλων και την αποφυγή υπερεκπαίδευσης (overfitting). Αντί να βασίζεται σε έναν μόνο διαχωρισμό εκπαίδευσης-δοκιμής (π.χ., 70% εκπαίδευση και 30% δοκιμή), το cross-validation διαχωρίζει τα δεδομένα έτσι ώστε να εξασφαλίσει αξιόπιστη αξιολόγηση. Η πιο διαδεδομένη μέθοδος είναι η k-fold cross-validation, όπου τα δεδομένα εκπαίδευσης χωρίζονται σε k ισάριθμα μέρη. Σε κάθε επανάληψη, ένα μέρος χρησιμοποιείται ως σύνολο επικύρωσης, ενώ τα υπόλοιπα k-1 μέρη χρησιμοποιούνται για εκπαίδευση. Αυτή η διαδικασία επαναλαμβάνεται k φορές, με κάθε μέρος να έχει λειτουργήσει ως σύνολο επικύρωσης μία φορά. Η απόδοση του μοντέλου καταγράφεται σε κάθε επανάληψη και ο τελικός δείκτης απόδοσης είναι ο μέσος όρος όλων των πτυχών. Παρόλο που η k-fold cross-validation προσφέρει πιο αξιόπιστη εκτίμηση της ικανότητας γενίκευσης ενός μοντέλου, μπορεί να είναι υπολογιστικά απαιτητική λόγω της επαναληπτικής φύσης της [78, 4].

2.4.2.1 Leave-One-Out Cross-Validation

Η Leave-One-Out Cross-Validation (LOOCV) είναι μια εξειδικευμένη τεχνική όπου κάθε μεμονωμένο δείγμα του συνόλου δεδομένων χρησιμοποιείται μία φορά μόνο του ως σύνολο επικύρωσης, ενώ τα υπόλοιπα δεδομένα σχηματίζουν το σύνολο εκπαίδευσης. Αυτή η διαδικασία επαναλαμβάνεται n φορές (για n συνολικά δείγματα), διασφαλίζοντας ότι κάθε δείγμα χρησιμοποιείται ακριβώς μία φορά για επικύρωση [62].

2.4.3 Ρύθμιση Υπερπαραμέτρων

Η ρύθμιση υπερπαραμέτρων περιλαμβάνει την επιλογή των βέλτιστων τιμών για τις υπερπαραμέτρους ενός μοντέλου μηχανικής μάθησης. Αυτή η διαδικασία περιλαμβάνει συνήθως τον καθορισμό ενός εύρους πιθανών τιμών για κάθε υπερπαράμετρο, την εκπαίδευση του μοντέλου με διαφορετικούς συνδυασμούς και την αξιολόγηση της απόδοσης σε ένα σύνολο επικύρωσης. Ο στόχος είναι να βρεθεί μια ισορροπία που αποφεύγει την υποεκπαίδευση και την υπερεκπαίδευση. Η ρύθμιση των υπερπαραμέτρων μπορεί να γίνει handcrafted, βασιζόμενη στη διαίσθηση και την παρατήρηση, ή αυτόματα με τη χρήση συστηματικών μεθόδων αναζήτησης. Οι καλύτερες στρατηγικές για τη ρύθμιση υπερπαραμέτρων είναι:

Το gridsearch είναι η μέθοδος που επιλέχθηκε για το πειραματικό μέρος αυτής της εργασίας. Το gridsearch είναι μια τεχνική ρύθμισης υπερπαραμέτρων που εκτελεί συστηματικά εξαντλητική αναζήτηση σε ένα προκαθορισμένο πλέγμα υπερπαραμέτρων. Σε αυτό το πλαίσιο, το πλέγμα αντιπροσωπεύει όλους τους δυνατούς συνδυασμούς υπερπαραμέτρων και των αντίστοιχων τιμών τους. Το grid search αξιολογεί κάθε συνδυασμό αυτών των υπερπαραμέτρων, που αντιστοιχεί σε ένα "σημείο" του πλέγματος, για να εντοπίσει το σύνολο που προσφέρει τη βέλτιστη απόδοση του μοντέλου, συνήθως μετρώντας τη χρήση διασταυρούμενης επικύρωσης [70].

2.4.4 Υποδειγματοληψία και Υπερδειγματοληψία

Τα ανισοκατανεμημένα σύνολα δεδομένων αποτελούν ένα διαχρονικό πρόβλημα στη μηχανική μάθηση, όπου η κατανομή των κλάσεων είναι έντονα λοξή, συχνά οδηγώντας σε μοντέλα με χαμηλή απόδοση στις μειοψηφικές κλάσεις [63]. Για την αντιμετώπιση αυτού, η τυχαία υπερδειγματοληψία και υποδειγματοληψία είναι δύο βασικές τεχνικές για την επαναπροσαρμογή της κατανομής των κλάσεων πριν από την εκπαίδευση των μοντέλων. Αυτές οι μέθοδοι στοχεύουν να μειώσουν την μεροληψία που εισάγουν τα ανισοκατανεμημένα δεδομένα, διασφαλίζοντας ότι τα μοντέλα μπορούν να μάθουν αποτελεσματικά από όλες τις κλάσεις [10].

2.4.4.1 Τυχαία Υποδειγματοληψία

Η τυχαία υποδειγματοληψία εχει ως στόχο τη μείωση του αριθμού των δεδομένων της πλειοψηφικής κλάσης με τυχαία αφαίρεση δειγμάτων μέχρι να επιτευχθεί η επιθυμητή ισορροπία κλάσεων [63]. Ο βαθμός υποδειγματοληψίας μπορεί να ρυθμιστεί για συγκεκριμένες αναλογίες κλάσεων. Για παράδειγμα, μια αναλογία 1:1 διασφαλίζει ότι η πλειοψηφική κλάση έχει τον ίδιο αριθμό περιπτώσεων με τη μειοψηφική, ενώ μια αναλογία 0,5 ορίζει την πλειοψηφική κλάση στο μισό μέγεθος της μειοψηφικής [63].

Η υποδειγματοληψία μειώνει την μεροληψία προς την πλειοψηφική κλάση, επιτρέποντας στο μοντέλο να εστιάζει πιο αποτελεσματικά στη μειοψηφική κλάση [63]. Μειώνοντας το συνολικό μέγεθος του συνόλου δεδομένων, η υποδειγματοληψία μπορεί να επιταχύνει σημαντικά την εκπαίδευση και να μειώσει τις απαιτήσεις σε υπολογιστικούς πόρους, ιδιαίτερα χρήσιμη για πολύ μεγάλα σύνολα δεδομένων ή περιορισμένη υπολογιστική ικανότητα [23].

2.4.4.2 Τυχαία Υπερδειγματοληψία

Η τυχαία υπερδειγματοληψία περιλαμβάνει την αύξηση του αριθμού των περιπτώσεων της μειοψηφικής κλάσης με αντιγραφή υπαρχόντων παραδειγμάτων μέχρι να επιτευχθεί η επιθυμητή ισορροπία κλάσεων. Όπως και στην υποδειγματοληψία, ο βαθμός υπερδειγματοληψίας μπορεί να ρυθμιστεί για συγκεκριμένες αναλογίες κλάσεων [63]. Σε αντίθεση με προηγμένες μεθόδους όπως το SMOTE, η τυχαία υπερδειγματοληψία δεν δημιουργεί νέα συνθετικά παραδείγματα αλλά βασίζεται αποκλειστικά σε αντιγραφή [10].

2.4.4.3 Υπερδειγματοληψία SMOTE

Το SMOTE (Synthetic Minority Oversampling Technique) είναι μια ευρέως χρησιμοποιούμενη μέθοδος υπερδειγματοληψίας για την αντιμετώπιση ανισοκατανεμημένων συνόλων δεδομένων [19]. Το SMOTE λειτουργεί συνθέτοντας νέα παραδείγματα μειοψηφικής κλάσης βασισμένα στον χώρο χαρακτηριστικών των υπαρχόντων δεδομένων. Ο αλγόριθμος επιλέγει τυχαία ενα δείγμα μειοψηφικής κλάσης και προσδιορίζει τους k πλησιέστερους γείτονές της (συνήθως k = 5). Ο πραγματικός αριθμός γειτόνων που χρησιμοποιούνται για τη δημιουργία συνθετικών δειγμάτων εξαρτάται από το απαιτούμενο ποσοστό υπερδειγματοληψίας, για παράδειγμα, για διπλασιασμό της μειοψηφικής κλάσης (100% υπερδειγματοληψία), χρησιμοποιείται ένας γείτονας ανά περίπτωση, ενώ υψηλότερα ποσοστά απαιτούν περισσότερους γείτονες, μερικές φορές με δειγματοληψία με αντικατάσταση εάν ο απαιτούμενος αριθμός υπερβαίνει το k [64]. Το επόμενο βήμα είναι η δημιουργία ενός συνθετικού παραδείγματος με παρεμβολή μεταξύ της επιλεγμένης περίπτωσης και ενός από τους γείτονές της, τοποθετώντας το νέο δείγμα κατά μήκος της γραμμής που τους συνδέει στον χώρο χαρακτηριστικών [64]. Αυτή η παρεμβολή πραγματοποιείται υπολογίζοντας τη διαφορά μεταξύ των διανυσμάτων χαρακτηριστικών της επιλεγμένης περίπτωσης και του γείτονά της, πολλαπλασιάζοντας αυτή τη διαφορά με έναν τυχαίο αριθμό μεταξύ 0 και 1 και προσθέτοντας το αποτέλεσμα στο διάνυσμα της αρχικής περίπτωσης. Αυτή η διαδικασία επαναλαμβάνεται μέχρι να επιτευχθεί η επιθυμητή ισορροπία κλάσεων [11].

2.5 Συνοπτική Επισκόπηση Σχετικής Βιβλιογραφίας

Η αυτόματη αναγνώριση κατάθλιψης έχει προσελκύσει έντονο ενδιαφέρον, καθώς η ομιλία αποτελεί έναν μη επεμβατικό, οικονομικό και απομακρυσμένο βιοδείκτη για την ανίχνευση της διαταραχής. Η έρευνα έχει εξελιχθεί από παραδοσιακές μεθόδους με handcrafted χαρακτηριστικά και κλασικούς αλγορίθμους μηχανικής μάθησης (όπως SVM, GMM, HMM) σε βαθιά νευρωνικά δίκτυα που μαθαίνουν αυτόματα χαρακτηριστικά από το ωμό σήμα ή το φασματογράφημα [55].

2.5.1 Handcrafted Χαρακτηριστικά και Παραδοσιακή Μηχανική Μάθηση

Στα πρώτα στάδια, η ανάλυση επικεντρώθηκε σε ακουστικά χαρακτηριστικά (π.χ. τονικότητα, ενέργεια, φασματικές ιδιότητες) και χρησιμοποιήθηκαν αλγόριθμοι όπως SVM, GMM και δέντρα απόφασης. Τα αποτελέσματα σε γνωστά σύνολα δεδομένων (π.χ. DAIC-WOZ, AVEC, Mundt-35) δείχνουν F1-score έως 0.63 για SVM και 0.81 για δέντρα απόφασης σε ταξινόμηση κατάθλιψης [55].

2.5.2 Βαθιά Μάθηση

Η χρήση βαθιών δικτύων (LSTM, CNN, GAN, Transformers) έχει βελτιώσει σημαντικά την απόδοση, με F1-score ως και 0.90 σε ορισμένες μελέτες. Επίσης, έχουν αναπτυχθεί end-to-end αρχιτεκτονικές που επεξεργάζονται απευθείας το ωμό ηχητικό σήμα, όπως τα DepAudioNet και EmoAudioNet, με F1-score από 0.52 έως 0.82 [55].

2.5.3 Βασικές Επιδόσεις

Το βασικό σημείο αναφοράς για το σύνολο δεδομένων DAIC-WOZ (AVEC 2016) με γραμμικό SVM και βασικά χαρακτηριστικά είναι F1-score 0.58 για το ηχητικό σήμα. Αυτό το αποτέλεσμα χρησιμοποιείται ως σημείο σύγκρισης για μελλοντικά πειράματα.

3 Εργαλεία και Μέθοδοι

3.1 Δήλωση του Προβλήματος

Όπως περιγράφεται στην Εισαγωγή, το πεδίο της διπλωματικής εργασίας είναι η αξιολόγηση και σύγκριση της απόδοσης διαφόρων τύπων χαρακτηριστικών και αλγορίθμων μηχανικής μάθησης για την πρόβλεψη της κατάθλιψης. Η μελέτη περιλαμβάνει δύο βασικές εργασίες: (1) τη δημιουργία συνόλων χαρακτηριστικών ήχου και κειμένου, και (2) την αξιολόγηση της απόδοσης διαφόρων μοντέλων μηχανικής μάθησης χρησιμοποιώντας αυτά τα χαρακτηριστικά. Ο στόχος αυτής της μελέτης είναι να εντοπιστεί ο πιο αποτελεσματικός συνδυασμός χαρακτηριστικών και αλγορίθμων για την πρόβλεψης κατάθλιψης.

3.2 Περιγραφή Συνόλου Δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν για το πεδίο αυτής της διπλωματικής εργασίας προέρχονται από τη βάση δεδομένων DAIC-WOZ, η οποία αποτελεί υποσύνολο του Distress Analysis Interview Corpus (DAIC). Το DAIC είναι μια πολυτροπική συλλογή ημιδομημένων κλινικών συνεντεύξεων σχεδιασμένων για να βοηθούν στη διάγνωση ψυχολογικών καταστάσεων δυσφορίας, όπως άγχος, κατάθλιψη και μετατραυματική διαταραχή στρες.

Το σύνολο δεδομένων DAIC-WOZ είναι μοναδικό λόγω της multimodal φύσης του, καθώς ενσωματώνει δεδομένα ήχου, βίντεο και κειμένου, καθώς και του κλινικού πλαισίου (ημιδομημένες συνεντεύξεις με εικονικό πράκτορα). Περιλαμβάνει επίσης σχολιασμένες κλινικές βαθμολογίες, όπως το PHQ-8, καθιστώντας το πολύτιμο πόρο για έρευνα στην ανίχνευση κατάθλιψης. Σημαντικό μέρος των πρόσφατων μελετών σε αυτόν τον τομέα έχει αξιοποιήσει τις multimodal δυνατότητες του συνόλου δεδομένων για την προώθηση του πεδίου.

3.2.1 Συνεντεύξεις Wizard-of-Oz

Όπως αναφέρθηκε παραπάνω, για αυτή τη διπλωματική εργασία χρησιμοποιήθηκε το σύνολο δεδομένων DAIC-WOZ, το οποίο περιλαμβάνει τις συνεντεύξεις Wizard-of-Oz, που διεξήχθησαν από τον κινούμενο εικονικό συνεντευκτή Ellie [40].

3.3 Σύνθεση Συνόλου Δεδομένων

Το σύνολο δεδομένων DAIC-WOZ αποτελείται από 189 συνεδρίες, καθεμία από τις οποίες περιλαμβάνει ένα ακατέργαστο αρχείο ήχου και την αντίστοιχη απομαγνητοφώνησή του. Για την πειραματική αξιολόγηση αυτής της διπλωματικής εργασίας, η βαθμολογία PHQ-8 χρησιμοποιήθηκε ως η «αλήθεια εδάφους» (ground truth) για τον προσδιορισμό της παρουσίας κατάθλιψης στους συμμετέχοντες. Η δυαδική ταξινόμηση πραγματοποιήθηκε χρησιμοποιώντας κατώφλι 10, σύμφωνα με τις οδηγίες βαθμολόγησης του PHQ-8, οι οποίες υποδηλώνουν ότι μια βαθμολογία 10 ή μεγαλύτερη είναι ενδεικτική της Μείζονος Καταθλιπτικής Διαταραχής [17].

3.3.1 Ήχος

Στην παρούσα έρευνα, η πρώτη προσέγγιση αφορά την προεπεξεργασία των ακατέργαστων αρχείων ήχου. Εντοπίστηκαν και διορθώθηκαν κατεστραμμένα δείγματα ήχου με τη μετατροπή της μορφής τους χρησιμοποιώντας τη βιβλιοθήκη ffmpeg.

Handcrafted χαρακτηριστικά ήχου εξήχθησαν χρησιμοποιώντας τη βιβλιοθήκη pyAudioAnalysis.

Η εξαγωγή handcrafted χαρακτηριστικών για ένα μεμονωμένο αρχείο ήχου περιλαμβάνει τα εξής βήματα:

- Εντοπισμός κάθε εκφώνησης (utterance) ήχου χρησιμοποιώντας τα αρχεία απομαγνητοφώνησης.
- Εξαγωγή μεσοπρόθεσμων χαρακτηριστικών (mid-term features) για κάθε εκφώνηση με τη χρήση της αντίστοιχης συνάρτησης της βιβλιοθήκης.
- Τέλος, για να ληφθεί μια συνολική αναπαράσταση σε κάθε αρχείου ήχου, υπολογίζεται ο μέσος όρος των μεσοπρόθεσμων χαρακτηριστικών σε όλες τις εκφωνήσεις, παράγοντας ένα μοναδικό διάνυσμα χαρακτηριστικών ανά αρχείο ήχου.

Τα embeddings εξήχθησαν με παρόμοια προσέγγιση: εντοπίστηκαν οι εκφωνήσεις, εξήχθησαν embeddings για κάθε εκφώνηση και στη συνέχεια τα embeddings αθροίστηκαν κατά μέσο όρο για να παραχθεί ένα μοναδικό αντιπροσωπευτικό διάνυσμα χαρακτηριστικών για ολόκληρο το αρχείο ήχου.

Επιπλέον, εφαρμόστηκε η προσέγγιση βασισμένη στους ρόλους, όπου τα χαρακτηριστικά εξήχθησαν με παρόμοιο τρόπο για κάθε εκφώνηση. Ωστόσο, για κάθε αρχείο δημιουργήθηκαν δύο διακριτά διανύσματα χαρακτηριστικών: ένα με τον μέσο όρο των χαρακτηριστικών των εκφωνήσεων της Ellie και ένα με τον μέσο όρο των χαρακτηριστικών των εκφωνήσεων του συμμετέχοντα.

3.3.1.1 Προσέγγιση Βασισμένη σε Ρόλους

Η εξαγωγή χαρακτηριστικών σε επίπεδο εκφώνησης διασφαλίζει συνέπεια στη διαδικασία εξαγωγής, ανεξάρτητα από την προσέγγιση βασισμένη στο ρόλο. Αυτή η συνέπεια μας επέτρεψε να συγκρίνουμε αποτελεσματικά τα αποτελέσματα. Επιπλέον, εξήχθησαν και χαρακτηριστικά κειμένου με παρόμοιο τρόπο, διευκολύνοντας την συγχώνευση διαφορετικών συνόλων δεδομένων, όπως αναλυτικά περιγράφεται στην παράγραφο 3.3.3.

Ο διαχωρισμός των χαρακτηριστικών ήχου με βάση τους ρόλους επιτρέπει στην ανάλυση να λαμβάνει υπόψη τις εγγενείς διαφορές στα πρότυπα ομιλίας και να δίνει μεγαλύτερη βαρύτητα στον έναν ή τον άλλο ομιλητή. Τα χαρακτηριστικά που εξάγονται από κάθε ρόλο μπορούν να αναδείξουν ενδείξεις ειδικές για τον ρόλο. Είναι επίσης σημαντικό να σημειωθεί ότι τα χαρακτηριστικά βασισμένα στο ρόλο επιτρέπουν πιο σαφή ερμηνεία του ποια συμπεριφορά ομιλητή οδηγεί σε ορισμένα αποτελέσματα, διευκολύνοντας πιο στοχευμένες παρεμβάσεις ή συμπεράσματα. Επιπλέον, με το να μοντελοποιούνται οι ρόλοι ξεχωριστά, τα μοντέλα μηχανικής μάθησης μπορούν να μάθουν πρότυπα ειδικά για κάθε ρόλο χωρίς σύγχυση, βελτιώνοντας την ακρίβεια ταξινόμησης ή παλινδρόμησης [39].

3.3.2 Κείμενο

Παρόμοια με τα χαρακτηριστικά ήχου, το πρώτο βήμα είναι η προεπεξεργασία των απομαγνητοφωνήσεων. Για το σκοπό αυτό εφαρμόστηκε μια βασική μέθοδος καθαρισμού κειμένου.

Τα χαρακτηριστικά κειμένου εξήχθησαν χρησιμοποιώντας το μοντέλο GloVe για embeddings λέξεων, καθώς και το SBERT για embeddings προτάσεων, για συμπληρωματικά πειράματα.

Τα embeddings λέξεων που χρησιμοποιήθηκαν σε αυτή τη μελέτη προέρχονται από το προεκπαιδευμένο μοντέλο GloVe, εκπαιδευμένο στα σώματα κειμένου Wikipedia 2014 και Gigaword 5, που περιλαμβάνουν 6 δισεκατομμύρια tokens και ένα λεξιλόγιο 400.000 λέξεων χωρίς διάκριση πεζών-κεφαλαίων [48]. Συγκεκριμένα, χρησιμοποιήθηκε η 50-διάστατη έκδοση των διανυσμάτων GloVe. Η διαδικασία εξαγωγής ακολουθεί παρόμοια προσέγγιση με αυτή που χρησιμοποιήθηκε για τα χαρακτηριστικά ήχου.

Ο μέσος όρος των embeddings λέξεων αποτελεί μια ευρέως χρησιμοποιούμενη και απλή μέθοδο για την απόκτηση αναπαράστασης σταθερού μήκους μιας απομαγνητοφώνησης. Ωστόσο, αυτή η προσέγγιση έχει σημαντικό περιορισμό: αντιμετωπίζει όλες τις λέξεις ισότιμα και αγνοεί τη σειρά των λέξεων και τη συντακτική δομή, κάτι που μπορεί να οδηγήσει σε απώλεια σημαντικών συμφραζόμενων πληροφοριών [18, 50].

Τέλος, ακολουθώντας την ίδια προσέγγιση που χρησιμοποιήθηκε για την εξαγωγή χαρακτηριστικών ήχου βασισμένη σε ρόλους, υπολογίζουμε επίσης ξεχωριστά τον μέσο όρο των χαρακτηριστικών κειμένου για τις εκφωνήσεις της Ellie και του συμμετέχοντα.

Όσον αφορά τα embeddings προτάσεων, χρησιμοποιήθηκε το SBERT κατά παρόμοιο τρόπο.

3.3.3 Συνδυασμοί Χαρακτηριστικών

Για κάθε συνεδρία του dataset, εξάγουμε τρία χαρακτηριστικά ήχου (pyAudioAnalysis), τρία embeddings ήχου (wav2vec 2.0), τρία χαρακτηριστικά κειμένου (GloVe) και τρία embeddings προτάσεων (SBERT). Επιπλέον, δημιουργούμε δύο επιπλέον χαρακτηριστικά με τη συγχώνευση (concatenation) των χαρακτηριστικών του συμμετέχοντα και της Ellie τόσο για τον ήχο όσο και για το κείμενο, που ονομάζονται «Concatenated Features».

Δημιουργήσαμε δύο σύνολα δεδομένων για τα πειράματά μας. Το κύριο σύνολο δεδομένων περιλαμβάνει όλα τα χαρακτηριστικά pyAudioAnalysis, τα embeddings GloVe και τον

συνδυασμό τους, και χρησιμοποιείται για τις βασικές αναλύσεις. Το δεύτερο σύνολο δεδομένων, που χρησιμοποιείται για επιπλέον πειράματα, περιέχει τα embeddings wav2vec 2.0 και SBERT μαζί με τον συνδυασμό τους.

3.4 Μετρικές Αξιολόγησης

Για να αναλυθούν διεξοδικά οι μετρικές αξιολόγησης που χρησιμοποιούνται σε αυτή τη διπλωματική εργασία, είναι απαραίτητο πρώτα να παρουσιαστούν κάποιες γενικές πληροφορίες σχετικά με την αξιολόγηση της δυαδικής ταξινόμησης.

Στη δυαδική ταξινόμηση, τα δείγματα ταξινομούνται είτε ως θετικά είτε ως αρνητικά. Ένα ταξινομημένο δείγμα ανήκει σε μία από τις ακόλουθες κατηγορίες:

- **Αληθώς Θετικό (True Positive, TP):** το δείγμα ταξινομείται σωστά ως θετικό
- **Ψευδώς Θετικό (False Positive, FP):** το δείγμα ταξινομείται λανθασμένα ως θετικό
- **Αληθώς Αρνητικό (True Negative, TN):** το δείγμα ταξινομείται σωστά ως αρνητικό
- **Ψευδώς Αρνητικό (False Negative, FN):** το δείγμα ταξινομείται λανθασμένα ως αρνητικό

Accuracy

Η κύρια μετρική που χρησιμοποιείται για την αξιολόγηση μοντέλων είναι συχνά η accuracy, η οποία περιγράφει τον αριθμό των σωστών προβλέψεων σε σχέση με το συνολικό αριθμό προβλέψεων. Ο τύπος για τον υπολογισμό της accuracy εκφράζεται με διάφορους τρόπους, αλλά όλοι αναπαριστούν την ίδια έννοια [13]:

$$Accuracy = \frac{TruePositives + TrueNegatives}{TruePositives + TrueNegatives + FalsePositives + FalseNegatives} \tag{1}$$

Ακρίβεια (Precision) και Ανάκληση (Recall)

Εναλλακτικές μετρικές που παρέχουν καλύτερη κατανόηση της απόδοσης ενός μοντέλου είναι η ακρίβεια και η ανάκληση. Αυτές οι μετρικές είναι ιδιαίτερα χρήσιμες όταν αντιμετωπίζουμε ανισόρροπα σύνολα δεδομένων [9].

Η ακρίβεια μετρά το ποσοστό των αληθώς θετικών προβλέψεων μεταξύ όλων των θετικών προβλέψεων. Ο τύπος για την ακρίβεια είναι [9]:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{2}$$

Η ανάκληση μετρά το ποσοστό των αληθώς θετικών περιπτώσεων που ταξινομήθηκαν σωστά, σε σχέση με όλες τις πραγματικά θετικές περιπτώσεις στο σύνολο δεδομένων. Ο τύπος για την ανάκληση είναι [9]:

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{3}$$

F1-Score

Το F1-Score είναι μια μετρική που συνδυάζει τόσο την ακρίβεια όσο και την ανάκληση, και ορίζεται ως ο αρμονικός μέσος τους. Ο τύπος για το F1-Score είναι [14]:

$$F_1 = \frac{Precision \times Recall}{Precision + Recall}$$

Σε αυτή τη διπλωματική εργασία, χρησιμοποιήθηκε η μέθοδος macro-averaging για το F1-scroe λόγω της σημαντικής ανισορροπίας που υπάρχει στο σύνολο δεδομένων DAIC-WOZ.

AUC

Το AUC-ROC (Area Under the Receiver Operating Characteristic Curve) είναι μια ευρέως χρησιμοποιούμενη μετρική για την αξιολόγηση δυαδικών ταξινομητών. Μετρά την ικανότητα ενός μοντέλου να διακρίνει μεταξύ θετικών και αρνητικών κλάσεων σε όλα τα πιθανά κατώφλια ταξινόμησης [87].

Η καμπύλη ROC είναι μια γραφική αναπαράσταση που σχεδιάζει τον Ρυθμό Αληθώς Θετικών (TPR) στον άξονα y έναντι του Ρυθμού Ψευδώς Θετικών (FPR) στον άξονα x για διάφορες τιμές κατωφλίου. Κάθε σημείο στην καμπύλη αντιστοιχεί σε ένα συγκεκριμένο κατώφλι, που καθορίζει πώς οι προβλέψεις ταξινομούνται ως θετικές ή αρνητικές [87].

$$TPR = \frac{TP}{TP + FN} \tag{4}$$

$$FPR = \frac{FP}{FP + TN} \tag{5}$$

Το AUC (Εμβαδόν κάτω από την καμπύλη) αναπαριστά το συνολικό εμβαδόν κάτω από αυτή την καμπύλη ROC και παρέχει μια μονοδιάστατη τιμή που συνοψίζει την απόδοση του μοντέλου:

- Ένα AUC ίσο με 1 υποδηλώνει τέλεια διάκριση μεταξύ των κλάσεων.
- Ένα AUC ίσο με 0.5 υποδηλώνει απόδοση ίση με τυχαία εικασία.
- Τιμές κοντά στο 1 υποδηλώνουν καλύτερη απόδοση μοντέλου, ενώ τιμές κοντά στο 0 υποδηλώνουν φτωχή απόδοση.

4 Πειραματική Αξιολόγηση

Σε αυτό το κεφάλαιο, συζητάμε τη δομή των πειραμάτων και τα αποτελέσματα που παράγουν. Όπως αναφέρθηκε στα προηγούμενα κεφάλαια, ο στόχος αυτής της διπλωματικής είναι η πρόβλεψη της κατάθλιψης βάση της ομιλίας ενός ατόμου. Για το σκοπό αυτό, χρησιμοποιήθηκε το σύνολο δεδομένων DAIC-WoZ για την εκτέλεση των ακόλουθων πειραμάτων.

4.1 Δομή Πειραμάτων και Αποτελέσματα

Τα σύνολα δεδομένων που χρησιμοποιούνται σε αυτή τη μελέτη περιλαμβάνουν τρεις βασικούς τύπους χαρακτηριστικών: χαρακτηριστικά βασισμένα στον ήχο, χαρακτηριστικά βασισμένα στο κείμενο και πολυτροπικά χαρακτηριστικά που σχηματίζονται με τη συγχώνευση των αναπαραστάσεων ήχου και κειμένου. Για κάθε τύπο εξήχθησαν τρεις παραλλαγές για να καλυφθούν διαφορετικά πεδία ομιλητών. Πρώτον, λήφθηκαν χαρακτηριστικά που αναπαριστούν ολόκληρο τον ήχο, συνδυάζοντας τόσο τον συμμετέχοντα όσο και την Ellie. Δεύτερον, εξήχθησαν χαρακτηριστικά που αντιστοιχούν αποκλειστικά στην ομιλία του συμμετέχοντα. Τρίτον, δημιουργήθηκε μια συγχώνευση ξεχωριστών χαρακτηριστικών για τον συμμετέχοντα και την Ellie, ώστε να διατηρηθούν πληροφορίες ειδικές για τον ρόλο. Αυτή η δομή επιτρέπει στην ανάλυση να λαμβάνει υπόψη τα ατομικά χαρακτηριστικά των ομιλητών καθώς και τις συνδυασμένες αλληλεπιδράσεις τους σε διάφορες μορφές δεδομένων.

Στα πειράματά μας χρησιμοποιήσαμε μοντέλα Support Vector Machines (SVM) και XGBoost. Αυτά τα μοντέλα επιλέχθηκαν λόγω της ισχυρής απόδοσής τους και των συμπληρωματικών πλεονεκτημάτων που έχουν επιδείξει σε προηγούμενες μελέτες σε εργασίες όπως η αναγνώριση συναισθήματος στην ομιλία και η ανάλυση συναισθήματος. Μελέτες έχουν δείξει ότι τα SVM συχνά επιτυγχάνουν ανταγωνιστική ή ανώτερη ακρίβεια, ενώ το XGBoost μπορεί να αξιοποιήσει την βελτιστοποίηση χαρακτηριστικών για να φτάσει σε επίπεδα ακρίβειας συγκρίσιμα με μοντέλα βαθιάς μάθησης, αλλά με μικρότερη υπολογιστική πολυπλοκότητα [45, 30, 32, 44].

Η ρύθμιση των υπερπαραμέτρων έγινε με χρήση GridSearch σε συνδυασμό με LOOCV, για να εξεταστούν διεξοδικά οι συνδυασμοί παραμέτρων ενώ μεγιστοποιείται η χρήση των δεδομένων [69].

Για την εκπαίδευση των μοντέλων, εφαρμόστηκε επίσης χειροκίνητη μέθοδος Leave-One-Out cross-validation (LOOCV). Η LOOCV περιλαμβάνει την επαναλαμβανόμενη εκπαίδευση του μοντέλου σε όλα τα δείγματα εκτός από ένα, το οποίο χρησιμοποιείται για δοκιμή. Αυτή η διαδικασία επαναλαμβάνεται για κάθε δείγμα, εξασφαλίζοντας αμερόληπτη εκτίμηση της απόδοσης χωρίς ξεχωριστό σύνολο επικύρωσης. Αυτή η προσέγγιση διασφαλίζει ότι το μοντέλο δοκιμάζεται σε δείγμα που δεν έχει δει ποτέ πριν. Είναι επίσης πολύ σημαντικό να σημειωθεί ότι, σε κάθε επανάληψη, δημιουργείται ένα νέο στιγμιότυπο του μοντέλου για να αποφευχθεί διαρροή πληροφορίας, δηλαδή να μην επηρεάζονται οι προβλέψεις του τρέχοντος μοντέλου από προηγούμενες επαναλήψεις, διατηρώντας έτσι την ακεραιότητα της διαδικασίας cross-validation.

Μετά την επιλογή των βέλτιστων παραμέτρων και την εκπαίδευση του μοντέλου, παράγουμε ταξινομήσεις για κάθε σύνολο δοκιμής, δηλαδή για κάθε δείγμα δοκιμής. Ορίζουμε τις συγκεντρωτικές προβλέψεις του συνόλου δοκιμής ως το σύνολο όλων των ταξινομήσεων, επιτρέποντάς μας να υπολογίσουμε μετρικές όπως το F1-macro, την ακρίβεια και το AUC χρησιμοποιώντας αυτές τις συγκεντρωτικές προβλέψεις.

Οι παρακάτω πίνακες παρέχουν τα αποτελέσματα των διεξαχθέντων πειραμάτων. Είναι σημαντικό να σημειωθεί ότι η Ακρίβεια, το F1-macro και το AUC υπολογίζονται συγκεντρωτικά σε όλα τα folds του LOOCV για να παρέχουν συνολικές εκτιμήσεις απόδοσης.

Σε αυτό το κεφάλαιο παρουσιάζουμε την απόδοση δύο εκπαιδευμένων μοντέλων μηχανικής

μάθησης στο κύριο σύνολο δεδομένων, χρησιμοποιώντας τα τρία διακριτά σύνολα χαρακτηριστικών. Συγκεκριμένα, οι πίνακες περιλαμβάνουν τις εξής κατηγορίες χαρακτηριστικών: χαρακτηριστικά που προέρχονται από ολόκληρα τα αρχεία ήχου *(Whole Audio)*, χαρακτηριστικά που προκύπτουν από τη συγχώνευση των δεδομένων της Ellie με αυτά των συμμετεχόντων *(Concat)* και χαρακτηριστικά που εξάγονται αποκλειστικά από τους συμμετέχοντες *(Participants)*.

Αποτελέσματα:

Είναι επίσης σημαντικό να σημειωθεί ότι, επειδή το σύνολο δεδομένων DAIC-WoZ είναι ανισόρροπο, χρησιμοποιούμε κυρίως τις τιμές AUC για τον προσδιορισμό της απόδοσης ενός μοντέλου. Επομένως, στους επόμενους πίνακες επισημαίνεται η καλύτερη τιμή AUC για κάθε υποσύνολο δεδομένων [31].

	SVM			**XGBoost**		
	Whole Audio	**Concat**	**Participant**	**Whole Audio**	**Concat**	**Participant**
acc	0.70	0.70	0.70	0.72	0.71	0.73
f1	0.41	0.41	0.41	0.61	0.45	0.59
auc	0.50	0.50	0.50	**0.61**	0.52	0.59

Table 1: Using the pyAudioAnalysis Audio Features.

	SVM			**XGBoost**		
	Whole Audio	**Concat**	**Participant**	**Whole Audio**	**Concat**	**Participant**
acc	0.77	0.78	0.73	0.74	0.74	0.71
f1	0.70	0.73	0.59	0.64	0.63	0.58
auc	0.69	**0.72**	0.59	0.63	0.62	0.58

Table 2: Using the Glove Word Embeddings.

	SVM			**XGBoost**		
	Whole Audio	**Concat**	**Participant**	**Whole Audio**	**Concat**	**Participant**
acc	0.74	0.78	0.70	0.74	0.74	0.73
f1	0.69	0.73	0.41	0.64	0.63	0.60
auc	0.69	**0.72**	0.50	0.63	0.62	0.60

Table 3: Using the Concatenated pyAudioAnalysis and GloVe Features.

Οι μετρικές στον Πίνακα 1 δείχνουν ότι όλα τα μοντέλα SVM αποτυγχάνουν να διακρίνουν μεταξύ των δύο κλάσεων, προβλέποντας σταθερά μόνο την πλειοψηφούσα κλάση σε όλα τα σύνολα δοκιμών. Η τιμή AUC 0.5, κατά τον ορισμό, υποδηλώνει έλλειψη διακριτικής ικανότητας [92]. Επιπλέον, τα μοντέλα XGBoost για το «Whole Audio» και τον «Participant» δίνουν ελαφρώς καλύτερα αποτελέσματα, ενώ το «Concat» είναι πιο κοντά

στην τυχαία επιλογή. Αυτές οι παρατηρήσεις υποδεικνύουν ότι τόσο τα SVM όσο και τα XGBoost δυσκολεύονται με τα handcrafted χαρακτηριστικά ήχου.

Όπως φαίνεται στον Πίνακα 2, τόσο τα μοντέλα SVM όσο και XGBoost αποδίδουν καλύτερα με τα embeddings GloVe, ωστόσο τα αποτελέσματα παραμένουν υποβέλτιστα. Η καλύτερη βαθμολογία μέχρι στιγμής δίνεται από το μοντέλο SVM «Concat» με AUC 72.

Επιπλέον, ο Πίνακας 3 δείχνει ότι η συγχώνευση χαρακτηριστικών ήχου και κειμένου δεν βελτιώνει την απόδοση σε σχέση με τη χρήση μόνο των embeddings κειμένου, ενισχύοντας το συμπέρασμα ότι τα μοντέλα δυσκολεύονται με τα χαρακτηριστικά pyAudioAnalysis.

4.1.1 Υποδειγματοληψία (Undersampling)

Για να αντιμετωπιστεί η ανισορροπία των κλάσεων, χρησιμοποιήθηκαν δύο εναλλακτικές προσεγγίσεις. Η πρώτη προσέγγιση περιλαμβάνει την χειροκίνητη εξισορρόπηση του συνόλου δεδομένων, που επιτεύχθηκε με τυχαία αφαίρεση συγκεκριμένου αριθμού δειγμάτων από την πλειοψηφούσα κλάση πριν από την εκτέλεση των πειραμάτων [7].

Αποτελέσματα:

	SVM			XGBoost		
	Whole Audio	**Concat**	**Participant**	**Whole Audio**	**Concat**	**Participant**
acc	0.59	0.59	0.59	0.68	0.63	0.60
f1	0.37	0.37	0.37	0.67	0.60	0.56
auc	0.50	0.50	0.50	**0.66**	0.60	0.57

Table 4: Using the Balanced Audio-Based Dataset.

	SVM			XGBoost		
	Whole Audio	**Concat**	**Participant**	**Whole Audio**	**Concat**	**Participant**
acc	0.70	0.75	0.60	0.68	0.70	0.61
f1	0.68	0.74	0.60	0.66	0.67	0.46
auc	0.68	**0.74**	0.60	0.66	0.67	0.53

Table 5: Using the Balanced Text-Based Dataset.

Όπως φαίνεται στον Πίνακα 4, το σύνολο δεδομένων βασισμένο στον ήχο δείχνει ότι η εξισορρόπηση δεν βελτιώνει την απόδοση των μοντέλων SVM. Αντίθετα, τα μοντέλα XGBoost παρουσιάζουν μικρή βελτίωση, με το καλύτερο μοντέλο («Whole Audio») να αυξάνει το AUC από 61 σε 66. Παρόμοια τάση παρατηρείται και στο σύνολο δεδομένων βασισμένο στο κείμενο, όπως φαίνεται στον Πίνακα 4.6. Εδώ, το καλύτερο μοντέλο από το αρχικό πείραμα («Concat» SVM) παρουσιάζει μικρή βελτίωση με τη χειροκίνητη εξισορρόπηση. Ωστόσο, τα αποτελέσματα για το συγχωνευμένο σύνολο δεδομένων

	SVM			XGBoost		
	Whole Audio	**Concat**	**Participant**	**Whole Audio**	**Concat**	**Participant**
acc	0.70	0.75	0.59	0.72	0.71	0.63
f1	0.69	0.74	0.37	0.70	0.70	0.58
auc	0.59	0.50	0.37	**0.7**	**0.7**	0.59

Table 6: Using the Balanced Concatenated Audio-Text Dataset.

ήχου-κειμένου, που παρουσιάζονται στον Πίνακα 4.7, αποκλίνουν από αυτά των προηγούμενων συνόλων. Σημαντικά, δεν υπάρχει συνολική βελτίωση · πριν την εξισορρόπηση, το καλύτερο μοντέλο είχε AUC 72, ενώ μετά την εξισορρόπηση το AUC του καλύτερου μοντέλου μειώθηκε σε 70.

4.1.2 Υπερδειγματοληψία SMOTE

Για περαιτέρω αντιμετώπιση της ανισορροπίας κλάσεων, χρησιμοποιήθηκε μια δεύτερη μέθοδος, η υπερδειγματοληψία SMOTE.

Αποτελέσματα:

	SVM			XgBoost		
	Whole Audio	**Concat**	**Participant**	**Whole Audio**	**Concat**	**Participant**
acc	0.70	0.70	0.70	0.69	0.63	0.60
f1	0.41	0.41	0.41	0.63	0.55	0.53
auc	0.50	0.50	0.50	**0.62**	0.55	0.53

Table 7: Using the Audio-Based SMOTE Dataset.

	SVM			XgBoost		
	Whole Audio	**Concat**	**Participant**	**Whole Audio**	**Concat**	**Participant**
acc	0.62	0.72	0.59	0.68	0.69	0.61
f1	0.59	0.69	0.56	0.61	0.65	0.54
auc	0.62	**0.70**	0.58	0.61	0.65	0.54

Table 8: Using the Text-Based SMOTE Dataset.

Όπως φαίνεται στην table 7, η υπερδειγματοληψία SMOTE δεν βοηθά το μοντέλο SVM που προβλέπει μόνο την πλειοψηφούσα κλάση στο σύνολο δεδομένων βασισμένο στον ήχο. Επιπλέον, τα αποτελέσματα που προέκυψαν με τη χρήση της SMOTE τόσο για το σύνολο δεδομένων βασισμένο στο κείμενο όσο και για το συγχωνευμένο σύνολο ήχου-κειμένου παραμένουν συνεπή με αυτά που επιτεύχθηκαν με την αρχική μέθοδο.

	SVM			XgBoost		
	Whole Audio	**Concat**	**Participant**	**Whole Audio**	**Concat**	**Participant**
acc	0.74	0.72	0.70	0.72	0.69	0.60
f1	0.71	0.69	0.41	0.68	0.65	0.52
auc	**0.73**	0.70	0.50	0.68	0.65	0.52

Table 9: Using the Concatenated Audio-Text SMOTE Dataset.

4.2 Συμπληρωματικά Πειράματα

Μετά την ανάλυση των αποτελεσμάτων που παρουσιάστηκαν παραπάνω, καταλήγουμε στο συμπέρασμα ότι, στο πλαίσιο των πειραμάτων μας, τα χαρακτηριστικά ήχου δεν παρέχουν ικανοποιητική απόδοση στην πρόβλεψη της κατάθλιψης χρησιμοποιώντας το σύνολο δεδομένων DAIC-WOZ, ακόμη και όταν εφαρμόζονται τεχνικές εξισορρόπησης δεδομένων που στοχεύουν στην αντιμετώπιση της ανισορροπίας κλάσεων. Αυτό υποδηλώνει ότι οι αναπαραστάσεις ήχου που χρησιμοποιήθηκαν ενδέχεται να μην έχουν επαρκή διακριτική ικανότητα για αυτή την εργασία.

Στην επόμενη ενότητα, θα εξερευνήσουμε την υλοποίηση εναλλακτικών τεχνικών embedding ήχου και κειμένου για να αξιολογήσουμε αν πιο προηγμένες μέθοδοι εξαγωγής χαρακτηριστικών μπορούν να βελτιώσουν τη διάκριση μεταξύ των δύο κλάσεων στο σύνολο δεδομένων DAIC-WOZ. Αυτές οι μέθοδοι μπορεί να αξιοποιούν βαθύτερες συμφραζόμενες πληροφορίες ή πιο εξελιγμένη μοντελοποίηση των χρονικών δυναμικών, ενδεχομένως βελτιώνοντας την ακρίβεια της πρόβλεψης.

Είναι επίσης σημαντικό να σημειωθεί ότι η προσέγγιση βασισμένη στους ρόλους, που διαχωρίζει τα δεδομένα με βάση τους ρόλους των ομιλητών, δεν βελτίωσε τη διάκριση μεταξύ καταθλιπτικών και μη καταθλιπτικών ατόμων. Δεδομένης της περιορισμένης συνεισφοράς της, θα αποκλείσουμε τα σύνολα δεδομένων βασισμένα σε ρόλους από τα επόμενα πειράματα, ώστε να απλοποιήσουμε την ανάλυση και να επικεντρωθούμε σε πιο υποσχόμενες αναπαραστάσεις χαρακτηριστικών.

Τέλος, τα πειράματα που παρουσιάζονται παρακάτω στοχεύουν στο να παράσχουν πληροφορίες σχετικά με τους περιορισμούς των αρχικών πειραμάτων και να εντοπίσουν πιθανές κατευθύνσεις για μελλοντική έρευνα. Μέσω συστηματικής αξιολόγησης εναλλακτικών τεχνικών ενσωμάτωσης και προσεγγίσεων μοντελοποίησης, ελπίζουμε να αποκαλύψουμε παράγοντες που συμβάλλουν σε βελτιωμένη ανίχνευση της κατάθλιψης και να υποστηρίξουμε την ανάπτυξη πιο αποτελεσματικών διαγνωστικών εργαλείων.

Όσον αφορά τα χαρακτηριστικά ήχου, χρησιμοποιούμε τα embeddings ήχου wav2vec 2.0. Αποτελέσματα:

Όπως μπορούμε να δούμε από τον πίνακα table 10, η αναπαράσταση wav2vec 2.0 οδηγεί στο ίδιο αποτέλεσμα με τα handcrafted χαρακτηριστικά pyAudioAnalysis, προβλέποντας μόνο την αρνητική κλάση. Αυτή η ομοιότητα στην απόδοση μπορεί να εξηγηθεί από τους εγγενείς περιορισμούς των wav2vec 2.0 embeddings όταν αντιμετωπίζουν σοβαρή ανισορροπία κλάσεων.

Μπορούμε να υποθέσουμε ότι η δυσκολία στην πρόβλεψη της ελάχιστης κλάσης δεν

	SVM	XGBoost
	Whole Audio	**Whole Audio**
acc	0.70	0.72
f1	0.41	0.61
auc	0.50	0.61

Table 10: Using the wav2vec 2.0 Dataset.

σχετίζεται με τον τύπο των χαρακτηριστικών που χρησιμοποιούνται, αλλά συσχετίζεται έντονα με την ανισορροπία των κλάσεων. Επομένως, για την αντιμετώπιση του προβλήματος της εκτίμησης της κατάθλιψης βάσει ομιλίας, οι καλύτερες προσεγγίσεις θα ήταν:

- Βελτιστοποίηση του wav2vec 2.0 end-to-end με σταθμισμένη απώλεια για την αντιμετώπιση της ανισορροπίας κλάσεων.
- Διατήρηση ή βελτίωση της απώλειας ποικιλομορφίας κατά τη διάρκεια της εκπαίδευσης για την πρόληψη της κατάρρευσης τρόπου στα embeddings.
- Χρήση επαύξησης δεδομένων και προ-εκπαίδευσης σε σχετικά σύνολα δεδομένων για τον εμπλουτισμό των χαρακτηριστικών της ελάχιστης κλάσης.
- Πειραματισμός με προηγμένους ταξινομητές ή προσεγγίσεις συνόλου για καλύτερη ανίχνευση της ελάχιστης κλάσης.

Όσον αφορά τα πειράματα που βασίζονται στο κείμενο, όπως αναφέρθηκε παραπάνω, χρησιμοποιούμε το Sentence-BERT (SBERT). Αποτελέσματα:

	SVM	XGBoost
	Whole Audio	**Whole Audio**
acc	0.77	0.79
f1	0.72	0.74
auc	0.71	0.73

Table 11: Using the SBERT Dataset.

Από τον Πίνακα table 1, η υψηλότερη τιμή AUC που επιτεύχθηκε χρησιμοποιώντας embeddings GloVe είναι 0.69 με το μοντέλο XGBoost. Ωστόσο, όπως φαίνεται στον Πίνακα table 4.12, η χρήση των SBERT embeddings με το μοντέλο XGBoost αποδίδει μια βελτιωμένη τιμή AUC 0.74, όπως αναμενόταν.

4.3 Αξιολόγηση Αποτελεσμάτων

Στα πειράματα που βασίζονται στους ρόλους, τα αποτελέσματα ήταν ασυνεπή και δεν έδειξαν σαφή βελτίωση στην πρόβλεψη της κατάθλιψης. Τα αρχικά πειράματα αποκάλυψαν ότι τα μοντέλα SVM δυσκολεύτηκαν με τα handcrafted χαρακτηριστικά ήχου (pyAudioAnalysis). Τα μοντέλα XGBoost επέδειξαν ελαφρώς καλύτερη απόδοση με τα χαρακτηριστικά ήχου, αλλά τα αποτελέσματα παρέμειναν μην ικανοποιητικά.

Τα χαρακτηριστικά που βασίζονται στο κείμενο (embeddings GloVe) απέδωσαν βελτιωμένα αποτελέσματα τόσο για τα SVM όσο και για τα XGBoost, με το μοντέλο SVM "Concat" να επιτυγχάνει το υψηλότερο AUC 72 (Πίνακας 2). Ωστόσο, η συγχώνευση των χαρακτηριστικών ήχου και κειμένου δεν βελτίωσε με συνέπεια την απόδοση, υποδηλώνοντας ότι τα μοντέλα δυσκολεύτηκαν να ενσωματώσουν αποτελεσματικά τα χαρακτηριστικά pyAudioAnalysis.

Για την αντιμετώπιση της ανισορροπίας κλάσεων, εφαρμόστηκαν χειροκίνητη εξισορρόπηση και υπερδειγματοληψία SMOTE. Η χειροκίνητη εξισορρόπηση παρείχε μια μικρή βελτίωση για το μοντέλο SVM "Concat" στο σύνολο δεδομένων που βασίζεται στο κείμενο, αλλά δεν βελτίωσε σταθερά τα αποτελέσματα σε όλα τα σύνολα δεδομένων και τα μοντέλα. Η υπερδειγματοληψία SMOTE δεν ενίσχυσε σημαντικά το μοντέλο SVM και παρείχε μόνο μια μικρή βελτίωση για το μοντέλο XGBoost.

Τα καλύτερα αποτελέσματα σε όλα τα πειράματα ήταν ένα AUC 0.66 για χαρακτηριστικά ήχου, 0.74 για χαρακτηριστικά κειμένου και 0.73 για συγχωνευμένα χαρακτηριστικά. Όσον αφορά τα χαρακτηριστικά ήχου, η υψηλότερη βαθμολογία F1-macro που επιτεύχθηκε ήταν 0.67, γεγονός που υποδηλώνει μια μικρή βελτίωση από τη baseline βαθμολογία 0.58.

Επιπλέον, τα πρόσθετα πειράματα που διεξήχθησαν έδειξαν ότι:

- Οι embeddings wav2vec 2.0 από μόνες τους δεν επιλύουν εγγενώς το πρόβλημα της ανισορροπίας κλάσεων, περιορίζοντας το πλεονέκτημά τους έναντι των handcrafted χαρακτηριστικών στην πρόβλεψη της ελάχιστης κλάσης.
- Επιπλέον, τα πειράματα που βασίζονται στο κείμενο χρησιμοποιώντας embeddings προτάσεων SBERT έδειξαν βελτιωμένες τιμές AUC.

5 Συμπεράσματα και Μελλοντικές Προεκτάσεις

Η παρούσα διπλωματική εξέτασε την αποτελεσματικότητα της εκτίμησης κατάθλιψης μέσω της ομιλίας με τη χρήση μηχανικής μάθησης, με στόχο την ανάπτυξη ενός αξιόπιστου αυτόματου συστήματος εκτίμησης κατάθλιψης. Με κίνητρο την ανάγκη για αντικειμενικά, δεδομενοστραφή εργαλεία που συμπληρώνουν τις κλινικές αξιολογήσεις, αξιοποιήθηκε το σύνολο δεδομένων DAIC-WOZ για την εξαγωγή και ανάλυση ηχητικών και χαρακτηριστικών κειμένου. Καινοτόμο στοιχείο αποτέλεσε η ανάλυση χαρακτηριστικών βάσει ρόλων, διαχωρίζοντας την ομιλία συμμετέχοντα και συνεντευκτή, με σκοπό την καλύτερη αποτύπωση της δυναμικής των κλινικών συνεντεύξεων. Οι βασικές συνεισφορές περιλαμβάνουν τον σχεδιασμό ενός multimodal συστήματος εξαγωγής χαρακτηριστικών, την εισαγωγή πλαισίου ανάλυσης βάσει ρόλων και τη συστηματική αξιολόγηση της επίδρασης των συνδυασμών χαρακτηριστικών και ρόλων στην απόδοση.

Τα πειραματικά αποτελέσματα έδειξαν ότι τα χαρακτηριστικά κειμένου υπερέχουν των ηχητικών, με το μέγιστο AUC να φτάνει το 0,74 έναντι 0,66 για τον ήχο. Η ανάλυση βάσει ρόλων υποδεικνύει πιθανά οφέλη στην αποτύπωση της αλληλεπίδρασης, αλλά με ασυνεπή αποτελέσματα, υπογραμμίζοντας την ανάγκη για πιο εξελιγμένα μοντέλα διαλόγου. Περιορισμοί περιλαμβάνουν το μικρό μέγεθος του DAIC-WOZ και τη χρήση παραδοσιακών αλγορίθμων. Μελλοντικές κατευθύνσεις αφορούν την ενσωμάτωση βαθιών νευρωνικών δικτύων (π.χ. RNN, LSTM, transformers), την επέκταση σε multimodal

δεδομένα (εκφράσεις προσώπου, φυσιολογικά σήματα), πιο σύνθετες τεχνικές σύντηξης χαρακτηριστικών και τη χρήση transfer μάθησης. Συνολικά, η εργασία προωθεί την αυτόματη εκτίμηση της κατάθλιψης, εισάγοντας νέες μεθόδους και επισημαίνοντας προκλήσεις και δυνατότητες για βελτίωση.

Chapter 1

Introduction

Depression is a leading cause of disability worldwide and represents a major challenge for public health systems. Depression (also major depression, major depressive disorder, or clinical depression) is the most prevalent mood disorder, characterized by a persistently low mood, diminished interest in activities, impaired cognitive function, and a range of symptoms that disrupt daily functioning [43, 47]. According to the World Health Organization, an estimated 280 million people, including 5% of all adults, have experienced depression [71].

Beyond its high prevalence, depression is associated with significant morbidity and mortality. It is closely linked to high rates of suicide, with approximately 50% of individuals who have committed suicide having a primary diagnosis of depression . Suicide remains one of the leading causes of death among young adults worldwide, highlighting the urgent need for effective identification and intervention strategies [68, 90]. Additionally, depression frequently coexists with other chronic diseases, leading to more severe health outcomes than either condition would cause on its own [53].

These findings, along with the high suicide rate, highlight the importance of prioritizing the diagnosis and treatment of depression as a public health issue in order to lower disability rates, reduce disease burden, and mitigate depression's major complications at the individual level. Early detection and intervention are critical, as untreated depression can lead to chronic disability, reduced productivity, and increased healthcare costs [82]. Despite this, many individuals with depression remain undiagnosed or undertreated due to stigma, lack of access to mental health services, and limitations in current diagnostic methods [53, 68, 90].

The current diagnosis of depression is based on a clinical examination, with the DSM-V criteria being the standard approach [81]. To be diagnosed with depression, a person must exhibit persistent symptoms for more than two weeks. According to the DSM-V, five or more of the following symptoms must be present during the same two-week period, indicating a change from previous functioning. At least one of these symptoms must be either a depressed mood or a loss of interest or pleasure to classify as depression. In total, there are nine equally important symptoms that assess the patient's mood, fatigue, loss of interest and concentration, as well as changes in sleep, agitation, and weight [24, 81].

This approach relies on the patient's ability to report their symptoms and respond to the physician's questions. However, these reports are often subjective and can be influenced by factors such as recall biases (e.g., downplaying or exaggerating symptoms), cognitive limitations (e.g., memory of episodes and environment, causal inference), and social stigma [66]. Additionally, subjective factors like patients' expressions, cultural background, and attitudes can complicate the diagnosis of depression, increasing the likelihood of misdiagnosis [66]. Another drawback to the current method is that clinicians may overlook depression unless the patient shows clear signs of sadness. Patients often emphasize physical symptoms, making it hard to tell if a symptom is due to depression or another condition [81]. Moreover, a major drawback of the current method is that clinical symptoms must persist for at least two weeks to confirm a diagnosis of depression, which can lead to limited care or treatment for patients during the early stages of the disorder [24].

These challenges underscore the urgent need for objective, data-driven methods to support and enhance the diagnosis of depression. Advances in digital health technologies and artificial intelligence offer promising avenues to complement traditional clinical assessments. For instance, machine learning algorithms can analyze complex patterns in speech, facial expressions, and language use that may be indicative of depression, potentially enabling earlier and more accurate detection [86]. Wearable sensors and mobile apps can continuously monitor behavioral and physiological signals, providing real-time data that can inform diagnosis and treatment [59]. Such objective measures could reduce reliance on subjective self-report and help overcome barriers related to stigma and cultural differences.

1.1 Motivation

Given the subjectivity and delays inherent in the current diagnostic practices, machine learning offers a promising avenue for developing more objective and timely tools for depression estimation. Machine learning is increasingly being utilized for depression estimation, with researchers focusing on the collection and analysis of mental-health related data and also the development of models for depression prediction. This thesis focuses on the development of a pipeline that utilizes a feature extraction mechanism. Specifically, the DAIC-WOZ dataset is utilized to extract features, and multiple versions of the dataset are created to evaluate the predictive performance of machine learning models in diagnosing depression.

In terms of feature extraction, this study focuses on extracting audio features using the pyAudioAnalysis library, as well as audio embeddings from interview recordings and text embeddings derived from their transcriptions. Another aspect of the feature extraction process involves role-based extraction, where the aforementioned types of features are separately extracted for both the participant and the interviewer. This approach is intended to assess the significance of each role in the depression prediction process. The objective not only to extract these features but also to evaluate whether their combination could enhance the accuracy of depression estimation. The combination involved integrating both the different types of features (audio and text) and the role-based aspect.

1.2 Thesis Contribution

This thesis makes several significant contributions to the field of automatic depression estimation through feature extraction and machine learning. The primary contributions are as follows:

- **Development of a Feature Extraction Pipeline:**

  Introduction of a novel pipeline for extracting audio features, audio embeddings, and text embeddings from the DAIC-WOZ dataset.

- **Role-Based Feature Analysis:**

  Implementation of a role-based extraction method that independently processes features for both the participant and the interviewer. This approach offers new insights into the significance of each role in depression estimation, highlighting the influence of interaction dynamics on predictive accuracy.

- **Creation of Different Dataset Versions:**

  Generation of multiple dataset versions, each incorporating different combinations of extracted features. This enables a comprehensive evaluation of how feature sets affect model performance and contributes to optimizing predictive accuracy.

- **Advancements in Automatic Depression Estimation (ADE):**

  Application of machine learning techniques to the extracted features to improve depression detection. The proposed methods support the development of reliable tools for the diagnosis and monitoring of depression.

These contributions not only advance the academic understanding of depression estimation but also have practical implications for the development of more accurate and accessible diagnostic tools in mental health care. In contrast to previous studies, this thesis introduces a role-based feature extraction approach, enabling a more nuanced analysis of interaction dynamics in depression estimation.

1.3 Thesis Outline

This thesis is organized into five chapters. The initial chapters provide the foundational theoretical knowledge essential for this thesis, covering both the topic of depression and the technical aspects such as feature extraction, machine learning techniques, and models. The latter part of the thesis is primarily focused on the experimental work, detailing the methodologies employed and the results obtained. More specifically:

- The present chapter (Chapter 1) provides a brief introduction, aiming to clarify the research problem and the motivation behind the study, while also outlining the overall structure of the thesis.
- Chapter 2 introduces an overview of depression and also the key concepts of artificial intelligence that are utilized for the scope of this thesis.

- Chapter 3 discusses the materials and methods used, provides an overview of the data, and describes the modeling approaches employed in the study.
- Chapter 4 details the experimental setup, presents the results obtained, and includes a discussion of the findings.
- Chapter 5 concludes the thesis and proposes directions for future research in automatic depression estimation.

Chapter 2

Background and Related Work

2.1 Depression Symptoms and Biomarkers

Depression, a complex mental health disorder, significantly impacts an individual's emotional state, cognitive functions, physical well-being and can also manifest in the individual's speech [34]. Altered speech patterns in individuals with psychiatric disorders are observed, noting characteristics such as monotone speech in depression [54]. Understanding how depression alters voice acoustics and identifying measurable features can provide valuable insights into diagnosis, severity assessment, and treatment monitoring [34, 54].

2.1.1 Symptoms of Depression and Their Impact on Speech

The Diagnostic and Statistical Manual of Mental Disorders (DSM-5) identifies psychomotor impairment as a key symptom of depression. This impairment often presents through reduced speech volume, inflection, lack of content variety, and can even be linked to muteness [54]. In addition, imbalances in serotonin, dopamine, and norepinephrine disrupt mood regulation and cognitive functions, which in turn affect vocal prosody and speech fluency. Neuroinflammation and dysregulation of the hypothalamic-pituitary-adrenal (HPA) axis influence the autonomic nervous system, leading to changes in vocal fold tension and respiratory patterns. These physiological alterations often result in flatter intonation, slower speech rate, and increased articulation errors among individuals with depression. Psychosocial factors, such as social withdrawal and reduced motivation, further compound these effects, making speech a rich source of biomarkers for depression detection [51, 77].

Moreover, the relationship between neurotransmitters and vocal characteristics is both complex and significant. For instance, gamma-aminobutyric acid (GABA), a neurotransmitter linked to increased susceptibility to depression and suicidality, has been shown to influence muscle tonicity. Alterations in muscle tension can affect the dynamics of the vocal tract, thereby restricting articulatory movements and further contributing to the speech abnormalities observed in depression [54].

In Summary the core symptoms of depression and their effects on speech are as follows:

1. **Emotional Changes:** Persistent sadness, anhedonia (loss of interest), and feelings of hopelessness. Positive emotions correlate with a higher-pitched, louder, and faster voice, while negative emotions are characterized by lower volume, slower speech, and longer pauses [54].
2. **Cognitive Impairment:** Difficulty concentrating and making decisions. Cognitive load increases pause frequency and duration in speech [34].
3. **Behavioral Changes:** Social withdrawal and reduced activity levels. This can contribute to reduced speech volume or monotone speech [54].
4. **Physical Symptoms:**
    - Psychomotor impairment manifests as reduced speech volume, inflection, lack of content variety, and can even lead to muteness [54].
    - Muscle tension, especially in the neck, shoulders, and jaw, affects the larynx and vocal cords, leading to a strained or constricted voice.
    - Reduced facial expression affects articulation, decreasing precision and clarity of speech [54].
    - Alterations in respiratory muscles affect subglottal pressure, impacting speech production [37].
    - Neurochemical imbalances (serotonin, dopamine, norepinephrine) disrupt mood regulation and cognitive functions, influencing vocal prosody and speech fluency.
    - Neuroinflammation and dysregulation of the hypothalamic-pituitary-adrenal (HPA) axis impact the autonomic nervous system, causing changes in vocal fold tension and respiratory patterns, resulting in flatter intonation, slower speech rate, and increased articulation errors.
    - Changes in gamma-aminobutyric acid (GABA) levels affect muscle tonicity, restricting articulatory movements and contributing to speech abnormalities [54].

2.1.2 Acoustic Features Affected by Depression

The symptoms of depression, though often subtle, can be systematically measured through acoustic features such as pitch variability, speaking rate, pause duration, and voice quality markers like jitter [52].

In order to determine the appropriate biomarkers, it is important to examine the process of speech. The brain organizes prosodic information, produces neuromuscular instructions that control the activities of muscles and tissues related to phonation movement. Next, the airflow stream out of the lungs either causes the vocal cords to vibrate (when the glottis is closed) or passes through the vocal cord smoothly (when the glottis is open). The oropharyngeal muscle forms the main channel of phonation, which is equivalent to a filter that can amplify or attenuate the sound of a specific frequency.

The vocal changes that occur due to depression and affect the aforementioned processed can be measured by characteristics of the speech signal. The most notable affected acoustic characteristics in depression are the following:

1. Speech production: Lower in depressed individuals.
2. Pitch and Intonation: Lower in depressed individuals
3. Intensity of sound: Lower in depressed individuals
4. Pause length and speech-to-pause ratio: Higher in depressed individuals
5. MFCCs: Analyze subtle differences in voice emotion.
6. Fundamental Frequency (F0): Generated by vocal cord vibrations, is lower in depressed individuals.
7. Zero-Crossing Rate (ZCR): Rate at which a signal crosses the zero amplitude axis. Used to differentiate voiced from voiceless sounds.
8. Harmonic-to-Noise Rarion (HNR): Reflects the strength of harmonic signals relative to noise.

In conclusion voice acoustic features hold promise as objective biomarkers for depression, with potential applications in monitoring treatment progress. Machine learning models can be used to predict depression severity and assess treatment response [58].

2.2 Audio Analysis Systems

Audio analysis systems are computational tools that extract meaningful patterns from audio signals by combining signal processing and machine learning. Audio analysis systems use feature extraction techniques to transform raw audio signals into higher-level representations, enabling applications such as emotion detection, speaker recognition, and event classification [84, 75].

Key components of audio analysis systems typically include signal acquisition, feature extraction, and machine learning integration. The feature extraction process involves analyzing temporal, spectral, and cepstral domains to capture a comprehensive representation of the audio signal. These extracted features are then used by machine learning algorithms to identify patterns and make predictions or classifications based on the audio content [84].

Audio features are categorized into three groups: high-level, mid-level, and low-level. High-level features describe abstract characteristics like rhythm, melody, tempo, and mood, while mid-level features focus on pitch, beat patterns, and MFCCs (Mel-Frequency Cepstral Coefficients). Low-level features include statistical measures such as amplitude envelope, energy, and zero-crossing rate. These features are extracted over different time frames: instantaneous (20-100 milliseconds), segment-level (2-20 seconds), and global (entire audio). Time-domain extraction reveals properties like signal energy and amplitude, while frequency-domain analysis uncovers spectral content and band energy.

Together, these features enable the manipulation of audio signals, such as noise reduction and balancing time-frequency ranges, making them essential for tasks like speech processing, music analysis, and sound classification [75].

2.2.1 Audio Features for Depression

As mentioned in section 2.1, there are several acoustic characteristics that act as indicators for depression, depression biomarkers. These characteristics can be directly related to specific audio features. This means that by using audio features, we can establish a clear correlation with depression and thus predict it using audio analysis systems [54, 58].

The following Table (table 2.1) shows the connection between depression symptoms and audio features.

Table 2.1: Depression Symptoms Related to Speech and Corresponding Acoustic Features for Tracking

Depression Symptoms (Speech)	**Acoustic Features for Tracking**
Reduced speech volume	Loudness: Measures the intensity of speech, which is often lower in depressed individuals.
Monotonous or flat speech prosody	Fundamental Frequency (F0): Tracks pitch variations, which tend to decrease in depression, leading to monotony.
Frequent pauses or slowed speech	Pause Duration and Variability: Quantifies the length and frequency of pauses, which are longer and more frequent in depression.
Limited variability in speech content	Mel-Frequency Cepstral Coefficients (MFCCs): Reflect subtle changes in vocal tract dynamics and emotional tone.
Hoarse or rough voice quality	Harmonic-to-Noise Ratio (HNR): Measures the ratio of harmonic sound to noise, which decreases in depression.
Reduced articulation or slurred speech	Zero-Crossing Rate (ZCR): Tracks transitions between voiced and voiceless sounds, often altered in depression.
Slowed speech rhythm	Speech Rate: Monitors the number of words spoken per minute, which tends to decrease in depression.

2.2.2 Audio Processing

2.2.2.1 Short-Term Audio Processing

Short-term processing is a technique in which the audio signal is divided into small overlapping or non-overlapping frames (or windows), typically lasting 20–100 milliseconds. The split into windows is important because sound signals are not static over time, on the contrary this segmentation assumes that the signal remains stationary within each frame, meaning its statistical properties do not change significantly during this short duration [85].

2.2.2.1.1 Mathematical Basis

Given a sound signal $x(n), n = 0, \ldots, N-1$ that is N samples long. At each processing step the signal is multiplied with a shifted version of a finate duration window function $w(n)$. The resulting signal at the ith step is the following:

$$x_i(n) = x(n)w(n - m_i), \quad i = 0, \ldots, K-1$$

where K is the number of frames and m_i is the number of samples by which the window is shifled in order to obtain the ith frame. Also important metrics are the window length W_L and the step size W_S.

The total number of segments in which the signal is divided is calculated as:

$$\frac{N - W_L}{W_S} + 1$$

[85]

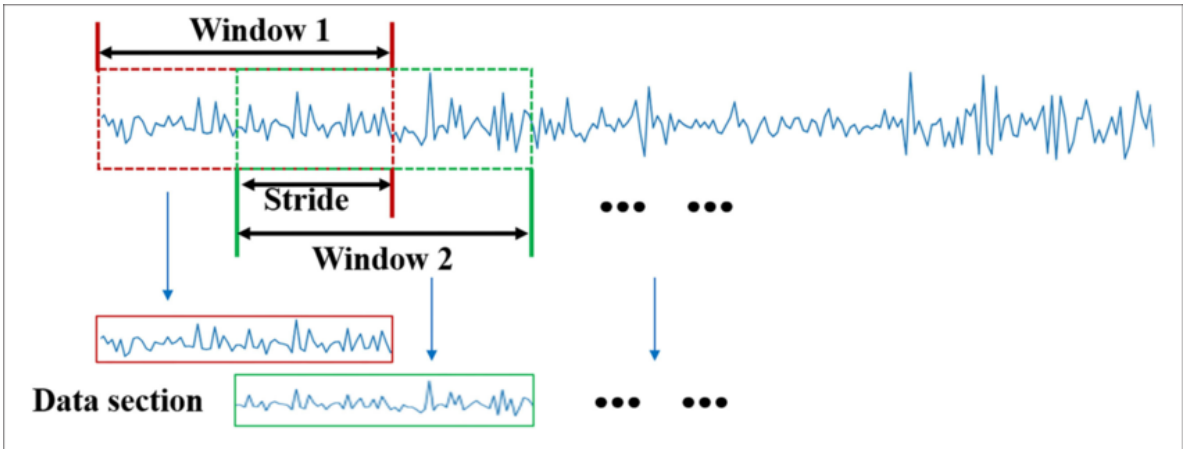

Figure 2.1: Signal Splitting in Windows [38].

2.2.2.2 Mid-Term Audio Processing

In mid-term processing, the audio signal is initially divided into larger segments, referred to as mid-term windows, which typically range in duration from 1 to 10 seconds. Each

mid-term segment undergoes short-term processing to extract features. These mid-term windows are characterized by homogeneity in their behavior, making it appropriate to compute statistical features on a segment-by-segment basis.

In many practical applications, mid-term feature extraction is employed to achieve a more comprehensive representation of the audio signal over extended periods. Instead of generating multiple feature vectors for each mid-term segment, it is often preferable to produce a single, consolidated feature vector that represents the entire sound. This is accomplished by averaging the statistical features derived from the mid-term segments, resulting in a unique vector [85].

2.2.3 Audio Features

An audio signal is a type of signal that carries information within the range of audio frequencies. Audio representation involves extracting key attributes or characteristics of an audio signal that reflect its acoustic composition—both in the time domain and frequency domain—as well as its behavior over time. This process is typically paired with feature selection, which identifies the most suitable features for the intended application of the audio signal. The primary objective is to extract features from audio data (such as speech) that can provide valuable information for training models.

The audio signal is divided into short-term windows, and specific characteristics are calculated for each window. From these, statistical values are computed over medium-term windows to summarize the signal's properties. There are numerous metrics that can be employed as features in audio analysis, and this section briefly describes some of the features used in system design.

Analyzing audio signals involves extracting various features that can be categorized into several types. These features are essential for understanding and processing audio data in applications like speech recognition, music classification, and audio event detection. Below is a comprehensive list of different types of audio features:

2.2.3.1 Time-Domain Features

- **Amplitude Envelope**: Captures the overall amplitude changes over time in an audio signal.
- **Root-Mean-Squared Energy (RMS)**: Measures the average energy level of the signal.
- **Zero Crossing Rate (ZCR)**: Indicates how often the signal crosses the zero amplitude.
- **Mean Absolute Value (MAV)**: Reflects the average absolute amplitude of the signal.
- **Standard Deviation (SD)**: Measures the variability of the signal's amplitude.
- **Kurtosis and Skewness**: Provide insights into the distribution of the signal's amplitude.

2.2.3.2 Frequency-Domain Features

- **Mel-Frequency Cepstral Coefficients (MFCCs)**: Capture the spectral envelope of an audio signal.
- **Spectral Centroid**: Measures the weighted mean of the frequency components.
- **Band Energy Ratio**: Compares the energy levels across different frequency bands.
- **Spectrogram**: A visual representation of the frequency content over time.
- **Spectral Flux**: Measures the rate of change of the spectral power distribution over time.
- **Spectral Rolloff**: The frequency below which a certain percentage of the signal's total energy is contained [67, 89, 73].

2.2.3.3 Perceptual Features

- **Gammatone-Frequency Cepstral Coefficients (GFCCs)**: Formed by passing the spectrum through a Gammatone filter bank.
- **Bark-Frequency Cepstral Coefficients (BFCCs)**: Based on the Bark scale.
- **Power-Normalized Cepstral Coefficients (PNCCs)**: Designed to improve robustness against noise and channel effects [73].

2.2.3.4 Dynamic Features

- **Delta Coefficients**: Represent the rate of change of static features over time.
- **Acceleration Coefficients**: Measure the rate of change of delta coefficients.
- **Temporal Trajectories**: Describe the evolution of features over longer intervals [67, 89].

2.2.3.5 Cepstral Features

- **Mel-Frequency Cepstral Coefficients (MFCCs)**: Capture the spectral envelope of an audio signal.
- **Gammatone-Frequency Cepstral Coefficients (GFCCs)**: Similar to MFCCs but use a Gammatone filter bank.
- **Bark-Frequency Cepstral Coefficients (BFCCs)**: Based on the Bark scale [73].

2.2.3.6 Prosodic Features

- **Pitch**: The perceived highness or lowness of a sound.
- **Intonation**: The rise and fall of pitch in speech.

- **Speed of Speech**: Influences the perception and understanding of spoken language [89, 73].

2.2.3.7 Chroma Features

- **Chroma Energy Normalized (CENS)**: Represents the distribution of energy across different musical notes.
- **Chroma STFT**: A short-time Fourier transform representation of chroma features [67].

2.2.3.8 Other Features

- **Autocorrelation**: Measures the similarity of the signal with itself at different time lags.
- **Cross-Correlation**: Measures the similarity between two different signals [89].

2.2.4 pyAudioAnalysis Features

pyAudioAnalysis is a Python library designed for audio analysis tasks such as feature extraction, segmentation, classification, and visualization. It implements both short-term and mid-term processing methodologies. The library supports various audio analysis tasks, including feature extraction from time and frequency domains, classification, regression, and segmentation [83].

The following Table contains the short-term audio features that the library extracts:

It is also important to note that there is a variety of audio feature extraction libraries are available for affective computing, each with distinct strengths and limitations. OpenSMILE is widely recognized for its comprehensive set of low-level descriptors, including pitch, energy, and spectral features, and is frequently used in emotion recognition challenges. Librosa, a Python-based library, offers a flexible framework for audio analysis, allowing for rapid prototyping and custom feature extraction, though it is less specialized for affective signals. pyAudioAnalysis, used in this thesis, provides a balance between ease of use and a robust set of features tailored for speech and music analysis, making it suitable for depression estimation tasks. However, each library presents challenges in terms of computational efficiency, scalability, and compatibility with real-world noisy data.

2.2.5 Pretrained Audio Embeddings

The process of audio analysis, as previously mentioned, encompasses a wide range of tasks such as speech recognition, speaker identification, and emotion recognition. Central to all these tasks is the need for effective audio representations. Traditionally handcrafted audio features, like those described in the earlier section, are used.

Although handcrafted features have achieved notable success, they come with significant limitations. First, they require a great deal of manual effort and fine-tuning to adapt to different audio tasks. Additionally, these features primarily capture low-level

Index	Name	Description
1	Zero Crossing Rate	The rate of sign-changes of the signal during the duration of a particular frame.
2	Energy	The sum of squares of the signal values, normalized by the respective frame length.
3	Entropy of Energy	The entropy of sub-frames' normalized energies. It can be interpreted as a measure of abrupt changes.
4	Spectral Centroid	The center of gravity of the spectrum.
5	Spectral Spread	The second central moment of the spectrum.
6	Spectral Entropy	Entropy of the normalized spectral energies for a set of sub-frames.
7	Spectral Flux	The squared difference between the normalized magnitudes of the spectra of the two successive frames.
8	Spectral Rolloff	The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.
9–21	MFCCs	Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale.
22–33	Chroma Vector	A 12-element representation of the spectral energy where the bins represent the 12 equal-tempered pitch classes of western-type music (semitone spacing).
34	Chroma Deviation	The standard deviation of the 12 chroma coefficients.

Table 2.2: Audio Features Extracted by pyAudioAnalysis [83].

acoustic details and often fail to represent higher-level information, such as emotions or environmental context. Furthermore, handcrafted features can be sensitive to background noise and recording conditions, which reduces their robustness.

To address these shortcomings, researchers have increasingly utilized pretrained audio embeddings. These embeddings are vector representations generated by deep neural networks trained on large and diverse collections of audio data, often using supervised or self-supervised learning methods. Unlike handcrafted features, these embeddings are learned automatically, allowing the model to discover optimal representations for capturing complex acoustic and semantic patterns [49].

A notable example of this approach is the work by Kong et al. (2019) on Pretrained Audio Neural Networks (PANNs). The PANNs architecture, particularly the Wavegram-Logmel-CNN model, combines learnable waveform-based features (wavegram) with traditional Mel-spectrogram inputs to leverage complementary information from raw audio and spectral representations.

PANNs achieved excellent results on audio tagging tasks, performing better than traditional feature-based methods and even other deep learning models that were trained from scratch. What makes PANNs especially valuable is that the audio embeddings they produce can be used for many different tasks. For example, these embeddings worked well for classifying different acoustic environments and detecting specific sound events. This flexibility shows that pretrained audio embeddings are powerful and can serve as general-purpose representations for a variety of audio analysis applications [42].

2.2.6 wav2vec 2.0

As discussed earlier, pretrained audio embeddings have revolutionized audio signal processing by providing transferable representations learned from large-scale data. Among these, wav2vec 2.0 stands out as a landmark model that learns speech representations directly from raw audio waveforms through self-supervised learning. Unlike traditional handcrafted features or earlier neural network embeddings, wav2vec 2.0 effectively captures both low-level acoustic patterns and higher-level linguistic structures, leading to improved results across a variety of speech-related tasks [33].

The wav2vec 2.0 framework is designed for self-supervised learning of speech representations by masking segments of latent audio features and training the model to solve a contrastive task over discrete, quantized speech units. More specifically, the model first learns powerful representations from large amounts of unlabeled speech audio. Subsequently, it is fine-tuned on smaller sets of transcribed speech data, outperforming previous semi-supervised approaches that relied heavily on labeled data. This is particularly important because, although neural networks generally benefit from large volumes of annotated data, such labeled datasets are often scarce and expensive to obtain.

The core idea behind wav2vec 2.0 is to encode raw speech audio using a multi-layer convolutional neural network, producing latent speech representations. Portions of these latent features are then masked, in a manner similar to masked language modeling in natural language processing. The masked latent representations are passed through

a Transformer network, which constructs contextualized embeddings by capturing dependencies across the entire sequence. The model is trained with a contrastive loss, where it must correctly identify the true latent representation from a set of distractors, thereby encouraging the learning of meaningful and discriminative features [33].

2.2.6.1 wav2vec 2.0 Architecture

The model architecture consists of several key components, as pictured in Figure 2.2:

- A multi-layer convolutional **feature encoder** defined as a function:

  $$f : \mathcal{X} \rightarrow \mathcal{Z}$$

  which takes raw audio input $\mathbf{X}$ and produces latent speech representations

  $$\mathbf{z}_1, \ldots, \mathbf{z}_T$$

  over T discrete time steps.

- These latent representations are then fed into a **Transformer network**

  $$g : \mathcal{Z} \rightarrow \mathcal{C}$$

  which generates contextualized outputs

  $$\mathbf{c}_1, \ldots, \mathbf{c}_T$$

  that capture information from the entire audio sequence.

- To enable the self-supervised learning objective, the continuous latent outputs of the feature encoder are discretized into quantized representations $\mathbf{q}_t$ through a **quantization module**.

  $$\mathcal{Z} \rightarrow \mathcal{Q}$$

The following provides a detailed analysis of each component of the model:

1. **Feature Encoder:** The feature encoder is composed of multiple blocks, each containing a temporal convolutional layer followed by layer normalization and a GELU activation function. Before processing, the raw waveform input is normalized to have zero mean and unit variance, which helps stabilize training. The encoder's total stride determines the temporal resolution of the latent representations, i.e., the number of time steps T that are subsequently input to the Transformer network.

2. **Contextualized Representations with Transformers:** The latent speech representations produced by the feature encoder are passed to a context network based on the Transformer architecture. Rather than using fixed positional embeddings that encode absolute positions, wav2vec 2.0 employs a convolutional layer to provide relative positional information. The output of this convolutional layer is combined with the input representations, followed by a GELU activation and layer normalization, which improves the model's ability to capture relative positional dependencies in the speech signal.

3. **Quantization Module:** For the self-supervised training objective, the continuous latent representations $\mathbf{z}$ from the feature encoder are discretized into a finite set of speech units using product quantization. This approach was shown to be effective in previous work that first learned discrete speech units and then trained contextualized models. Product quantization works by selecting quantized vectors from multiple codebooks and concatenating them to form a discrete representation. Specifically, given G codebooks (also called groups), each containing V entries $\mathbf{e} \in \mathbb{R}^{V \times d/G}$, the model selects one entry from each codebook. These selected vectors $\mathbf{e}_1, \ldots, \mathbf{e}_G$ are concatenated and passed through a linear transformation

$$\mathbb{R}^d \rightarrow \mathbb{R}^f$$

to produce the final quantized representation $\mathbf{q} \in \mathbb{R}^f$.

To allow the selection of discrete codebook entries to be differentiable, wav2vec 2.0 employs the Gumbel softmax technique. Using the straight-through estimator, the model performs G hard Gumbel softmax operations. The feature encoder output $\mathbf{z}$ is first mapped to logits $\mathbf{l} \in \mathbb{R}^{G \times V}$, and the probability of selecting the v-th entry in the g-th codebook is calculated as:

$$p_{g,v} = \frac{\exp\left(\frac{l_{g,v}+n_v}{\tau}\right)}{\sum_{k=1}^{V} \exp\left(\frac{l_{g,k}+n_k}{\tau}\right)}, \tag{2.1}$$

where τ is a non-negative temperature parameter controlling the softness of the distribution, $n = -\log(-\log(u))$ represents Gumbel noise, and u are samples drawn uniformly from the interval $U(0, 1)$. During the forward pass, the discrete codeword i is selected as

$$i = \arg\max_j p_{g,j},$$

while during the backward pass, gradients are propagated through the softmax outputs using the straight-through estimator, enabling end-to-end differentiable training.

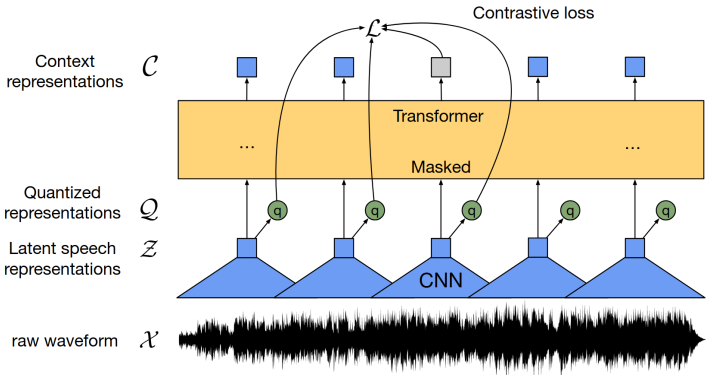

Figure 2.2: Illustration of our framework which jointly learns contextualized speech representations and an inventory of discretized speech units [33]

2.2.6.2 wav2vec 2.0 Training

Self-Supervised Pre-Training

The core innovation in wav2vec 2.0 is it's self-supervised pre-training objective, which is inspired by masked language modeling (as used in BERT for text). The pre-training process involves the following steps:

- **Masking:** A certain proportion of the latent representations output by the feature encoder are randomly masked. These masked representations are replaced with a learned mask vector before being fed into the context network.
- **Contrastive Learning Objective:** For each masked time step, the model is tasked with identifying the correct quantized latent audio representation from a set of distractors. This is a contrastive task: the model must distinguish the true quantized representation from several negative samples. The contrastive loss encourages the model to learn representations that are predictive of the underlying speech content, even in the absence of labels.
- **Joint Learning:** The model jointly learns the quantization codebooks and the contextualized speech representations. This end-to-end approach is more effective than previous methods that learned discrete units and contextual representations in separate steps.

Fine-Tuning for Speech Recognition

After pre-training on large amounts of unlabeled speech, wav2vec 2.0 can be fine-tuned for downstream speech recognition tasks using a relatively small amount of labeled data. The pre-trained model is connected to a linear layer and trained with Connectionist Temporal Classification (CTC) loss on transcribed speech. Fine-tuning adapts the learned representations for the specific task of speech-to-text conversion.

In conclusion, wav2vec 2.0 represents a major step forward in self-supervised learning for speech. By learning powerful and transferable representations from raw audio, it greatly reduces the need for large labeled datasets and opens up new possibilities for speech recognition in low-resource languages and domains. The model's architecture—combining a CNN encoder, Transformer context network, and quantization module—enables it to capture both local and global patterns in speech, making it a versatile foundation for a wide range of speech processing tasks [33].

2.3 Text Analysis Systems

Similarly to Audio Analysis Systems, there are two main tasks with the Text Analysis Systems. The first is Text Preprocessing and Generating Text Embeddings.

2.3.1 Text Preprocessing

Text preprocessing is a crucial step that transforms raw, unstructured text into a structured format suitable for analysis and modeling. This process includes several

techniques to remove noise and inconsistencies, making the data more uniform and manageable for NLP models. Key techniques include:

- **Segmentation**: Breaking down text into sentences or smaller units.
- **Tokenization**: Splitting text into individual words or tokens.
- **Lowercasing**: Converting all text to lowercase for uniformity.
- **Removal of Punctuation and Special Characters**: Removing non-alphanumeric characters.
- **Stopword Removal**: Removing common words like "the," "and," etc., that do not add much value to the meaning.
- **Stemming/Lemmatization**: Reducing words to their base form to reduce vocabulary size. Lemmatization preserves the meaning by converting words to their base form using dictionaries, unlike stemming, which may lose meaning.
- **Text Normalization**: Standardizing words to their canonical form (e.g., "real time" to "realtime") [27].

2.3.2 Text Embeddings

After preprocessing, the next step is the embeddings extraction, which can be done using either Word Embeddings or Sentence Embeddings.

- **Word Embeddings**: Techniques like Word2Vec and GloVe create word-level embeddings based on word contexts. Word2Vec uses windowed text sampling to create embeddings for individual words, while GloVe uses global matrix factorization.
- **Sentence Embeddings**: Models like BERT, SBERT and OpenAI's text-embedding models generate embeddings for entire sentences or documents, capturing contextual semantics effectively.

2.3.3 GloVe Embeddings

2.3.3.1 Introduction

GloVe (Global Vectors for Word Representation) is a widely used method for learning word embeddings—dense vector representations of words—by leveraging global word co-occurrence statistics from a corpus. Unlike earlier methods that focus either on global matrix factorization (like LSA) or local context window predictions (like word2vec), GloVe effectively combines the strengths of both approaches to yield embeddings that capture both the global statistical information and the linear substructure of word meaning.

Traditional approaches to word embeddings fall into two main categories:

- **Global Matrix Factorization Methods:** Techniques such as Latent Semantic Analysis (LSA) decompose large matrices (e.g., term-document or term-term matrices) to capture statistical information about word occurrences in a corpus.

While these methods efficiently use global statistics, they often fail to capture certain linguistic regularities, such as analogical relationships (e.g., "king" − "man" + "woman" ≈ "queen").

- **Local Context Window Methods:** Models like skip-gram and CBOW (continuous bag-of-words) predict words based on their local context windows. These methods excel at capturing fine-grained semantic and syntactic relationships but do not fully exploit the global co-occurrence statistics present in the corpus.

GloVe was designed to bridge this gap by constructing word vectors that directly encode the ratios of co-occurrence probabilities, which are shown to be particularly informative for distinguishing word meanings.

2.3.3.2 Theoretical Foundation

The core insight behind GloVe is that word meaning can be captured by examining how frequently words co-occur with other words across a large corpus. Specifically, GloVe models the ratios of co-occurrence probabilities. Consider two target words, i and j, and a context word k. The ratio of probabilities that k appears in the context of i versus j (P_{ik}/P_{jk}) can highlight aspects of meaning that differentiate i from j.

For example, the ratio of the probability that "solid" appears with "ice" versus "steam" is much greater than one, indicating a strong association with "ice." Conversely, the ratio for "gas" is much less than one, indicating a strong association with "steam." Ratios close to one (e.g., for "water") indicate words equally related to both.

2.3.3.3 Model Formulation

Let $\mathbf{X}$ be the word-word co-occurrence matrix, where X_{ij} is the number of times word j appears in the context of word i. The probability that word k appears in the context of word i is:

$$P_{ik} = \frac{X_{ik}}{X_i}$$

where

$$X_i = \sum_k X_{ik}$$

is the total number of context word appearances for word i.

GloVe seeks to find word vectors $\mathbf{w}_i$ and context word vectors $\tilde{\mathbf{w}}_k$ such that their dot product, plus bias terms, approximates the logarithm of the co-occurrence count:

$$\mathbf{w}_i^\top \tilde{\mathbf{w}}_k + b_i + \tilde{b}_k \approx \log(X_{ik})$$

This formulation is motivated by the desire for the model to be invariant to the exchange of word and context word roles, and for the vector space to encode the linear relationships found in word analogies.

2.3.3.4 Weighted Least Squares Objective

A naïve approach would factorize the log co-occurrence matrix directly. However, this would treat all co-occurrences equally, including rare or zero co-occurrences, which are noisy and less informative. To address this, GloVe introduces a weighted least squares regression objective:

$$J = \sum_{i,j=1}^{V} f(X_{ij}) \left(\mathbf{w}_i^\top \tilde{\mathbf{w}}_j + b_i + \tilde{b}_j - \log(X_{ij}) \right)^2$$

where V is the vocabulary size, and $f(x)$ is a weighting function that controls the influence of each co-occurrence pair. The function $f(x)$ is designed to:

- Be zero when $x = 0$ (so pairs that never co-occur do not contribute).
- Increase with x, so frequent co-occurrences are given more weight.
- Saturate for very large x, to prevent extremely frequent pairs from dominating.

A common choice for the weighting function is:

$$f(x) = \begin{cases} \left(\frac{x}{x_{\max}} \right)^\alpha & \text{if } x < x_{\max} \\ 1 & \text{otherwise} \end{cases}$$

with typical values $\alpha = 0.75$ and $x_{\max} = 100$.

2.3.3.5 Training Procedure

The GloVe model is trained by minimizing the objective function J over all word and context word vectors and bias terms, using stochastic gradient descent or similar optimization methods. The training process iterates over the nonzero entries of the co-occurrence matrix, updating the vectors and biases to best fit the observed log co-occurrence counts.

After training, the final embedding for each word can be taken as the sum (or average) of its word and context word vectors.

2.3.3.6 Properties and Advantages

- **Linear Substructure:** GloVe embeddings capture linear relationships, making them suitable for analogy tasks (e.g., "king" − "man" + "woman" ≈ "queen").
- **Efficient Use of Statistics:** By focusing on nonzero co-occurrences and weighting them appropriately, GloVe efficiently leverages the vast statistical information present in large corpora.
- **Scalability:** The model can be trained on very large corpora, producing high-quality embeddings for extensive vocabularies.

GloVe embeddings are a powerful and efficient way to learn word representations that encode both semantic and syntactic regularities. By directly modeling the global co-occurrence statistics of words, GloVe produces vector spaces with meaningful substructure, outperforming many previous methods on tasks such as word similarity and analogy. The resulting embeddings have become a standard tool in modern NLP pipelines [48].

2.3.4 SBERT Embeddings

2.3.4.1 Introduction to BERT Embeddings

BERT stands for Bidirectional Encoder Representations from Transformers. It is a groundbreaking model introduced by Google in 2018 that revolutionized natural language processing (NLP) by providing deep, contextualized word embeddings. Unlike previous models that read text either left-to-right or right-to-left, BERT reads text bidirectionally, meaning it considers the entire context of a word by looking at the words before and after it simultaneously. This bidirectional approach allows BERT to capture more nuanced meanings and relationships in language.

BERT uses a transformer architecture that employs self-attention mechanisms to weigh the importance of each word in a sentence relative to others. It generates word representations that change depending on the context, unlike static embeddings such as Word2Vec or GloVe. BERT is first pre-trained on large corpora with unsupervised tasks like masked language modeling and next sentence prediction, and then fine-tuned on specific NLP tasks such as question answering, sentiment analysis, or named entity recognition.

However, while BERT excels at understanding word-level context, it was not originally designed to produce fixed-size sentence embeddings that can be directly compared. For tasks like semantic similarity or clustering, a single vector representation per sentence is needed. The problem arises because to compare two sentences using standard BERT, both sentences must be input together as a pair, which leads to high computational costs. For example, comparing 10,000 sentences pairwise would require processing about 50 million pairs, making real-time or large-scale applications impractical.

2.3.4.2 Introduction to SBERT Embeddings

Sentence-BERT (SBERT) addresses this limitation by modifying the original BERT architecture to create a dual (or twin) network structure where each sentence is encoded independently into a fixed-size vector embedding, as shown in fig. 2.4. This means each sentence is passed through the network separately, producing a fixed-length vector that represents the sentence's meaning. These sentence embeddings can then be compared efficiently using simple similarity measures like cosine similarity.

The benefits of SBERT include scalability, enabling fast, large-scale semantic search and clustering; efficiency, by drastically reducing computational costs through avoiding pairwise input processing; and practicality, making it suitable for real-world applications such as information retrieval, duplicate detection, and question answering.[79].

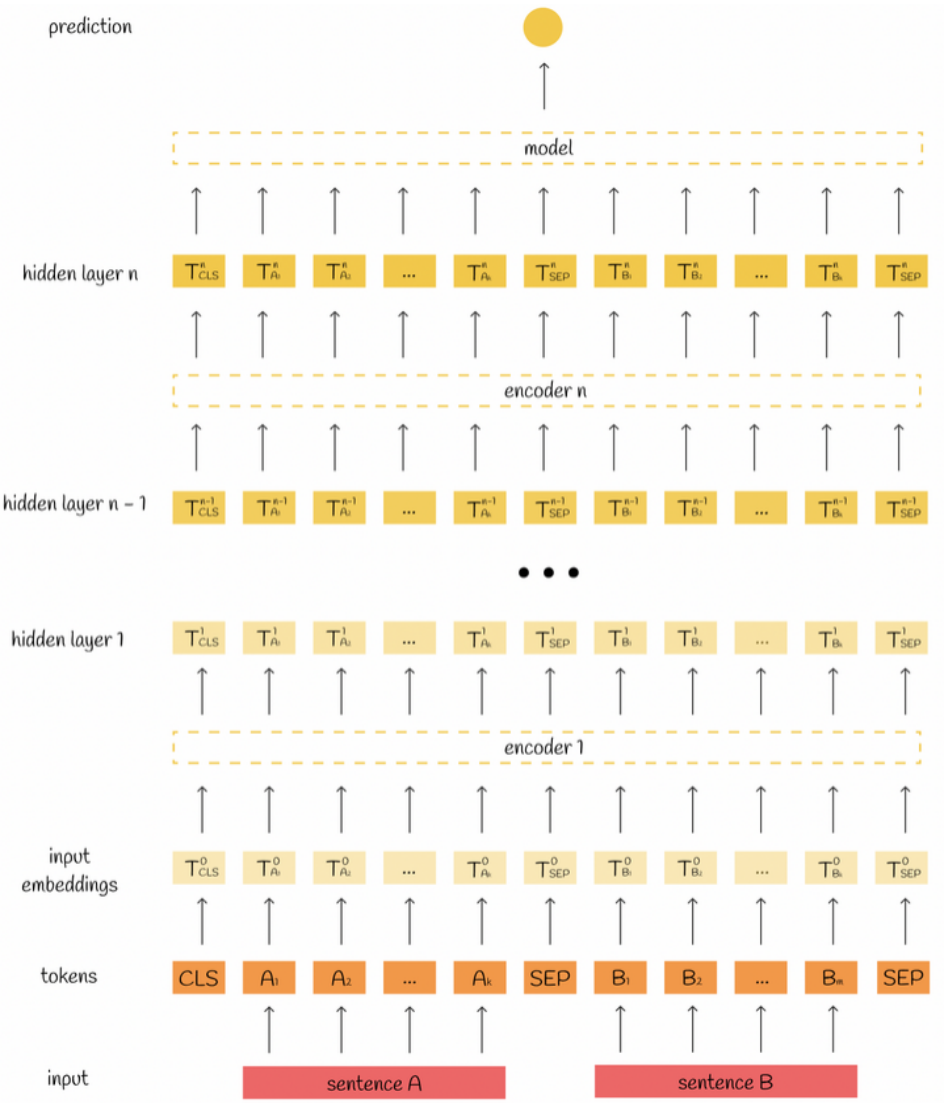

Figure 2.3: BERT Architecture [88].

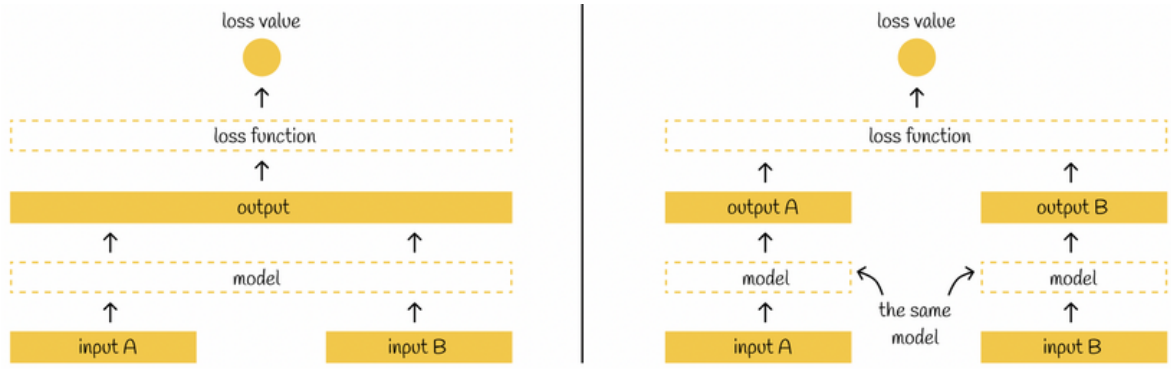

Figure 2.4: The non-Siamese (cross-encoder) model, shown on the left, processes both inputs simultaneously. In contrast, the Siamese (bi-encoder) model on the right handles inputs independently and in parallel, so each output is generated without depending on the other [88].

2.3.4.3 SBERT Architecture and Training

SBERT is fine-tuned on datasets such as the Stanford Natural Language Inference (SNLI) and Semantic Textual Similarity (STS) datasets. The main training objectives are:

1. Classification Loss: Used for natural language inference tasks, where the model learns to classify the relationship between sentence pairs (e.g., entailment, contradiction, neutral).
2. Regression Loss: Used for semantic similarity tasks, where the model learns to predict a similarity score between pairs of sentences.

These objectives ensure that semantically similar sentences are mapped close together in the embedding space, while dissimilar sentences are mapped further apart.

Since BERT outputs a sequence of token embeddings, SBERT applies a pooling operation to generate a single fixed-size vector for each sentence. The most effective pooling strategy found is mean pooling, which averages the token embeddings. Alternatives like max pooling and using the [CLS] token were also explored, but mean pooling generally yields the best results [79].

2.3.4.4 SBERT Use Cases

SBERT embeddings are widely used for:

- Semantic Search: Retrieving documents or passages based on semantic similarity to a query.
- Clustering: Grouping similar sentences or documents for tasks like topic modeling.
- Paraphrase Identification: Detecting duplicate or near-duplicate questions in forums or databases.
- Information Retrieval: Matching user queries to relevant answers or documents [79, 20].

2.3.4.5 Advantages and Limitations of SBERT

SBERT allows precomputing embeddings for large corpora, enabling rapid similarity searches using vector operations. As far as semantic quality, SBERT embeddings capture deeper semantic relationships than simple word averaging or the [CLS] token approach. Additionally SBERT's architecture supports large-scale applications that would be infeasible with vanilla BERT.

However SBERT's performance is dependent on the quality and diversity of its training data. For highly nuanced or context-specific tasks, further fine-tuning may be necessary.

In conclusion SBERT represents a significant advancement in sentence embedding technology, addressing the limitations of BERT for semantic similarity tasks. Its efficiency, semantic richness, and scalability make it a crucial tool in both research and industry, powering applications [79, 20, 88].

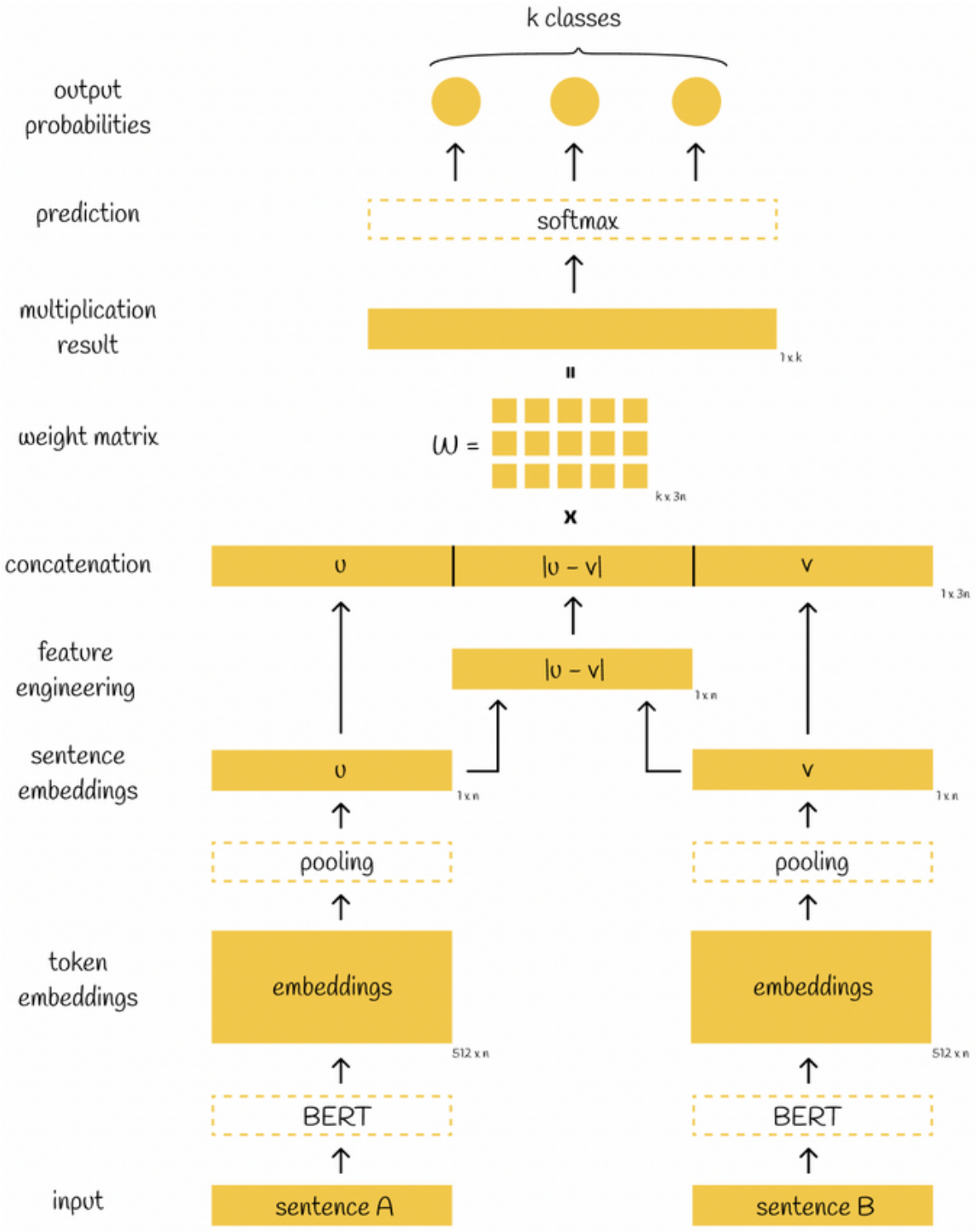

Figure 2.5: The SBERT classification architecture uses embeddings of size n and outputs k labels, where k is the number of classes [88].

2.4 Machine Learning Theoretical Background

This section provides a comprehensive theoretical overview of artificial intelligence and machine learning, for understanding the methodologies employed in this thesis.

2.4.1 Algorithms

a. Support Vector Machines

Support Vector Machines (SVMs) in machine learning are supervised learning models that are used for data-driven modeling and classification.

The Support Vector Machine (SVM) algorithm operates by creating a decision boundary, referred to as a hyperplane, to separate data points into distinct classes within a high-dimensional feature space [76]. Specifically, in a two-dimensional space, the hyperplane is a line that partitions data points into two classes. Extending this concept, in an N-dimensional space, a hyperplane is defined as having (N-1) dimensions, facilitating the classification process [74].

For a given classification problem, multiple hyperplanes may be viable. However, the objective of the SVM is to identify the hyperplane that maximizes the distance between the hyperplane and the closest data points from each class, i.e. the margin [76]. A larger margin increases confidence in classification because it shows a clear separation between the decision boundary and the nearest data points. Thus, the margin indicates how well the classes are separated within the feature space [74].

The data points closest to the hyperplane are known as support vectors. These support vectors determine the hyperplane's position and orientation, greatly affecting the SVM's classification accuracy. In fact, SVMs are named after these support vectors because they "support" or define the decision boundary. Support vectors are essential for calculating the margin [74].

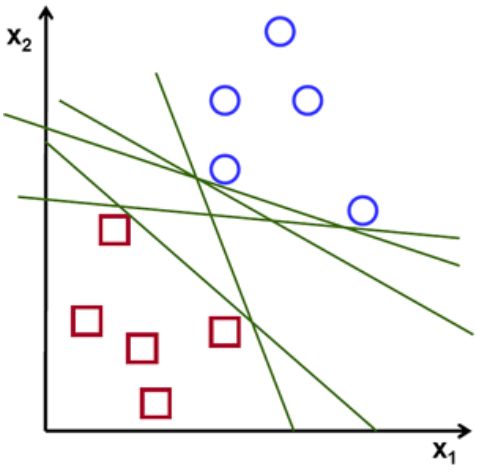

Figure 2.6: Possible hyperplanes

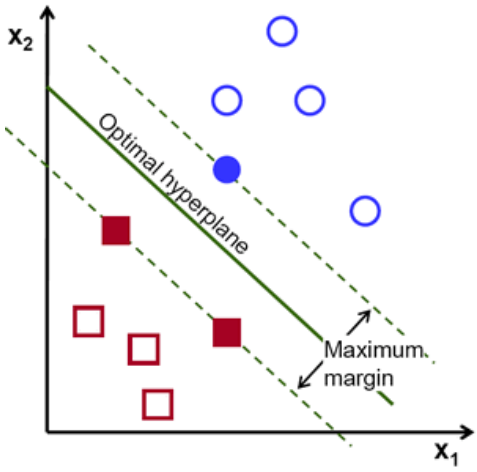

Figure 2.7: Optimal hyperplane

2.4.1.1 Hard and Soft Margin

- Hard Margin SVM: This is the ideal scenario where a decision boundary perfectly separates all data points, with no misclassifications.
- Soft Margin SVM: This approach allows some data points to be misclassified or lie within the margin, balancing margin maximization and error minimization.

2.4.1.2 Mathematical Formulation

For a linearly separable dataset, the model can be defined mathematically as follows:

$function = sign(\mathbf{wx} + b)$

where the hyperplane that we mentioned above is $\mathbf{wx} + b = 0$.

For every data-point the following constraints are enforced:

- $\mathbf{wx} - b \geq 1$, for $y_i = +1$
- $\mathbf{wx} - b \leq 1$, for $y_i = -1$

The SVM optimization problem is to minimize the norm of $\mathbf{w}$, $\|\mathbf{w}\| = \sqrt{\sum_{j=1}^{D} w_j^2}$ and enforcing the constrain $y_i(\mathbf{w}x_i - b) \geq 1$ for every data-point.

2.4.1.3 Parameters

When using Support Vector Machines (SVMs), several key parameters need to be tuned for optimal performance. The parameters we tuned in our experiments are the following:

- Kernel Function: Most commonly used 'linear', 'poly', 'rbf', 'sigmoid'. Different kernels allow SVMs to handle different types of data. For example, linear kernels are suitable for linearly separable data, while RBF kernels are more versatile for non-linear data.
- Regularization Parameter (C): Controls the trade-off between margin maximization and misclassification error. A higher value of C means a higher penalty for misclassifications, potentially leading to overfitting. Increasing C can lead to more complex models that fit the training data better but may not generalize well to new data. Often set to 1.0, but needs tuning based on the dataset.
- Kernel Coefficient (Gamma): Used in RBF, polynomial, and sigmoid kernels. It determines how much influence a single data point has on the decision boundary. A smaller gamma value means a larger influence of each data point, potentially leading to overfitting. Often set to 'scale' or 'auto', which automatically computes gamma based on the data.

b. Gradient Boosting

Gradient boosting is a machine learning technique that builds a strong predictive model by combining multiple weak models, typically decision trees, into a stronger ensemble model. As weak models can be defined models that are only slightly better than random

guessing [65]. This method focuses on correcting errors of earlier models by optimizing a loss function through gradient descent [25, 60].

Key components of gradient boosting include:

- Weak Learners: Decision trees are commonly used due to their simplicity and ability to model non-linear relationships effectively [65].
- Additive Model: Predictions from all weak learners are summed to form the final output, with each new learner trained on residuals from previous steps [60].
- Loss Function: The choice of loss function depends on the problem type and must be differentiable to facilitate optimization [22].

Gradient boosting has become a cornerstone in machine learning due to its flexibility and effectiveness in handling structured data [22].

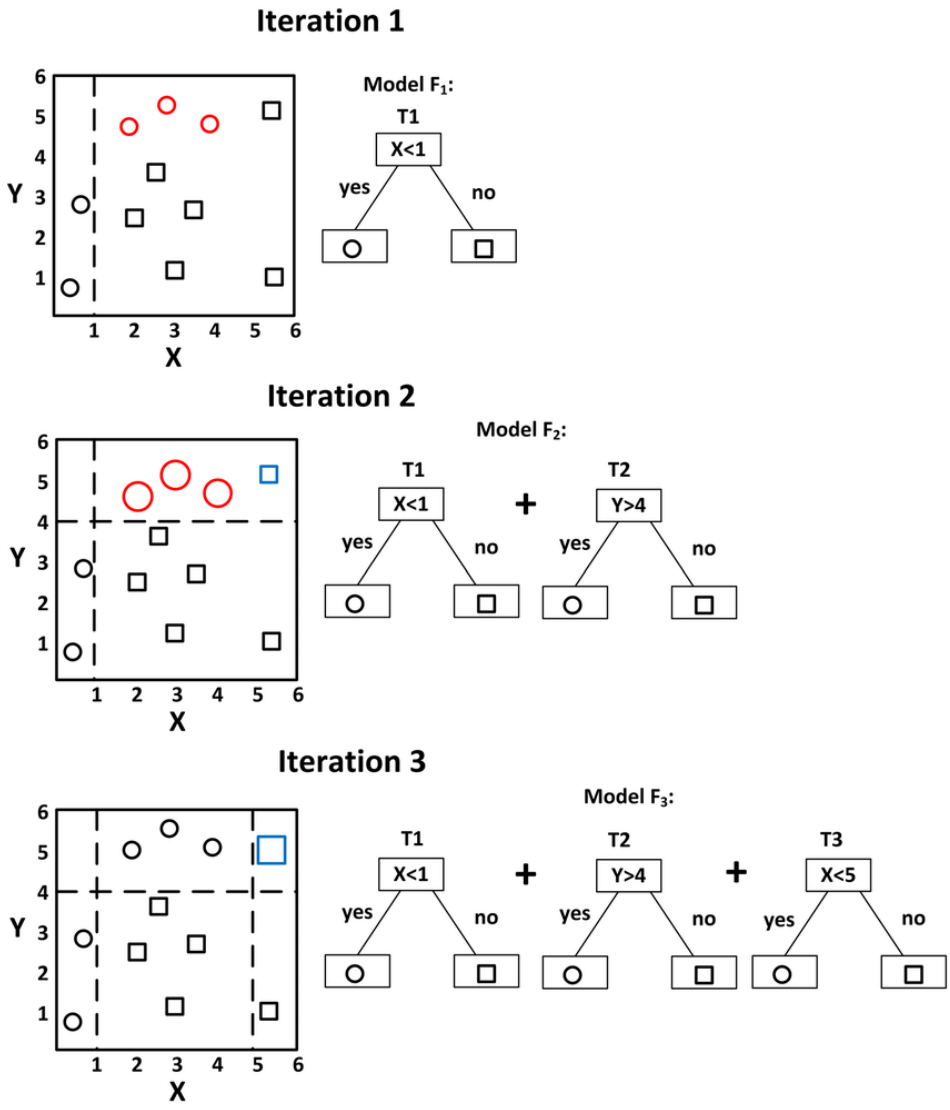

Figure 2.8: Simple gradient boosting example [57].

c. XGBoost

XGBoost, short for eXtreme Gradient Boosting, is a powerful machine learning library that utilizes gradient boosted decision trees to enhance predictive performance.

XGBoost has become renowned for its scalability and efficiency in handling large datasets. It incorporates several key features that distinguish it from other gradient boosting algorithms, such as its ability to handle sparse data efficiently using the weighted quantile sketch algorithm. Additionally, XGBoost offers robust regularization techniques, including L1 and L2 penalties, to prevent overfitting. Its parallel tree structure and cache-aware block design enable fast computation on multicore systems, making it highly suitable for complex data analysis tasks. Overall, XGBoost's combination of speed, scalability, and robust performance has made it a popular choice in machine learning competitions and real-world applications.

2.4.2 Cross-Validation

Cross-validation is a fundamental technique in machine learning for evaluating model performance and avoiding overfitting. Instead of relying on a single train-test split (e.g., 70% training and 30% testing), cross-validation systematically partitions the data to ensure robust evaluation. The most widely used method is k-fold cross-validation, where the training data is divided into k equal partitions. In each iteration, one partition is used as the validation set, while the remaining k-1 partitions are used for training. This process repeats k times, with each partition serving as the validation set once. The model's performance is recorded for each iteration, and the final performance metric is the average across all folds. While k-fold cross-validation provides a more reliable estimate of a model's generalization ability, it can be computationally intensive due to its iterative nature. Despite this, it remains a cornerstone of model evaluation, especially when integrated with techniques like grid search for hyperparameter tuning [78, 4].

2.4.2.1 Leave-One-Out Cross-Validation

Leave-One-Out Cross-Validation (LOOCV) is a specialized cross-validation technique where each individual observation in the dataset serves as the validation set once, while the remaining observations form the training set. This process is repeated n times (for n total samples), ensuring every data point is used exactly once for validation [62].

LOOCV is ideal for small datasets because it minimizes the use of available training data, unlike k-fold validation, which reserves a portion of data for validation. This minimizes bias in performance estimation and ensures robust hyperparameter tuning during gridsearch [1, 91]. By averaging results across all iterations, LOOCV provides a nearly unbiased estimate of model generalization error [62, 91]. Additionally, gridsearch with LOOCV mitigates the overfitting risk by evaluating hyperparameters across diverse training-validation splits and thus reducing the likelihood of overfitting to a specific subset. This is crucial for small datasets, where random splits might disproportionately affect model evaluation [1]

A key limitation of LOOCV is its potential for high variance in error estimation because each iteration trains the model on nearly identical datasets (n-1 samples), the resulting performance metrics are highly correlated and and may fail to reliably capture the model's true generalization ability [1].

2.4.3 Hyperparameter Tuning

Hyperparameters are essential external variables in machine learning that optimize model training. They are set manually before training and control how the model learns from data. By defining these hyperparameters, the model's development can be tailored to achieve specific goals, influencing its structure and behavior. This optimization is crucial for improving model performance and ensuring it meets desired outcomes. Hyperparameters are often associated with a model's architecture, learning rate, and complexity. In contrast, parameters are internal values that a model learns and continually updates during training to find the optimal settings that best fit the data [12, 61].

Identifying hyperparameters requires understanding the specific machine learning algorithm being used, as each model has its unique set of configurations [61].

Hyperparameter tuning involves selecting the best values for a machine learning model's hyperparameters. This process typically includes setting a range of possible values for each hyperparameter, training the model with different combinations, and evaluating performance on a validation set. The aim is to find a balance that avoids underfitting and overfitting. Hyperparameter tuning can be done manually, relying on intuition and observation, or automatically using systematic search methods. The best strategies for hyperparameter tuning are:

- Gridsearch: Gridsearch is the method chosen for the experimental part of this thesis. Gridsearch is a hyperparameter tuning technique that systematically performs an exhaustive search over a predefined grid of hyperparameters. In this context, the grid represents all possible combinations of hyperparameters and their corresponding values. Grid search evaluates each combination of these hyperparameters, that correspond to a "grid" point, to identify the set that yields the best model performance, typically measured using cross-validation [70].
- Randomized Search
- Bayesian Optimization[61]

2.4.3.1 Hyperparameters in Support Vector Machines (SVMs)

Support vector machines are highly dependent on hyperparameters like the kernel type, regularization parameter and gamma.

2.4.3.2 Hyperparameters in XGBoost

In XGBoost, there are two main types of hyperparameters: tree-specific and learning task-specific.

Hyperparameter	Description and Effect
Kernel Type	Defines the type of decision boundary (e.g., linear, polynomial, radial basis function).
Regularization Parameter (C)	Controls the trade-off between fitting the training data closely and maintaining a smooth decision boundary. - Small C: More generalized model - Large C: Focuses on fitting training data
Gamma (γ)	Determines the influence of a single training point. - High γ: Captures fine-grained patterns, risks overfitting - Low γ: Smoother, more generalized decision boundary

Table 2.3: Summary of SVM Hyperparameters [26]

- Tree-specific hyperparameters control the construction and complexity of the decision trees.
- Learning task-specific hyperparameters control the overall behavior of the model and the learning process [8].

Parameter	Description and Effect
max_depth	Maximum depth of a tree. Deeper trees can capture more complex patterns in the data, but may also lead to overfitting.
min_child_weight	Minimum sum of instance weight needed in a child. This parameter can be used to control the complexity of the decision tree by preventing the creation of too small leaves.
subsample	Percentage of rows used for each tree construction. Lowering this value can prevent overfitting by training on a smaller subset of the data.
colsample_bytree	Percentage of columns used for each tree construction. Lowering this value can prevent overfitting by training on a subset of the features.

Table 2.4: Decision Tree Parameters [8].

Parameter	Description and Effect
eta (Learning Rate)	Step size shrinkage used in updates to prevent overfitting. Lower values make the model more robust by taking smaller steps.
gamma	Minimum loss reduction required to make a further partition on a leaf node of the tree. Higher values increase the regularization.
lambda	L2 regularization term on weights. Higher values increase the regularization.
alpha	L1 regularization term on weights. Higher values increase the regularization.

Table 2.5: Learning Parameters [8].

2.4.4 Under and Oversampling

Imbalanced datasets are a pervasive issue in machine learning, where the distribution of classes is heavily skewed, often leading to models that perform poorly on minority classes [63]. To address this, random oversampling and undersampling are two fundamental techniques used to re-balance class distributions before training machine learning models. These methods aim to mitigate the bias introduced by imbalanced data, ensuring that models can learn effectively from all classes [10].

2.4.4.1 Random Undersampling

Random undersampling involves reducing the number of majority-class instances by randomly removing examples until a desired class balance is achieved [63], without considering the importance or informativeness of the examples [10]. The degree of undersampling can be adjusted to achieve specific class ratios. For example, a 1:1 ratio ensures that the majority class has the same number of instances as the minority class, while a 0.5 ratio sets the majority class at half the size of the minority [63].

Undersampling reduces bias toward the majority class, enabling the model to focus more effectively on the minority class [63]. By reducing the overall dataset size, undersampling can significantly speed up training and reduce computational resource requirements, which is especially beneficial for very large datasets or limited computational capacity [23].

However there are a few disadvantages to this method. Random removal of majority instances may discard valuable data, leading to underfitting or reduced model performance [23]. Also if critical boundary examples are removed, the model may struggle to learn decision boundaries accurately [63].

Random undersampling is most effective when the majority class contains redundant or less informative examples. It is often used in scenarios where computational efficiency is a priority or when the dataset is too large to process in its entirety [63].

2.4.4.2 Random Oversampling

Random oversampling involves increasing the number of minority-class instances by duplicating existing examples until a desired class balance is achieved. Like undersampling, the degree of oversampling can also be adjusted to achieve specific class ratios [63]. Unlike advanced methods like SMOTE, random oversampling does not create new synthetic examples but relies solely on replication [10].

Random oversampling offers several advantages. By increasing the representation of the minority class, it enhances the model's ability to learn patterns from these critical examples, improving minority-class performance [63]. Additionally, it retains all original data points, ensuring that no critical information is discarded during the process [10]. However, there are notable disadvantages to this method. Repeated duplication of minority examples can lead to overfitting, where the model memorizes the duplicated instances rather than generalizing from them [23]. The larger dataset size resulting from oversampling may also increase computational costs, slowing down training, particularly for computationally intensive algorithms [63]. Furthermore, since no new examples are generated, the model may not learn diverse patterns within the minority class, limiting its ability to generalize effectively [10].

Random oversampling is particularly effective when the minority class is small but contains critical information. It is often used as a baseline method before exploring more advanced techniques like SMOTE or ADASYN [63].

2.4.4.3 SMOTE Oversampling

SMOTE (Synthetic Minority Oversampling Technique) is a widely used oversampling method designed to address class imbalance in datasets. SMOTE overcomes the limitation of random oversampling by generating synthetic examples, thereby improving the model's ability to learn decision boundaries for the minority class [19].

SMOTE operates by synthesizing new minority class examples based on the feature space of existing data. The algorithm selects a minority class instance at random and identifies its k nearest neighbors (typically $k = 5$). The actual number of neighbors used to generate synthetic samples depends on the oversampling percentage required; for example, to double the minority class (100% oversampling), one neighbor per instance is used, while higher oversampling rates involve more neighbors, sometimes sampled with replacement if the required number exceeds k [64].

The next step is to create a synthetic example by interpolating between the selected instance and one of its neighbors, placing the new instance along the line connecting them in the feature space [64]. This interpolation is performed by calculating the difference between the feature vectors of the selected instance and its neighbor, multiplying this difference by a random number between 0 and 1, and adding the result to the original instance's vector. This process is repeated until the desired class balance is achieved [11].

The steps in SMOTE are: first the random selection of an instance of the minority class. The next step is the identification of the k nearest neighbors of said instance. Lastly, a synthetic example is created by interpolating between the selected instance and one of

its neighbors [80]. SMOTE is often integrated into machine learning pipelines, where it is applied only to training data during cross-validation to prevent data leakage and ensure realistic performance evaluation [36].

Generalization: The synthetic examples encourage the model to create larger, more general decision regions for the minority class, reducing overfitting [35]. However, SMOTE may generate synthetic samples in overlapping class regions, potentially introducing noise. To address this, variants such as Borderline-SMOTE and ADASYN have been developed to focus synthetic sample generation on harder-to-learn or borderline cases, improving robustness [41].

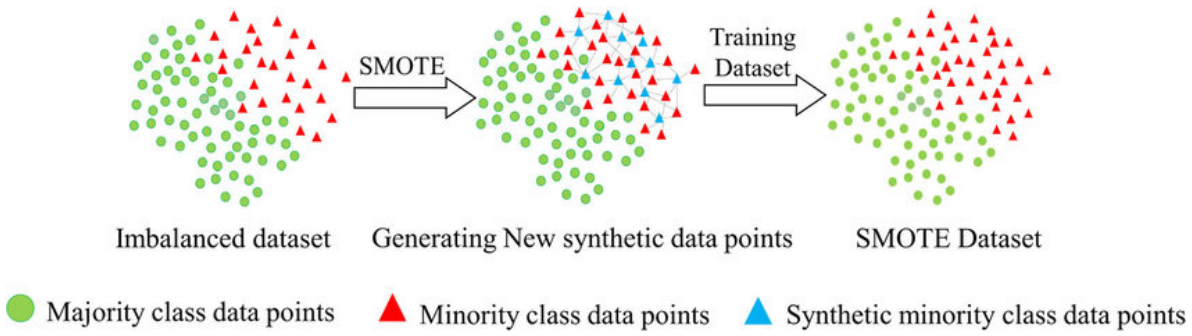

Figure 2.9: SMOTE Oversampling [29].

2.5 Related Work

Automatic depression recognition has gained significant attention in medicine, psychology, and computer science. This growing interdisciplinary interest reflects the urgent need to develop objective, efficient, and scalable methods for detecting depression, a mental health disorder that affects millions globally and often goes undiagnosed or untreated. Researchers have concentrated on analyzing differences in facial expressions, body postures, speech patterns, physiological signals, and audio cues between depressed individuals and the general population to predict depression levels. These behavioral and physiological markers provide critical insights into the subtle manifestations of depression, enabling more accurate and timely diagnosis.

As discussed in section section 2.1, speech has been established as an effective biomarker for depression, validating the pursuit of speech depression recognition (SDR) research. The unique advantages of speech signals include their non-intrusive nature, the possibility of remote data collection, and cost-effectiveness compared to other diagnostic modalities such as neuroimaging or biochemical tests. This makes SDR particularly suitable for continuous monitoring and large-scale screening, especially in resource-limited settings or telehealth applications.

SDR has evolved significantly over time, transitioning from traditional approaches relying on hand-crafted features to advanced deep learning architectures. This evolution has also seen a shift from focusing solely on acoustic features to incorporating multiple complementary features, thereby enhancing the accuracy and robustness of depression

detection systems. The progression of Speech Depression Recognition can be broadly categorized into the following stages:

- Early Stage: Hand-crafted Features and Traditional Machine Learning: In the initial phase of SDR research, the focus was on identifying and extracting acoustic features that correlate with depressive symptoms. Researchers experimented with various feature sets, including prosodic features (e.g., pitch, energy), spectral features, and voice quality measures, to improve predictive performance. Traditional machine learning algorithms such as Support Vector Machine (SVM), Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), and K-means clustering were extensively utilized. These methods depended heavily on domain expertise to manually engineer relevant features and were foundational in establishing the relationship between speech patterns and depression.
- Shift to Deep Learning: Deep neural networks were employed both as classifiers using hand-crafted features and as end-to-end architectures that automatically learned high-level features from raw audio signals or spectrograms. Various neural network architectures, including Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory networks (LSTMs), and Transformers, were explored for their potential in improving SDR accuracy [55].

2.5.1 Hand-Crafted Features and Traditional ML

The most frequently used datasets in SDR research include the following.

Dataset	Modality	Label	Number of Subjects	Number of Clips	Duration
Mundt-35	Audio	HAMD QIDS	35 patients	-	-
AVEC2013	Audio/Video	BDI-II	84 patients	150 chips	20–50m
AVEC2014	Audio/Video	BDI-II	84 patients	300 clips	6s-4m
DAIC-WOZ	Audio/Video/ECG/GSR	PHQ-8	189 patients	189 clips	Wizard-of-Oz 5–20m
E-DAIC	Audio/Video	PHQ-8	351 patients	275 clips	-
Bipolar corpus	Audio/Video	YMRS MADRS	46 depressed, 49 control	218 clips	At most 3.7m
MODMA	Audio/EEG	HRSD DSM-IV	23 depressed, 29 control	1508 clips	At most 2.45m

Table 2.6: Summary of Datasets Used in Speech Depression Recognition [55].

The following tables summarize the existing research on depression, highlighting the key findings. Notably, the studies by Nasir, Gong, and Pampouchidou are particularly relevant to this investigation, as they all utilize the DAIC-WOZ dataset and utilize Support Vector Machines (SVM) and Decision Trees, respectively [55].

Method	Paper	Dataset	Performance
GMM	Helfer et al. 2013	Mundt-35	AUC 0.76
	Williamson et al. 2013	AVEC2013	MAE/RMSE 5.75/7.42
	Williamson et al. 2014	AVEC2014	MAE/RMSE 6.52/8.50
SVM	Cummins et al. 2013	Mundt-35	Accuracy 66.9%
	Nasir et al. 2016	DAIC-WOZ	F1 0.63
	Gong et al. 2017	DAIC-WOZ	MAE/RMSE 3.96/4.99
LR	Jan et al. 2017	AVEC2014	MAE/RMSE 6.14/7.43
	Jayawardena et al. 2020	DAIC-WOZ	RMSE 6.84
Decision Tree	Pampouchidou et al. 2016	DAIC-WOZ	F1 (D/N) 0.52/0.81

Table 2.7: Some traditional classification and regression algorithms applied in speech depression recognition (SDR) [55].

2.5.2 Deep Learning

The following Table also summarizes the results from studies that have used Deep Classifiers and hand-crafted features.

Method	Paper	Dataset	Performance
LSTM	Alhanai et al. 2018	DAIC-WOZ	MAE/RMSE 4.97/6.27
	Du et al. 2018	BD	UAR/UAP/Accuracy 0.651/0.678/65.0%
	Salekin et al. 2018	DAIC-WOZ	F1/Accuracy 0.901/90%
CNN	Yang et al. 2017	DAIC-WOZ	MAE/RMSE 5.163/5.974
	Huang et al. 2020	DAIC-WOZ	F1/Accuracy 0.700/82.9%
GAN	Yang et al. 2020	DAIC-WOZ	MAE/RMSE 4.634/5.520
Transformer	Sun et al. 2021	E-DAIC	RMSE 3.783

Table 2.8: Deep classifiers applied in SDR and their performance

Lastly there are studies that create an end to end deep architecture that pushes raw signal or spectrogram into its model to learn and give results. Two notable studies are:

- DepAudioNet [56]: Uses DCNN to extract high-level feature from raw wave and LSTM to learn the temporal change of Mel scale filter feature. The F1 score is 52%.
- EmoAudioNet [46]: Uses DCNN and MFCC-based CNN for spectrum analysis. The F1 score is 82% [55].

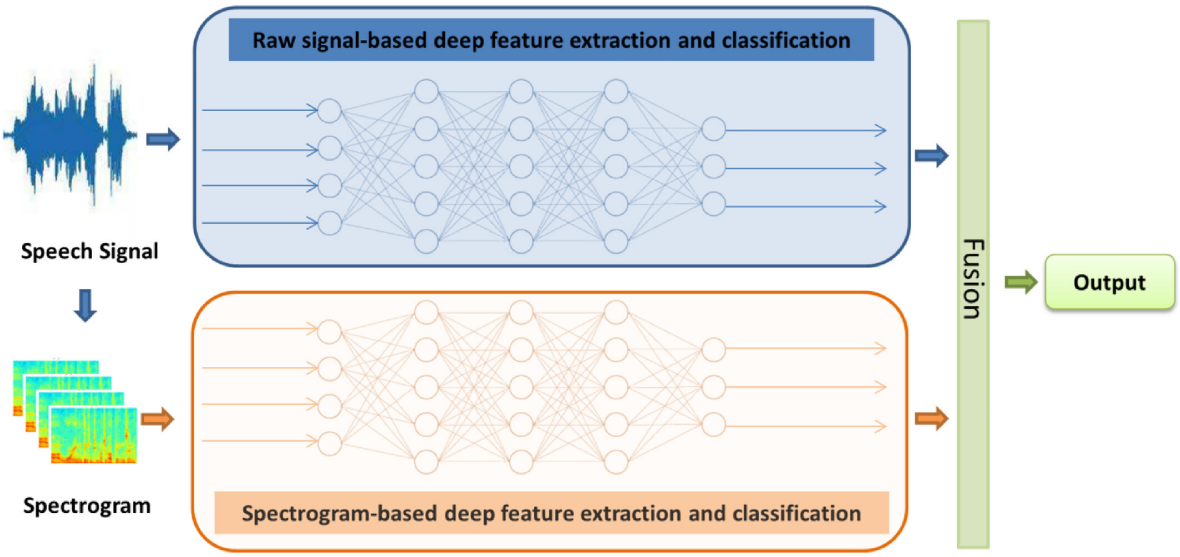

Figure 2.10: End to End Deep Architecture Framework [55].

2.5.3 Baseline Results

The baseline using the DAIC-WOZ dataset was set by the 2016 Audio-Visual Emotion Challenge and Workshop (AVEC 2016), a competition event aimed at comparison of multimedia processing and machine learning methods for automatic audio, video, and physiological analysis of emotion and depression. Specifically the results for depression classification are as follows. Performance is measured in F1 score for depressed and not depressed classes as reported through the PHQ-8.

A linear support vector machine was trained using stochastic gradient descent, where the loss is calculated one sample at a time and the model is updated sequentially. The model was validated on the development set, and a grid search was performed for optimal hyperparameters for both audio and video data separately. The features used were taken from the baseline features provided for both modalities.

Partition	Modality	F1 Score	Precision	Recall
Development	Audio	.462 (.682)	.316 (.938)	.857 (.540)
Development	Video	.500 (.896)	.600 (.867)	.428 (.928)
Development	Ensemble	.500 (.896)	.600 (.867)	.428 (.928)
Test	Audio	.410 (.582)	.267 (.941)	.889 (.421)
Test	Video	.583 (.851)	.467 (.938)	.778 (.790)
Test	Ensemble	.583 (.857)	.467 (.938)	.778 (.790)

Table 2.9: Performance metrics for different modalities on Development and Test partitions.

The baseline performance for the audio feature set is established with an **F1-Score of 0.58**, which will serve as the reference point for comparing the results of subsequent experiments.

Chapter 3

Materials and Methods

3.1 Problem Statement

As described in Section 1.1, Motivation, the scope of this thesis is to evaluate and compare the performance of various feature types and machine learning algorithms for predicting depression. The study involves two main tasks: (1) generating audio and text feature sets, and (2) assessing the performance of various machine learning models using these features. This study aims to identify the most effective combination of features and algorithms that can enhance the accuracy and reliability of depression prediction models.

3.2 Dataset Description

The data used for the scope of this thesis is the DAIC-WOZ database that is subset of the Distress Analysis Interview Corpus (DAIC). The DAIC is a multimodal collection of semi-structured clinical interviews designed to aid in diagnosing psychological distress conditions, including anxiety, depression, and post-traumatic stress disorder. Each interview session includes synchronized audio recordings, transcriptions, and expert-annotated depression scores based on standardized clinical assessments. The dataset encompasses a diverse participant pool in terms of age, gender, and ethnicity, although some demographic groups remain underrepresented, potentially limiting generalizability.

The annotation process for the DAIC is carried out by trained clinicians who assign depression severity ratings to each session. This rigorous approach ensures a high degree of reliability in the annotations; however, it also introduces an element of subjectivity, as clinical judgment can vary between annotators. Additional known limitations of the dataset include its relatively modest sample size and the somewhat artificial nature of the interview setting, which may not fully capture the complexity and variability of real-world clinical interactions [40].

The DAIC is structured around four distinct interview formats, each designed to elicit different types of participant responses:

- **Face-to-Face:** Direct interaction between the participant and a human interviewer conducted in person.
- **Teleconference:** Remote interaction between the participant and a human interviewer conducted via a teleconferencing system.
- **Wizard-of-Oz:** Interaction with an animated virtual interviewer named Ellie, whose responses and behaviors are controlled in real-time by a human operator located in a separate room.
- **Automated:** Interaction with Ellie functioning as a fully automated agent, engaging with the participant without any human intervention.

For the interviews, participants were selected from two distinct populations within the greater Los Angeles metropolitan area:

- Veterans: Individuals were recruited on-site at a US Vets reintegration facility in Southern California.
- General Public: Additional participants were sourced via online advertisements posted on Craigslist.org, broadening the demographic representation within the dataset [40].

All participants were fluent speakers of English, and all interviews were conducted in English to maintain consistency. Each participant underwent assessments for depression, PTSD, and anxiety, utilizing standardized psychiatric questionnaires to ensure reliable and comparable measurements across the dataset [40].

The semi-structured interviews followed a set progression; the initial phase involved neutral questions used to establish rapport and ensure participant comfort. The interviews then transitioned to a symptom exploration phase, where specific questions about symptoms and events related to depression and post-traumatic stress disorder (PTSD) were asked. Lastly, the conclusion phase involved a "cool-down" session, designed to ensure that participants did not leave in a distressed state [40].

The DAIC-WOZ dataset is unique due to its multimodal nature, incorporating audio, video, and text data, as well as its clinical context (semi-structured interviews conducted with a virtual agent). It also features annotated clinical scores, such as the PHQ-8, making it a valuable resource for research in depression detection. A significant portion of recent studies in this area have leveraged the dataset's multimodal capabilities to advance the field.

3.2.1 Wizard-of-Oz Interviews

As stated above, for this thesis, the DAIC-WOZ dataset was utilized, which includes the Wizard-of-Oz interviews, conducted by the animated virtual interviewer, Ellie. In these interviews, participants were situated alone in a room with a large computer screen displaying Ellie. Ellie's interactions were controlled by two operators: one managed non-verbal cues, such as nods and facial expressions, while the other handled verbal

responses. Ellie utilized a predefined set of utterances, incorporating pre-recorded audio and pre-animated gestures and facial expressions, which were based on behaviors observed in face-to-face interviews [40].

An example excerpt from a Wizard-of-Oz interview is shown below:
Ellie: Who's someone that's been a positive influence in your life?
Participant: Uh my father.
Ellie: Can you tell me about that?
Participant: Yeah, he is a uh
Participant: He's a very he's a man of few words
Participant: And uh he's very calm
Participant: Slow to anger
Participant: And um very warm very loving man
Participant: Responsible
Participant: And uh he's a gentleman has a great sense of style and he's a great cook.
Ellie: Uh huh
Ellie: What are you most proud of in your life?

Figure 3.1: Wizard-of-Oz Interview Sample [40]

3.2.2 Dataset Composition

The DAIC-WOZ dataset consists of 189 sessions, each comprising a raw audio file and its corresponding transcription. Although pre-extracted feature files are available for all sessions, they were not utilized in this thesis. This decision was made because existing studies have already explored these features extensively. Instead, the focus of this research was to investigate how alternative types of features, extracted through novel methods, might contribute to improving the accuracy of depression estimation. Additionally, three CSV files are provided, which collectively contain participant IDs, PHQ-8 binary labels and scores, as well as gender information for both the training and test sets.

For the experimental evaluation of this thesis, the PHQ-8 score was used as the ground truth for determining the presence of depression in participants. Binary classification was performed using a threshold of 10, in accordance with the PHQ-8 scoring guidelines, which indicate that a score of 10 or higher is suggestive of Major Depressive Disorder [17]. This approach ensures that the classification of depression aligns with established clinical standards.

3.3 Data Pre-processing and Feature Extraction

Data pre-processing and feature extraction are crucial steps in preparing raw data for analysis and machine learning tasks. For audio data, pre-processing involves techniques such as resampling and filtering, while feature extraction transforms raw audio signals into meaningful representations like spectrograms or Mel-frequency cepstral coefficients (MFCCs) [3]. In contrast, text data pre-processing typically includes tokenization and

normalization, with feature extraction methods like bag-of-words or embeddings to represent text in a format suitable for modeling.

3.3.1 Audio

In this research, the first approach is the pre-processing of the raw audio files. Corrupted audio samples were identified and corrected by converting their format using the ffmpeg library.

Handcrafted audio features were extracted using the pyAudioAnalysis library, while more advanced features were obtained, for additional experiments, by extracting Wav2-Vec2.0 embeddings.

The handreafted feature extraction for a single audio file involves the following steps:

- Identify every utterance in the audio using the transcription file.
- Extract mid-term features for each utterance using the mid-term function provided by the library.
- Finally average the mid-term features across all utterances; thus obtain a single feature vector as a comprehensive representation per audio file.

The embeddings were extracted using a similar approach: each utterance within the audio file was identified, embeddings were extracted for each utterance, and then these embeddings were averaged to obtain a single representative feature vector for the entire audio file.

Additionally, a role-based approach was implemented where features were similarly extracted for each utterance. However, for each file, two distinct feature vectors were generated: one by averaging the features of Ellie's utterances and one by averaging the features of the participant's utterances. This approach allowed for the creation of separate feature vectors for Ellie and the participants.

3.3.1.1 Role-Based Approach:

Extracting features at the utterance level ensures consistency in the feature extraction process, independent of the role-based approach. This consistency enabled us to compare results effectively. Furthermore, we also extracted text features in a similar manner, which facilitated the concatenation of different datasets, as detailed in section 3.3.3.

Separating audio features based on roles enables the analysis to account for the inherent differences in speech patterns and also to give more significance to one or the other speaker. Features extracted from each role can highlight role-specific cues. In this case, participant speech features might be more indicative of emotional or psychological states, while interviewer features could reflect conversational control or elicitation strategies. It is also important to note that role-based features allow clearer interpretation of which speaker's behavior drives certain outcomes, facilitating more targeted interventions or insights. Additionally by modeling roles separately, machine learning

models can learn role-specific patterns without confusion, enhancing classification or regression accuracy [39].

3.3.2 Text

Similar to the audio features, the first step is the pre-processing of the transcriptions. To that end, a basic text cleaning method was implemented. Briefly the text cleaning entails removal of punctuation, lower-casing, removing filler words like "uhm". Additionally based on the provided transcriptions, the cleaning also entailed removing phrases or words inside parentheses. These parentheses are not part of the utterance but serve as descriptors, for example "(welcome)" when the utterance contains greeting ect.

Text features were extracted using the GloVe model for word embeddings, as well as SBERT for contextualized sentence embeddings, for additional experiments.

The word embeddings used in this study were obtained from the pre-trained GloVe model trained on the Wikipedia 2014 and Gigaword 5 corpora, which includes 6 billion tokens and a vocabulary of 400,000 uncased words [48]. Specifically, the 50-dimensional version of the GloVe vectors was utilized. The extraction process follows a similar approach to the one used for audio features. For each transcription, we extracted embeddings for every word in each utterance. These word embeddings were then averaged across the entire utterance. Finally, to obtain the text feature vector for the entire transcription, we averaged the feature vectors across all utterances.

Averaging word embeddings is a widely used and straightforward method for obtaining a fixed-length representation of a transcription. However, this approach has a significant limitation: it treats all words equally and disregards word order and syntactic structure, which can result in the loss of important contextual information. To overcome these drawbacks, more advanced techniques have been developed. Contextualized embeddings, such as those generated by models like BERT, produce dynamic word vectors that change according to the surrounding context. Furthermore, methods like weighted averaging or attention mechanisms can assign varying importance to different words during aggregation, enhancing the quality of the resulting representation. Other alternatives include sentence- or document-level embedding models, such as Doc2Vec or transformer-based encoders, which explicitly capture word order and contextual relationships. For the score of this thesis we only use the averaging technique [18, 50].

Lastly, following the same approach used for role-based extraction of audio features, we also average the text features separately for Ellie's and the participant's utterances. This process generates separate feature vectors for each session, one for Ellie's utterances and one for the participant.

Regarding sentence embeddings, SBERT was used to generate contextualized embeddings for each utterance, which were then averaged across the entire transcription. Mirroring the approach applied to audio features, embeddings were averaged separately for the participant's and Ellie's utterances within each session, producing distinct feature vectors that effectively capture the semantic content of their speech. This averaging strategy yields fixed-length representations while retaining rich contextual information.

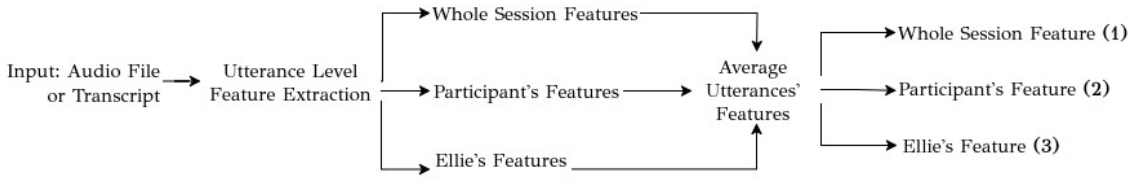

Figure 3.2: Feature extraction process either from audio file or transcription file.

3.3.3 Feature Combinations

As depicted in fig. 3.2, for each session, we extract three audio features (pyAudioAnalysis), three audio embeddings (wav2vec 2.0), three text features (GloVe) and three sentence embeddings (SBERT). Additionally, we create two additional features by concatenating the participant's and Ellie's features for both audio and text, resulting in 'Concatenated Features'. This approach aims to investigate whether combining features from both roles improves model performance.

We created two datasets for our experiments. The main dataset includes all pyAudioAnalysis features, GloVe embeddings, and their combination, and is used for the primary analyses. The second dataset, used for additional experiments, contains wav2vec 2.0 and SBERT embeddings along with their combination.

To create the afformentioned combination of audio and text features, we concatenate the full set of audio features with the full set of text features, resulting in a comprehensive multimodal feature set. This fusion aims to explore whether integrating both audio and textual information can enhance model performance.

It is noteworthy that Ellie's features are not used independently, as each experiment focuses on assessing the participant's depression, making the inclusion of the participant's data necessary.

The distinct datasets that are created through the aforementioned extraction process are summarized in table 3.1.

3.4 Evaluation Metrics

To thoroughly analyze the evaluation metrics used in this thesis, it is essential to first introduce some general information on the evaluation of binary classification.

In binary classification instances are classified as either positive or negative. A classified instance belongs in one of the following categories:

- True Positive (TP): the instance is correctly classified as positive
- False Positive (FP): the instance is incorrectly classified as positive
- True Negative (TN): the instance is correctly classified as negative
- False Negative (FN): the instance is incorrectly classified as negative

Feature Type	Second Level	Third Level	Feature Vector Length
Audio	pyAudioAnalysis	Whole Audio Features	136
		Participant's Features	136
		Ellie's Features	136
		Concatenated Features	136
	wav2vec 2.0	Whole Audio Features	768
		Participant's Features	768
		Ellie's Features	768
		Concatenated Features	768
Text	GloVe	Whole Audio Features	50
		Participant's Features	50
		Ellie's Features	50
		Concatenated Features	50
	SBERT	Whole Audio Features	384
		Participant's Features	384
		Ellie's Features	384
		Concatenated Features	384
Concatenation	pyAudioAnalysis and GloVe	Whole Audio Features	186
		Participant's Features	186
		Ellie's Features	186
		Concatenated Features	186
	wav2vec 2.0 and SBERT	Whole Audio Features	1152
		Participant's Features	1152
		Ellie's Features	1152
		Concatenated Features	1152

Table 3.1: Summary of Datasets.

Accuracy

The primary metric used for model evaluation is often, accuracy, which describes the number of correct predictions over the total number of predictions. The formula for calculating accuracy is expressed in various ways, but they all represent the same concept [13].

$$Accuracy = \frac{TruePositives + TrueNegatives}{TruePositives + TrueNegatives + FalsePositives + FalseNegatives} \tag{3.1}$$

One of the main drawbacks of accuracy is its failure to consider class distribution in the data. This means that even if a model struggles to predict the minority class, it

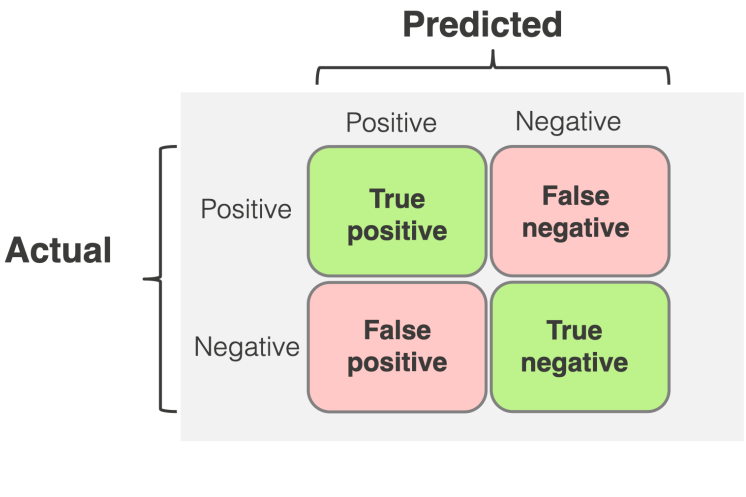

Figure 3.3: Confusion Matrix [16].

can still achieve a high accuracy if the majority class is sufficiently large [6].

Precision and Recall

Alternative metrics that provide a better understanding of a model's performance are precision and recall. These metrics are especially useful when dealing with imbalanced datasets [9].

Precision measures the proportion of true positive predictions among all positive predictions. The formula for precision is [9]:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{3.2}$$

Recall, measures the proportion of true positive cases correctly classified, over all actual positive cases in the dataset. The formula for recall is [9]:

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{3.3}$$

F1-Score

The F1-Score is a metric that combines both precision and recall, and is defined as their harmonic mean. The harmonic mean is considered more appropriate for ratios, such as precision and recall, compared to the arithmetic mean. The principle behind this score is to create a metric that weighs the ratios in a balanced way, requiring both to have a higher value for the F1-score to rise. The formula for the F1-Score is [14]:

$$F_1 = \frac{Precision \times Recall}{Precision + Recall}$$

There are different averaging methods for calculating F1-scores. These methods differ based on whether they consider each class's support value, which refers to the number

Method Name	Description	Usage
Macro-Averaging	Computes the arithmetic mean of all per-class F1-scores, treating all classes equally.	Useful for imbalanced datasets.
Weighted Averaging	Computes the mean of all per-class F1-scores, weighted by the frequency of each class in the dataset.	Accounts for the proportion of each class's occurrences.
Micro-Averaging	Sums the true positives (TP), false negatives (FN), and false positives (FP) to determine the global average F1-score.	Closely aligns with overall accuracy.

Table 3.2: Averaging Methods for calculating F1-Score

of occurrences of a class in the dataset. Depending on the dataset, a different method may be preferred [21].

In this thesis, the macro-averaging method was utilized due to the significant imbalance present in the DAIC-WOZ dataset.

AUC

AUC-ROC (Area Under the Receiver Operating Characteristic Curve) is a widely used metric for evaluating binary classifiers. It measures a model's ability to distinguish between positive and negative classes across all classification thresholds [87].

The ROC curve is a graphical representation that plots the True Positive Rate (TPR) on the y-axis against the False Positive Rate (FPR) on the x-axis for various threshold values. Each point on the curve corresponds to a specific threshold, which determines how predictions are classified as positive or negative [87].

$$TPR = \frac{TP}{TP + FN} \tag{3.4}$$

$$FPR = \frac{FP}{FP + TN} \tag{3.5}$$

The AUC (Area Under the Curve) represents the total area under this ROC curve and provides a single scalar value to summarize the model's performance:

- An AUC of 1 indicates perfect discrimination between classes.
- An AUC of 0.5 suggests no better performance than random guessing.
- Values closer to 1 indicate better model performance, while values closer to 0 indicate poor performance.

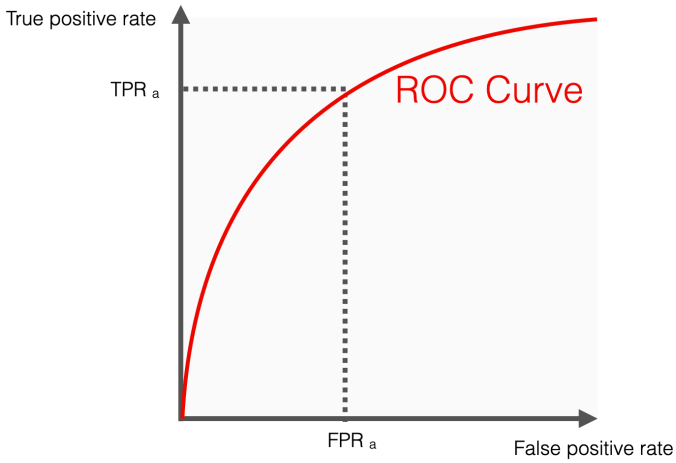

Figure 3.4: Receiver Operating Characteristic (ROC) Curve [15].

AUC-ROC is valuable because it measures model performance independently of any threshold and remains robust with imbalanced class distributions. By computing the area under the ROC curve, it summarizes a model's ability to distinguish positive from negative classes across all thresholds [2].

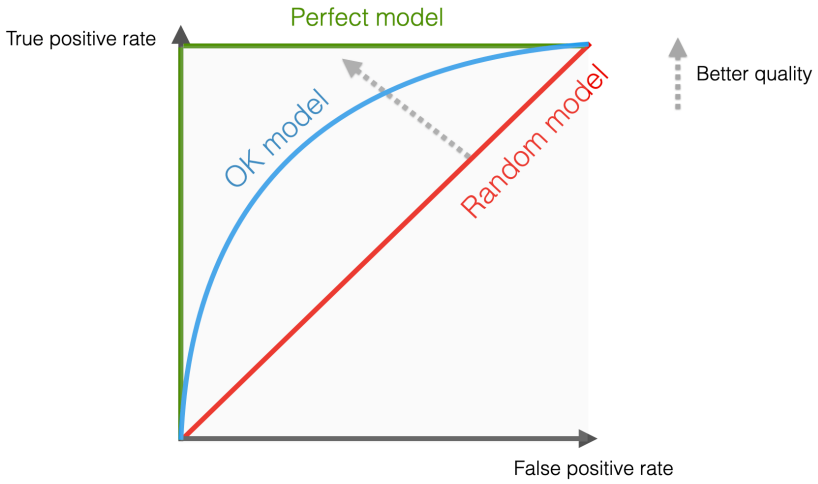

Figure 3.5: ROC Curve Interpretation [15].

Chapter 4

Experimental Evaluation

In this chapter, we discuss the experimental setup and the results related to the classification problem we are addressing. As mentioned in earlier chapters, the goal of this thesis is to predict whether a person can be diagnosed with depression based on their speech. For this, the DAIC-WoZ dataset was used to perform the following experiments.

4.1 Experimental Setup and Results

In table 3.1, the distinct datasets derived from the original DAIC-WOZ dataset are presented. The same experiments are performed across all feature set of the **primary dataset (pyAudioAnalysis and Glove)** in order to extract useful information, such as whether audio and text contribute equally to the estimation of depression or to examine whether a role-based approach is beneficial in yielding more accurate results.

The datasets used in this study consist of three main types of features: audio-based, text-based, and multimodal features formed by concatenating audio and text representations. For each type, three variants were extracted to capture different speaker scopes. First, features representing the entire audio, combining both the participant and Ellie, were obtained. Second, features corresponding exclusively to the participant's speech were extracted. Third, a concatenation of separate features for the participant and Ellie was created to preserve role-specific information. This structure allows the analysis to consider individual speaker characteristics as well as their combined interactions across different modalities.

In our experiments, we used Support Vector Machines (SVM) and XGBoost models. These models were chosen due to their strong performance and complementary strengths demonstrated in prior research across tasks such as speech emotion recognition and sentiment analysis. Studies have shown that SVM often achieves competitive or superior accuracy, while XGBoost can leverage feature optimization to reach accuracy levels comparable to deep learning models but with less computational complexity [45, 30, 32, 44].

Hyperparameter tuning was performed using GridSearch combined with LOOCV to exhaustively explore parameter combinations while maximizing data utilization [69].

Model	Parameter Values
SVM	• C: [0.1, 1, 10, 100] • gamma: [0.1, 1, 10] • kernel: [linear, rbf] These parameters are used to tune the SVM model for optimal performance.
XGBoost	• max depth: [3, 5, 7] • min child weight: [3, 5, 7] • learning rate: [0.1, 0.01, 0.001] • subsample: [0.5, 0.7, 1] These parameters are crucial for adjusting the complexity and performance of the XGBoost model.

Table 4.1: Parameter Values for GridSearch

For model training, a manual Leave-One-Out cross-validation method is also implemented. LOOCV involves iteratively training the model on all samples except one, which is used for testing. This process repeats for each sample, ensuring unbiased performance estimation without a separate validation set. This approach ensures that the model is tested on a sample that has never been seen before. It is also very important to note that, in each iteration, a new instance of the model is created to prevent information leakage, i.e., prevent information from previous iterations influencing the predictions of the current model, thus maintaining the integrity of the cross-validation process.

After selecting the optimal model parameters and training the model, we generate classifications for each test set, i.e. for each test instance. We define the aggregated test set predictions as the combination of all classifications, allowing us to calculate metrics such as F1-macro, accuracy, and AUC using these aggregated predictions.

The tables 5.1-5.9 provide detailed results of the experiments conducted. It is important to note that the Accuracy, F1-macro and AUC are computed by aggregating predictions across all LOOCV folds to provide overall performance estimates.

In this chapter, we present the performance of two trained machine learning models across the primary dataset, utilizing the three distinct feature sets. The results are organized in tables that illustrate the features extracted from the audio files. Specifically, the tables include the following feature categories: features derived from the entire audio files *(Whole Audio)*, features obtained from concatenating Ellie's data with the participants' *(Concat)*, and features extracted solely from the participants *(Participants)*.

Results:

It is also important to note that because the DAIC-WoZ dataset is imbalanced, we predominantly use the AUC values to determine the performance of a model, thus in the following tables the best AUC value for each Sub-Dataset is highlighted. As mentioned previously, Accuracy is not an ideal metric when dealing with an imbalanced dataset. The same is true for F1-score, which balances precision and recall, it focuses on the positive class and ignores true negatives, which can be problematic when both classes are important or when the dataset is severely imbalanced [31]. The metrics

	SVM			**XGBoost**		
	Whole Audio	**Concat**	**Participant**	**Whole Audio**	**Concat**	**Participant**
acc	0.70	0.70	0.70	0.72	0.71	0.73
f1	0.41	0.41	0.41	0.61	0.45	0.59
auc	0.50	0.50	0.50	**0.61**	0.52	0.59

Table 4.2: Using the pyAudioAnalysis Audio Features.

	SVM			**XGBoost**		
	Whole Audio	**Concat**	**Participant**	**Whole Audio**	**Concat**	**Participant**
acc	0.77	0.78	0.73	0.74	0.74	0.71
f1	0.70	0.73	0.59	0.64	0.63	0.58
auc	0.69	**0.72**	0.59	0.63	0.62	0.58

Table 4.3: Using the Glove Word Embeddings.

	SVM			**XGBoost**		
	Whole Audio	**Concat**	**Participant**	**Whole Audio**	**Concat**	**Participant**
acc	0.74	0.78	0.70	0.74	0.74	0.73
f1	0.69	0.73	0.41	0.64	0.63	0.60
auc	0.69	**0.72**	0.50	0.63	0.62	0.60

Table 4.4: Using the Concatenated pyAudioAnalysis and GloVe Features.

in 4.2 indicate that all the SVM models fail to distinguish between the two classes, consistently predicting only the majority class across all test sets. The AUC value of 0.5, by definition, indicates no discriminative ability [92]. Additionally, the XGBoost models for "the Whole Audio" and "Participant" give slightly better results while "Concat" is closer to random choice. These observations indicate that both SVM and XGBoost struggle with the handcrafted audio features.

As shown in 4.3, both the SVM and XGBoost models perform better with GloVe embeddings, yet the results remain suboptimal. The best score so far is given by the SVM Model "Concat" with AUC of 0.72.

Additionally, 4.4 shows that concatenation of audio and text does not improve upon using text embeddings alone, reinforcing the conclusion that the models struggle with the pyAudioAnalysis features.

4.1.1 Undersampling

In order to account for class imbalance, two alternative approaches were employed. The first approach involves manually balancing the dataset, that was achieved by randomly removing a specified number of instances from the majority class prior to conducting experiments. This method aims to create a balance between the number of samples in the two classes, a process commonly referred to as random undersampling. However, a limitation of this approach is that our original dataset was already limited in size, so balancing by reducing the number of instances in an already small dataset can lead to significant loss of valuable information, loss of data diversity or insufficient minority class representation [7]. As a result undersampling did not substantially improve the results of the models.

Results:

	SVM			**XGBoost**		
	Whole Audio	**Concat**	**Participant**	**Whole Audio**	**Concat**	**Participant**
acc	0.59	0.59	0.59	0.68	0.63	0.60
f1	0.37	0.37	0.37	0.67	0.60	0.56
auc	0.50	0.50	0.50	**0.66**	0.60	0.57

Table 4.5: Using the Balanced Audio-Based Dataset.

	SVM			**XGBoost**		
	Whole Audio	**Concat**	**Participant**	**Whole Audio**	**Concat**	**Participant**
acc	0.70	0.75	0.60	0.68	0.70	0.61
f1	0.68	0.74	0.60	0.66	0.67	0.46
auc	0.68	**0.74**	0.60	0.66	0.67	0.53

Table 4.6: Using the Balanced Text-Based Dataset.

As illustrated in table 4.5, the Audio-Based Dataset reveals that balancing does not enhance the performance of SVM models. In contrast, XGBoost models exhibit a slight improvement, with the best-performing model ("Whole Audio") increasing its AUC from 0.61 to 0.66. A similar trend is observed in the Text-Based dataset, as shown in table 4.6. Here, the best-performing model from the original experiment ("Concat" SVM) demonstrates a slight improvement with manual balancing. However, the results for the Concatenated Audio-Text Dataset, presented in table 4.7, diverge from those of the previous datasets. Notably, there is no overall improvement; before balancing, the

	SVM			XGBoost		
	Whole Audio	**Concat**	**Participant**	**Whole Audio**	**Concat**	**Participant**
acc	0.70	0.75	0.59	0.72	0.71	0.63
f1	0.69	0.74	0.37	0.70	0.70	0.58
auc	0.59	0.50	0.37	**0.7**	**0.7**	0.59

Table 4.7: Using the Balanced Concatenated Audio-Text Dataset.

best model achieved an AUC of 0.72, whereas after balancing, the best model's AUC decreased to 0.7.

4.1.2 SMOTE Oversampling

To further address class imbalance, a second method, SMOTE oversampling, is utilized. As outlined in the theoretical section of this thesis, SMOTE generates additional instances for the minority class, thereby increasing the total number of samples. Notably, to prevent information leakage, SMOTE is applied within the Leave-One-Out framework to ensure that oversampling occurred after removing the test instance from the dataset.

Results:

	SVM			XgBoost		
	Whole Audio	**Concat**	**Participant**	**Whole Audio**	**Concat**	**Participant**
acc	0.70	0.70	0.70	0.69	0.63	0.60
f1	0.41	0.41	0.41	0.63	0.55	0.53
auc	0.50	0.50	0.50	**0.62**	0.55	0.53

Table 4.8: Using the Audio-Based SMOTE Dataset.

	SVM			XgBoost		
	Whole Audio	**Concat**	**Participant**	**Whole Audio**	**Concat**	**Participant**
acc	0.62	0.72	0.59	0.68	0.69	0.61
f1	0.59	0.69	0.56	0.61	0.65	0.54
auc	0.62	**0.70**	0.58	0.61	0.65	0.54

Table 4.9: Using the Text-Based SMOTE Dataset.

As shown in table 4.8, SMOTE oversampling does not help with the SVM model only predicting the majority class in the Audio-Based Dataset. Additionally, the results obtained using SMOTE for both the Text-Based and Concatenated Audio-Text Datasets remain consistent with those achieved by the original method. The theory of SMOTE

	SVM			XgBoost		
	Whole Audio	**Concat**	**Participant**	**Whole Audio**	**Concat**	**Participant**
acc	0.74	0.72	0.70	0.72	0.69	0.60
f1	0.71	0.69	0.41	0.68	0.65	0.52
auc	**0.73**	0.70	0.50	0.68	0.65	0.52

Table 4.10: Using the Concatenated Audio-Text SMOTE Dataset.

indicates that in cases where the classifier is biased towards the majority class, as observed with the SVM models, SMOTE may not significantly enhance prediction accuracy [5, 72], which is consistent with our results. It is also of importance to note that SMOTE is less effective in high-dimensional feature spaces, because it generates synthetic samples by interpolating between neighboring examples and can fail to capture more complex patterns [28].

4.2 Additional Experiments

After analyzing the results presented above, we conclude that, within the scope of our experiments, the audio features do not provide satisfactory performance in predicting depression using the DAIC-WOZ dataset, even when employing data-balancing techniques aimed at mitigating class imbalance. This suggests that the audio representations used may lack sufficient discriminatory power for this task. In contrast, the text embeddings employed, yielded slightly better outcomes, indicating that linguistic features may capture more relevant information related to depressive states in this dataset.

In the following section, we will explore the implementation of alternative audio and text embedding techniques to assess whether more advanced feature extraction methods can improve the discrimination between the two classes in the DAIC-WOZ dataset. These methods may leverage deeper contextual information or more sophisticated modeling of temporal dynamics, potentially enhancing predictive accuracy.

It is also important to note that the role-based approach, which differentiates data based on speaker roles, did not enhance the differentiation between depressed and non-depressed individuals. Given its limited contribution, we will exclude role-based datasets from subsequent experiments to streamline the analysis and focus on more promising feature representations.

Finally, the experiments presented below aim to provide insights into the limitations of the original experiments and to identify potential directions for future research. By systematically evaluating alternative embedding techniques and modeling approaches, we hope to uncover factors that contribute to improved depression detection and inform the development of more effective diagnostic tools.

Regarding audio features, we utilize wav2vec 2.0 audio embeddings, which represent a significant advancement over handcrafted features by leveraging self-supervised learning on large-scale speech data. To enable direct comparison with previous results, we

employ both Support Vector Machine (SVM) and XGBoost models, which are well-established classifiers in this domain. Additionally, we will directly use the wav2vec 2.0 model in an end-to-end manner to evaluate whether it can achieve improved performance by capturing richer audio representations.

	SVM	**XGBoost**
	Whole Audio	**Whole Audio**
acc	0.70	0.72
f1	0.41	0.61
auc	0.50	0.61

Table 4.11: Using the wav2vec 2.0 Dataset.

As we can see from the table table 4.11, the wav2vec 2.0 representation leads to the same result as the pyAudioAnalysis handcrafted features in predicting only the negative class. This similarity in performance can be explained by intrinsic limitations of wav2vec 2.0 embeddings when handling severe class imbalance.

Specifically, wav2vec 2.0 embeddings suffer from mode collapse, where the learned representations focus on a limited subset of modes in the feature space, causing reduced expressiveness for minority classes. This problem is worsened by highly skewed codebook distributions during training, where dominant modes are over-represented and minority class features are under-represented or neglected. As a result, the embeddings fail to capture the distinctive characteristics of the positive class, making it difficult for classifiers to differentiate it despite data-level balancing techniques like oversampling or undersampling. Therefore, the wav2vec 2.0 embeddings do not inherently solve the class imbalance problem because their representation learning is biased toward majority class modes, limiting their ability to improve minority class prediction beyond what handcrafted features achieve.

Thus, we can assume that the difficulty in predicting the minority class is not related to the type of features used but is strongly correlated with class imbalance. Therefore, to address the problem of speech-based depression estimation, the best approaches would be to:

To address class imbalance in wav2vec 2.0-based models, the approach involves fine-tuning the model end-to-end using a weighted loss function that prioritizes under-represented classes during training. This is combined with maintaining or enhancing diversity loss mechanisms to prevent mode collapse in embedding spaces, ensuring the model captures nuanced acoustic variations across all classes. Data augmentation techniques like speed perturbation or noise injection are applied alongside pre-training on domain-related datasets to amplify minority class features and improve generalization. Finally, the framework incorporates experiments with advanced classifiers – such as XGBoost or SVM ensembles – to optimize decision boundaries for better minority class detection while maintaining overall performance.

Regarding the text-based experiments, as mentioned above, we employ sentence embeddings instead of word embeddings, as sentence embeddings capture the contextual

meaning of entire sentences rather than isolated words. Specifically, we utilize Sentence-BERT (SBERT), as mentioned before, is designed to generate semantically rich and context-aware sentence representations.

	SVM	**XGBoost**
	Whole Audio	**Whole Audio**
acc	0.77	0.79
f1	0.72	0.74
auc	0.71	0.73

Table 4.12: Using the SBERT Dataset.

From Table table 4.3, the highest AUC value achieved using GloVe embeddings is 0.69% with the XGBoost model. However, as shown in Table table 4.12, using SBERT sentence embeddings with the XGBoost model yields an improved AUC value of 0.74, as was expected.

4.3 Result Discussion

Regarding the role-based approach in the experiments, the results are inconsistent and do not provide clear evidence on whether separating features by speaker role improves depression prediction.

The initial experiments revealed that SVM models struggled with handcrafted audio features (pyAudioAnalysis), often failing to outperform random guessing, as indicated by AUC scores close to 0.50(Table 4.2). XGBoost models demonstrated slightly better performance with audio features, but the results remained suboptimal.

Text-based features (GloVe embeddings) yielded improved results for both SVM and XGBoost, with the "Concat" SVM model achieving the highest AUC of 0.72. However, concatenating audio and text features did not consistently enhance performance, suggesting that the models struggled to effectively integrate the pyAudioAnalysis features.

To address class imbalance, manual balancing and SMOTE oversampling were implemented. Manual balancing provided a slight improvement for the "Concat" SVM model in the text-based dataset, increasing AUC to 0.74, but it did not consistently improve results across all datasets and models. SMOTE oversampling did not significantly enhance the SVM model and only provided a marginal improvement for the XGBoost model. This finding is consistent with the theory that SMOTE may not be effective when the classifier is heavily biased towards the majority class.

The best results across all experiments were an AUC of 0.66 for audio features, 0.74 for text features, and 0.73 for concatenated features. Regarding audio features, the highest F1-macro score achieved was 0.67, which indicates a slight improvement from the baseline score of 0.0.58.

Furthermore the additional experiments that were conducted showed that:

- wav2vec 2.0 embeddings alone do not inherently resolve the class imbalance problem, limiting their advantage over handcrafted features in minority class prediction.
- Moreover, the text-based experiments using SBERT sentence embeddings showed improved AUC values compared to previous word embedding approaches (e.g., GloVe), indicating that semantically rich, context-aware sentence representations are more effective for depression detection in this dataset.

In summary, the findings suggest that text-based features are more informative than audio features for depression detection in this dataset. While balancing techniques can offer marginal improvements, the choice of features appears to be the most critical factor influencing model performance.

Chapter 5

Conclusion and Future Work

This thesis investigated the efficacy of speech-based depression estimation using machine learning techniques, focusing on developing a robust pipeline for automatic depression assessment. The motivation behind this work stems from the urgent need for objective, data-driven tools that can complement traditional clinical evaluations. Conventional assessments often suffer from subjectivity, recall bias, and the pervasive social stigma surrounding mental health disorders, which may hinder accurate diagnosis and timely intervention.

By leveraging the DAIC-WOZ dataset, a widely recognized resource in depression estimation, this study aims to mitigate these challenges through the systematic extraction and analysis of both audio and text features. A novel aspect of this research was the incorporation of a role-based feature analysis, distinguishing between participant and interviewer contributions in speech, thereby aiming to capture the dynamics of clinical interviews.

The primary contributions of this thesis are threefold:

- First, a comprehensive feature extraction pipeline was designed, encompassing both handcrafted audio features and advanced text embeddings, facilitating a multimodal approach to depression detection. This pipeline enabled the exploration of complementary information contained in speech acoustics and linguistic content.
- Second, the introduction of a role-based analysis framework allowed for the separate extraction and evaluation of features corresponding to the participant and the interviewer. This separation was intended to provide deeper insights into how each interlocutor's behavior influences predictive modeling, an area that remains relatively underexplored in existing literature.
- Third, by generating multiple dataset variants reflecting different combinations of features and speaker roles, the study systematically assessed the impact of these factors on classification performance, providing a better understanding of their relative importance.

Experimental results, detailed in Chapter 4, revealed that text-based features generally

outperformed audio features, with the highest area under the curve (AUC) reaching 0.74 for text models compared to 0.66 for audio models. This finding underscores the rich semantic information captured by text embeddings, which may provide more information than acoustic cues alone. The role-based analysis further suggested potential benefits in modeling the interactive aspects of clinical interviews, although results were inconsistent, indicating that more sophisticated approaches—such as modeling sequential dialogue dynamics—may be necessary to fully exploit this dimension.

Despite these advances, several limitations must be acknowledged.

- The DAIC-WOZ dataset, while valuable, is relatively small and may not adequately represent the full variability of speech and behavioral patterns seen in diverse populations. This limitation constrains the generalization of the findings and highlights the need for larger, more heterogeneous datasets.
- Additionally, the study primarily employed traditional machine learning algorithms, such as SVM and XGBoost, which, while effective, may lack the capacity to capture complex temporal and contextual dependencies inherent in speech and language data. Incorporating deep learning architectures could address this gap by enabling end-to-end learning and richer feature representations.
- Furthermore, the role-based analysis was limited to feature-level separation without explicitly modeling the temporal interplay between speaker turns, which could be critical for understanding conversational dynamics relevant to depression.

Building on the related work discussed in Chapter 2, several promising avenues for future research emerge:

- Integrating advanced deep learning models—such as recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and transformer-based architectures—could significantly enhance the modeling of sequential and contextual information in both speech and text modalities. These models have demonstrated success in related domains and hold promise for improving depression estimation accuracy.
- Expanding the scope of analysis to include additional modalities, such as facial expressions, physiological signals (e.g., heart rate variability, galvanic skin response), and behavioral cues, could improve the robustness and ecological validity of automatic depression detection systems. Multimodal fusion techniques, especially those leveraging attention mechanisms and hierarchical modeling, offer a powerful means to capture the complex interplay of verbal and nonverbal signals in mental health assessment.
- Given that simple concatenation of audio and text features did not consistently enhance performance, future work could explore more sophisticated fusion strategies, such as late fusion methods (e.g., ensemble voting) or cross-modal attention mechanisms, to better integrate complementary information from different modalities.
- Transfer learning and domain adaptation techniques represent a promising solution to the challenge posed by limited dataset sizes, enabling models to leverage

knowledge from related tasks or larger corpora to improve generalization.

In conclusion, this thesis advances the field of automatic depression estimation by introducing a novel feature extraction pipeline and exploring the impact of role-based analysis on model performance. The results demonstrate the potential of machine learning to support and enhance the diagnosis of depression, while also identifying key challenges and opportunities for future research.

Bibliography

[1] 3.1. cross-validation: evaluating estimator performance. https://scikit-learn.org/stable/modules/cross_validation.html.

[2] Auc-roc. https://h2o.ai/wiki/auc-roc/.

[3] Audio feature extraction. https://docs.edgeimpulse.com/docs/concepts/data-engineering/audio-feature-extraction.

[4] Gridsearchcv: How to tune hyperparameters and improve model performance. https://www.mygreatlearning.com/blog/gridsearchcv/.

[5] Is the result produced after smote reliable? https://stackoverflow.com/questions/28958200/is-the-result-produced-after-smote-reliable.

[6] Why accuracy fails in imbalanced datasets. https://www.restack.io/p/automated-feature-selection-techniques-answer-accuracy-imbalanced-datasets-cat-ai.

[7] What is undersampling?, 2022. https://www.mastersindatascience.org/learning/statistics-data-science/undersampling/.

[8] Optimizing xgboost: A guide to hyperparameter tuning, January 2023. https://medium.com/@rithpansanga/optimizing-xgboost-a-guide-to-hyperparameter-tuning-77b6e48e289d.

[9] Precision-recall curve in ml, March 2023. https://www.geeksforgeeks.org/precision-recall-curve-ml/.

[10] Handling imbalanced data for classification, January 2024. https://www.geeksforgeeks.org/handling-imbalanced-data-for-classification/.

[11] Smote, August 2024. https://learn.microsoft.com/en-us/azure/machine-learning/component-reference/smote?view=azureml-api-2.

[12] What are hyperparameters?, 2024. https://www.coursera.org/articles/what-are-hyperparameters.

[13] Evaluation metrics in machine learning, April 2025. https://www.geeksforgeeks.org/metrics-for-machine-learning-model/.

[14] F1 score in machine learning, March 2025. https://www.geeksforgeeks.org/f1-score-in-machine-learning/.

[15] How to explain the roc auc score and roc curve, 2025. https://www.evidentlyai.com/classification-metrics/explain-roc-curve.

[16] How to interpret a confusion matrix for a machine learning model, 2025. https://www.evidentlyai.com/classification-metrics/confusion-matrix.

[17] Patient health questionnaire-8 (phq-8) scoring guide, 2025. https://adhdtreat.caddra.ca/wp-content/uploads/2025/01/Patient-Health-Questionnaire-8-PHQ-8-Scoring-Guide.pdf.

[18] Arseniev-Koehler A. Theoretical foundations and limits of word embeddings: what types of meaning can they capture? *arXiv preprint arXiv:2107.10413*, 2021. https://arxiv.org/abs/2107.10413.

[19] Awan A. A. An introduction to smote, November 2022. https://www.kdnuggets.com/2022/11/introduction-smote.html.

[20] Chaturvedi A. Sentence embedding with sentence transformers (sbert): Part 1, 2022. https://medium.com/@aryan_c/sentence-embedding-with-sentence-transformers-sbert-part-1-1542ba42a65d.

[21] Kumar A. Micro-average, macro-average, weighting: Precision, recall, f1-score, December 2023. https://vitalflux.com/micro-average-macro-average-scoring-metrics-multi-class-classification-python/.

[22] Natekin A. and Knoll A. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7, December 2013. https://doi.org/10.3389/fnbot.2013.00021.

[23] Al-Serw N. A.-R. Undersampling and oversampling: An old and a new approach, Febuary 2021. https://medium.com/analytics-vidhya/undersampling-and-oversampling-an-old-and-a-new-approach-4f984a0e8392.

[24] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association, 5th ed. edition, 2013. https://doi.org/10.1176/appi.books.9780890425596.

[25] Li C. Gradient boosting, 2014. https://www.chengli.io/tutorials/gradient_boosting.pdf.

[26] Sampaio C. Understanding svm hyperparameters. https://stackabuse.com/understanding-svm-hyperparameters/.

[27] Jurafsky D. and James H. M. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Stanford University, third edition draft edition, 2024. https://web.stanford.edu/~jurafsky/slp3/.

[28] Kumar D. How to handle unbalanced data with smote?, 2025. https://www.baeldung.com/cs/synthetic-minority-oversampling-technique.

[29] Gomede E. Synthetic minority over-sampling technique (smote): Empowering ai through imbalanced data handling, July 2023. https://pub.aimind.so/synthetic-minority-over-sampling-technique-smote-empowering-ai-through-imbalanced-data-handling-d86f4de32ea3.

[30] Aakanksha J. et al. Speech emotion recognition using classifiers and xgboost algorithm. *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)*, 2(2), 2022.

[31] Abdelhamid M. et al. Balancing the scales: A comprehensive study on tackling class imbalance in binary classification. *arXiv preprint arXiv:2409.19751*, 2024. https://arxiv.org/abs/2409.19751.

[32] Aditi A. et al. Comparative analysis of speech emotion recognition models and technique. In *2023 International Conference on Computational Intelligence, Communication Technology and Networking (CICTN)*, pages 499–505. IEEE, 2023.

[33] Baevski A. et al. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pages 12449–12460, 2020. https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html.

[34] Brown A. D. et al. Speech-based markers for posttraumatic stress disorder in us veterans. *Depression and Anxiety*, 36(7):607–616, 2019. https://doi.org/10.1002/da.22890.

[35] Chawla N. V. et al. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002. http://dx.doi.org/10.1613/jair.953.

[36] Cheng W.-C. et al. From smote to mixup for deep imbalanced classification. *arXiv preprint arXiv:2308.15457*, 2023.

[37] Cummins N. et al. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49, July 2015. https://www.sciencedirect.com/science/article/pii/S0167639315000369.

[38] Fulai P. et al. Gesture recognition by ensemble extreme learning machine based on surface electromyography signals. *Frontiers in Human Neuroscience*, 16:911204, 06 2022. https://doi.org/10.3389/fnhum.2022.911204.

[39] Goldstein A. et al. A unified acoustic-to-speech-to-language embedding space captures the neural basis of natural language processing in everyday conversations. *Nature Human Behaviour*, 9(5):1041–1055, 2025. https://doi.org/10.1038/s41562-025-02105-9.

[40] Gratch J. et al. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), May 2014. http://www.lrec-conf.org/proceedings/lrec2014/pdf/508_Paper.pdf.

[41] Haibo H. et al. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pages 1322–1328. IEEE, 2008. https://sci2s.ugr.es/keel/pdf/algorithm/congreso/2008-He-ieee.pdf.

[42] Kong Q. et al. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

[43] Liu L. et al. Diagnostic accuracy of deep learning using speech samples in depression: a systematic review and meta-analysis. *Journal of the American Medical Informatics Association*, 31(10):2394–2404, 2024. https://doi.org/10.1093/jamia/ocae189.

[44] Mai E. E. et al. Bag-of-words from image to speech: a multi-classifier emotions recognition system. *International Journal of Engineering & Technology*, 9:770, 08 2020.

[45] Oktafia O. et al. Comparison of support vector machine(svm), xgboost and random forest for sentiment analysis of bumble app user comments. *Proxies : Jurnal Informatika*, 6(1), 2024. https://doi.org/10.24167/proxies.v6i1.12453.

[46] Othmani A. et al. Towards robust deep neural networks for affect and depression recognition from speech. In *Pattern Recognition. ICPR International Workshops and Challenges*, pages 5–19, Cham, 2021. Springer International Publishing. https://doi.org/10.1007/978-3-030-68790-8_1.

[47] Otte C. et al. Major depressive disorder. *Nature reviews. Disease primers*, 2:16065, September 2016. https://doi.org/10.1038/nrdp.2016.65.

[48] Pennington J. et al. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics, 2014. https://nlp.stanford.edu/projects/glove/.

[49] Phukan O. C. et al. A Comparative Study of Pre-trained Speech and Audio Embeddings for Speech Emotion Recognition. *IEEE Access*, 11:123456–123468, 2023. https://doi.org/10.48550/arXiv.2304.11472.

[50] Rücklé A. et al. Concatenated power mean word embeddings as universal cross-lingual sentence representations. *arXiv preprint arXiv:1803.01400*, 2018. https://arxiv.org/abs/1803.01400.

[51] Sălcudean A. et al. Unraveling the complex interplay between neuroinflammation and depression: A comprehensive review. *International Journal of Molecular Sciences*, 26(4):1645, 2025. https://doi.org/10.3390/ijms26041645.

[52] Taguchi T. et al. Major depressive disorder discrimination using vocal acoustic features. *Journal of Affective Disorders*, 225:214–220, 2018. https://doi.org/10.1016/j.jad.2017.08.038.

[53] Wang J. et al. Acoustic differences between healthy and depressed people: a cross-situation study. *BMC Psychiatry*, 19(300), 2019. https://doi.org/10.1186/s12888-019-2300-7.

[54] Wang Y. et al. Fast and accurate assessment of depression based on voice acoustic features: a cross-sectional and longitudinal study. *Frontiers in Psychiatry*, 14, June 2023. https://doi.org/10.3389/fpsyt.2023.1195276.

[55] Wu P. et al. Automatic depression recognition by intelligent speech signal processing: A systematic survey. *CAAI Transactions on Intelligence Technology*, 8(3):701–711, June 2022. https://doi.org/10.1049/cit2.12113.

[56] Xingchen M. et al. Depaudionet: An efficient deep model for audio based depression classification. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, AVEC '16, page 35–42. Association for Computing Machinery, 2016. https://doi.org/10.1145/2988257.2988267.

[57] Zhang Z. et al. Exploring the clinical features of narcolepsy type 1 versus narcolepsy type 2 from european narcolepsy network database with machine learning. *Scientific Reports*, 8(1):10628, July 2018. https://www.nature.com/articles/s41598-018-28840-w.

[58] Zhao Q. et al. Vocal acoustic features as potential biomarkers for identifying/diagnosing depression: A cross-sectional study. *Frontiers in Psychiatry*, 13, 2022. https://doi.org/10.3389/fpsyt.2022.815678.

[59] Maria Faurholt-Jepsen, Michael Bauer, and Lars Vedel Kessing. Smartphone data as objective measures of bipolar disorder symptoms. *BMC Psychiatry*, 19(1):1–10, 2019.

[60] Kaur G. Boosting algorithms in machine learning part ii: Gradient boosting, November 2024. https://towardsdatascience.com/boosting-algorithms-in-machine-learning-part-ii-gradient-boosting-c155ae505fe9/.

[61] Myrianthous G. What is hyperparameter tuning?, December 2024. https://builtin.com/articles/hyperparameter-tuning.

[62] Brownlee J. Loocv for evaluating machine learning algorithms, August 2020. https://machinelearningmastery.com/loocv-for-evaluating-machine-learning-algorithms/.

[63] Brownlee J. Random oversampling and undersampling for imbalanced classification, January 2021. https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/.

[64] Brownlee J. Smote oversampling for imbalanced classification, January 2021. https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/.

[65] Brownlee J. Strong learners vs. weak learners for ensemble learning, April 2021. https://machinelearningmastery.com/strong-learners-vs-weak-learners-for-ensemble-learning/.

[66] Low D. M., Bentley K. H., and Ghosh S. S. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope investigative otolaryngology*, 5(1):96–116, 2020. https://doi.org/10.1002/lio2.354.

[67] Müller M. *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Springer, 2015. https://link.springer.com/book/10.1007/978-3-319-21945-5.

[68] Reddy M.S. Depression: The disorder and the burden. *Indian Journal of Psychological Medicine*, 32(1):1–2, 2010.

[69] Khadka N. How leave-one-out cross validation (loocv) improve's model performance. https://dataaspirant.com/leave-one-out-cross-validation-loocv/.

[70] Piepenbreier N. Scikit-learn gridsearchcv: Hyper-parameter tuning in machine learning, February 2022. https://datagy.io/sklearn-gridsearchcv/.

[71] World Health Organization. Depression, 2024. https://www.who.int/health-topics/depression/#tab=tab_2.

[72] Ahirwar P., Deen A. J., and M. K. Ahirwar. Analysis of machine learning algorithm: Smote with svm. *International Journal of Novel Research and Development*, 8:g256–g259, 2023. https://www.ijnrd.org/papers/IJNRD2304631.pdf.

[73] Rao P. Audio signal processing. In *Speech, Audio, Image and Biomedical Signal Processing using Neural Networks*, 2008. https://api.semanticscholar.org/CorpusID:15210084.

[74] Tibrewal T. P. Support vector machines (svm): An intuitive explanation. https://medium.com/low-code-for-advanced-data-science/support-vector-machines-svm-an-intuitive-explanation-b084d6238106, July 2023.

[75] Agashe R. Building intelligent audio systems- audio feature extraction using machine learning. https://www.einfochips.com/blog/building-intelligent-audio-systems-audio-feature-extraction-using-machine-learning/, October 2021.

[76] Gandhi R. Support vector machine -introduction to machine learning algorithms.

https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47, June 2018.

[77] Kumar R. and Bhattacharyya P. Nlp for mental health: A survey. Technical report, Computation for Indian Language Technology (CFILT), Indian Institute of Technology Bombay, 2024. https://www.cfilt.iitb.ac.in/resources/surveys/2024/Survey_Raja_NLPforMentalHealth_2024%20(1).pdf.

[78] Shah R. Tune hyperparameters with gridsearchcv. https://www.analyticsvidhya.com/blog/2021/06/tune-hyperparameters-with-gridsearchcv/.

[79] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[80] Galli S. Overcoming class imbalance with smote, March 2023. https://www.blog.trainindata.com/overcoming-class-imbalance-with-smote/.

[81] Goldman L. S., Nielsen N. H., and Champion H. C. Awareness, diagnosis, and treatment of depression. *Journal of general internal medicine*, 14(9):569–580, 1999. https://doi.org/10.1046/j.1525-1497.1999.03478.x.

[82] K Smith et al. Economic burden of depression: A systematic review. *Pharmacoeconomics*, 37(6):653–665, 2019.

[83] Giannakopoulos T. pyaudioanalysis: An open-source python library for audio signal analysis. *PLoS ONE*, 10(12):1–17, December 2015. https://doi.org/10.1371/journal.pone.0144610.

[84] Giannakopoulos T. Intro to audio analysis: Recognizing sounds using machine learning, September 2020. https://hackernoon.com/intro-to-audio-analysis-recognizing-sounds-using-machine-learning-qy2r3ufl.

[85] Giannakopoulos T. and Pikrakis A. *Introduction to Audio Analysis: A MATLAB Approach*. Academic Press, 2014. https://books.google.gr/books/about/Introduction_to_Audio_Analysis.html?id=zbHVAQAAQBAJ&redir_esc=y.

[86] John Torous, Jukka-Pekka Onnela, and Matcheri Keshavan. Machine learning for mental health: A systematic review of methods and applications. *Psychiatric Clinics of North America*, 43(3):465–477, 2020.

[87] Chugh V. Which metric should i use? accuracy vs. auc, October 2022. https://www.kdnuggets.com/2022/10/metric-accuracy-auc.html.

[88] Efimov V. Sbert - sentence-bert, 2020. https://towardsdatascience.com/sbert-deb3d4aef8a4/.

[89] Oppenheim A. V. and Schafer R. W. *Discrete-Time Signal Processing*. Pearson Prentice Hall, 3rd edition, 2010. https://www.pearson.com/en-us/subject-catalog/p/discrete-time-signal-processing/P200000003226/9780137549771.

[90] World Health Organization. *Suicide worldwide in 2019: Global health estimates*. World Health Organization, Geneva, 2021. ISBN 978-92-4-002664-3 (electronic version).

[91] Bobbitt Z. A quick intro to leave-one-out cross-validation (loocv), November 2020. https://www.statology.org/leave-one-out-cross-validation/.

[92] Bobbitt Z. What is a good auc score?, September 2021. https://www.statology.org/what-is-a-good-auc-score/".