NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
DIVISION OF INFORMATION TRANSMISSION SYSTEMS &
MATERIAL TECHNOLOGY

Interpretable Transformer-Based Longitudinal Modeling of Alzheimer's Disease Diagnosis and Progression

THESIS

Georgios-Efraim Kanellopoulos

Supervisor: Konstantina Nikita
Professor NTUA

Athens, October 2025

NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
DIVISION OF INFORMATION TRANSMISSION SYSTEMS &
MATERIAL TECHNOLOGY

Interpretable Transformer-Based Longitudinal Modeling of Alzheimer's Disease Diagnosis and Progression

THESIS

Georgios-Efraim Kanellopoulos

Supervisor: Konstantina Nikita
Professor NTUA

Approved by the examination committee on 27th October 2025

Konstantina Nikita	Giorgos Stamou	Athanasios Voulodimos
Professor, NTUA	*Professor, NTUA*	*Assistant Professor, NTUA*

Athens, October 2025

. . . . . .

Georgios-Efraim Kanellopoulos
Graduate of Electrical and Computer Engineering, National Technical University of Athens

Περίληψη

Η νόσος Alzheimer αποτελεί την πιο συχνή αιτία άνοιας παγκοσμίως, αντιπροσωπεύοντας το 60-80% των περιπτώσεων. Πρόκειται για μια προοδευτική νευροεκφυλιστική διαταραχή που χαρακτηρίζεται από απώλεια μνήμης, γνωστική απώλεια και λειτουργική ανεπάρκεια, οδηγώντας τελικά σε απώλεια αυτονομίας και θάνατο. Παρά τις δεκαετίες ερευνητικών προσπαθειών, η αξιόπιστη και έγκαιρη διάγνωση και η πρόβλεψη της εξέλιξης της νόσου παραμένουν μια κρίσιμη πρόκληση, ιδιαίτερα στο στάδιο της ήπιας γνωστικής διαταραχής (MCI), όπου τα κλινικά συμπτώματα είναι ήπια αλλά ο κίνδυνος μετάβασης σε AD αυξημένος. Η παρούσα διπλωματική εργασία αναπτύσσει προσεγγίσεις βαθιάς μάθησης, χρησιμοποιώντας Transformers για τη μοντελοποίηση της κατάστασης και της εξέλιξης της νόσου, αξιοποιώντας διαχρονικά δεδομένα από το Alzheimer's Disease Neuroimaging Iniative (ADNI). Η έρευνα επικεντρώνεται σε δύο προβλήματα: (i) τη διάγνωση της τρέχουσας επίσκεψης (AD vs MCI/CN, AD vs CN) και (ii) την πρόβλεψη της μετάβασης στην επόμενη επίσκεψη από MCI σε AD. Τα αποτελέσματα δείχνουν υψηλή απόδοση για τη διάγνωση της τρέχουσας επίσκεψης (AUC ROC > 0.90 για AD vs CN), ενώ η πρόβλεψη της μετάβασης παραμένει ένα πιο απαιτητικό πρόβλημα. Για την ερμηνευσιμότητα του μοντέλου χρησιμοποιήθηκε η μέθοδος Integrated Gradients για την απόδοση των προβλέψεων του μοντέλου σε επιμέρους χαρακτηριστικά, αναδεικνύοντας βιοδείκτες της νόσου. Πραγματοποιήθηκε επίσης ανάλυση υποομάδων, από την οποία προέκυψαν μια υποομάδα με άτομα υψηλότερου κινδύνου και μία χαμηλότερου. Συνολικά, τα ευρημάτα αυτά αναδεικνύουν την αξία των Transformer για την πρόβλεψη της εξέλιξης της νόσου με διαχρονικά δεδομένα.

Λέξεις κλειδιά: Γήρανση Εγκεφάλου, Νόσος Alzheimer, Ερμηνεύσιμη Τεχνητή Νοημοσύνη, Βαθιά Μάθηση, Transformer, Πρόβλεψη Διάγνωσης

Abstract

Alzheimer's disease (AD) is the most common cause of dementia worldwide, accounting for 60-80% of cases. It is a progressive neurodegenerative disorder characterized by memory loss, cognitive decline and functional impairment, ultimately leading to loss of independence and death. Despite decades of research, reliable early diagnosis and prediction of disease progression remain a critical challenge, particularly at the stage of mild cognitive impairment (MCI), when clinical symptoms are subtle but risk of conversion to AD is elevated. This thesis develops and evaluates deep learning Transformer-based approaches for modeling disease status and progression using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). The work focuses on two predictive tasks: (i) current-visit diagnosis (AD vs. mild cognitive impairment (MCI) and cognitively normal (CN)), and (ii) next-visit conversion prediction, from MCI to AD. Results show strong performance for current-visit diagnosis (mean ROC-AUC > 0.90 for AD vs CN), while conversion prediction remains more challenging. To enhance intepretability, Integrated Gradients were employed to attribute model predictions to individual features. This analysis consistently highlights established AD biomarkers. Subgroup analysis was also conducted by clustering patient embeddings in the learned representation space. This revealed patterns of risk: one subgroup characterized by more high-risk individuals, and the other more low-risk. Together, these findings underscore the potential of Transformer-based models for longitudinal prediction in Alzheimer's disease.

Keywords: Brain Aging, Alzheimer's Disease, Interpretable Artificial Intelligence, Deep Learning, Transformer, Longitudinal Analysis, Diagnosis Prediction

Στη μητέρα μου, Ελένη
και στη μνήμη του πατέρα μου, Γιάννη

Ευχαριστίες

Θα ήθελα να ευχαριστήσω την καθηγήτρια κ. Κωνσταντίνα Νικήτα, που μου έδωσε την ευκαιρία να εκπονήσω αυτή την διπλωματική στο εργαστήριο Βιοϊατρικών Προσομοιώσεων και Απεικονιστικής Τεχνολογίας (BIOSIM) και την κ. Μαρία Αθανασίου για την πολύτιμη βοήθεια και καθοδήγηση που μου προσέφερε καθ' όλη την διάρκεια. Τέλος, ευχαριστώ την Μέλπω, που με στήριξε στην πορεία εκπόνησης αυτής της διπλωματικής.

Contents

Chapter 1

Ελληνική Περίληψη

1.1 Εισαγωγή

Η νόσος Alzheimer (AD) αποτελεί την πιο συχνή αιτία άνοιας παγκοσμίως, ευθυνόμενη για περίπου το 60-80% των περιστατικών. Πρόκειται για μια νευροεκφυλιστική διαταραχή που χαρακτηρίζεται από απώλεια μνήμης, γνωστική και λειτουργική ανεπάρκεια, οδηγώντας τελικά σε απώλεια της αυτονομίας του ατόμου και θάνατο. Παρά τις δεκαετίες ερευνών, η αξιόπιστη και έγκαιρη διάγνωση και πρόβλεψη της πορείας της νόσου παραμένουν κρίσιμες προκλήσεις, ιδιαίτερα στο στάδιο της ήπιας γνωστικής διαταραχής (MCI), όπου τα κλινικά συμπτώματα είναι ήπια αλλά ο κίνδυνος μετατροπής σε Alzheimer είναι αυξημένος.

Η παρούσα διπλωματική εργασία αναπτύσσει και αξιολογεί προσεγγίσεις βαθιάς μάθησης με τη χρήση Transformer για τη μοντελοποίηση της κατάστασης και της εξέλιξης της νόσου, χρησιμοποιώντας δεδομένα από το Alzheimer's Disease Neuroimaging Initiative (ADNI). Η μελέτη εστιάζει σε δύο προβλήματα: (i) τη διάγνωση της τρέχουσας επίσκεψης (AD έναντι ατόμων ήπιας γνωστικής διαταραχής (MCI) και γνωστικά φυσιολογικών ατόμων (CN)), και (ii) την πρόβλεψη μετατροπής από MCI σε AD, στην *επόμενη* επίσκεψη.

1.2 Μεθοδολογία

1.2.1 Δεδομένα

Όπως αναφέρεται παραπάνω, τα δεδομένα που χρησιμοποιήθηκαν στην διπλωματική προέρχονται από το ADNI και περιλαμβάνουν δύο σύνολα: (α) ένα στατικό (cross-sectional) σύνολο δεδομένων ασθενών, το οποίο περιλαμβάνει μετρήσεις εγκεφαλικών περιοχών (ROIs) και γενετικές πληροφορίες (SNPs), και (β) ένα διαχρονικό (longitudinal) σύνολο δεδομένων, το οποίο περιέχει πληροφορίες για τους όγκους εγκεφαλικών περιοχών ανά επίσκεψη.

Το διαχρονικό σύνολο χρησιμοποιήθηκε για την πρόβλεψη της διάγνωσης και της εξέλιξης της νόσου, ενώ το στατικό χρησιμοποιήθηκε για διερευνητική άναλυση εντός υποομάδων. Οι όγκοι των εγκεφαλικών περιοχών εξήχθησαν από εικόνες μαγνητικής τομογραφίας (MRI) μετά από προεπεξεργασία, η οποία περιλάμβανε αφαίρεση του κρανίου (skull-stripping) ώστε να απομονωθεί ο εγκεφαλικός ιστός.

Το κύριο διαχρονικό σύνολο δεδομένων περιλαμβάνει 2398 ασθενείς και συνολικά 10805 εξετάσεις. Ένα μικρό ποσοστό επισκέψεων που περιείχαν μη έγκυρες τιμές όγκων αποκλείστηκε, με αποτέλεσμα 2391 ασθενείς να συμπεριληφθούν τελικά. Για κάθε ασθενή είναι διαθέσιμα στατικά δημογραφικά χαρακτηριστικά (όπως φύλο και φυλή), καθώς και διαχρονικές μετρήσεις για 145 εγκεφαλικές περιοχές (ROIs) και την ηλικία κατά την επίσκεψη.

Τα χρονικά διαστήματα μεταξύ των επισκέψεων δεν είναι ομοιόμορφα, γεγονός που απαιτεί περαιτέρω ανάλυση. Όπως φαίνεται στα σχετικά διαγράμματα, ο συνολικός αριθμός επισκέψεων και η συνολική χρονική διάρκεια ανά ασθενή ποικίλουν, επηρεάζοντας τον τελικό αριθμό επισκέψεων που χρησιμοποιήθηκε για κάθε πείραμα.

Επιπλέον, πολλοί ασθενείς διαθέτουν μόνο μια επίσκεψη, γεγονός που είναι χρήσιμο για το πρόβλημα της διάγνωσης αλλά όχι και για την πρόβλεψη της μετατροπής από MCI σε AD. Για τον λόγο αυτό, το σύνολο δεδομένων χωρίστηκε σε δύο υποσύνολα:

- το υποσύνολο διάγνωσης (diagnosis cohort), που περιλαμβάνει το πλήρες δείγμα, και
- το υποσύνολο μετάβασης (conversion cohort), που δημιουργήθηκε αποκλείοντας ασθενείς με μόνο μια επίσκεψη, όσους είχαν ήδη διάγνωση AD στην αρχική τους επίσκεψη, καθώς και όλες τις CN καταστάσεις.

1.2.2 Επεξεργασία Διαχρονικών Δεδομένων

Υποσύνολο Διάγνωσης

Το αρχικό σύνολο δεδομένων περιλαμβάνει συνολικά 2391 ασθενείς και 10730 εξετάσεις. Σε επίπεδο επίσκεψης, οι διαγνωστικές κλάσεις κατανέμονται ως εξής: 3700 υγιείς (CN), 4781 με ήπια γνωστική διαταραχή (MCI) και 2249 με νόσο Alzheimer (AD), με την κλάση MCI να αποτελεί την πιο συχνή κατάσταση.

Η πλειοψηφία των συμμετεχόντων διαθέτει πολλαπλές επισκέψεις, με την κατανομή να μειώνεται όσο αυξάνεται ο αριθμός των επισκέψεων. Συγκεκριμένα, από τους 2391 συμμετέχοντες, 1907 έχουν τουλάχιστον δύο επισκέψεις, 1594 έχουν τρεις και μόλις 267 άτομα διαθέτουν εννέα επισκέψεις. Η πτώση αυτή στον αριθμό των διαθέσιμων ασθενών στις μεταγενέστερες επισκέψεις περιορίζει τη χρησιμότητα τους, καθώς το δείγμα γίνεται στατιστικά αδύναμο.

Η αρχική διαγνωστική κατανομή είναι σχετικά ισορροπημένη — περίπου 36% CN, 46% MCI και 18% AD στους συμμετέχοντες της πρώτης επίσκεψης — ωστόσο, μετά την έβδομη επίσκεψη ο αριθμός διαθέσιμων παρατηρήσεων μειώνεται απότομα. Για

τον λόγο αυτό, η κύρια διαχρονική (longitudinal) ανάλυση περιορίστηκε σε μέγιστο μήκος ακολουθίας επτά επισκέψεων (T = 7), εξασφαλίζοντας ικανοποιητική κάλυψη της χρονικής εξέλιξης με επαρκές μέγεθος δείγματος.

Για κάθε ασθενή i, τα δεδομένα οργανώθηκαν ως σειρά επισκέψεων σε χρονολογική σειρά από τη βασική (baseline) εξέταση έως την έβδομη. Οι ασθενείς με λιγότερες από επτά επισκέψεις «συμπληρώθηκαν» (right-padding) ώστε να διατηρηθεί σταθερό το μήκος της ακολουθίας, με χρήση μάσκας ώστε οι ανύπαρκτες χρονικές στιγμές να μη συνεισφέρουν στην εκπαίδευση.

Σε κάθε επίσκεψη t, το διάνυσμα εισόδου $x_{i,t}$ περιλάμβανε:

- 145 όγκους εγκεφαλικών περιοχών (ROIs), οι οποίοι εξήχθησαν από δομικές μαγνητικές τομογραφίες (MRI) μετά από προεπεξεργασία, και
- την ηλικία του ασθενούς τη συγκεκριμένη στιγμή

Όλες οι συνεχείς μεταβλητές (ηλικία και όγκοι ROI) κανονικοποιήθηκαν μέσω z-score ως προς το μέσο όρο και την τυπική απόκλιση των φυσιολογικών (CN) ατόμων.

Το επόμενο στάδιο της προεπεξεργασίας που εφαρμόστηκε στα αριθμητικά δεδομένα των VOI είναι η γραμμική διόρθωση συμμεταβλητών που είχε ως στόχο την εξάλειψη της επίδρασης της ηλικίας, του φύλου και του συνολικού όγκου του εγκεφάλου στα δεδομένα. Αναλυτικότερα, επειδή η ανατομία του εγκεφάλου και ο όγκος της λευκής και της φαιάς ουσίας κάθε περιοχής διαφέρει ανάλογα με την ηλικία ή το φύλο ενός ασθενούς, ήταν απαραίτητη η προσαρμογή των δεδομένων, ώστε να διατηρηθούν μόνο οι σχετιζόμενες με τις υπό μελέτη νόσους νευροανατομικές διαφοροποιήσεις του όγκου των περιοχών ενδιαφέροντος.

Για τον σκοπό αυτό ένα μοντέλο Γραμμικής Παλινδρόμησης (Linear Regression) εκπαιδεύτηκε με τις συμμεταβλητές της ηλικίας, του φύλου και του συνολικού όγκου του εγκεφάλου των 449 υγιών ατόμων ως προβλεπτικούς παράγοντες και τα VOIs του εγκεφάλου ως έξοδο. Πραγματοποιήθηκε προσαρμογή του μοντέλου στους προβλεπτικούς παράγοντες και στη συνέχεια το μοντέλο της Γραμμικής Παλινδρόμησης εφαρμόστηκε σε ολόκληρο το σύνολο δεδομένων, ώστε να προκύψει η πρόβλεψη για τον όγκο των περιοχών ενδιαφέροντος κάθε ατόμου. Στη συνέχεια, η τιμή πρόβλεψης για κάθε άτομο αφαιρέθηκε από την αρχική τιμή για καθένα από τα VOIs και προέκυψε η ζητούμενη τιμή υπολοίπου στην οποία είχε εξαλειφθεί η επίδραση της ηλικίας και του φύλου.

Συνοψίζοντας, κάθε ασθενής αναπαρίσταται ως μια ακολουθία επτά χρονικών στιγμών που περιλαμβάνουν τυποποιημένα συνεχή χαρακτηριστικά (ηλικία και όγκους ROI) και τη διαγνωστική ετικέτα ανά επίσκεψη. Η μορφοποίηση αυτή αποτέλεσε τη βάση για όλα τα μοντέλα διαχρονικής πρόβλεψης που ακολούθησαν στην παρούσα εργασία.

Υποσύνολο Μετάβασης

Η ανάλυση της μετάβασης ορίστηκε ως το πρόβλημα της πρόβλεψης της εξέλιξης από την ήπια γνωστική διαταραχή (MCI) προς τη νόσο Alzheimer (AD), λαμβάνοντας

υπόψη τη διάγνωση της επόμενης επίσκεψης κάθε ασθενούς. Οι υγιείς (CN) συμμετέχοντες αποκλείστηκαν από αυτή την ανάλυση, καθώς και τα άτομα που είχαν διαγνωστεί με AD από την πρώτη επίσκεψη. Επιπλέον, ασθενείς με μία μόνο επίσκεψη δεν διαθέτουν επόμενη χρονική παρατήρηση και συνεπώς αποκλείστηκαν.

Μετά την εφαρμογή αυτών των κριτηρίων παρέμειναν 980 ασθενείς με συνολικά 3867 εξετάσεις. Από αυτούς, 340 χαρακτηρίστικαν ως converters, δηλαδή εμφάνισαν αλλαγή διάγνωσης από MCI σε AD σε κάποια μελλοντική επίσκεψη, ενώ οι υπόλοιποι 640 παρέμειναν σταθεροί στην κατάσταση MCI.

Για κάθε ασθενή i, τα δεδομένα οργανώθηκαν ως μια ακολουθία επισκέψεων $t = 0, 1, \ldots, T_i - 1$. Για κάθε ζεύγος (i, t), ορίζουμε τη δυαδική τιμή

$$z_{i,t} = \begin{cases} 1, & \text{if } y_{i,t} = \text{MCI and } y_{i,t+1} = \text{AD}, \\ 0, & \text{if } y_{i,t} = \text{MCI and } y_{i,t+1} = \text{MCI}. \end{cases} \tag{1.1}$$

Το διάνυσμα χαρακτηριστικών εισόδου κάθε επίσκεψης περιλαμβάνει:

- **145 όγκους εγκεφαλικών περιοχών (ROIs)** που εξήχθησαν από MRI,
- την **ηλικία του ασθενούς** κατά την επίσκεψη, και
- το **χρονικό διάστημα (Δt)** μέχρι την επόμενη επίσκεψη.

Το μέγιστο πλήθος επισκέψεων ανά ασθενή ορίστηκε ως $K = 7$, ίδιο με αυτό της διάγνωσης. Η απώλεια και οι μετρικές υπολογίζονται μόνο σε έγκυρα χρονικά βήματα, εξαιρώντας τα σημεία με padding ή μη επιλέξιμες καταστάσεις. Επίσης εφαρμόστηκε z-score κανονικοποίηση, χρησιμοποιώντας μέσους όρους και τυπικές αποκλίσεις του φυσιολογικού πληθυσμού (CN), σύμφωνα με το πρότυπο του πλήρους συνόλου δεδομένων.

Σε επίπεδο ασθενών, το ποσοστό μετατροπής ήταν περίπου 34.7% (340/980). Για τη δημιουργία των ετικετών, χρησιμοποιήθηκε χρονική μετατόπιση μίας επίσκεψης, ενώ τα δεδομένα χωρίστηκαν σε σύνολα εκπαίδευσης και ελέγχου με αναλογία 80%-20%.

Τέλος, εφαρμόστηκε γραμμική διόρθωση (linear correction) στα χαρακτηριστικά των ROIs, με σκοπό την αφαίρεση των γραμμικών επιδράσεων των επιμέρους μεταβλητών (ηλικία, φύλο και baseline DLICV). Για κάθε ROI εκπαιδεύτηκε ένα γραμμικό μοντέλο στα δεδομένα εκπαίδευσης και χρησιμοποιήθηκαν τα υπολειπόμενα (residuals) ως διορθωμένες τιμές. Η διαδικασία αυτή εφαρμόστηκε στα προβλήματα AD vs CN και MCI σε AD, αλλά όχι στην AD vs Rest, όπου λόγω της κλινικής πραγματικότητας του προβλήματος, προτιμήθηκαν τα raw δεδομένα.

1.2.3 Επεξεργασία Στατικών Δεδομένων

Για τη διαδικασία ανάλυσης υποομάδων χρησιμοποιήθηκε ένα στατικό σύνολο δεδομένων 1463 ατόμων, και αυτό προερχόμενο από τη βάση δεδομένων ADNI, το οποίο περιλαμβάνει δομικά (MRI), δημογραφικά και γενετικά χαρακτηριστικά. Από τους συμμετέχοντες, 449 ήταν υγιείς (CN), 740 είχαν ήπια γνωστική διαταραχή (MCI)

και 274 διαγνώστηκαν με νόσο Alzheimer (AD), με ηλικίες που κυμαίνονταν από 60 έως 86 ετών. Το ποσοστό συμμετεχόντων ανά κατηγορία ήταν 30.7% CN, 50.6% MCI και 18.7% AD.

Τα δημογραφικά χαρακτηριστικά περιλαμβάνουν την ηλικία και το φύλο, ενώ τα κλινικά χαρακτηριστικά προέρχονται από δομικές μαγνητικές τομογραφίες εγκεφάλου (T1 MRI) και περιλαμβάνουν τον συνολικό όγκο εγκεφάλου καθώς και 145 όγκους περιοχών ενδιαφέροντος, όπως ο ιππόκαμπος και η αμυγδαλή. Τα γενετικά δεδομένα αποτελούνται από 54 μονονουκλεοτιδικούς πολυμορφισμούς (SNPs) που έχουν συσχετιστεί με τη νόσο Alzheimer, με τιμές που εκφράζουν τον αριθμό των αλληλόμορφων (0, 1 ή 2) που έχει κάθε άτομο. Όλα τα αριθμητικά χαρακτηριστικά κανονικοποιήθηκαν με z-score.

Επίσης, εφαρμόστηκε γραμμική διόρθωση (linear correction) στα δεδομένα των εγκεφαλικών όγκων για να εξαλειφθούν οι επιδράσεις της ηλικίας, του φύλου και του ενδοκρανιακού όγκου, παράγοντες που επηρεάζουν σημαντικά τη μορφολογία του εγκεφάλου. Για τον σκοπό αυτό, όπως και στα διαχρονικά δεδομένα, εκπαιδεύτηκε ένα γραμμικό μοντέλο γραμμικής παλινδρόμησης χρησιμοποιώντας μόνο τους CN συμμετέχοντες ως δείγμα αναφοράς. Οι υπολλειματικές τιμές (residuals) που προέκυψαν μετά την αφαίρεση των προβλεπόμενων τιμών χρησιμοποιήθηκαν ως τα τελικά δεδομένα.

1.2.4 Μοντέλο Πρόβλεψης Διάγνωσης με χρήση Transformer

Για την πρόβλεψη της διάγνωσης της νόσου σχεδιάστηκε το μοντέλο *Tralzformer*, μία παραλλαγή βασισμένη στην αρχιτεκτονική των Transformers [1]. Το μοντέλο αξιοποιεί τον μηχανισμό *self-attention* ώστε να μάθει αναπαραστάσεις που ενσωματώνουν τη χρονική εξάρτηση της νόσου.

Κάθε ασθενής αναπαρίσταται ως μια ακολουθία T επισκέψεων, όπου για κάθε χρονικό σημείο t δίνεται ένα σύνολο χαρακτηριστικών $\mathcal{F}_t$. Κάθε χαρακτηριστικό $f \in \mathcal{F}_t$ προβάλλεται σε ένα κοινό χώρο διάστασης d (*embedding*), παράγοντας το διάνυσμα $z^{(t,f)} \in \mathbb{R}^d$. Έπειτα, υπολογίζοντας τον μέσο όρο των embeddings των εγκεφαλικών περιοχών (ROIs) κάθε επίσκεψης προκύπτει μια συνολική αναπαράσταση $\bar{z}^{(t)}$.

Στη συνέχεια, προστίθεται η πληροφορία της ηλικίας του ασθενούς $e^{(t)}_{\text{age}}$ και ο αντίστοιχος συντελεστής w_{age}, παράγοντας το διάνυσμα επίσκεψης $\hat{z}^{(t)} = \bar{z}^{(t)} + w_{\text{age}} e^{(t)}_{\text{age}}$. Η χρονική πληροφορία εισάγεται μέσω sinusoidal positional encoding, δίνοντας την τελική είσοδο $\tilde{z}^{(t)}$ για κάθε επίσκεψη.

Η ακολουθία των embeddings $\tilde{Z} = \tilde{z}^{(1)}, \dots, \tilde{z}^{(T)}$ τροφοδοτείται σε έναν κωδικοποιητή (*Transformer Encoder*) με ένα επίπεδο *multi-head self-attention*, ο οποίος υπολογίζει τις αναπαραστάσεις $H = h^{(1)}, \dots, h^{(T)} \in \mathbb{R}^{T \times d}$.

Κάθε χρονικό βήμα t αντιστοιχεί σε μια πρόβλεψη μέσω ενός *classification head* τριών κλάσεων, $y^{(t)} = \text{Head}^{(t)}(h^{(t)})$, με χρήση *softmax*. Οι προβλέψεις και οι απώλειες υπολογίζονται μόνο για τα έγκυρα χρονικά βήματα, σύμφωνα με τη μάσκα.

Το μοντέλο εκπαιδεύεται με τον βελτιστοποιητή *Adam*, χρησιμοποιώντας διάσταση χώρου $d = 128$.

1.2.5 Μοντέλο Πρόβλεψης Μετάβασης σε κατάσταση AD

Για την εκτίμηση της πιθανότητας κλινικής μετατροπής σε μελλοντικές επισκέψεις, αναπτύχθηκε μια αρχιτεκτονική βασισμένη σε Transformers, ακολουθώντας την προσέγγιση που χρησιμοποιήθηκε και για την πρόβλεψη διάγνωσης. Το μοντέλο σχεδιάστηκε έτσι ώστε να μαθαίνει τις *χρονικές εξαρτήσεις* των ογκομετρικών χαρακτηριστικών των ασθενών και να προβλέπει, σε κάθε χρονικό βήμα, την πιθανότητα μετατροπής σε Alzheimer στην επόμενη επίσκεψη.

Τα δεδομένα εισόδου αποτελούν ακολουθίες ROIs ανά ασθενή, όπου κάθε χρονικό βήμα αντιστοιχεί σε μία κλινική επίσκεψη. Κάθε χαρακτηριστικό προβάλλεται αρχικά σε μια διάσταση d μέσω ενός γραμμικού στρώματος (Linear Layer) με *Batch Normalization*, όπως και στο μοντέλο πρόβλεψης διάγνωσης. Για την ενσωμάτωση του χρονικού παράγοντα, χρησιμοποιήθηκαν *sinusoidal positional encodings*, επιτρέποντας στο μοντέλο να αναγνωρίζει τη σειρά των επισκέψεων εντός της κάθε ακολουθίας. Η ακολουθία των embeddings τροφοδοτείται σε έναν *κωδικοποιητή* με ένα επίπεδο και τέσσερα attention heads.

Για να αντιμετωπιστεί η ανομοιομορφία στα χρονικά διαστήματα παρακολούθησης μεταξύ ασθενών, ενσωματώθηκαν τρεις τύποι χρονικών πληροφοριών: (i) η ηλικία, (ii) το sinusoidal positional encoding και (iii) το χρονικό διάστημα μεταξύ διαδοχικών επισκέψεων (Δt). Έτσι, το τελικό διάνυσμα εισόδου για κάθε επίσκεψη ορίζεται ως:

$$\tilde{z}^{(t)} = \bar{z}^{(t)} + p^{(t)} + \phi(\Delta t^{(t)}),$$

Σύμφωνα με τη συνήθη πρακτική στη βιβλιογραφία, το πρόβλημα μετατροπής ορίστηκε αποκλειστικά για άτομα που είχαν διάγνωση *ήπιας γνωστικής διαταραχής (MCI)* στην αρχή ή κατά τη διάρκεια της παρακολούθησης, ενώ η *μετατροπή* ορίζεται ως η μετάβαση από MCI σε *νόσο Alzheimer (AD)* στην επόμενη επίσκεψη. Με τον τρόπο αυτό, το μοντέλο εστιάζει στην πρόβλεψη της εξέλιξης από MCI σε AD, αντανακλώντας το ουσιαστικό κλινικό ενδιαφέρον για τον εντοπισμό ατόμων με τον υψηλότερο κίνδυνο ανάπτυξης της νόσου Alzheimer.

1.2.6 Ανάλυση Υποομάδων Ασθενών

Παραγωγή Embeddings

Για τη παραγωγή embeddings από τα δεδομένα, χρησιμοποιήθηκε ένα *Multilayer Perceptron*, με στόχο την εξαγωγή *embeddings* από τα στατικά δεδομένα απεικόνισης εγκεφάλου και γενετικών χαρακτηριστικών. Ο σκοπός του μοντέλου ήταν να συμπιέσει τα χαρακτηριστικά αυτά σε έναν χαμηλότερης διάστασης χώρο

(*latent space*), διατηρώντας παράλληλα τη διακριτική πληροφορία που σχετίζεται με την κατάσταση της νόσου.

Η εκπαίδευση του μοντέλου έγινε με στόχο τη διάκριση *φυσιολογικών ατόμων* (CN) από *μη φυσιολογικά* (Not-CN). Η συνολική συνάρτηση απώλειας ορίστηκε ως το άθροισμα της *δυαδικής διασταυρούμενης εντροπίας (binary cross-entropy)* για το πρόβλημα ταξινόμησης CN vs Not-CN και ενός *αντιθετικού όρου απώλειας (contrastive loss)*, ο οποίος ωθούσε τις ενσωματώσεις των CN δειγμάτων να πλησιάζουν μεταξύ τους στο χώρο, ενώ απομάκρυνε τις Not-CN. Με αυτόν τον τρόπο, οι παραγόμενες αναπαραστάσεις είναι δομημένες με τρόπο που ενισχύει τη διακριτότητα μεταξύ φυσιολογικών και παθολογικών ομάδων.

Έστω ότι το διάνυσμα χαρακτηριστικών του δείγματος i είναι $x_i \in \mathbb{R}^d$. Το MLP αποτελείται από L πλήρως συνδεδεμένα στρώματα, καθένα από τα οποία ακολουθείται από μη γραμμική ενεργοποίηση ReLU:

$$h^{(0)} = x_i, \quad h^{(l)} = \mathrm{ReLU}(W^{(l)} h^{(l-1)} + b^{(l)}), \quad l = 1, \dots, L$$

Το τελικό κρυφό στρώμα παράγει το διάνυσμα ενσωμάτωσης $z_i \in \mathbb{R}^k$:

$$z_i = h^{(L)}$$

Το MLP εκπαιδεύτηκε σε 1.463 ασθενείς, χρησιμοποιώντας *stratified 5-πλή διασταυρούμενη επικύρωση (5-fold stratified cross-validation)*. Για την ποιοτική αξιολόγηση του διαχωρισμού μεταξύ ομάδων, οι ενσωματώσεις οπτικοποιήθηκαν σε δύο διαστάσεις χρησιμοποιώντας τις τεχνικές UMAP και t-SNE.

Αλγόριθμος Ομαδοποίησης

Μετά την εξαγωγή των embeddings z_i από το MLP, εφαρμόστηκε *ομαδοποίηση (clustering)* με στόχο τον εντοπισμό ομάδων ατόμων με παρόμοια πρότυπα στο χώρο. Η ομαδοποίηση στο χώρο των embeddings, αντί απευθείας στα αρχικά χαρακτηριστικά, επιτρέπει τον σχηματισμό ομάδων με βάση αναπαραστάσεις ανώτερων διαστάσεων που πιθανόν αντικατοπτρίζουν καλύτερα τη δομή της νόσου.

Ο κύριος αλγόριθμος που χρησιμοποιήθηκε ήταν ο *k-Means*, ενώ για σύγκριση δοκιμάστηκαν και οι *Agglomerative clustering* και *DBScan*. Ο αλγόριθμος k-Means χωρίζει το χώρο των embeddings σε K ομάδες:

$$\mathcal{L}_{\text{k-Means}} = \sum_{j=1}^{K} \sum_{z_i \in C_j} \|z_i - \mu_j\|^2$$

όπου C_j είναι το σύνολο των embeddings που ανήκουν στην ομάδα j και μ_j το κεντροειδές αυτής της ομάδας.

Η ποιότητα της ομαδοποίησης αξιολογήθηκε μέσω των δεικτών *Silhouette Score*, *Davies–Bouldin Index* και *Adjusted Rand Index (ARI)*. Επιπλέον, πραγματοποιήθηκε ανάλυση σε επίπεδο χαρακτηριστικών για να περιγραφούν οι προκύπτουσες υποομάδες.

Οι αλγόριθμοι ομαδοποίησης δοκιμάστηκαν για τιμές $k = 2$ έως $k = 5$, με την τιμή $k = 2$ να επιλέγεται βάση κλινικής ερμηνευσιμότητας, αλλά και μετρικών.

Για την ερμηνεία των υποομάδων, εφαρμόστηκαν στατιστικές δοκιμές για τον εντοπισμό χαρακτηριστικών που διαφοροποιούν τις υποομάδες. Χρησιμοποιήθηκαν τόσο παραμετρικές (ANOVA) όσο και μη παραμετρικές (Kruskal–Wallis) δοκιμές σε όλους τους διαθέσιμους εγκεφαλικούς όγκους και SNPs.

1.3 Αποτελέσματα και Ερμηνευσιμότητα

1.3.1 Αποτελέσματα Μοντέλων

AD vs MCI/CN

Table 1.1: AUC μετρικές ανά επίσκεψη

Visit	AUC (ROC)	AUC (PR)
0	0.866	0.622
1	0.855	0.636
2	0.844	0.600
3	0.866	0.618
4	0.836	0.548
5	0.802	0.452
6	0.802	0.548
Mean	**0.853**	**0.575**

Table 1.2: Μετρικές ανά επίσκεψη (AD vs MCI/CN)

Visit	Confusion Matrix	Acc.	Bal. Acc.	Prec.	Recall	Spec.	F1	MCC	AUC (ROC)	AUC (PR)
0	[[385, 11], [58, 26]]	0.86	0.64	0.71	0.32	0.97	0.44	0.41	0.87	0.62
1	[[301, 9], [46, 23]]	0.86	0.67	0.68	0.36	0.97	0.46	0.42	0.86	0.64
2	[[241, 13], [42, 23]]	0.82	0.65	0.63	0.35	0.95	0.45	0.38	0.84	0.59
3	[[208, 16], [28, 28]]	0.85	0.72	0.64	0.51	0.93	0.56	0.48	0.87	0.62
4	[[163, 12], [26, 13]]	0.82	0.64	0.54	0.34	0.93	0.42	0.33	0.84	0.53
5	[[118, 11], [19, 8]]	0.81	0.61	0.44	0.31	0.91	0.36	0.26	0.80	0.45
6	[[79, 7], [15, 11]]	0.80	0.66	0.60	0.40	0.92	0.49	0.38	0.80	0.55

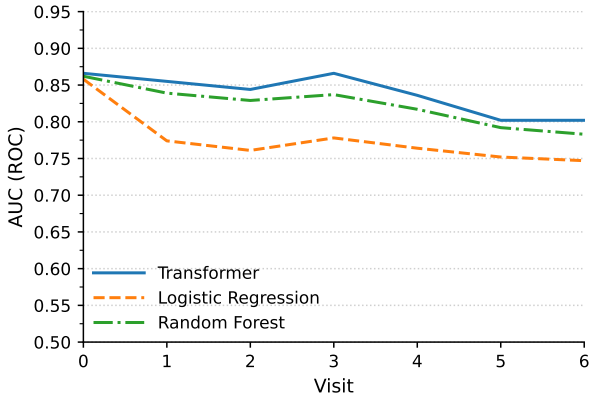


Figure 1.1: AUC (ROC)

AD vs CN

Table 1.3: AUC μετρικές ανά επίσκεψη (AD vs CN)

Visit	AUC (ROC)	AUC (PR)
0	0.930	0.912
1	0.932	0.920
2	0.920	0.910
3	0.940	0.926
4	0.910	0.858
5	0.902	0.818
6	0.876	0.820
Mean	**0.916**	**0.879**

Table 1.4: Μετρικές ανά επίσκεψη (AD vs CN)

Visit	Confusion Matrix	Acc.	Bal. Acc.	Prec.	Recall	Spec.	F1	MCC	AUC (ROC)	AUC (PR)
0	[[172, 5], [16, 68]]	0.92	0.89	0.93	0.81	0.97	0.86	0.81	0.93	0.91
1	[[123, 3], [15, 53]]	0.91	0.88	0.95	0.78	0.98	0.86	0.80	0.93	0.92
2	[[92, 6], [11, 55]]	0.89	0.88	0.90	0.83	0.94	0.86	0.78	0.92	0.91
3	[[82, 7], [8, 48]]	0.90	0.89	0.88	0.85	0.93	0.86	0.78	0.94	0.93
4	[[69, 8], [8, 31]]	0.86	0.85	0.80	0.79	0.90	0.79	0.69	0.91	0.86
5	[[56, 6], [8, 19]]	0.85	0.81	0.77	0.72	0.90	0.74	0.63	0.90	0.82
6	[[42, 8], [8, 18]]	0.80	0.78	0.71	0.71	0.85	0.71	0.56	0.87	0.82

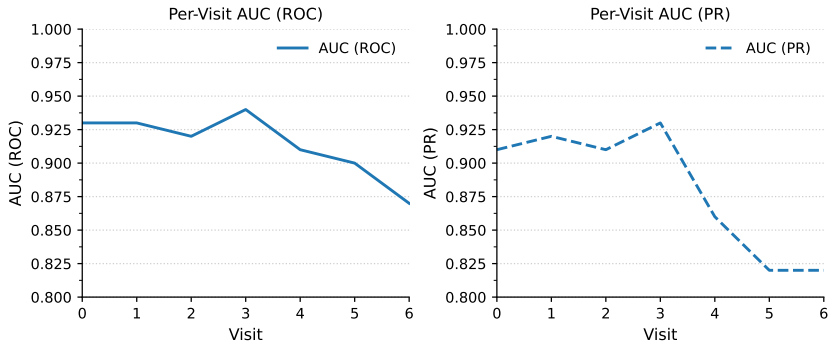


Figure 1.2: AUC ROC και AUC PR για AD vs CN

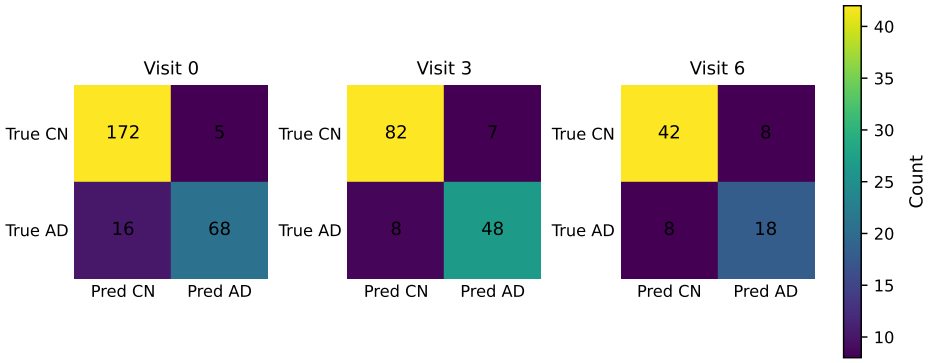


Figure 1.3: AD vs CN Confusion Matrices για 1, 4 και 7 επισκέψεις

MCI to AD

Table 1.5: AUC μετρικές ανά επίσκεψη (πρόβλεψη μετάβασης)

Visit	AUC (ROC)	AUC (PR)
0	0.774	0.186
1	0.770	0.258
2	0.578	0.176
3	0.808	0.450
4	0.674	0.262
5	0.630	0.324
6	0.838	0.628
Mean	**0.724**	**0.326**

1.3.2 Αποτελέσματα Ερμηνευσιμότητας

AD vs MCI/CN

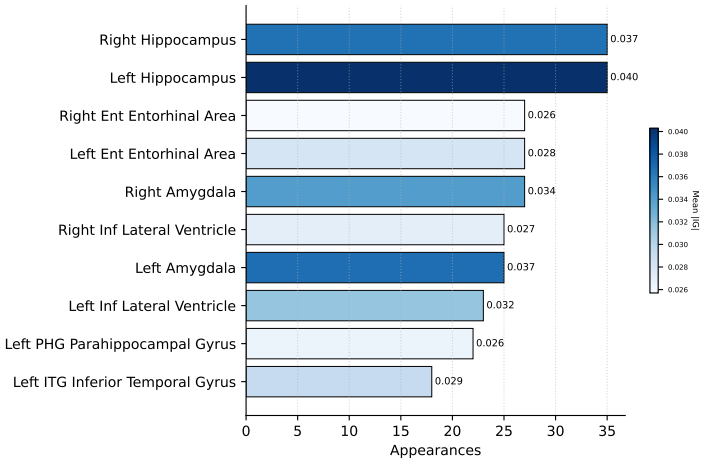


Figure 1.4: AD vs MCI/CN ROIs

AD vs CN

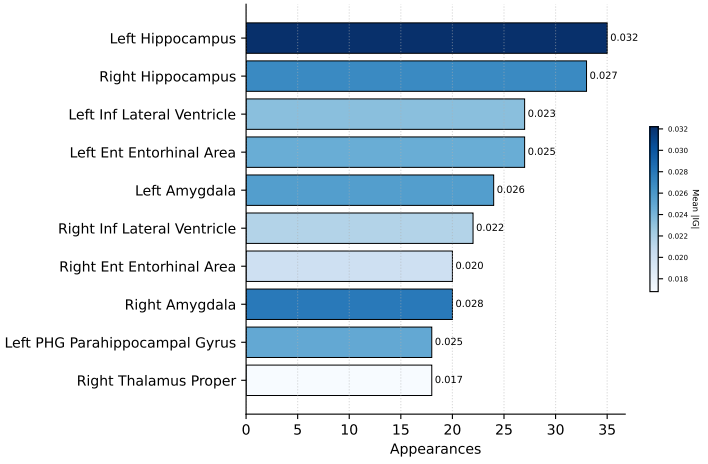


Figure 1.5: AD vs CN ROIs

MCI to AD

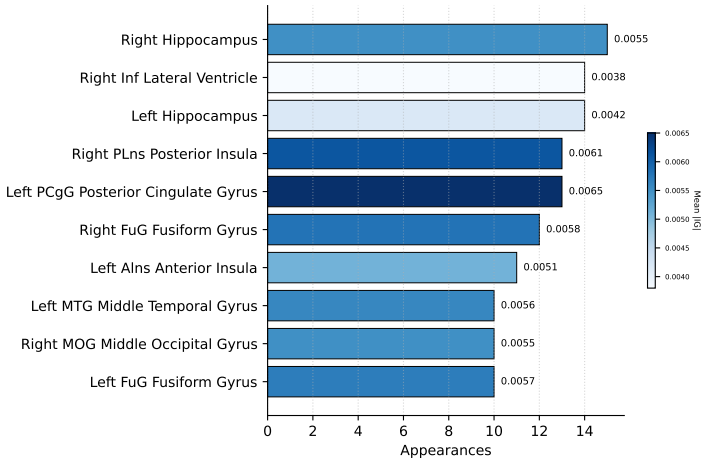


Figure 1.6: MCI to AD ROIs

1.3.3 Αποτελέσματα Ανάλυσης των Υποομάδων

Βιοδείκτες ανά Υποομάδα

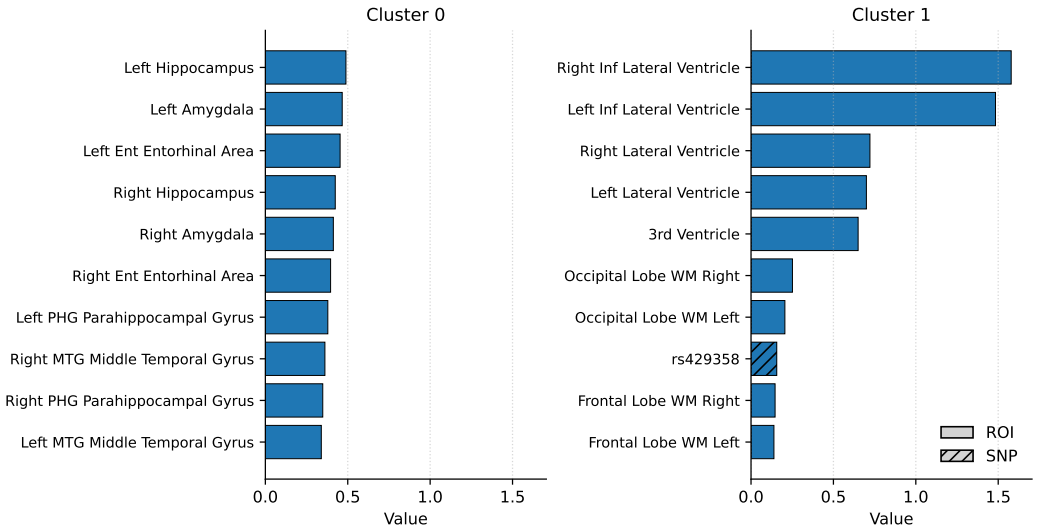


Figure 1.7: Αυξημένα features ανά υποομάδα

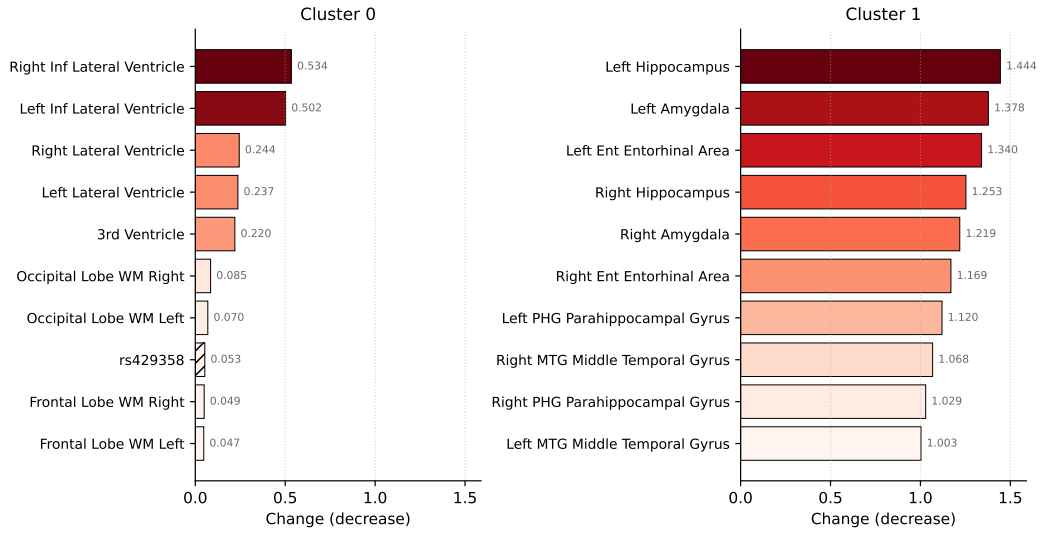


Figure 1.8: Μειωμένα features ανά υποομάδα

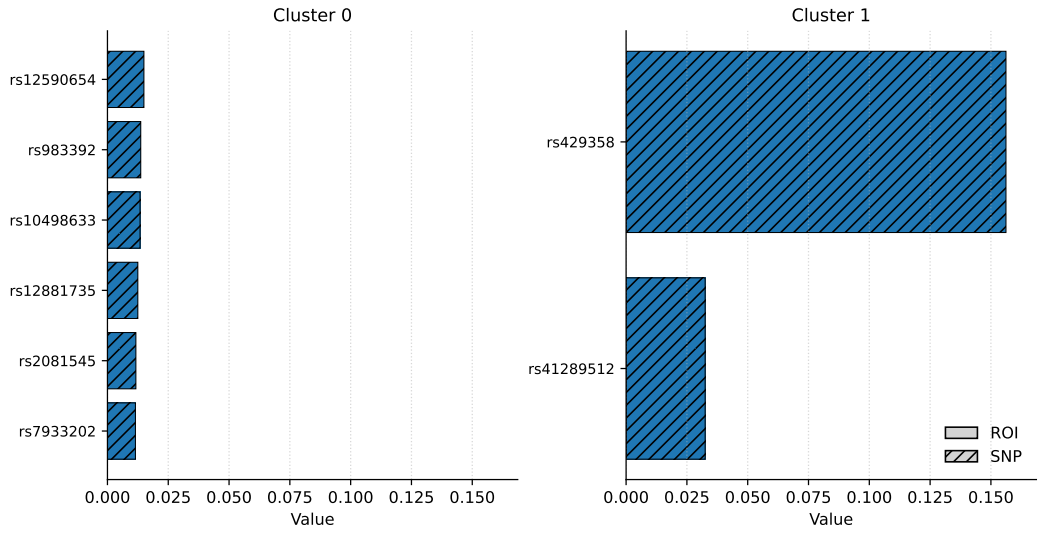


Figure 1.9: Αυξημένα SNPs ανά υποομάδα

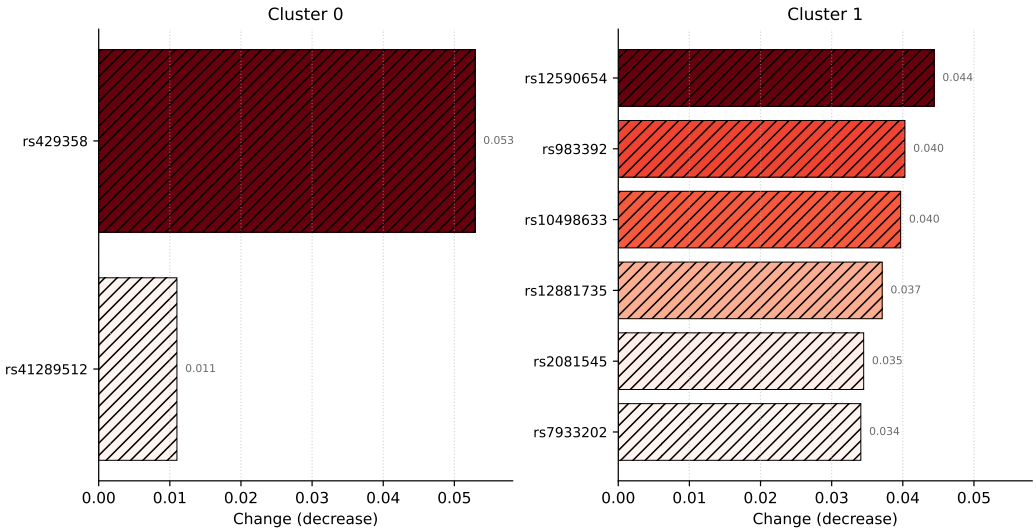


Figure 1.10: Μειωμένα SNPs ανά υποομάδα

AD vs MCI/CN

Table 1.6: AUCs για AD vs MCI/CN ανά υποομάδα

Cluster 0		**Cluster 1**	
AUC (ROC)	**AUC (PR)**	**AUC (ROC)**	**AUC (PR)**
0.846 ± 0.036	0.488 ± 0.106	0.655 ± 0.167	0.678 ± 0.101

Table 1.7: Top ROIs ανά υποομάδα

Subgroup 0		**Subgroup 1**	
Region	**Appearances**	**Region**	**Appearances**
Left Hippocampus	35	Left Hippocampus	35
Right Hippocampus	35	Right Hippocampus	35
Right Ent Entorhinal Area	31	Left Amygdala	34
Right Amygdala	27	Left Ent Entorhinal Area	33
Left Amygdala	24	Right Amygdala	31
Left Ent Entorhinal Area	23	Left Inf Lateral Ventricle	30
Right Inf Lateral Ventricle	21	Right Inf Lateral Ventricle	29
Left PHG Parahippocampal Gyrus	20	Left PHG Parahippocampal Gyrus	27
Left Inf Lateral Ventricle	18	Right Ent Entorhinal Area	26
Left ITG Inferior Temporal Gyrus	15	Left ITG Inferior Temporal Gyrus	26

AD vs CN

Table 1.8: AUC μετρικές ανά επίσκεψη (AD vs CN)

Subgroup 0		
Visit	**AUC (ROC)**	**AUC (PR)**
0	0.898	0.708
1	0.884	0.662
2	0.860	0.694
3	0.922	0.764
4	0.896	0.708
5	0.912	0.750
6	0.866	0.706
Mean	**0.891**	**0.713**

Table 1.9: Top ROIs για την υποομάδα 0 (AD vs CN)

Region	**Mean |IG|**	**Appearances**
Left Hippocampus	0.0373	33
Right Hippocampus	0.0311	33
Left Inf Lateral Ventricle	0.0268	21
Left Amygdala	0.0290	21
Right Inf Lateral Ventricle	0.0250	21
Left Ent Entorhinal Area	0.0293	20
Right Thalamus Proper	0.0240	20
Right Ent Entorhinal Area	0.0261	20
Right Amygdala	0.0317	19
Left Thalamus Proper	0.0276	19

MCI to AD

Table 1.10: AUCs για την πρόβλεψη μετάβασης ανά υποομάδα

Cluster 0		**Cluster 1**	
AUC (ROC)	**AUC (PR)**	**AUC (ROC)**	**AUC (PR)**
0.742 ± 0.029	0.347 ± 0.025	0.671 ± 0.014	0.536 ± 0.014

Table 1.11: Top ROIs για την πρόβλεψη μετάβασης στην υποομάδα 0

Region	**Mean |IG|**	**Appearances**
Left PCgG Posterior Cingulate Gyrus	0.0071	15
Left Hippocampus	0.0048	15
Right Hippocampus	0.0060	13
Right FuG Fusiform Gyrus	0.0067	13
Right Plns Posterior Insula	0.0060	12
Left PP Planum Polare	0.0078	11
Right IOG Inferior Occipital Gyrus	0.0057	11
Left Plns Posterior Insula	0.0058	10
Right MOG Middle Occipital Gyrus	0.0060	10
Left FRP Frontal Pole	0.0059	10

Table 1.12: Top ROIs για την πρόβλεψη μετάβασης στην υποομάδα 1 (high-risk)

Region	**Mean |IG|**	**Appearances**
Left Amygdala	0.0079	23
Left Hippocampus	0.0063	21
Right Inf Lateral Ventricle	0.0051	21
Left Inf Lateral Ventricle	0.0085	20
Right Amygdala	0.0054	19
Right Hippocampus	0.0078	17
Left Plns Posterior Insula	0.0070	16
Left Alns Anterior Insula	0.0060	13
Left FuG Fusiform Gyrus	0.0087	13
Right Plns Posterior Insula	0.0065	12

Chapter 2

Introduction

Over 50 million people have some form of dementia, while this number is bound to increase to 152 million until 2050. Cognitive malfunction is affected by a number of factors. Mild Cognitive Impairment (MCI) and dementia consist of the most representative neurodegenrative diseases. Until now, there is no cure that halts dementia's progression. Therefore, it is crucial for the medical science to focus on the early stages of the disease. The clinical diagnosis of dementia is based on the detailed medical history of the patients or their family, as well as the neuropsychological examinations and brain imaging.

Brain aging is related with complex changes in the structure and function of the brain. There are signs of the disease that can be identified during an imaging exam, like magnetic resonance imaging (MRI) and positron emission tomography (PET). While brain atrophy is a critical sign in neurodegenerative diseases, the individual stays asymptomatic for a long time before the diagnosis. Therefore, the development of diagnosis tools for the early identification of neurodegenerative diseases is essential for the management of these diseases.

Artificial intelligence has acquired a major role in our everyday lives, apparent or not. Healthcare is an area where machine learning has and can deeply affect. This research aims on developing an artificial intelligence model for the diagnosis of the progression of Alzheimer's Disease, using numerical data of brain volume measurements, derived from MRIs, as well as genetical data that consist of single nucleotide polymorhpisms. The model consist of two layers: the first layer acts as a clustering process, dividing patients into two clusters based on their static features and the second layer consist of a transformer predicting the individual's diagnosis on each examination through time.

Chapter 3

Human Brain and Neurodegenerative Diseases

In this chapter, we will briefly discuss on the anatomical areas of the human brain, and the single nucleotide polymorphisms (SNP), which consist the features included in our datasets. Then, we will expand on neurodegenerative diseases and Alzheimer's disease.

3.1 Brain Anatomy

Human nervous system, is the most complex organ found in a living organism, after 600 million years of evolution. The nervous system is composed of two parts, the central nervous system (CNS) and the peripheral nervous system (PNS). The peripheral nervous system consists of the spinal and cranial nevres, while the central nervous system is represented by the brain and spinal cord. The human brain is a relatively small structure weighing about 1400 g and consituting about 2 percent of total body weight. The brain is regarded as the organ solely concerned with thought, memory, and consciousness, but these are only a few of its complex and varied functions. All information we have concerning the world about us is conveyed centrally to the brain by an elaborate sensory system.

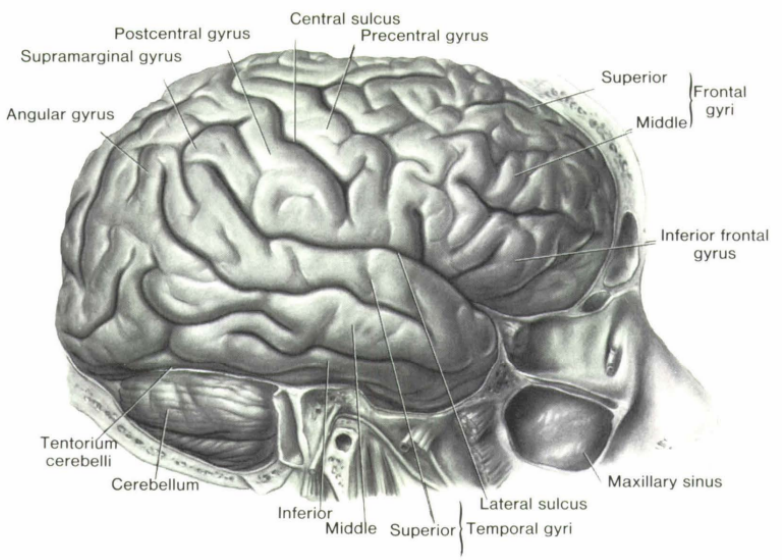


Figure 3.1: Lateral view of the brain exposed in the skull

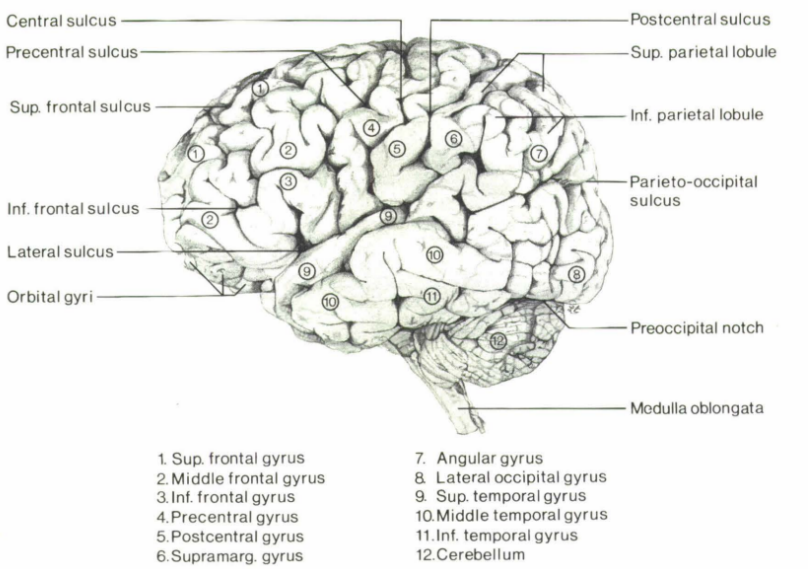


Figure 3.2: Lateral surface of the brain

The brain consists of four subdivisions, the cerebral hemispheres, the brainstem, the diencephalon, and the cerebellum. The paired cerebral hemispheres consist of a highly convoluted gray cortex, an underlying white matter of considerable magnitude and a collection of deeply located neuronal masses, known as the basal ganglia. Each cerebral hemisphere is subdivided in lobes, most of which are named

after the bones of the skull overlying them. The gray cellular mantle of the cerebral cortex in humans is highly convoluted. The crest of a single convolution is referred to as a gyrus. Sulci (fissures) separate the various gyri, producing a pattern with more or less constant features. On the basis of the more constant sulci and gyri, the cerebrum is divided into six so-called lobes: (a) frontal, (b) temporal, (c) parietal, (d) occipital, (e) insular, and (f) limbic. Neither the insula nor the limbic lobe is a true lobe. The insula is a cortical area buried in the depths of the lateral sulcus. The limbic lobe is a highly heterogeneous entity on the medial aspect of the hemisphere consisting of portions of the frontal, parietal, occipital, and temporal lobes which surround the upper part of the brainstem [2].

Diencephalon The diencephalon contains the *thalamus*, the *subthalamus*, and the *hypothalamus*.

Cerebellum The cerebellum is composed of the left and right *cerebellar hemispheres* and midline *vermis* which unites them.

Brainstem The brainstem is subdivided into the *midbrain*, the *pons*, and the *medulla*.

Cortical Areas The cerebral cortex is composed of three areas: the lateral, medial and the inferior, which is also named ventral. Moreover, the transitional areas form the *frontal*, *temporal*, and *occipital poles*.

Lateral Surface Four lobes are visible on the lateral surface of the cerebral hemispheres: the frontal, temporal, parietal, and occipital lobes.

The lateral surface of the **frontal lobe** is subdivided by three sulci—the superior frontal sulcus, inferior frontal sulcus, and precentral sulcus—into four distinct gyri:

- Superior frontal gyrus
- Middle frontal gyrus
- Inferior frontal gyrus
- Precentral gyrus

The lateral surface of the **temporal lobe** is divided by two sulci—the superior temporal sulcus and the inferior temporal sulcus—into three gyri:

- Superior temporal gyrus
- Middle temporal gyrus
- Inferior temporal gyrus

The lateral surface of the **parietal lobe** is subdivided by the intraparietal sulcus into three main gyri:

- Postcentral gyrus
- Superior parietal gyrus (lobule)
- Inferior parietal gyrus (lobule)
  - Supramarginal gyrus
  - Angular gyrus

The lateral surface of the **occipital lobe** is divided by two sulci—the superior occipital sulcus and inferior occipital sulcus—into three gyri:

- Superior occipital gyrus
- Middle occipital gyrus
- Inferior occipital gyrus

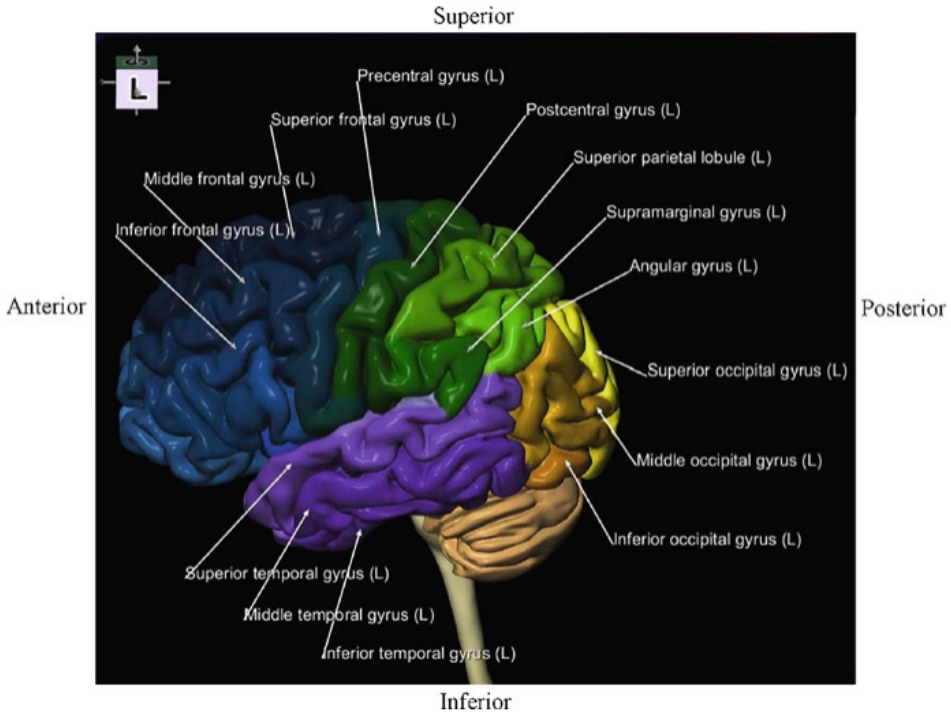


Figure 3.3: Lateral view of the cortical areas of the left hemisphere. Each gyrus is assigned a unique color.

Medial Surface The medial surface of the cerebral hemisphere includes the frontal, parietal, occipital, and limbic lobes. The **limbic lobe** comprises the gyri located along the inner margin (*limbus*) of the hemisphere:

- Subcallosal gyrus (areas)
- Cingulate gyrus
- Isthmus (of the cingulate gyrus)
- Parahippocampal gyrus

The **superior frontal gyrus**, which is separated from the limbic lobe by the cingulate sulcus, occupies most of the medial surface of the frontal lobe. The **parietal lobe** includes the *precuneus*, which is separated from the occipital lobe by the parieto-occipital fissure. The **occipital lobe** consists of the *cuneus* and the *lingual gyrus*.

Inferior Surface

Deep Gray Nuclei The deep gray nuclei are paired gray matter structures.

- **Basal ganglia (nuclei)**
  - Caudate nucleus
  - Lentiform nuclei
    - Putamen
    - Globus pallidus
      - Lateral (outer) segment
      - Medial (inner) segment (see also Sect. **??**)
- **Thalamus**
- **Hippocampus**
- **Amygdala (amygdaloid body)**

The lentiform nuclei together with the caudate nucleus form the *striatum*.

Ventricular System The ventricular system consists of four interconnected cerebral ventricles (cavities) filled with cerebrospinal fluid (CSF):

- Left and right lateral ventricles
- Third ventricle

- Fourth ventricle

CSF is secreted primarily by the choroid plexus, a network of blood vessels located within the ventricles. It circulates from the lateral ventricles through the paired interventricular foramina (of Monro) into the third ventricle, and subsequently passes through the cerebral aqueduct to reach the fourth ventricle.

The lateral ventricles are the largest of the four and each includes the following regions:

- Body (or central portion)
- Atrium (or trigone)
- Horns
  - Frontal (anterior)
  - Occipital (posterior)
  - Temporal (inferior)

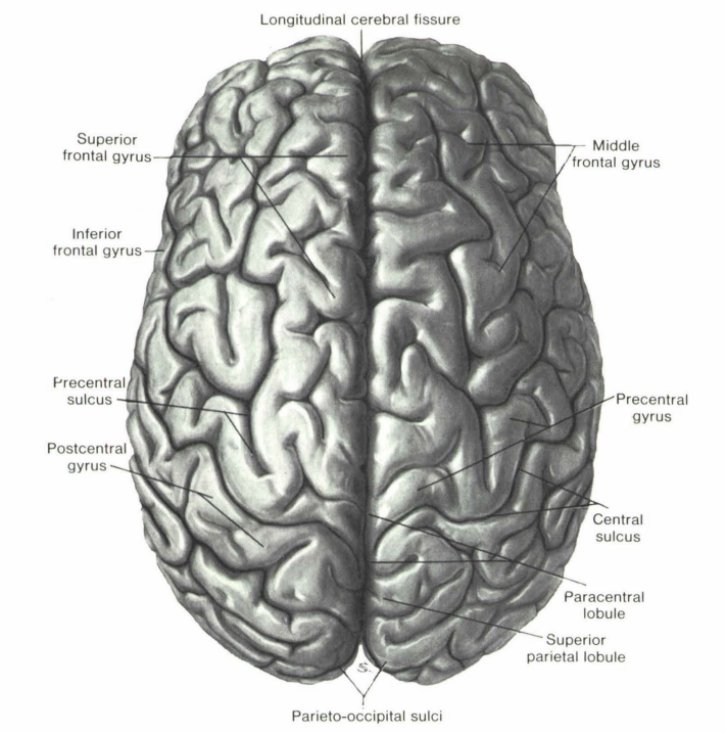


Figure 3.4: Superior view of the brain

3.2 Single Nucleotide Polymorphisms

A single nucleotide polymorphism (SNP, produced snip) is a genomic variant at a single base position in the DNA. For example, two sequenced DNA fragments from different individuals, AAGCCTA to AAGCTTA, contain a difference in a single nucleotide. In this case we say there are two alleles: C and T. Almost all common SNPs have only two alleles.

Single nucleotides may be changed (substitution), removed (deletions) or added (insertion) to a polynucleotide sequence. Single nucleotide polymorphisms may fall within coding sequences of genes, non-coding regions of genes, or in the intergenic regions between genes. SNPs within a coding sequence will not necessarily change the amino acid sequence of the protein that is produced, due to degeneracy of the genetic code.

A SNP in which both forms lead to the same polypeptide sequence is termed synonymous (sometimes called a silent mutation) - if a different polypeptide sequence is produced they are nonsynonymous. A nonsynonymous change may either be missense or nonsense, where a missense change results in a different amino acid, while a nonsense change results in a premature stop codon. SNPs that are not in protein-coding regions may still have consequences for gene splicing, transciption factor binding, or the sequence of non-conding ribonucleic acid (RNA).

Variations in the DNA sequences of humans can affect how humans develop diseases and respond to pathogens, chemicals, medication, vaccines, and other agents. SNPs are also thought to be key enablers in realizing the concept of personalized medicine. However, their greatest importance in biomedical research is for comparing regions of the genome between cohorts (such as matched cohorts with and without a disease).

3.3 Brain Aging

Brain aging is complex biological process, which is characterized by the accumulation of molecular and cellular damage during one's life. The body's inability of restoring this damage, leads to the loss of certain body functions, like feeling, moving and cognitive functions. Aging also consist a basic risk factor for a number of disease, including cancer, cardiovascular disease, as well as neurodegenerative ones.

The brain is specially sensitive in aging's effect, changing its structure and cognitive processes. The most common changes related to aging is brain atrophy (the decrease in volume of gray matter), the decline in quality and volume of the white matter and the abnormal connectivity between them.

3.4 Disease Related to Brain Aging

3.4.1 Mild Congitive Impairment

Mild cognitive impairment (MCI) is an early stage of memory loss or other cognitive ability loss (such as language or visual/spatial perception) in individuals who maintain the ability to independently perform most activities of daily living. MCI can develop for multiple reasons, and individuals with MCI may go on to develop dementia; others will not. For neurodegenerative diseases, MCI can be an early stage of the disease continuum including for Alzheimer's. In some individuals, MCI reverts to normal cognition or remains stable. In other cases, such as when a medication causes cognitive impairment, MCI is mistakenly diagnosed.

Mild cognitive impairment is classified based on the thinking skills affected:

- Amnestic MCI: MCI that primarily affects memory. A person may start to forget important information that he or she would previously have recalled easily.

- Nonamnestic MCI: MCI that affects thinking skills other than memory, including the ability to make sound decisions, judge the time or sequence of steps needed to complete a complex task, or visual perception.

Diagnosis The main criteria defined for the identification of Mild Cognitive Impairment (MCI) are the following: (1) the individual does not present normal cognitive function, but also does not meet the criteria for dementia; (2) there is evidence of a decline in cognitive abilities, either objectively measured over time or reported by the individual or an informant, accompanied by demonstrable cognitive dysfunction; and (3) daily activities are preserved, with complex functions either unaffected or only minimally impaired. These criteria aim to broaden the concept of MCI to include cognitive domains beyond memory and to recognize it as a prodromal stage of various types of dementia [3].

MCI does not reflect a long-standing state of reduced cognitive function, but rather a change in an individual's cognitive abilities. For this reason, having longitudinal information on the patient's cognitive history is essential, allowing the clinician to focus on the specific nature and timing of cognitive changes that have occurred. For example, if memory dysfunction is the primary symptom, the clinician should concentrate on events or difficulties that have arisen recently—typically within the past six to twelve months [4].

Progression to dementia The amnestic single-domain and multi-domain subtypes of MCI with a degenerative etiology are indicative of a possible progression to Alzheimer's disease. In contrast, the non-amnestic subtypes of MCI, where the impairment involves cognitive domains other than memory, are more often associated with the future development of non-Alzheimer dementias, such as frontotemporal dementia or dementia with Lewy bodies. Furthermore, individuals with amnestic multi-domain MCI are more likely to convert to dementia at a faster

rate (typically 10–15% per year) compared to those with amnestic single-domain MCI. The combination of the clinical subtype and the presumed underlying etiology can therefore be particularly useful in predicting the eventual type of dementia into which these syndromes may progress [4]. The ability to identify individuals with MCI who are likely to progress to dementia or Alzheimer's disease more rapidly than others remains an important area of ongoing research within the field of MCI. Recent advances in machine learning and neuroimaging have enabled data-driven approaches to this problem, aiming to model disease trajectories and predict future conversion risk.

In recent years, neuroimaging has proven particularly valuable for predicting the progression of MCI to Alzheimer's disease. Several modalities are sensitive to MCI-related changes, including magnetic resonance imaging (MRI), positron emission tomography (PET), and electroencephalography (EEG). The neuroimaging and electrophysiological assessments used for the study of MCI often overlap with those applied in the early stages of dementia. Hippocampal atrophy has been shown to be a strong predictor of the conversion of amnestic MCI to Alzheimer's disease, while other structural indicators—such as total brain volume and ventricular enlargement—also contribute to predictive performance. These findings highlight the utility of both structural MRI and FDG-PET imaging. Moreover, medial temporal lobe atrophy on MRI and glucose hypometabolism on FDG-PET have been observed in patients with MCI, and the presence of these alterations has a high predictive value for subsequent conversion to dementia [16]. In parallel, molecular imaging techniques that allow in vivo visualization of pathological processes have also been considered particularly promising for understanding and tracking disease progression.

It is also worth noting that carriage of the apolipoprotein E ε4 (APOE-ε4) allele represents a well-established genetic risk factor, as it has been shown to contribute to the prediction of MCI progression to Alzheimer's disease. Mutations or allelic variations in the APOE gene markedly increase the risk of conversion from amnestic MCI to Alzheimer's disease by altering cholesterol transport and synaptic plasticity. Moreover, the presence of the APOE-ε4 allele has been associated with a more rapid rate of hippocampal atrophy on MRI, even in cognitively normal individuals.

3.4.2 Dementia

Dementia is a general term for loss of memory, language, problem-solving and other thinking abilities that are severe enough to interfere with daily life. Alzheimer's is the most common cause of dementia. Dementia is not a single disease. It's an overall term to describe a collection of symptoms that one may experience if they are living with a variety of diseases, including Alzheimer's disease. Diseases grouped under the general term "dementia" are caused by abnormal brain changes. Dementia symptoms trigger a decline in thinking skills, also known as

cognitive abilities, severe enough to impair daily life and independent function. They also affect behavior, feelings and relationships.

Alzheimer's disease accounts for 60 percent - 80 percent of cases. Vascular dementia, which occurs because of microscopic bleeding and blood vessel blockage in the brain, is the second most common cause of dementia. Those who experience the brain changes of multiple types of dementia simultaneously have mixed dementia. There are many other conditions that can cause symptoms of cognitive impairment but that aren't dementia, including some that are reversible, such as thyroid problems and vitamin deficiencies. Dementia symptoms are progressive, which means that the signs of cognitive impairment start out slowly and gradually get worse over time, leading to dementia.

Causes Dementia is caused by a variaty of diseases that cause damage to brain cells. This damage interferes with the ability of brain cells to communicate with each other. When brain cells cannot communicate normally, thinking, behavior and feelings can be affected.

Different types of dementia are associated with particular types of brain cell damage in particular regions of the brain. For example, in Alzheimer's disease, high levels of certain proteins inside and outside brain cells make it hard for brain cells to stay healthy and to communicate with each other. The brain cells in the hippocampus are often the first to be damaged. That's why memory loss is often one of the earliest symptoms of Alzheimer's.

Diagnosis The evaluation of dementia requires a concise medical history, as well as cognitive and neurological examination. The medical history remains the most important diagnostic tool and should be obtained both from the patient and from an informant, since while some patients may recognize their memory loss, others may not recall relevant details or may present with anosognosia—that is, lack of awareness of their condition. The history should focus on medical conditions that affect cognitive function, including vascular risk factors (such as hypertension and diabetes), pre-existing neurological disorders (such as stroke, Parkinson's disease, or traumatic brain injury), and current medications that may impair cognition (for example, anxiolytics such as benzodiazepines or analgesics containing codeine).

Cognitive testing helps determine the presence, severity, and nature of cognitive impairment, while neurological examination can identify objective signs of neurocognitive dysfunction such as aphasia, apraxia, and agnosia. Physical examination is also necessary to detect systemic vascular disease or other systemic findings that may be associated with rarer causes of dementia. Routine laboratory evaluation typically includes blood tests (for instance, vitamin B12 and thyroid-stimulating hormone [TSH]) and neuroimaging to identify cortical and hippocampal atrophy—commonly seen in Alzheimer's disease—or neuropathology suggestive of potentially treatable causes of dementia [5].

For patients whose diagnosis is uncertain or inconsistent with Alzheimer's

disease, clinicians may consider referral to a specialist and the performance of additional diagnostic tests. Functional neuroimaging, such as positron emission tomography (PET), can reveal metabolic patterns suggestive of Alzheimer's disease—typically showing bilateral yet asymmetric temporoparietal hypometabolism using standard tracers such as fluorodeoxyglucose (FDG).

In cases of frontotemporal dementia (FTD), FDG-PET typically demonstrates reduced and asymmetric frontal lobe metabolism in patients with the behavioral variant, and anterior temporal hypometabolism in those with the language variant. In some cases, cerebrospinal fluid (CSF) analysis may also be necessary to detect biomarkers indicative of Alzheimer's disease (e.g., low amyloid-β and elevated tau protein levels), other neurodegenerative disorders, or secondary causes of dementia. Finally, genetic testing can be useful, particularly in younger patients with a strong family history or first-degree relatives affected by dementia [5].

Chapter 4

Theoretical Background

4.1 Introduction

This chapter outlines the theoretical background that underpins the methodology of this thesis. It first introduces the main principles of machine learning, with particular focus on approaches and model families relevant to biomedical data analysis. The discussion then moves to deep learning, where commonly used architectures are reviewed, with an emphasis on Transformers, which serve as the central framework in this study. The chapter also includes a section on interpretability, describing techniques such as integrated gradients that are employed to better understand model predictions. Overall, the goal is to provide the reader with the conceptual foundations necessary to follow the design choices and experimental strategy developed in the remainder of this work.

4.2 Machine Learning Problems

Until quite recently, most computer programs encountered in daily life were built as fixed sets of rules, explicitly defined by programmers to govern how the software should respond in every situation. However, many tasks we wish to automate cannot be fully captured by handcrafted instructions. For example, designing a rule-based program that identifies every person in an image and draws a bounding outline around them is extremely challenging. Although such recognition feels effortless to humans, the exact sequence of cognitive steps involved is not consciously accessible, making it difficult to encode manually.

Machine learning addresses this limitation by developing algorithms that improve automatically through experience. Here, experience is typically provided in the form of data or interactions with an environment, and performance is measured by how effectively the algorithm generalizes from that experience to new situations.

One of the domains where machine learning is expected to have a profound societal impact is healthcare [6]. With the increasing availability of large-scale

medical data, ranging from electronic medical records (EMRs) to clinical notes and imaging studies, new opportunities have emerged for data-driven decision support. Health data can be broadly categorized into structured information, such as laboratory values or demographic variables, and unstructured information, such as free-text physician notes or diagnostic images. Since most medical data are unstructured, the ability to analyze and extract knowledge from such sources is especially important for deploying machine learning effectively.

Applied appropriately, machine learning has the potential to assist clinicians across a wide range of tasks: improving diagnostic accuracy, recommending personalized treatments, identifying patients at high risk for adverse outcomes, and ultimately enhancing patient care while reducing costs. By leveraging both structured and unstructured data, machine learning can support more informed decision-making and strengthen the patient–doctor relationship.

There are four primary types of learning: supervised, unsupervied, semi-supervied and reinforcement learning.

4.2.1 Supervised Learning

Supervised learning refers to tasks in which a dataset contains both input variables (features) and their corresponding outputs (labels), and the objective is to learn a model that can predict the correct label given new input features. Each pair of inputs and labels is referred to as an example. From a probabilistic perspective, supervised learning typically involves estimating the conditional probability of a label given the input features. Among the various paradigms in machine learning, supervised learning has produced the majority of practical successes, largely because many real-world problems can be framed as predicting an unknown quantity based on available data. Examples include classifying medical images as cancerous or not, diagnosing Alzheimer's disease from MRI scans, or predicting the correct translation of a sentence from English into French. In essence, supervised learning can be summarized as the task of "predicting labels from input features."

The process of supervised learning can be described as follows. First, a collection of examples with known features is obtained, and a subset of them is paired with ground-truth labels. These labels may already exist (e.g., from clinical records) or may need to be generated by human annotators. Together, the features and their corresponding labels constitute the training set. A supervised learning algorithm is then applied to this dataset. The algorithm takes the training data as input and outputs a predictive function, commonly referred to as the learned model. Finally, the model can be evaluated on new, unseen data by providing it with input features and comparing its predicted outputs to the true labels. The full process is shown below:

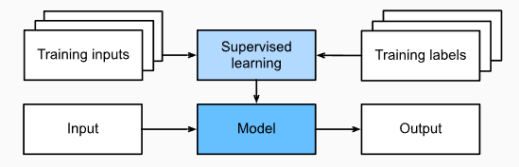


Figure 4.1: Supervised learning

4.2.2 Unsupervised and Self-Supervised Learning

In supervised learning, models are trained on datasets where both input features and corresponding labels are available, providing direct guidance on the desired output for each example. In contrast, *unsupervised learning* deals with datasets that contain only features, without associated labels. The objective in this setting is to uncover hidden patterns, structures, or groupings within the data. A common example is clustering, where the goal is to group similar observations together. For instance, one might cluster images into categories such as landscapes, animals, or people, or group users with similar browsing behaviors based on their activity logs.

A recent extension of this paradigm is *self-supervised learning*, which leverages inherent structure in unlabeled data to create auxiliary prediction tasks that provide supervisory signals. In natural language processing, a common strategy is to mask certain words in a sentence and train the model to predict them from surrounding context. In computer vision, examples include predicting the relative position of cropped image patches, reconstructing occluded regions, or determining whether two samples are augmented views of the same image. These pretext tasks encourage models to learn rich representations that can later be fine-tuned for downstream supervised tasks, often leading to significant performance improvements with limited labeled data.

4.3 Classification Methods

4.3.1 Machine Learning Algorithms

In supervised classification, the objective is to assign a class label y' to a previously unseen data point x', given a dataset $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ of examples with known labels. For simplicity, we focus on the binary case, where $y \in \{0, 1\}$. Each observation x_i is typically represented as an m-dimensional feature vector describing the covariates of interest.

In most real-world applications, there is no deterministic functional mapping $y = f(x)$. Instead, the relationship between inputs and outputs is modeled probabilistically by the joint distribution $P(x, y)$, from which D is assumed to be sampled. According to statistical decision theory, the Bayes-optimal classifier assigns y' by maximizing the posterior probability $P(y \mid x')$.

Machine learning algorithms differ in how they approximate this posterior distribution. Broadly, some models provide only a discrete decision boundary between the two classes (e.g., support vector machines), while others aim to explicitly model $P(y \mid x)$ and thus yield both a predicted label and an estimated probability of class membership. The latter category includes logistic regression, artificial neural networks, k-nearest neighbors, and decision trees. These methods vary in flexibility, interpretability, and computational complexity, and are compared in the subsections that follow.

Random Forests

Random Forests are ensemble learning methods used for both regression and classification. Their core idea is to build a collection of decision trees, each trained on a randomized version of the data (and/or features), and to combine their outputs—averaging for regression or voting for classification—to produce a final prediction.

Formally, assume we have an input random vector $X \in \mathcal{X} \subset \mathbb{R}^p$ and a real-valued target (response) Y. The goal is to estimate the regression function

$$m(x) = \mathbb{E}[Y \mid X = x],$$

using a training sample

$$D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$$

of independent realizations of (X, Y).

A random forest predictor consists of M randomized regression trees. Each tree is constructed using randomness both in sampling the dataset (often via bootstrapping) and in selecting which features to consider at each split. Let $\Theta_1, \dots, \Theta_M$ be i.i.d. random variables determining the randomization in each tree (for example, which examples to sample and which split directions/criteria at nodes). Then the jth tree gives a prediction

$$m_n(x; \Theta_j, D_n),$$

and the forest aggregates these via

$$m_{M,n}(x; \Theta_1, \dots, \Theta_M, D_n) \;=\; \frac{1}{M} \sum_{j=1}^{M} m_n(x; \Theta_j, D_n).$$

In regression trees, each tree's prediction can be written as an average of the training responses Y_i among those data points that fall into the same leaf (cell) as x. Concretely, defining $A_n(x; \Theta_j, D_n)$ to be the leaf cell that contains x, and $N_n(x; \Theta_j, D_n)$ its number of training examples,

$$m_n(x; \Theta_j, D_n) = \frac{1}{N_n(x; \Theta_j, D_n)} \sum_{i \in D_n(\Theta_j)} \mathbf{1}_{X_i \in A_n(x; \Theta_j, D_n)} \, Y_i.$$

Random Forests are consistent under quite general conditions. For example, Breiman's original formulation ([7]) shows they converge (in classification error or regression risk) as the number of trees grows, under some assumptions about strength of individual trees and correlation among them. More recent theoretical work (e.g. [8]) has provided L^2-consistency proofs even in additive regression settings.

In practice, choices like how many trees M, how many features to try at each node (often called 'mtry'), and how large each leaf is will affect bias–variance tradeoffs. Random Forests are popular in biomedical applications because of their flexibility, resistance to overfitting, and ability to handle high-dimensional or mixed-type datasets.

Logistic Regression

Logistic Regression (LR) is one of the most widely used statistical methods for predicting the occurrence of a binary outcome from one or more independent variables [9], [10]. The dependent variable Y takes the value 1 if the event of interest occurs and 0 otherwise. Each predictor variable is assigned a coefficient that quantifies its independent contribution to variation in the dependent variable.

The model is expressed through the natural logarithm of the odds ratio:

$$\ln\left(\frac{P(Y)}{1-P(Y)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k, \tag{1}$$

where $\ln\left(\frac{P(Y)}{1-P(Y)}\right)$ is the log-odds (logit) of the outcome, $X_1, X_2, \ldots, X_k$ are the predictor variables, β_0 is the intercept, and $\beta_1, \ldots, \beta_k$ are the regression coefficients.

Rearranging, the odds can be written as

$$\frac{P(Y)}{1-P(Y)} = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}, \tag{2}$$

and the probability of Y as

$$P(Y) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}}. \tag{3}$$

The parameters β are typically estimated via maximum likelihood estimation, which identifies the coefficients that maximize the probability of the observed data. Each coefficient reflects the expected change in the log-odds of the outcome for a one-unit increase in the corresponding predictor, holding the others constant. Exponentiating a coefficient yields the *odds ratio*, a common measure of effect size in biomedical applications.

Logistic regression is valued for its interpretability, probabilistic outputs, and efficiency. However, it assumes a linear relationship between predictors and the log-odds, which may limit its applicability in more complex data settings.

Support Vector Machines

Support Vector Machines (SVMs) are supervised learning models designed for classification and regression tasks. The key idea is to find a separating hyperplane that maximizes the margin between different classes, thereby improving the model's ability to generalize beyond the training data [11]. In the linearly separable case, this corresponds to identifying the hyperplane

$$w^T x + b = 0,$$

such that the distance (margin) to the nearest training points is maximized. These critical training points are known as support vectors, as they alone determine the decision boundary.

Formally, the optimization problem can be written as

$$\min_{w,b} \ \frac{1}{2}\|w\|^2 \quad \text{subject to } y_i(w \cdot x_i + b) \geq 1 \ \ \forall i,$$

which seeks the hyperplane with maximal margin. In practice, kernel functions extend SVMs to non-linear problems by implicitly mapping the data into higher-dimensional feature spaces where linear separation becomes feasible.

SVMs have been widely applied in biomedical data analysis and medical imaging due to their robustness in high-dimensional spaces and relatively small sample settings. However, their reliance on kernel selection and computational cost in large-scale problems has limited their adoption compared to modern deep learning approaches.

Gradient Boosting

Gradient Boosting (GB) is an ensemble method that builds a strong predictive model by iteratively combining multiple weak learners, typically shallow decision trees [12]. The method frames supervised learning as an optimization problem: given training data $\{(x_i, y_i)\}_{i=1}^N$, the goal is to approximate the underlying function $f(x)$ by minimizing a chosen loss function $L(y, f(x))$.

Instead of fitting $f(x)$ directly, gradient boosting constructs it in stages:

$$f_0(x) = \text{constant}, \quad f_t(x) = f_{t-1}(x) + \rho_t h(x; \theta_t),$$

where $h(x; \theta_t)$ is a base learner fitted to the negative gradient of the loss with respect to the current model $f_{t-1}(x)$, and ρ_t is the optimal step size. By sequentially adding functions aligned with the steepest descent direction, the model gradually reduces the loss.

This approach generalizes to different response types through the choice of loss function: squared error for regression, logistic loss for binary classification, and other tailored losses for specific distributions (e.g., Poisson counts). Gradient boosting is widely used due to its flexibility, high accuracy, and interpretability through feature importance, though it can be computationally expensive and prone to overfitting without careful regularization.

XGBoost

Extreme Gradient Boosting (XGBoost) is an efficient and scalable implementation of gradient boosting that has gained widespread popularity due to its speed, regularization, and strong empirical performance [13], [14]. Like gradient boosting (GB), XGBoost builds an additive model of decision trees, where each new tree is fit to the negative gradients (residual errors) of the loss function from the previous stage.

Formally, given a dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$, with features x_i and labels y_i, the prediction for instance i at boosting step m can be written as the sum over T regression trees:

$$\hat{y}_i^{(m)} = \sum_{t=1}^{T} f_t(x_i), \tag{1}$$

where each f_t corresponds to a tree structure with leaves weighted by parameters w_j.

The learning objective of GB is to minimize a differentiable loss function $L^{(t)}$:

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(m)}), \tag{2}$$

where $l(\cdot)$ measures the difference between the prediction and the true value. To prevent overfitting, common hyperparameters such as subsampling rate, maximum tree depth, and learning rate are employed.

XGBoost extends this formulation by introducing an explicit regularization term $\Omega(f_t)$ that penalizes model complexity. The objective becomes

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(m)}) + \Omega(f_t), \quad \Omega(f_t) = \gamma T + \tfrac{1}{2}\lambda \|w\|^2, \tag{3}$$

where T is the number of leaves in a tree, γ controls the minimum loss reduction required for further partitioning, and λ is an L_2 regularization coefficient on leaf weights. An additional hyperparameter α can be introduced for L_1 regularization, further controlling tree sparsity.

Compared with standard GB, XGBoost also employs column subsampling (random selection of features at each split), which has been shown to reduce overfitting more effectively than row subsampling alone. Together, these improvements make XGBoost highly accurate, robust, and well suited for large-scale structured data.

4.3.2 Deep Learning Algorithms

Introduction to Deep Learning

While many modern deep learning techniques have emerged only in recent decades, the fundamental idea of learning patterns from data has much deeper historical roots. For centuries, scientists and mathematicians have sought methods

to analyze observations and make predictions about the future. This long-standing pursuit of understanding and forecasting is at the core of both natural science and statistics.

A key milestone in the history of artificial intelligence was Alan Turing's famous paper *Computing Machinery and Intelligence*, in which he posed the question "Can machines think?". He introduced what is now known as the Turing test, arguing that a machine could be considered intelligent if its responses in a text-based conversation could not be reliably distinguished from those of a human.

Insights from neuroscience and psychology also shaped the early development of learning algorithms. Since humans clearly exhibit intelligent behavior, many researchers asked whether it might be possible to explain or replicate this ability computationally. One early biologically inspired principle was introduced by Donald Hebb in his book *The Organization of Behavior*, where he suggested that neurons strengthen their connections through repeated co-activation—a concept now referred to as Hebbian learning. This notion inspired subsequent work, including Rosenblatt's perceptron algorithm, and ultimately laid part of the groundwork for modern optimization approaches such as stochastic gradient descent.

The term *neural networks* itself reflects this biological inspiration. Researchers have long attempted to design computational architectures that mimic, at least abstractly, the interconnected networks of neurons in the brain. Over time, the link to biological realism became less literal, yet the terminology persisted. At their core, neural networks share a set of key principles that remain central today: - The alternation of linear transformations and nonlinear activation functions, organized into layers. - The use of the chain rule (backpropagation) to update all network parameters simultaneously during training.

Multilayer Perceptron

Multilayer Perceptrons (MLPs) are among the most widely used neural network architectures for supervised learning tasks [15]. They extend the original Perceptron model introduced by Rosenblatt in the 1950s by stacking multiple layers of neurons and allowing for non-linear decision boundaries through hidden layers.

An MLP is a feed-forward neural network consisting of an input layer, one or more hidden layers, and an output layer. The number of input neurons corresponds to the dimensionality of the feature vector, while the number of output neurons corresponds to the number of target classes in a classification problem. Information propagates through the network layer by layer: input features are linearly combined using weights, transformed by activation functions, and passed forward until an output is produced. The predicted class is typically the one associated with the output neuron of highest activation.

The flexibility of an MLP comes from its architecture: too few neurons or layers may lead to underfitting, while too many may cause overfitting. Choosing the number of layers, neurons, and connections—known as the architecture problem—is therefore central to building effective models.

Training an MLP involves adjusting its weights to minimize the discrepancy between predicted and true outputs. This is typically done via the backpropagation algorithm, which computes gradients of the loss function with respect to the weights using the chain rule and updates them through gradient descent. Backpropagation, together with non-linear activation functions, enables MLPs to approximate complex non-linear mappings and perform well on classification and regression tasks.

Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are specialized architectures designed to process data with a grid-like topology, such as images or time series [15]. Instead of fully connecting all neurons between layers, CNNs employ convolutional layers, where local receptive fields (filters) are learned and shared across spatial locations. This weight sharing reduces the number of parameters, enabling efficient learning and capturing translation-invariant features.

A typical CNN consists of convolutional layers, non-linear activation functions, and pooling layers that downsample feature maps. These components progressively extract hierarchical representations: lower layers detect simple features (e.g., edges), while deeper layers capture more complex patterns. CNNs are trained using backpropagation, with the convolution operation making gradient computation efficient through the chain rule.

CNNs have become the dominant method for tasks involving images and spatial data, achieving state-of-the-art performance in recognition, detection, and segmentation problems.

Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are designed for sequential data, where observations are ordered and dependencies exist across time [15]. Unlike feed-forward networks, RNNs include recurrent connections that allow information to persist across time steps. Formally, the hidden state h_t at time t is updated as

$$h_t = f(W_h h_{t-1} + W_x x_t + b),$$

where x_t is the input at time t, W_h and W_x are weight matrices, b is a bias, and f is a non-linear activation. The hidden state thus serves as a memory that encodes past information relevant to the current prediction.

RNNs are commonly used for language modeling, speech recognition, and other temporal sequence tasks. However, standard RNNs suffer from vanishing and exploding gradients, which limit their ability to model long-range dependencies. Variants such as Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) address this by introducing gating mechanisms that regulate information flow, enabling effective learning of long-term dependencies.

4.4 Transformers

Transformers

Attention mechanisms have become a central component in modern sequence modeling, as they enable the capture of dependencies between elements regardless of their distance in the input or output. In particular, self-attention (or intra-attention) relates different positions within the same sequence to derive richer contextual representations. This approach has demonstrated strong performance across a wide range of tasks, including reading comprehension, summarization, entailment recognition, and general-purpose sentence embeddings.

Many state-of-the-art sequence transduction models adopt an encoder–decoder framework, where the encoder maps an input sequence of symbols $(x_1, \dots, x_n)$ into a sequence of continuous embeddings $z = (z_1, \dots, z_n)$. The decoder then generates the output sequence $(y_1, \dots, y_m)$ in an autoregressive manner, producing each token step by step while conditioning on the previously generated outputs.

The Transformer architecture builds on this encoder–decoder structure but replaces recurrent and convolutional components with stacked layers of self-attention and position-wise feed-forward networks. As illustrated in the original paper [1], the encoder and decoder are structurally similar, differing primarily in how they apply masking and cross-attention mechanisms.

Encoder and Decoder Stacks **Encoder:** The encoder is composed of a stack of identical layers, each consisting of two primary components: a multi-head self-attention mechanism and a position-wise feed-forward network. To stabilize training and improve information flow, residual connections are added around each sub-layer, and the outputs are normalized using layer normalization. This ensures that every sub-layer operates on and returns vectors of consistent dimensionality, typically denoted as d_{model}.

Decoder: The decoder mirrors the encoder's layered structure but includes an additional cross-attention module in each layer. Alongside self-attention and feed-forward sub-layers, the cross-attention mechanism allows the decoder to attend to the encoder's outputs, integrating source-sequence information during generation. As in the encoder, residual connections and layer normalization are applied to each sub-layer. Furthermore, the self-attention in the decoder employs causal masking, which prevents each position from attending to future tokens. This enforces autoregressive generation, ensuring that predictions at step i depend only on outputs from earlier positions.

Attention In general terms, an attention mechanism takes as input a query vector together with a set of key–value pairs and produces an output representation. The output is a weighted combination of the values, where the weights reflect the similarity between the query and each key as measured by a compatibility function.

The most widely used variant in Transformer models is the *scaled dot-product attention*. Here, queries and keys have dimensionality d_k, while values have dimensionality d_v. The attention weights are obtained by computing the dot product between the query and each key, scaling by $\sqrt{d_k}$ to control for vector dimensionality, and applying the softmax function. This yields a normalized distribution that determines how much focus is placed on each value.

For efficiency, multiple queries are processed simultaneously by stacking them into a matrix Q, with corresponding matrices K and V for keys and values. The resulting operation can be written compactly as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V.$$

Two classical forms of attention mechanisms are commonly used: additive attention and dot-product (multiplicative) attention. In additive attention, a feed-forward network with a hidden layer is employed to compute the similarity between queries and keys. By contrast, dot-product attention measures compatibility directly through the inner product of query and key vectors. Although both methods have similar theoretical complexity, dot-product attention is computationally more efficient in practice because it can be implemented with optimized matrix multiplications.

For small query/key dimensions (d_k), both approaches tend to yield comparable results. However, when d_k becomes large, unscaled dot products can grow in magnitude and push the softmax function into regions where gradients are very small. To mitigate this, the dot products are scaled by a factor of $1/\sqrt{d_k}$, which stabilizes the computation and improves training.

The Transformer further extends this idea through *multi-head attention*. Instead of applying a single attention function with full d_{model}-dimensional queries, keys, and values, the model first linearly projects them into lower-dimensional spaces (d_k for queries and keys, d_v for values) using separate learned weight matrices. The attention operation is then carried out in parallel across multiple heads (h in total), each producing its own representation. The outputs of these heads are concatenated and passed through another linear transformation to produce the final attention output. This design allows the model to jointly capture information from different representation subspaces at multiple positions, as illustrated in Figure 2.

Multi-Head Attention Multi-head attention extends the basic attention mechanism by applying it in parallel across multiple learned projections of the queries, keys, and values. Instead of relying on a single attention head, which may average information across all representation subspaces, multiple heads enable the model to capture different types of relationships at different positions simultaneously.

Formally, the multi-head mechanism is defined as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O,$$

where each head is computed as

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V).$$

The learnable projection matrices are given by

$$W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}, \quad W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}, \quad W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}, \quad W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}.$$

Here, h denotes the number of attention heads, while d_k and d_v represent the dimensionalities of the projected key/query and value spaces, respectively. Each head produces an output of dimension d_v, and their concatenation is mapped back into the original model space of size d_{model}. Because the dimensionality of each head is reduced relative to the full model dimension, the computational cost of multi-head attention remains comparable to that of single-head attention, while providing richer representational capacity.

Position-wise Feed-Forward Networks In addition to the attention sub-layers, each encoder and decoder layer also includes a position-wise feed-forward network. This component is applied independently to each position in the sequence, using the same parameters for all positions within a layer. The network consists of two linear transformations separated by a non-linear activation function, typically the rectified linear unit (ReLU):

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2.$$

Although the same transformation is shared across sequence positions, the parameters are distinct for each layer in the stack. Conceptually, this operation can also be viewed as a pair of convolutions with kernel size equal to 1.

The input and output of the feed-forward network match the model dimension d_{model}, while the intermediate hidden dimension, often denoted d_{ff}, is chosen to be larger in order to increase representational capacity.

Positional Encoding Because the Transformer architecture does not rely on recurrence or convolution, it requires an explicit mechanism to incorporate information about token order. This is achieved by adding *positional encodings* to the input embeddings at the bottom of the encoder and decoder stacks. The encodings are defined to have the same dimensionality d_{model} as the embeddings, allowing them to be combined through element-wise addition. Several types of positional encodings have been proposed, including both fixed functions and learned embeddings [1].

A widely used approach employs sinusoidal functions of varying frequencies:

$$\text{PE}(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right),$$
$$\text{PE}(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right),$$

where *pos* denotes the position in the sequence and i indexes the embedding dimension. This construction yields encodings where each dimension corresponds to a sinusoid with a wavelength forming a geometric progression from 2π up to $10000 \cdot 2\pi$. The resulting representation facilitates learning of relative positions, since a shifted position $pos + k$ can be expressed as a linear function of PE(pos).

Alternatively, positional information can be introduced through learned embeddings, in which case the encoding vectors are optimized jointly with the rest of the model parameters. While both approaches have been shown to achieve similar empirical performance, fixed sinusoidal encodings have the additional advantage of enabling extrapolation to sequence lengths beyond those observed during training.

4.5 Metrics

Performance metrics are essential for evaluating and comparing classification models. They allow us to quantify how well a model distinguishes between classes, to compare the performance of different algorithms, and to analyze how a model behaves under different parameter settings ([16]). Most classification metrics are derived from the *confusion matrix*, which compactly summarizes correct and incorrect predictions.

Confusion Matrix

The confusion matrix is a contingency table that records the relationship between actual and predicted class labels. Rows represent the true labels, while columns correspond to model predictions. Correct classifications appear on the main diagonal, while off-diagonal entries indicate misclassifications. This structure provides the basis for a wide range of performance measures.

In what follows, we begin with binary classification metrics, which extend naturally to the multi-class case.

Precision and Recall

Precision measures the proportion of predicted positives that are actually positive:

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{4.1}$$

where TP denotes true positives and FP false positives. Precision therefore quantifies the reliability of positive predictions.

Recall, or sensitivity, measures the proportion of actual positives that are correctly identified:

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{4.2}$$

where FN denotes false negatives. Recall reflects the model's ability to detect all positive cases in the dataset.

These two metrics form the foundation for more advanced indicators such as the F1-score and balanced accuracy.

Accuracy

Accuracy is one of the most widely used metrics and represents the overall proportion of correctly classified instances:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{4.3}$$

Intuitively, accuracy gives the probability that a randomly selected instance will be correctly classified. It is simple and easy to interpret but can be misleading in imbalanced datasets, since the performance on minority classes may be hidden by the majority class. In such cases, accuracy reflects the dominance of larger classes rather than the true model performance across all categories.

Nevertheless, accuracy remains a useful and intuitive measure, bounded between 0 and 1, with the complement often referred to as the misclassification rate.

Balanced Accuracy

Balanced accuracy addresses the shortcomings of standard accuracy by averaging recall across all classes:

$$\text{Balanced Accuracy} = \frac{1}{C} \sum_{i=1}^{C} \frac{TP_i}{TP_i + FN_i}, \tag{4.4}$$

where C is the number of classes.

This metric gives equal weight to each class, regardless of class size, making it more informative for imbalanced datasets. When the class distribution is approximately uniform, accuracy and balanced accuracy converge to similar values. Differences between them become more pronounced as class imbalance increases.

F1-Score

The F1-score combines precision and recall into a single measure by computing their harmonic mean:

$$\text{F1-Score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{4.5}$$

The F1-score ranges from 0 to 1, with higher values indicating a better balance between precision and recall. Because it uses the harmonic mean, the F1-score penalizes large disparities between the two components, giving greater weight to the

smaller value. For example, a model with precision and recall both at 80% achieves an F1-score of 0.80, whereas a model with precision of 60% and recall of 100% achieves only 0.75. This property makes the F1-score particularly useful when the goal is to balance false positives and false negatives, rather than optimizing one at the expense of the other.

In practice, the F1-score is widely used in both binary and multi-class settings, especially in imbalanced classification problems where accuracy alone is insufficient.

Area Under the ROC Curve (AUC–ROC)

Receiver Operating Characteristic (ROC) analysis is a widely used method for evaluating classification performance, particularly in settings with imbalanced data or varying decision thresholds [17], [18]. An ROC curve is a two-dimensional plot with the true positive rate (sensitivity) on the vertical axis and the false positive rate (1–specificity) on the horizontal axis. Each point on the curve corresponds to a different decision threshold of the classifier, and the curve therefore illustrates the trade-off between detecting positives and avoiding false alarms.

While ROC curves provide a visual tool for comparing classifiers, it is often useful to summarize their performance with a single scalar. The most common summary statistic is the *area under the ROC curve* (AUC–ROC). Since the ROC curve lies within the unit square, the AUC is bounded between 0 and 1. A classifier with no discriminative ability (random guessing) corresponds to the diagonal line from $(0, 0)$ to $(1, 1)$, with an AUC of 0.5. Thus, practical classifiers should achieve AUC values above 0.5, with values closer to 1.0 indicating stronger performance.

AUC–ROC has an important probabilistic interpretation: it is equivalent to the probability that the classifier assigns a higher score to a randomly chosen positive instance than to a randomly chosen negative instance [19]. This interpretation links AUC directly to the Wilcoxon rank-sum statistic. Furthermore, AUC is related to other measures, such as the Gini coefficient, where $\text{Gini} = 2 \cdot \text{AUC} - 1$ [20].

The main advantages of AUC–ROC are its threshold-independence and scale-invariance. Unlike accuracy, which depends on a fixed classification threshold, AUC evaluates the model's ability to rank positive examples above negative ones across all thresholds. This makes it especially valuable in biomedical applications, where class imbalance is common and where the choice of decision threshold may vary depending on the clinical context.

Despite its popularity, AUC is not without limitations. It summarizes ranking ability but does not reflect the actual probabilities predicted by a model or the consequences of false positives versus false negatives in specific applications. In cases where precision is critical, the area under the precision–recall curve (AUC–PR) may provide complementary insights. Nonetheless, AUC–ROC remains one of the most widely used and robust performance indicators in machine learning and medical decision-making.

Area Under the Precision–Recall Curve (AUC–PR)

Precision–Recall (PR) curves are commonly used in information retrieval and increasingly in machine learning, particularly for evaluating classifiers on imbalanced datasets [21]. A PR curve plots precision on the vertical axis against recall on the horizontal axis, across all possible classification thresholds. The ideal region of performance lies in the upper-right corner, corresponding to high precision and high recall simultaneously.

Compared with ROC analysis, PR curves often provide a more informative view when the positive class is rare. This is because ROC curves can appear deceptively optimistic under severe class imbalance, as the false positive rate accounts for the large number of negative examples. In contrast, PR curves directly reflect the trade-off between identifying true positives and avoiding false positives, which better highlights differences between models in such settings.

The area under the PR curve (AUC–PR) summarizes the overall performance into a single scalar. Like AUC–ROC, values range between 0 and 1, with higher values indicating stronger performance. However, the baseline for AUC–PR is determined by the prevalence of the positive class: in a dataset where positives make up $p\%$ of the total, random guessing yields an expected AUC–PR of $p/100$. This dependence on class distribution makes AUC–PR particularly sensitive to class imbalance and therefore highly relevant for biomedical applications where positive cases are rare.

In practice, AUC–PR complements AUC–ROC by providing a clearer picture of performance in imbalanced datasets. While ROC curves capture the ability to rank positive examples above negatives across thresholds, PR curves focus on how many of the predicted positives are correct and how well the model captures all true positives. For this reason, AUC–PR is often reported alongside AUC–ROC to give a more complete assessment of classification performance.

4.6 Clustering Algorithms & Validation Metrics

Clustering is a basic process in data analysis. It aims to partition a set of objects into groups called clusters such that, ideally, objects in the same group are similar and objects in different groups are dissimilar to each other [22]. Unlike supervised learning, clustering operates in an unsupervised setting, where no ground-truth labels are available. It is therefore widely used for exploratory analysis, pattern discovery, and data summarization across diverse fields such as biology, image analysis, natural language processing and neuroscience. In practice, clustering can reveal hidden structures within complex, high-dimensional datasets, helping to identify meaningful subgroups or phenotypes in medical data. In this thesis, clustering is applied to the learned feature representations of subjects, with the aim

of uncovering potential subgroups relevant to disease progression. The following subsections provide an overview of the algorithms and metrics used for evaluating the quality of these clusters.

4.6.1 Algorithms

k-Means

The k-means algorithm is one of the most widely used clustering methods due to its simplicity and efficiency. It solves the problem of clustering by minimizing the sum of squared errors (SSE) [23].

In this problem, we are given a set of points $P \subset \mathbb{R}^d$ in a Euclidean space, and the goal is to find a set $C \subset \mathbb{R}^d$ of k points (not necessarily included in P) such that the sum of the squared distances of the points in P to their nearest center in C is minimized.

Thus, the objective function to be minimized is:

$$\text{cost}(P, C) := \sum_{p \in P} \min_{c \in C} \|p - c\|^2$$

where $\|\cdot\|^2$ denotes the squared Euclidean distance.

In order to solve the SSE problem heuristically, the k-means algorithm starts with an initial candidate solution $\{c_1, \ldots, c_k\} \subset \mathbb{R}^d$, which can be chosen arbitrarily (often, it is chosen as a random subset of P). Then, two steps are alternated until convergence: First, for each c_i, the algorithm calculates the set P_i of all points in P that are closest to c_i (where ties are broken arbitrarily). Then, for each $1 \leq i \leq k$, it replaces c_i by the mean of P_i. Because of this calculation of the "means" of the sets P_i, the algorithm is also called the k-means algorithm.

The k-Means Algorithm **Input:** Point set $P \subseteq \mathbb{R}^d$, number of centers k

1. Choose initial centers $c_1, \ldots, c_k$ from $\mathbb{R}^d$
2. **repeat**
   (a) $P_1, \ldots, P_k \leftarrow \emptyset$
   (b) For each $p \in P$ do:
      i. Let $i = \arg\min_{i=1,\ldots,k} \|p - c_i\|^2$
      ii. $P_i \leftarrow P_i \cup \{p\}$
   (c) For $i = 1$ to k do:
      i. If $P_i \neq \emptyset$ then $c_i = \frac{1}{|P_i|} \sum_{p \in P_i} p$
3. **until** the centers do not change

Agglomerative

Agglomerative Clustering is a hierarchical, bottom-up clustering method. Each data point starts as its own cluster, and pairs of clusters are successively merged based on a defined similarity or distance metric until all points belong to a single cluster or a stopping criterion is met [24].

The algorithm proceeds as follows:

1. Initialize each data point as its own cluster.
2. Compute a distance matrix between all clusters using a chosen metric (e.g., Euclidean, Manhattan).
3. Merge the two closest clusters according to a linkage criterion:
   - **Single linkage:** minimum distance between points of the two clusters.
   - **Complete linkage:** maximum distance between points of the two clusters.
   - **Average linkage:** average distance between points of the two clusters.
   - **Ward's method:** minimizes the increase in total within-cluster variance.
4. Update the distance matrix and repeat until the desired number of clusters is reached or all points are merged.

Agglomerative clustering produces a *dendrogram*, which is a tree-like diagram showing the hierarchical merging process. It is particularly useful when the number of clusters is not known in advance or when the cluster structure is hierarchical [25].

Spectral Clustering

Spectral Clustering is a graph-based clustering method that uses the eigenvalues (spectrum) of a similarity matrix to perform dimensionality reduction before clustering. Unlike traditional clustering algorithms such as k-means, spectral clustering is particularly effective for identifying non-convex clusters or clusters connected by complex shapes [26], [27].

The algorithm typically follows these steps:

1. Construct a similarity graph G from the data points, where nodes represent points and edges encode similarity (e.g., using a Gaussian kernel or k-nearest neighbors).
2. Compute the *graph Laplacian* L:

   $$L = D - W$$

   where W is the similarity (adjacency) matrix and D is the degree matrix.

3. Compute the first k eigenvectors of L, corresponding to the smallest eigenvalues (or the largest for the normalized Laplacian).

4. Treat each data point as a vector in the k-dimensional eigenspace and apply a standard clustering algorithm, typically k-means, to these vectors.

Spectral clustering can capture clusters that are not well separated in the original feature space and is widely used in image segmentation, social network analysis, bioinformatics, and other applications involving complex cluster shapes.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN is a density-based clustering algorithm that groups together points that are closely packed, while marking points in low-density regions as outliers. It does not require specifying the number of clusters in advance and can find arbitrarily shaped clusters ([28]).

The algorithm works as follows:

1. Define two parameters:
   - ε (epsilon): the maximum distance between two points to be considered neighbors.
   - *minPts*: the minimum number of points required to form a dense region.
2. Classify points into three categories:
   - **Core points:** have at least *minPts* neighbors within ε.
   - **Border points:** have fewer than *minPts* neighbors but are within ε of a core point.
   - **Noise points:** neither core nor border points.
3. Form clusters by connecting core points and including their reachable border points.

DBSCAN is especially useful when clusters have irregular, non-convex shapes or when the dataset contains noise. Its main advantage is that it does not require specifying the number of clusters a priori.

4.6.2 Validation Metrics

Silhouette Score

The *Silhouette Score*, introduced by [29], is an internal validation metric for clustering that measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation).

For a given sample i, let $a(i)$ be the average distance between i and all other points in the same cluster, and let $b(i)$ be the minimum average distance between i and all points in any other cluster. The silhouette value $s(i)$ is then defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

The silhouette value ranges from -1 to 1:

- Values close to 1 indicate that the sample is well matched to its own cluster and poorly matched to neighboring clusters.
- Values around 0 suggest that the sample lies between two clusters.
- Negative values indicate that the sample may have been assigned to the wrong cluster.

The overall Silhouette Score for a clustering is the mean of $s(i)$ across all samples. Higher values indicate better-defined clusters, making it a useful too for comparing the quality of different clustering configurations.

Davies–Bouldin Index

The *Davies–Bouldin Index* (DBI), introduced by Davies and Bouldin [30] is an internal clustering evaluation metric that quantifies the average similarity between each cluster and its most similar counterpart. For two clusters C_i and C_j, similarity is defined as the ratio of the sum of their average within-cluster scatter (S_i and S_j) to the distance between their centroids (M_{ij}). Formally, the index is given by:

$$\text{DBI} = \frac{1}{k} \sum_{i=1}^{k} \max_{\substack{j=1 \\ j \neq i}}^{k} \frac{S_i + S_j}{M_{ij}}$$

where k is the number of clusters. Lower DBI values indicate more compact and well-separated clusters, with 0 representing ideal separation and cohesion. Unlike external validation measures, DBI does not require ground truth labels and is sensitive to both intra-cluster variance and inter-cluster separation. However, its reliance on centroid-based distances makes it less effective for clusters with non-convex shapes or highly variable densities.

Adjusted Rand Index

The *Adjusted Rand Index* (ARI), introduced by Hubert and Arabie **hubert1985comparing** is an external evaluation metric for clustering that measures the similarity between a clustering result and a given ground truth partition. It is based on counting pairs of samples that are either assigned to the same or different clusters in both partitions. The ARI corrects the original Rand Index for chance, ensuring that a

score close to zero corresponds to random labeling, regardless of the number of clusters.

Given a contingency table where n_{ij} represents the number of objects that are in cluster i in the predicted partition and in cluster j in the ground truth partition, the ARI is defined as:

$$\text{ARI} = \frac{\sum_{ij} \binom{n_{ij}}{2} - \frac{\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \frac{\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}}{\binom{n}{2}}}$$

where:

- n is the total number of samples,
- $a_i = \sum_j n_{ij}$ is the sum over row i of the contingency table,
- $b_j = \sum_i n_{ij}$ is the sum over column j.

The ARI ranges from -1 (complete disagreement) to 1 (perfect agreement), with 0 indicating a level of agreement expected by random chance. Because of its chance correction, ARI is more reliable than the unadjusted Rand Index when comparing clustering results across datasets with varying cluster counts.

4.7 Feature Importance

4.7.1 Analysis of Variance (ANOVA)

Analysis of Variance (ANOVA) is a statistical method used to determine whether there are significant differences between the means of three or more independent groups. It decomposes the total variability in the data into variability *between groups* and *within groups*, and evaluates whether the between-group variance is large relative to the within-group variance.

The test statistic, called the *F-ratio*, is defined as:

$$F = \frac{\text{MS}_{\text{between}}}{\text{MS}_{\text{within}}}$$

where:

- $\text{MS}_{\text{between}} = \frac{\text{SS}_{\text{between}}}{k-1}$ is the mean square between groups,
- $\text{MS}_{\text{within}} = \frac{\text{SS}_{\text{within}}}{N-k}$ is the mean square within groups,
- $\text{SS}_{\text{between}}$ and $\text{SS}_{\text{within}}$ are the sum of squares between and within groups, respectively,

- k is the number of groups, and N is the total number of observations.

A large F value indicates that the group means are significantly different. ANOVA assumes independence of observations, normality within groups, and homogeneity of variances. When assumptions are violated, non-parametric alternatives such as the Kruskal–Wallis test can be used.

4.7.2 Kruskal–Wallis Test

The *Kruskal–Wallis test* is a non-parametric alternative to one-way ANOVA, used to determine whether there are statistically significant differences between the medians of three or more independent groups. Unlike ANOVA, it does not assume normality of the data and is suitable for ordinal or non-normally distributed continuous data.

The test is based on ranking all observations across groups. Let R_{ij} be the rank of the j-th observation in group i, and n_i the number of observations in group i. The Kruskal–Wallis statistic H is computed as:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \left(\bar{R}_i - \frac{N+1}{2} \right)^2$$

where:

- k is the number of groups,
- N is the total number of observations across all groups,
- $\bar{R}_i$ is the average rank of group i.

Under the null hypothesis of equal group distributions, H approximately follows a chi-squared distribution with $k-1$ degrees of freedom. A large value of H indicates that at least one group median is significantly different from the others.

4.8 Interpretability

4.8.1 Integrated Gradients

Deep neural networks achieve state-of-the-art performance in a wide range of domains but are often criticized for their lack of interpretability, which poses challenges in sensitive fields such as medicine and neuroscience. Model interpretability aims to provide insights into the internal decision-making process of a network, enabling researchers and clinicians to better trust and understand the outputs. Among the existing approaches, attribution methods are widely used to assign an importance score to each input feature with respect to a given prediction.

Integrated Gradients (IG) [31] is a widely adopted attribution method designed for differentiable models. It addresses some of the limitations of gradient-based

saliency maps, which often suffer from noise and saturation. The central idea of IG is to compute feature attributions by integrating the gradients of the model's output with respect to the input along a path from a baseline input (often chosen as a zero vector or mean reference) to the actual input. Formally, for a model $F : \mathbb{R}^n \to \mathbb{R}$, input $x \in \mathbb{R}^n$, and baseline x', the attribution for feature i is defined as:

$$\mathrm{IG}_i(x) = (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} \, d\alpha. \tag{4.6}$$

This formulation satisfies two desirable axioms: *sensitivity*, which ensures that features influencing the prediction receive non-zero attribution, and *implementation invariance*, which guarantees that functionally equivalent models yield identical attributions [31]. In practice, the path integral is approximated using a finite sum over discrete steps, which provides a tractable estimation of feature importance.

Integrated Gradients has been applied extensively in healthcare and neuroscience applications, including the interpretation of models for disease diagnosis and progression prediction ([32], [33]). Its ability to highlight input features that drive model decisions makes it particularly valuable for biomarker discovery and for assessing whether models rely on clinically meaningful signals.

4.8.2 SHAP (SHapley Additive Explanations)

SHAP is a model-agnostic explanation method based on cooperative game theory. It assigns each feature an importance value for a particular prediction, ensuring a fair distribution of contribution across all features.

The main idea is to consider each feature as a "player" in a cooperative game where the "payout" is the model's prediction. The *Shapley value* for a feature represents its average contribution to the prediction over all possible subsets of features. Formally, the Shapley value ϕ_i for feature i is computed as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$$

where:

- F is the set of all features.
- S is a subset of features excluding i.
- $f_S(x_S)$ is the model prediction using features in subset S.

SHAP provides both local explanations (for individual predictions) and global explanations (aggregated across the dataset), making it widely used for interpreting complex models such as tree ensembles and neural networks.

4.8.3 LIME (Local Interpretable Model-agnostic Explanations)

LIME is a model-agnostic explanation technique designed to provide local interpretability for individual predictions. Instead of explaining the entire model globally, LIME approximates the model's behavior in the neighborhood of a specific instance with a simpler, interpretable surrogate model (typically linear regression).

The core idea is to perturb the input data around the instance of interest and observe how the model's predictions change. Using these perturbed samples and their corresponding predictions, LIME fits a locally weighted interpretable model that mimics the complex model's decision boundary near that instance. Formally, LIME aims to minimize the following objective:

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

where:

- f is the original (black-box) model.
- $g \in G$ is a simple, interpretable model (e.g., linear or decision tree).
- $\mathcal{L}(f, g, \pi_x)$ measures the fidelity of g in approximating f around instance x, weighted by π_x, a proximity measure defining the neighborhood of x.
- $\Omega(g)$ penalizes the complexity of g to ensure interpretability.

LIME thus provides insight into how individual features influence a specific prediction by showing the local, linear approximation of the model's decision surface. Although it does not guarantee global faithfulness, it is highly valuable for understanding complex, non-linear models at the level of individual observations.

4.9 Related Work

Transformer architectures, originally introduced for sequence modeling in natural language processing, have increasingly been adopted in the medical domain for their ability to capture long-range dependencies and integrate heterogeneous data sources. In clinical prediction tasks, Transformers have been applied to longitudinal electronic health records, imaging and multimodal datasets, where their self-attention mechanism enables the modeling of temporal patterns and interactions across diverse modalities. This capability is particularly valuable in neurodegenerative diseases such as Alzheimer's, where disease progression unfolds over time and involves complex interactions between cognitive assessments, imaging biomarkers, and clinical variables. Recently, [34] demonstrated the potential of Transformer-based multimodal fusion in a large-scale study, achieving state-of-the-art differential diagnosis.

Recent advances have also applied Transformer models in Alzheimer's disease prediction. [35] use a transformer-based framework to fuse multimodal data and predict Aβ and tau burder with high accuracy. [36] propose an interpretable transformer for PET/MRI-based AD prediction-providing both accuracy and transparency.

ROI-based analysis in the field of computational neuroscience can be traced back to around the early 2000s.

In recent years, deep learning models utilizing ROI-based analysis have made significant progress in predicting Alzheimer's disease.

Recent studies highlight the potential of machine learning in advancing Alzheimer's disease diagnosis and prognosis [37], [38], [39], [40]. More recently, [41] developed an AI model for differential diagnosis across multiple dementia etiologies using a large multimodal dataset spanning over 50000 participants, achieving state-of-the-art accuracy and showing clinical utility in augmenting neurologist assessments.

[42] proposed a Transformer-based framework for predicting Alzheimer's disease progression using longitudinal multi-modal data, demonstrating the value of temporal modeling for forecasting future conversion.

[43] also extend the approach by comparing conventional 3D convolutional neural networks with vision Transformers for AD classification, showing that Transformer-based architectures can better capture long-range dependencies in structural brain imaging. Their results emphasize the advantages of attention mechanisms over purely convolutional approaches in handling high-dimensional neuroimaging data.

Building upon this body of work, the present thesis focuses on a Transformer-based framework designed to jointly model multimodal longitudinal data for both diagnosis and disease progression prediction. In addition, special emphasis is placed on interpretability and subgroup analysis, with the goal of providing clinically meaningful insights alongside predictive accuracy.

Chapter 5

Methods

5.1 Dataset Overview & Preprocessing

The data used for this experiment were acquired from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and comprise two datasets: one containing static cross-sectional data for patients (including brain region measures and SNPs), and one with longitudinal data focusing on brain region volumes per visit. The latter was used for predicting disease diagnosis and disease conversion tasks, while the cross-sectional subset was used for exploratory analysis within subgroups. In order to extract the numerical ROI volumes, brain MRI images were preprocessed by applying a skull-stripping algorithm [44], [45].

5.1.1 Longitudinal Dataset

Dataset Split

As stated above, the first dataset used for the transformer stage contains longitudinal data. More specifically, there are 2398 patients with a total number of 10805 examinations. A small number of visits (<1% of all examinations) contained invalid ROI volumes (all-zero or missing values). These visits were excluded before modeling, leaving behind 2391 patients. The dataset contains static data about the patient (PHASE, SITE, Sex, and Race), as well as longitudinal information for 145 brain ROIs and the age at the time of each visit. However, time intervals between visits are not guaranteed to be homogeneous, therefore further analysis is required. As shown in Figures 5.5 and 5.6 the total number of visits and the total timespan per patient also vary, which influences the number of visits chosen for the final tasks, both diagnosis and conversion. Many patients also seem to have only a single visit, which, while proving useful for the diagnosis task, seems redundant for the conversion task, motivating a split into two subsets: (i) one containing the full cohort, which we will call the *diagnosis cohort*, and (ii) one tailored to the conversion task, which we will call the *conversion cohort*. In order to construct the

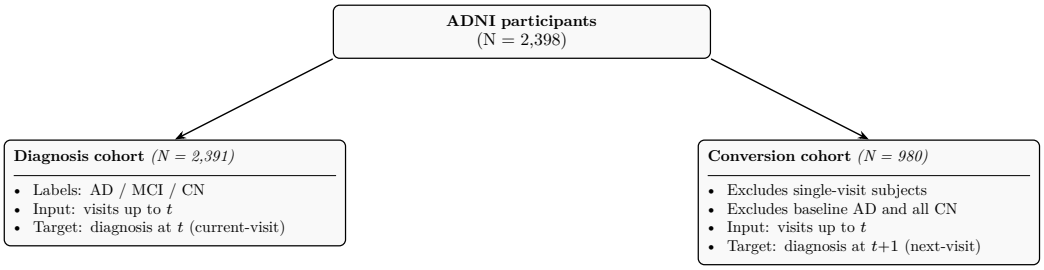


Figure 5.1: Cohort derivation from the ADNI sample.

conversion cohort, we exclude patients with only one visit, those diagnosed with AD at baseline, as well as all CN diagnoses; therefore the resulting binary task is next-visit MCI→AD vs MCI→MCI.

Diagnosis Cohort

Across the observation window, the dataset consists of 2391 patients contributing 10730 examinations. At the visit level, diagnostic labels were distributed as 3700 CN (34.5%), 4781 MCI (44.6%) and 2249 AD (21.0%) occurrences, indicating that MCI is the most frequent state in this cohort.

Table 5.1: Dataset summary statistics.

Statistic	Count
Total patients	2,391
Total examinations	10,730
CN occurrences	3,700
MCI occurrences	4,781
AD occurrences	2,249

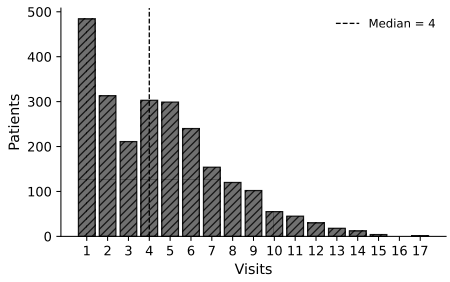


Figure 5.2: Number of visits per patient

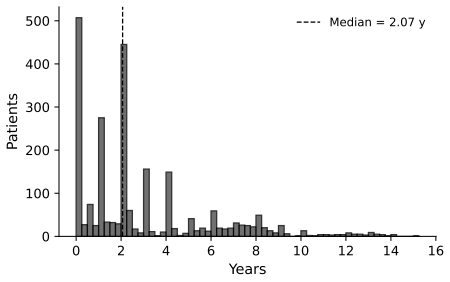


Figure 5.3: Total timespan per patient.

Table 5.2: Number of participants with at least k examinations.

Visits	Total participants
1	2391
2	1907
3	1594
4	1383
5	1080
6	781
7	541
8	387
9	267
10	165

Even though the class ratios remain relatively stable until later visits, the number of available patients drops sharply. This makes those later visits less informative for modeling, as the sample size shrinks too much. As shown in Table 5.3, the baseline distribution is balanced, with CN representing 36%, MCI 46%, and AD 18% of participants. Beyond $k = 7$, however, the sample size declines rapidly (e.g., only 267 participants remain at $k = 9$). These patterns motivated the choice to restrict the primary longitudinal analysis to a window of up to seven visits, which balances sequence coverage with sample size given the distribution of available visits, as well as the total number of visits and timespan per patient as shown above. Therefore, the maximum sequence length was set to T=7 visits.

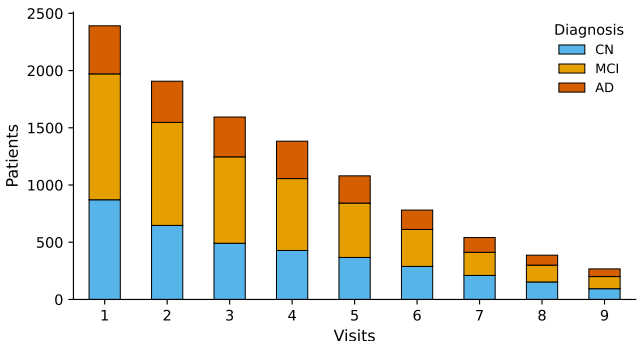


Figure 5.4: Diagnosis distribution per visit

Table 5.3: Class distribution at the k-th visit

k	**AD**	**CN**	**MCI**	**Total**	**AD (%)**	**CN (%)**	**MCI (%)**
1	420	871	1100	2391	17.6	36.4	46.0
2	360	647	900	1907	18.9	33.9	47.2
3	348	491	755	1594	21.8	30.8	47.4
4	327	428	628	1383	23.6	30.9	45.4
5	238	367	475	1080	22.0	33.9	43.9
6	169	289	323	781	21.6	37.0	41.4
7	129	210	202	541	23.8	38.8	37.3
8	87	153	147	387	22.5	39.5	37.9
9	67	94	106	267	25.1	35.2	39.7
10	41	54	70	165	24.8	32.7	42.4
11	31	43	36	110	28.2	39.1	32.7
12	16	26	23	65	24.6	40.0	35.4
13	11	14	10	35	31.4	40.0	28.6
14	4	8	5	17	23.5	47.1	29.4
15	1	3	1	5	20.0	60.0	20.0
16	0	1	0	1	0.0	100.0	0.0
17	0	1	0	1	0.0	100.0	0.0

For longitudinal modeling, the dataset was organized *per patient* as a fixed-length sequence of visits. Let i index patients and t index visits in chronological order from baseline. For each patient i we retained the first K visits, where $K = 7$. Patients with fewer than K visits were right-padded, and a padding mask was applied to ensure that padded time steps did not contribute to loss or metrics.

At each visit t, the input feature vector included the subject's *age at visit* and a set of *brain-region volumes (ROIs)* extracted from MRI (145 ROIs in our setup). The corresponding *diagnosis label* at visit t was kept in its original 3-class form: cognitively normal (CN), mild cognitive impairment (MCI), or Alzheimer's disease (AD). Labels were encoded as integers for training (CN=0, MCI=1, AD=2) and downstream task definitions (e.g., conversion vs non-conversion) were derived from these labels as described below.

Therefore, for each visit (i, t) we assemble the input feature vector $x_{i,t}$ as follows:

- Structural MRI ROI volumes: 145 regional brain volumes extracted after MRI preprocessing (see below).
- Age at visit

Aside from the exclusions discussed at the start of this section, the remaining ROI features were treated as observed, since missingness at the feature level was negligible in our full cohort subset.

Subsequently, a linear covariates adjustment was applied to the numerical ROI data in order to reduce the effect of age, sex and DLICV in our data. Since brain anatomy and the volumes of white and gray matter regions vary substantially with age and sex, linear correction helped retain only Alzheimer's-related neuroanatomical differences in ROI volumes.

For this purpose, a linear regression model was trained using age, sex and DLICV from 449 CN individuals as predictors, with the brain VOIs as output. The model was fitted on these predictors and applied on the entire dataset, to obtain predicted ROI values for each individual. The predicted value for each ROI was subtracted from the observed value, yielding residuals with the effects of age, sex, and DLICV removed [37].

Z-score standardization was applied to age and each ROI volume for all the individuals, using the following formula:

$$z\_\text{score}_i = \frac{x_i - \mu_{\text{CN}}}{\sigma_{\text{CN}}} \tag{5.1}$$

where x_i is the feature's value for which z-score is calculated, μ_{CN} is the mean of the CN individuals for that same feature and σ_{CN} is the standard deviation of CN individuals for that same feature.

In summary, each patient is represented as a length-K sequence of standardized continuous features (age + ROI volumes) with an accompanying per-visit diagnosis label in {CN, MCI, AD} and explicit masks for padding and eligibility. Finally, the data was split into training/validation and test set with a ratio of 0.8 and 0.2 respectively.

Conversion Cohort

The conversion analysis is framed as a *next-visit progression* task from mild cognitive impairment (MCI) to Alzheimer's disease (AD). Cases diagnosed as cognitively normal (CN) do not contribute for this specific task, and subjects diagnosed with AD at baseline cannot "convert" by definition. Moreover, single-visit subjects offer no next-visit signal. Accordingly, we construct a conversion-specific subset (the *conversion cohort*) via the following exclusions:

- Exclude patients with only one visit (no next-visit labels can be defined).
- Exclude patients diagnosed with AD at their first (baseline) visit.
- Exclude all CN diagnoses and CN-stable trajectories (conversion is defined only within the MCI spectrum).

After applying these criteria, 980 patients remain with a total of 3867 examinations; among them, 340 are converter subjects (i.e., they experience MCI→AD at some point during follow-up), yielding a binary next-visit task: MCI→AD (positive) versus MCI→MCI (negative).

Table 5.4: Conversion cohort summary statistics.

Statistic	**Count**
Total patients	980
Total examinations	3867
MCI-stable	640
AD-converters	340

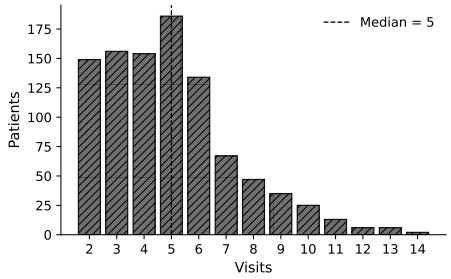


Figure 5.5: Number of visits per patient

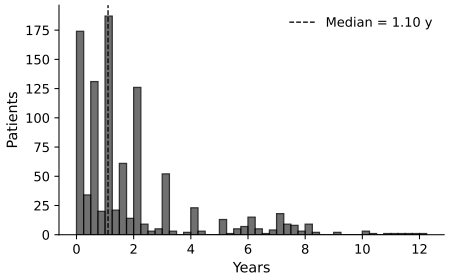


Figure 5.6: Total timespan per patient.

We organize the longitudinal data per patient i as an ordered sequence of visits $t = 0, 1, \ldots, T_i - 1$. Let $y_{i,t} \in \{\text{CN}, \text{MCI}, \text{AD}\}$ denote the clinical diagnosis at visit t. The conversion task is cast as a *next-visit* binary prediction conditioned on the current state being MCI:

$$\text{eligible at } (i,t) \iff y_{i,t} = \text{MCI and } t+1 < T_i.$$

For each eligible pair (i,t) we define the binary label

$$z_{i,t} = \begin{cases} 1, & \text{if } y_{i,t} = \text{MCI and } y_{i,t+1} = \text{AD}, \\ 0, & \text{if } y_{i,t} = \text{MCI and } y_{i,t+1} = \text{MCI}. \end{cases} \tag{5.2}$$

By construction, last visits $t = T_i - 1$ are *never* eligible (no $t+1$ label exists) and are removed prior to modeling. Post-conversion visits with $y_{i,t} = \text{AD}$ are not eligible either because the conditioning $y_{i,t} = \text{MCI}$ fails. In practice, each converter contributes (at most) one positive event ($z_{i,t} = 1$), whereas non-converters may contribute multiple negative MCI→MCI events ($z_{i,t} = 0$).

For each visit (i,t) we assemble the input feature vector $x_{i,t}$ as follows:

- Structural MRI ROI volumes: 145 regional brain volumes extracted after MRI preprocessing (see below).

- Age at visit
- Time interval Δt to next visit: Concatenated or embedded as appropriate.

We represent each patient with a fixed maximum number of visits K as in the diagnosis task($K = 7$). During training and evaluation, losses and metrics are computed only over eligible, non-padded time steps. Original multiclass diagnoses are retained as integers $\{\text{CN} = 0,\ \text{MCI} = 1,\ \text{AD} = 2\}$ at the sequence level. For the conversion task we derive a *binary* target at eligible steps, $z_{i,t} \in \{0, 1\}$, using the rule described above. Linear correction was applied to remove the linear effects of confounding variables (Age, Sex, and baseline DLICV) from ROI features as in the diagnosis cohort. For each ROI, a linear model was fitted using these covariates on the training set, and the residuals were used as corrected feature values. The last step of preprocessing is data standardization, for which we calculate the z-score standardization for all the inviduals, as in the full cohort. At the patient level, the share of converters is $340/980 \approx 34.7\%$. Labels are defined via a one-step shift over eligible MCI visits, with last visits removed. The data was split into training/validation and test set with a ratio of 0.8 and 0.2 respectively.

Table 5.5: Next-visit diagnosis distribution at the k-th visit in the *conversion cohort*

k	**AD**	**MCI**	**Total at k**	**AD ratio**	**MCI ratio**
1	36	944	980	0.036735	0.963265
2	62	769	831	0.074609	0.925391
3	66	609	675	0.097778	0.902222
4	80	441	521	0.153551	0.846449
5	44	291	335	0.131343	0.868657
6	27	174	201	0.134328	0.865672
7	15	119	134	0.111940	0.888060
8	7	80	87	0.080460	0.919540
9	4	48	52	0.076923	0.923077

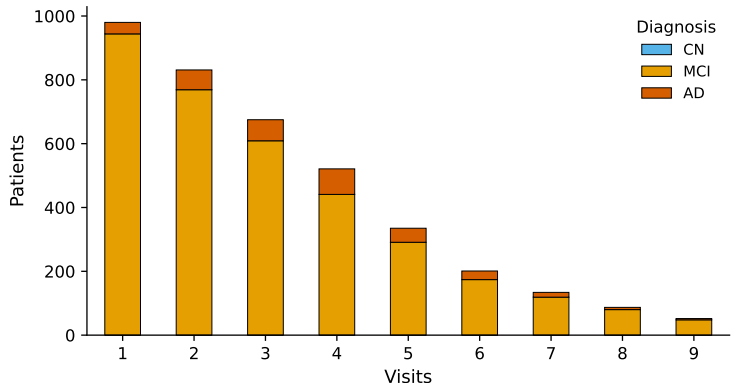


Figure 5.7: Next-visit diagnosis distribution

5.1.2 Cross-Sectional ROIs & SNPs Dataset

The dataset (1463 individuals) used for the subgroup analysis process, is also derived from ADNI, and includes cross-sectional data and SNPs for each patient. Among them, 449 participants were cognitively normal (CN), 740 were diagnosed with mild cognitive impairment (MCI), and 274 were diagnosed with Alzheimer's disease (AD). Paricipant ages ranged from 60 to 86 years.

The dataset contains demographic, clinical and genetic features. The demographic features are age and sex, the clinical features have been obtained from T1 MRI brain images, while genetic features are single nucleotide polymorphisms (SNPs) related to Alzheimer's disease. More specifically, the clinical data are the brain's total volume along with 145 volumes of interest (VOIs), from regions such as the hippocampus and amygdala, while the genetic data consist of 54 SNPs indicating the number of alleles each individual carries from the corresponding SNP/loci.

All the clinical data are numeric (tabular), with continuous numerical values, while sex is a categorical variable encoded as "1" for female and "0" for male. Moreover, the genetic data were standardized and express as discrete numerical values in the range [0, 1], indicating whether an individual has 0, 1 or 2 alleles of the corresponding SNP. Cognitively normal individuals are the 30.69% of the participants, MCI are 50.58% indiviuals and individuals with dementia are 18.73%.

Next, linear correction was applied to the baseline ROI features, similar to the longitudinal dataset. The final preprocessing step was data standardization, performed using z-score standardization across all inviduals. We first carry out longitudinal modeling of Alzheimer's diagnosis and then extend the same framework to conversion prediction by predicting the next-visit diagnosis.

5.2 Longitudinal Modeling using Transformers

5.2.1 Diagnosis Prediction Model

We introduce a Transformer-based model, *Tralzformer*, for longitudinal diagnosis prediction. The model uses self-attention [1] to learn representations across time.

Let the dataset consist of T visits per patient and a set of features $\mathcal{F}_t$ observed at each visit $t \in \{1, \dots, T\}$.

Each feature $f \in \{1, \dots, F\}$ at time $t \in \{1, \dots, T\}$ is represented by a vector:

$$x^{(t,f)} \in \mathbb{R}^{d_f}.$$

Then, each feature is embedded into a shared latent dimension d as

$$z^{(t,f)} = \text{Embed}_f\big(x^{(t,f)}\big) \in \mathbb{R}^d \tag{5.3}$$

where $\text{Embed}(\cdot)$ is a linear projection for numerical features.

To obtain a single representation per visit, the *ROI* feature embeddings present at that visit are averaged:

$$\bar{z}^{(t)} = \frac{1}{|\mathcal{F}_t'|} \sum_{f \in \mathcal{F}_t'} z^{(t,f)} \in \mathbb{R}^d, \qquad \mathcal{F}_t' = \{f \in \mathcal{F}_t : f \text{ is ROI}\}. \tag{5.4}$$

We further incorporate Age at visit t, denoted $e_{\text{age}}^{(t)} \in \mathbb{R}^d$ and a learnable scalar gate $w_{\text{age}} \in \mathbb{R}$:

$$\hat{z}^{(t)} = \bar{z}^{(t)} + w_{\text{age}}\, e_{\text{age}}^{(t)}. \tag{5.5}$$

Temporal information is injected with a sinusoidal positional encoding over the visit index:

$$\tilde{z}^{(t)} = \hat{z}^{(t)} + p^{(t)}, \tag{5.6}$$

where $p^{(t)}$ is the standard sinusoidal encoding.

The collection of all time-step embeddings forms the input sequence to the Transformer encoder:

$$\tilde{Z} = \{\tilde{z}^{(1)}, \dots, \tilde{z}^{(T)}\}.$$

This sequence is processed by a Transformer encoder consisting of L ($L = 1$) stacked layers of multi-head self-attention:

$$H^{(l)} = \text{TransformerLayer}(H^{(l-1)}), \quad l = 1, \dots, L, \quad H^{(0)} = \tilde{Z} \tag{5.7}$$

The encoder outputs a representation:

$$H = \{h^{(1)}, h^{(2)}, \dots, h^{(T)}\} \in \mathbb{R}^{T \times d}.$$

Finally, each prediction target is associated with its corresponding time-step representation, which is passed through a classification head (3 classes):

$$y^{(t)} = \mathrm{Head}^{(t)}\big(h^{(t)}\big), \quad \mathrm{Head}^{(t)} : \mathbb{R}^d \to \mathbb{R}^3,$$

followed by a softmax at evaluation. Predictions are computed only on valid (non-padded) time steps according to the mask.

The model is trained using the Adam optimizer, with the d dimension equal to 128.

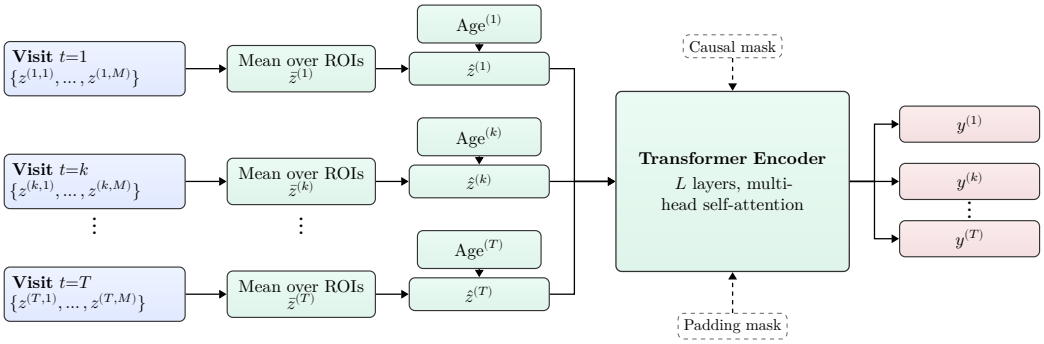


Figure 5.8: Diagnosis model: per-visit mean over feature embeddings, residual addition of Age$^{(t)}$, sinusoidal positional encoding, causal Transformer, and per-time classification heads.

5.2.2 Conversion Prediction Model

To model the probability of clinical conversion at future visits, we developed a longitudinal Transformer-based architecture, following the approach used for diagnosis prediction. The model was designed to capture temporal dependencies in volumetric patient features and to predict, at each time step, the likelihood of conversion at the subsequent visit.

The input consisted of sequential features per patient, with each time step corresponding to a clinical visit. Each feature type was first embedded into a fixed-dimensional representation using a linear layer with batch normalization, as in the diagnosis prediction setup.

To incorporate temporal context, sinusoidal positional encodings were applied to ensure that the model could distinguish the order of visits within each sequence. The sequence of embeddings was then processed by an encoder consisting of a layer with four attention heads. On top of the Transformer representations, a shared binary classification head (linear layer with output dimension 1) was applied across all time steps to predict conversion at the *next* visit. The model outputs a vector of logits with length T for each sample. This architecture enabled patient-level, per-visit prediction of conversion risk while explicitly modeling temporal dynamics and enforcing causality.

As seen before, follow-up intervals are irregular across patients. To address this, the model incorporates (i) *age* as a covariate, (ii) a sinusoidal positional encoding and (iii) the *time interval* Δt between each visit. Thus, to represent temporal information, two types of embeddings are added:

$$\tilde{z}^{(t)} = \bar{z}^{(t)} + p^{(t)} + \phi(\Delta t^{(t)}), \tag{5.8}$$

where $p^{(t)}$ is a sinusoidal positional encoding for the time index, and $\phi(\Delta t^{(t)})$ is a learned projection of the time gap (in months) to the next visit.

As noted in data overview, an important design choice in defining the conversion prediction task concerns which diagnostic groups are included as eligible for conversion. Following common practice in the literature, we restricted the conversion cohort to individuals diagnosed with mild cognitive impairment (MCI) at baseline or during follow-up, and defined conversion as the transition from MCI to Alzheimer's disease (AD) at the next visit. This definition ensures that the model estimates the risk of imminent MCI-to-AD progression, which aligns with clinical interest in identifying subjects at the highest risk of developing dementia.

5.3 Training Setup & Evaluation Metrics

The Transformer model was implemented with a hidden dimensionality of 128 ($d_{\text{model}} = 128$) and four attention heads ($n_{\text{head}} = 4$). Training was carried out for 64 epochs using a batch size of 64 and binary cross-entropy loss. Optimization was performed with the Adam optimizer, with a learning rate of 5×10^{-4} and weight decay set to 1×10^{-5}. To address class imbalance, we employed focal loss and performance was evaluated at each time step using the area under the receiver operating characteristic curve (AUC-ROC) and the area under the precision–recall curve (AUPR). These metrics are threshold-independent and thus provide a robust evaluation of the model's ability. The model was trained and validated using 5-fold cross-validation, and final performance was assessed on a held-out test set.

As stated before, preprocessing was tailored per task. For the diagnosis prediction, two binary subproblems were studied, the AD vs MCI/CN, which is the most clinically relevant problem, as the diagnostic uncertainty is higher, and the AD vs CN, which is biologically clearer. For the conversion prediction, the task examined was the transition from MCI to AD. For each task, results are presented using the linearly corrected features to minimize demographic and volumetric effects, as isolating disease-specific effects was of primary interest. Results with the alternative pipeline, using the raw ROI data, are provided in the Appendix.

5.4 Subgroup Analysis

5.4.1 Representation Learning

In this part of the experiment, a Multilayer Perceptron was used to extract meaningful embeddings from the input data. The goal of the embedding model was to compress heterogeneous, high-dimensional brain imaging and genetic features into a lower-dimensional latent space that preserves discriminative information relevant to disease status.

The embeddings were optimized for Controls vs non-Controls separation (CN vs Not-CN). The total loss was defined as the sum of binary cross-entropy for the CN vs Not-CN classification task and a contrastive loss term that pulled CN samples closer together in the embedding space while pushing them away from Not-CN samples. This ensured that the latent representations were not only predictive but also structured in a way that highlights the separation between disease and control groups.

Let $x_i \in \mathbb{R}^d$ denote the feature vector of the i-th sample. The MLP consists of L fully connected layers, each followed by a non-linear activation function (ReLU) and optional batch normalization:

$$h^{(0)} = x_i, \quad h^{(l)} = \text{ReLU}(W^{(l)} h^{(l-1)} + b^{(l)}), \quad l = 1, \ldots, L \tag{5.9}$$

The final hidden layer produces the embedding vector $z_i \in \mathbb{R}^k$:

$$z_i = h^{(L)}$$

These embeddings aim to capture the most informative representation of the input data, preserving similarity relationships between samples. The MLP was trained on 1463 samples, using 5-fold stratified cross-validation. To qualitatively assess separation between clusters, the embeddings were projected into two dimensions using UMAP and t-SNE, two widely used non-linear dimensionality reduction techniques that preserve local neighborhood structure. These visualizations provide an intuitive view of class separation in the latent space.

5.4.2 Clustering Algorithm

After obtaining the embeddings z_i from the MLP, clustering was performed to identify groups of similar samples in the laten space. Operating in the embedding space, rather than directly on the raw features, allows subjects to be grouped according to higher-level representations learned by the model, which may better capture disease-related structure.

The primary algorithm employed was k-Means clustering, with Agglomerative clustering and DBScan also explored for comparison. The k-Means algorithm partitions the embedding space into K clusters by minimizing the within-cluster sum of squares:

$$\mathcal{L}_{\text{k-Means}} = \sum_{j=1}^{K} \sum_{z_i \in C_j} \|z_i - \mu_j\|^2 \tag{5.10}$$

where C_j is the set of embeddings assigned to cluster j and μ_j is the centroid of cluster j.

Clustering quality was evaluated using the Silhouette Score, the Davies-Bouldin Index, and the Adjusted Rand Index (ARI). Furthermore, feature-level analysis was conducted to characterize the resulting clusters.

Clustering algorithms were tested for $k = 2$ to $k = 5$, with $k = 2$ selected as a trade-off between internal validation metrics and clinical interpretability. To interpret the resulting subgroups, univariate statistical tests were applied to identify discriminative features between clusters. Both parametric (ANOVA) and non-parametric (Kruskal–Wallis) tests were applied across all available brain volumes and SNPs, followed by false discovery rate (FDR) correction to account for multiple comparisons.

5.5 Interpretability Methods

To better understand the decision-making process of our models, we employed attribution and subgroup-based interpretability analyses.

Integrated Gradients

For interpretability, we employed Integrated Gradients (IG) [31], a widely used attribution method for differentiable models, to quantify the contribution of each input feature to the model predictions. IG attributes importance scores by integrating the gradients of the output with respect to the input along a linear path from a baseline reference to the actual input and is computationally efficient in deep architectures such as Transformers. This method satisfies desirable axioms such as sensitivity and implementation invariance, and has been widely adopted in biomedical machine learning applications. For each prediction, we computed IG scores and aggregated them at the feature and region levels to highlight the most influential biomarkers.

Alternative methods such as SHAP [32], approximate Shapley values and are popular in clinical machine learning, but are computationally more demanding and less straightforward to apply in sequential neural networks. We therefore selected IG as a suitable and well-established choice for attribution in this setting.

Subgroup Evaluation

In addition to feature-level attributions, we assessed model performance within patient subgroups. These subgroups were obtained by clustering patients in the

learned representation space, yielding clinically meaningful strata with mixed CN, MCI, and AD profiles. Each model was evaluated separately across these subgroups to examine whether predictive performance was consistent across different patient populations, and to identify potential biases or failure modes.

Chapter 6

Results

This chapter presents the results obtained from the above methodology. First, results and comparisons with baselines and current literature are presented for the current time-step diagnosis prediction task. Next, results are presented in a similar manner about the next time-step conversion prediction task. Then both prediction task results are interpreted, and finally the clustering process is evaluated and combined with the models produced by the experiments.

6.1 Diagnosis and Conversion Prediction Results

6.1.1 Current-Visit Diagnosis Prediction

AD vs MCI/CN Problem

As outlined in previous sections, the objective of this experiment is to examine and interpret the progression of Alzheimer's disease over time, emphasizing the importance of longitudinal modeling and temporal dependency over purely cross-sectional analysis. The Transformer is first evaluated on *current-visit* diagnosis, framed as a binary classification task between AD and MCI/CN subjects. For each time step $t \in \{0, \dots, 6\}$, the model processes all available visits up to t and predicts the probability of AD at the same visit t.

Table 6.1: Per-visit AUC metrics for AD vs MCI/CN

Visit	AUC (ROC)	AUC (PR)
0	0.866	0.622
1	0.855	0.636
2	0.844	0.600
3	0.866	0.618
4	0.836	0.548
5	0.802	0.452
6	0.802	0.548
Mean	**0.853**	**0.575**

Table 6.2: Per-visit metrics (AD vs MCI/CN)

Visit	Confusion Matrix	Acc.	Bal. Acc.	Prec.	Recall	Spec.	F1	MCC	AUC (ROC)	AUC (PR)
0	[[385, 11], [58, 26]]	0.86	0.64	0.71	0.32	0.97	0.44	0.41	0.87	0.62
1	[[301, 9], [46, 23]]	0.86	0.67	0.68	0.36	0.97	0.46	0.42	0.86	0.64
2	[[241, 13], [42, 23]]	0.82	0.65	0.63	0.35	0.95	0.45	0.38	0.84	0.59
3	[[208, 16], [28, 28]]	0.85	0.72	0.64	0.51	0.93	0.56	0.48	0.87	0.62
4	[[163, 12], [26, 13]]	0.82	0.64	0.54	0.34	0.93	0.42	0.33	0.84	0.53
5	[[118, 11], [19, 8]]	0.81	0.61	0.44	0.31	0.91	0.36	0.26	0.80	0.45
6	[[79, 7], [15, 11]]	0.80	0.66	0.60	0.40	0.92	0.49	0.38	0.80	0.55

Overall, the results indicate stable performance across visits, with mean AUC of 0.85 and 0.58. The model performs well already at baseline, indicating that the available features are highly informative for distinguishing diagnostic categories. Table 6.1 summarizes the predictive performance of the model across consecutive visits, reported in terms of AUC (ROC) and AUC (PR) from the results obtained on the held-out test set.

Table 6.3: Comparison of mean AUC (ROC) and AUC (PR) across visits between our Transformer model and benchmark models

Model	Mean AUC (ROC)	Mean AUC (PR)
Transformer	0.853	0.575
Random Forest	0.820	0.561
Logistic Regression	0.777	0.517

Table 6.4: Per visit comparison of AUC (ROC) and AUC (PR) across models.

Visit	Transformer		Logistic Regression		Random Forest	
	AUC (ROC)	**AUC (PR)**	**AUC (ROC)**	**AUC (PR)**	**AUC (ROC)**	**AUC (PR)**
0	0.866	0.622	0.858	0.600	0.862	0.557
1	0.855	0.636	0.774	0.454	0.839	0.585
2	0.844	0.600	0.761	0.459	0.829	0.552
3	0.866	0.618	0.778	0.548	0.837	0.612
4	0.836	0.548	0.764	0.427	0.817	0.539
5	0.802	0.452	0.752	0.417	0.792	0.513
6	0.802	0.548	0.747	0.418	0.783	0.540

When contrasting the Transformer model with two classical models, logistic regression and random forest, the Transformer consistently achieves the highest AUC ROC, with values reach 0.87, while the benchmark models generally remain in the 0.74-0.87 range. These results indicate the transformer is learning temporal progression patterns and longitudinal information that LR/RF cannot capture. Overall, the transformer outperforms both logistic regression and random forest, while the consistency across time suggests that the Tralzformer provides a meaningful advantage in modeling Alzheimer's disease progression.

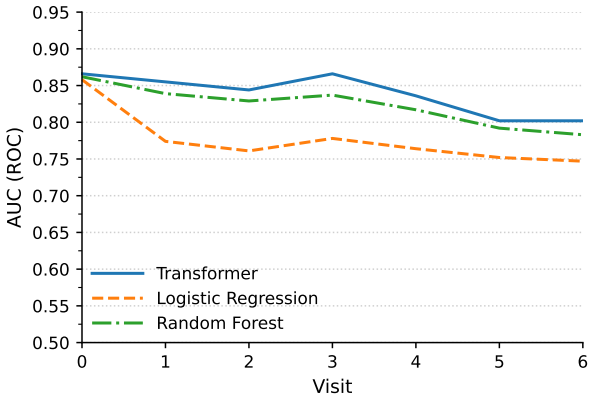


Figure 6.1: AUC (ROC) across models

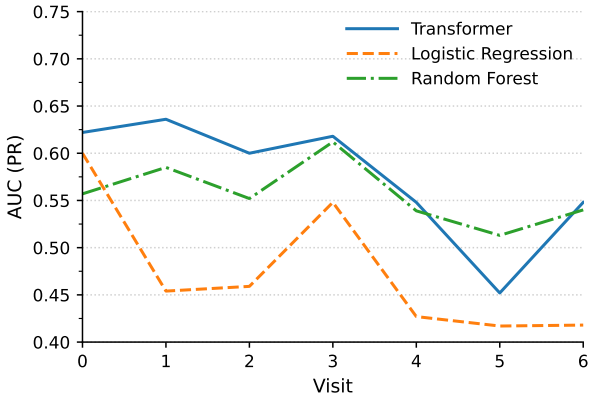


Figure 6.2: AUC PR across models

As most of current literature focuses on the AD vs CN task, comparison was left to those benchmark models. Future work could include comparison with time-aware models such as RNNs. Nonetheless, AD vs MCI/CN defines a more realistic approach to clinical diagnosis, therefore performance is expected to be lower here.

AD vs CN Problem

Even though the AD vs (CN/MCI) setup is closer to real clinical progression, AD vs CN was also tested, as the AD vs CN problem is biologically clearer and usually easier.

Table 6.5: Per time step AUC metrics (AD vs CN)

Visit	AUC (ROC)	AUC (PR)
0	0.930	0.912
1	0.932	0.920
2	0.920	0.910
3	0.940	0.926
4	0.910	0.858
5	0.902	0.818
6	0.876	0.820
Mean	**0.916**	**0.879**

Table 6.6: Per time step metrics (AD vs CN)

Visit	Confusion Matrix	Acc.	Bal. Acc.	Prec.	Recall	Spec.	F1	MCC	AUC (ROC)	AUC (PR)
0	[[172, 5], [16, 68]]	0.92	0.89	0.93	0.81	0.97	0.86	0.81	0.93	0.91
1	[[123, 3], [15, 53]]	0.91	0.88	0.95	0.78	0.98	0.86	0.80	0.93	0.92
2	[[92, 6], [11, 55]]	0.89	0.88	0.90	0.83	0.94	0.86	0.78	0.92	0.91
3	[[82, 7], [8, 48]]	0.90	0.89	0.88	0.85	0.93	0.86	0.78	0.94	0.93
4	[[69, 8], [8, 31]]	0.86	0.85	0.80	0.79	0.90	0.79	0.69	0.91	0.86
5	[[56, 6], [8, 19]]	0.85	0.81	0.77	0.72	0.90	0.74	0.63	0.90	0.82
6	[[42, 8], [8, 18]]	0.80	0.78	0.71	0.71	0.85	0.71	0.56	0.87	0.82

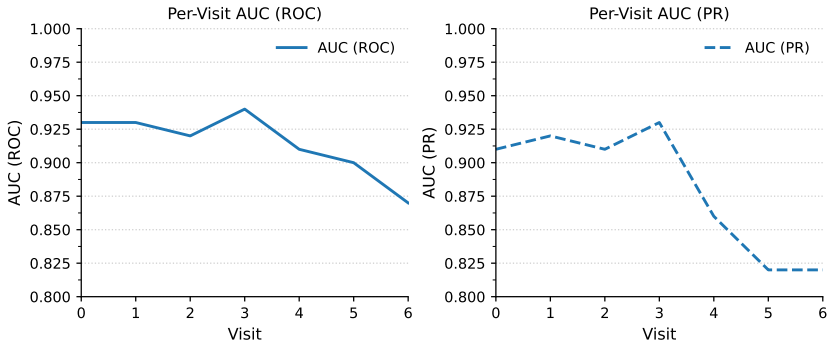


Figure 6.3: AUC ROC and AUC PR for AD vs CN

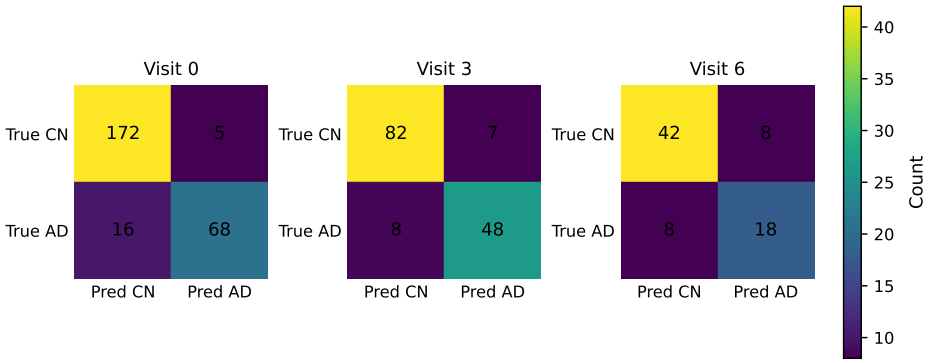


Figure 6.4: AD vs CN Confusion Matrices for 1, 4 and 7 visits

The confusion matrices show that, under a fixed classification threshold of 0.5, the model maintains a consistent balance between sensitivity and specificity across visits (Figure **??**). The AUC (ROC) and AUC (PR) metrics provide a more comprehensive view of performance independent of the chosen threshold. The results yielded in this setup show better performance than in the AD vs MCI/CN setup, with consistently high performance across visits, mantaining AUC (ROC) values above 0.9 and achieving a mean of 91.6% (Table **??**). This indicates that the model effectively distinguishes AD from CN subjects Accuracy and balanced

accuracy remain consistently high ($\geq$0.80), although a moderate decline in recall is observed at later visits, reflecting the reduced number of available samples. Overall, these results indicate that the model retains reliable diagnostic performance and generalizes well under the linearly corrected feature representation.

As stated in the previous chapter, to account for potential demographic and volumetric influences, ROI features were linearly corrected for Age, Sex, and DLICV prior to model training. The linearly corrected features produced results comparable to, and in several cases slightly more stable than, those obtained with uncorrected data across time steps and folds. This stability suggests that the correction effectively reduced demographic and structural bias, allowing the model to focus on disease-related variability while maintaining strong predictive performance. Consequently, the linearly corrected representation was adopted as the main configuration for the analysis. However, the results on the unprocessed data, reported in the Appendix, serve as a complementary comparison.

6.1.2 Next-Visit Conversion Prediction (MCI→AD)

Finally, the model was evaluated on the task of predicting conversion from MCI to AD at the subsequent visit. Performance was considerably lower than for current-visit diagnosis, reflecting the increased difficulty of forecasting disease progression and the limited sample size of converters available at each time step. While current-visit classification benefits from direct access to the diagnostic state, next-visit prediction requires extrapolation into the future, making the task inherently harder and more sensitive to noise and imbalance. Using the held-out test set across seven visits, the model achieved an average AUC of 0.724 (ROC) and 0.326 (PR), as summarized in Table 6.7. Other threshold-dependent metrics such as precision and F1 score were unstable due to class imbalance, with a small number of positive cases per fold. These results suggest that while the framework can capture early signals of conversion, robust prediction of future progression will likely require larger cohorts and additional modalities.

Table 6.7: Per time step AUC metrics (Conversion Prediction)

Visit	**AUC (ROC)**	**AUC (PR)**
0	0.774	0.186
1	0.770	0.258
2	0.578	0.176
3	0.808	0.450
4	0.674	0.262
5	0.630	0.324
6	0.838	0.628
Mean	**0.724**	**0.326**

6.2 Interpretability Results

6.2.1 Feature Attributions

Diagnosis Prediction

To identify important regions, features were ranked by their frequency of appearance among the top attributions across subjects, rather than by mean absolute IG alone. Age was consistently among the most infuential features, which aligns with clinical knowledge. For interpretability, we therefore focus on features beyond age.

Hippocampus and Amygdala regions seem highly important in both experiments. In the early stages of AD [46], the hippocampus shows rapid loss of its tissue, which is associated with the functional disconnection with other parts of the brain. In the progression of AD, atrophy of medial temporal and hippocampal regions are the structural markers in magnetic resonance imaging (MRI). This proves that our model is in line with neuropathology, since the hippocampus is a stable biomarker in the model.

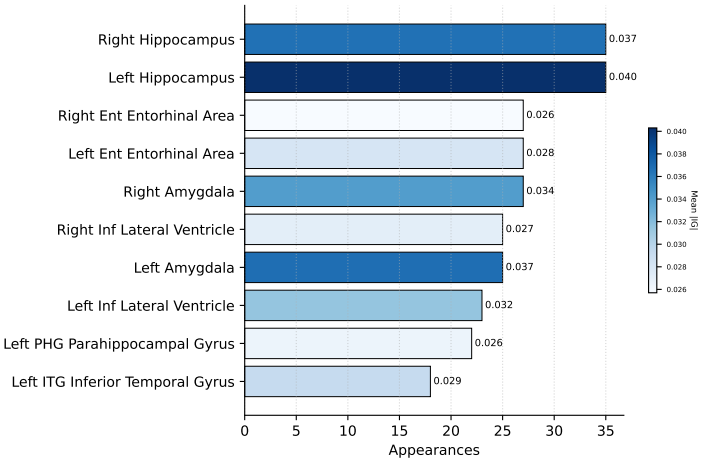


Figure 6.5: AD vs MCI/CN most frequent ROIs

Most important regions in the AD vs MCI/CN setup:

- Hippocampus
- Amygdala
- Entorhinal Area
- Lateral Ventricles

Findings also suggest that the magnitude of amygdala atrophy [47] - related to aberrant motor behavior and irritability - is comparable to that of the hippocampus in the earliest clinical stages of AD. Furthermore, both left and right entorhinal areas also appear prominently in both lists. This matches expectations [48], since the entorhinal cortex is the brain region that often exhibits the earliest histological alterations in Alzheimer's disease, including the formation of neurofibrillary tangles and cell death.

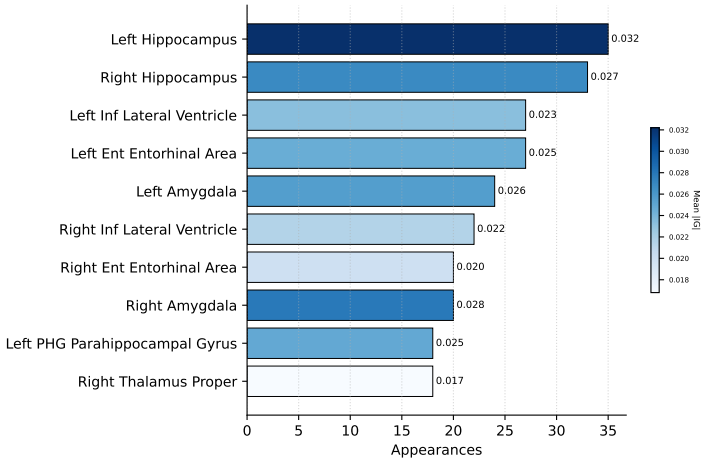


Figure 6.6: AD vs CN most frequent ROIs

Most important regions in the AD vs CN setup:

- Hippocampus
- Entorhinal Area
- Amygdala
- Lateral Ventricles
- Parahippocampal Gyrus

In addition to that, before applying linear correction, the lateral ventricles appeared among the most important features in the interpretability analysis for AD vs MCI/CN, reflecting their strong correlation with overall brain atrophy and age. However, after correction for Age, Sex, and DLICV, their relative importance decreased, while more disease-specific regions such as the thalamus and hippocampus became more prominent. This shift suggests that linear correction reduced the influence of global volumetric confounds, highlighting structural alterations more directly associated with Alzheimer's pathology. The correspondent interpretation tables can be found in the appendix.

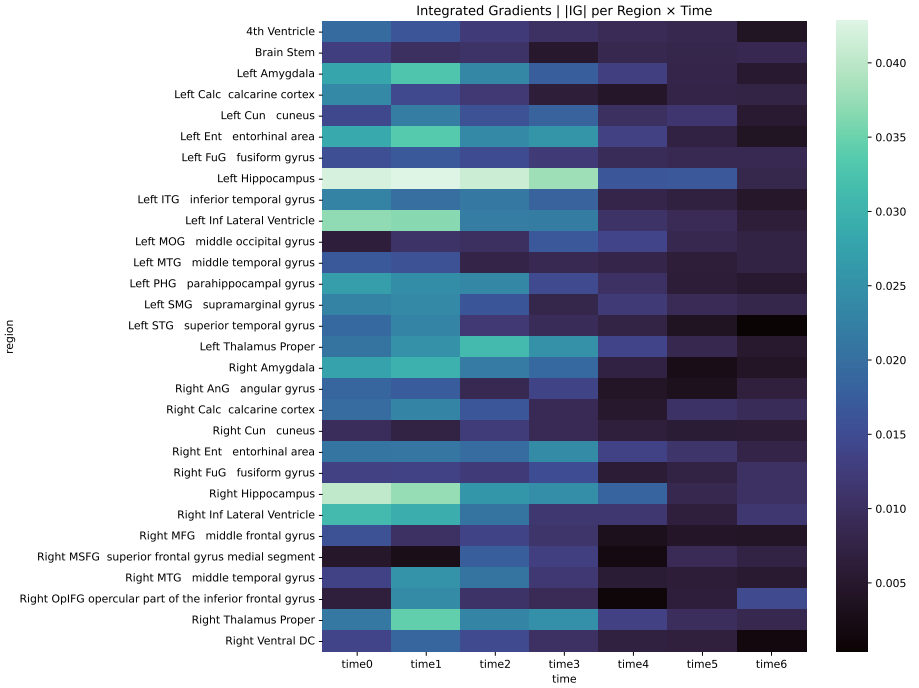


Figure 6.7: AD vs CN ROIs heatmap per visit

Conversion Prediction

In conversion prediction the integrated gradients methodology yields results comparable to those of the diagnosis, even though focus is on areas not found on the diagnosis task attributions. As in diagnosis prediction, hippocampal atrophy is among the most consistent predictors of conversion from MCI to AD. Posterior cingulate gyrus (PCgG) is a core element in the default mode network (DMN). Hypometabolism and structural changes here are consistently linked to prodromal AD. As for inferior lateral ventricles (both left and right), ventricular enlargement is a marker of adjacent tissue loss and often appears in progression studies. On the other hand, Fusiform gyrus (FuG) is a temporal lobe structure, involved in higher-level visual processing, which could indicate association with AD caused atrophy. Same goes for Posterior insula (both left and right). Middle Temporal Gyrus (MTG) and Middle Occipital Gyrus (MOG) are often associated with earlier stages of AD and executive dysfunction [49], [50].

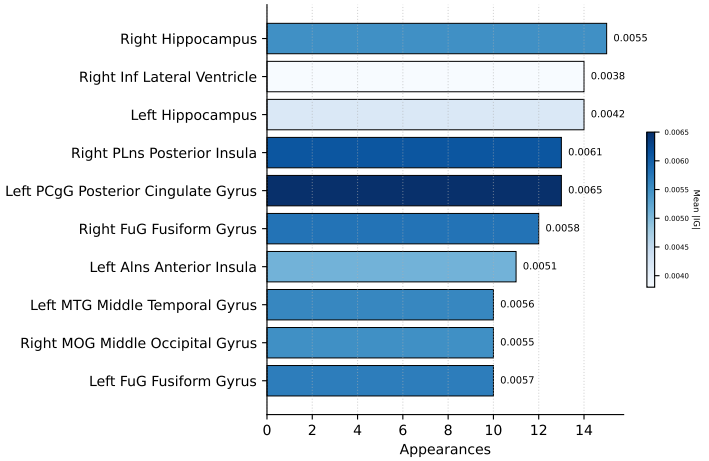


Figure 6.8: MCI to AD most frequent ROIs

Table 6.8: Top 10 regions by mean |IG| across folds and time

Region	**Mean |IG|**	**Appearances**
Right Hippocampus	0.0055	15
Left Hippocampus	0.0042	14
Right Inf Lateral Ventricle	0.0038	14
Left PCgG Posterior Cingulate Gyrus	0.0065	13
Right PLns Posterior Insula	0.0061	13
Right FuG Fusiform Gyrus	0.0058	12
Left Alns Anterior Insula	0.0051	11
Left FuG Fusiform Gyrus	0.0057	10
Right MOG Middle Occipital Gyrus	0.0055	10
Left MTG Middle Temporal Gyrus	0.0056	10

Most important regions in the MCI to AD conversion prediction setup:

- Hippocampus
- Posterior Cingulate Gyrus (PcgG)
- Posterior Insula
- Fusiform Gyrus
- Anterior Insula
- Middle Temporal Gyrus
- Middle Occipital Gyrus
- Right Inferior Lateral Ventricle

6.3 Patient Subgroup Analysis

6.3.1 UMAP Visualization of Representations

As noted before, the MLP was trained on 1463 samples, using 5-fold stratified cross-validation, achieving a performance of 0.8237 ROC AUC and 0.9194 AUPR. The high-dimensional embeddings generated from the MLP were projected into two dimensions using UMAP and t-SNE. As shown in Table 6.9, the dataset is imbalanced, which results in the minority class (CN) forming compact subclusters, while the non-CN occupies a more diffuse region. Figure 6.9 presents the resulting projections obtained with both UMAP and t-SNE.

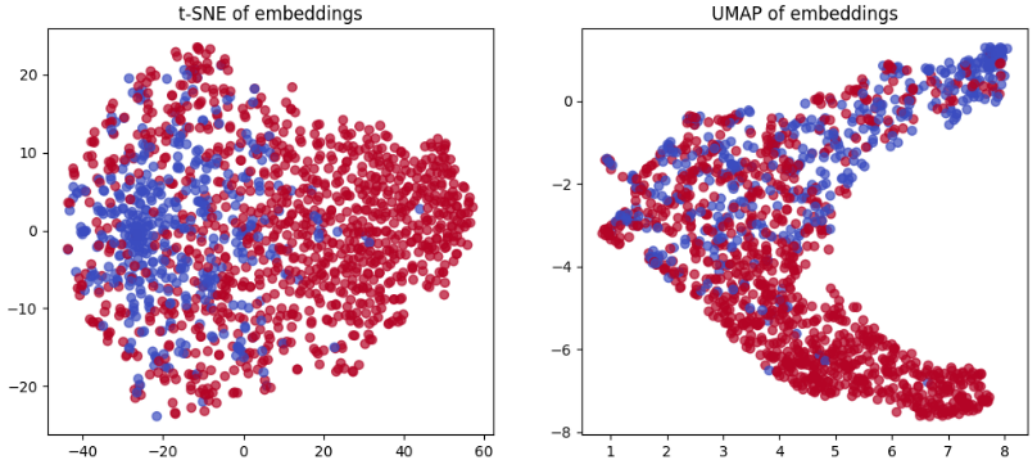


Figure 6.9: t-SNE & UMAP Embeddings' Projections

Table 6.9: Distribution of subjects by class

Class	Percentage
Non-controls	69.31%
Controls	30.69%

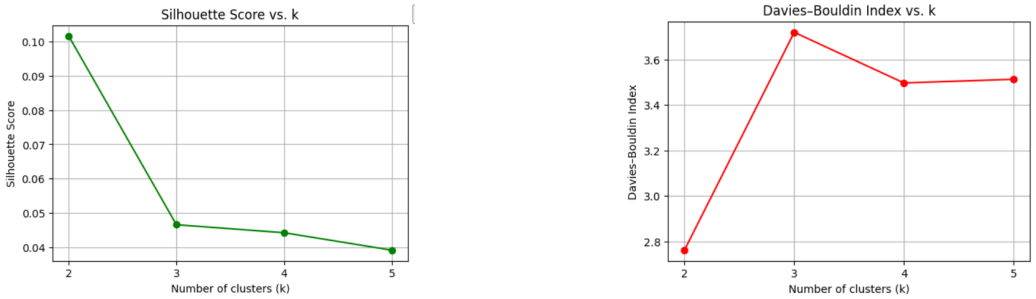
In the t-SNE projection, the blue points (class CN) cluster mostly on the left, fairly localized. The red points (class non-CN) spread widely around the rest of the space, encircling the blue cluster. This suggests the model is able to learn a representation where the minority class forms a distinct "core" cluster, but the majority class (red) occupies a broader, more diffuse area.

In the UMAP projection, the blue points concentrate on the top part of the boomerang shape, which again shows grouping of the minority class in a localized region. The red points occupy the lower parts and are distributed more broadly along the lower sections. This indicates partial grouping of the control individuals,

but also overlap with the non-controls class, reflecting that the classes are not perfectly separable in the embedding space. These patterns are consistent with the 30/70 class imbalance and suggest that while the model learns discriminative features, some overlap remains due to shared features and inherent variability.

6.3.2 Clustering Algorithm & Subgroup Characterization

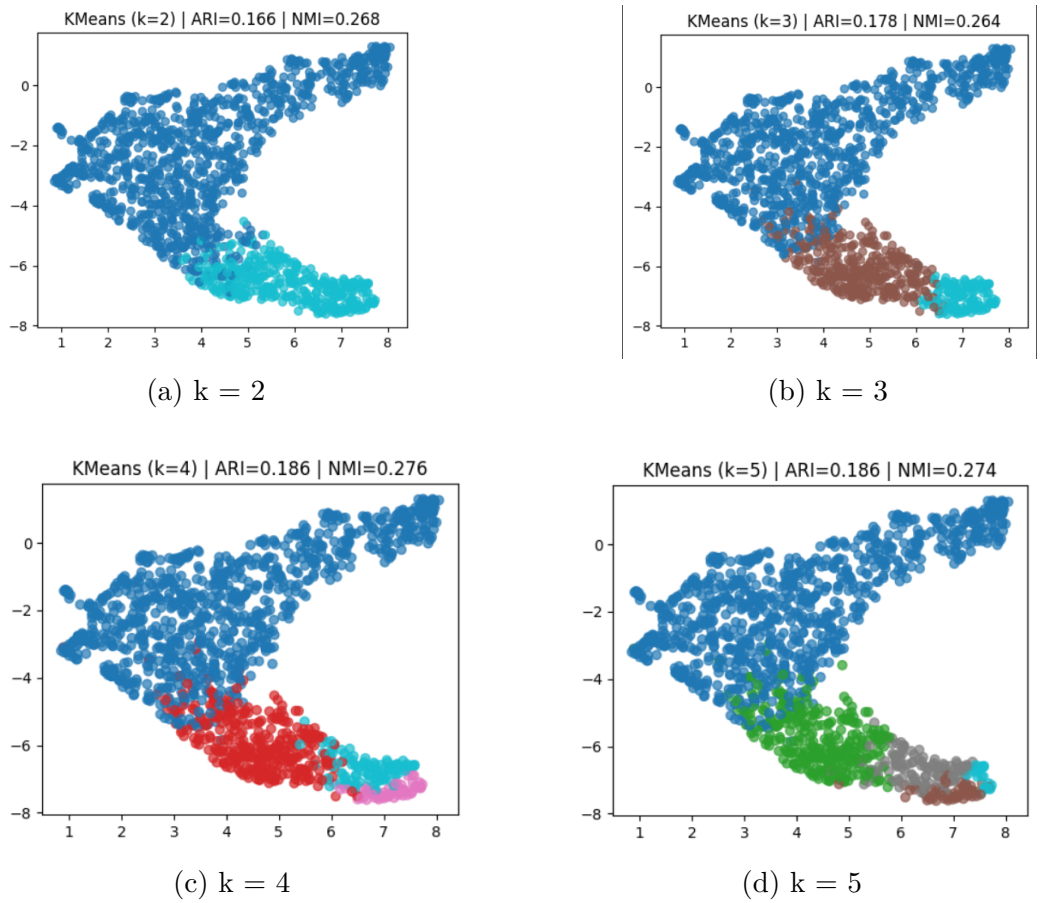
We applied k-Means for the clustering process, first applying a k-sweep from *k=2 to k=5*, as shown in the UMAP projections. Although ARI/NMI scores increase slightly with higher k, the improvement is marginal and remains below 0.3, indicating limited alignment with diagnostic labels across all k values. k=2 was selected because it yields a clear, interpretable division of the cohort into two groups: one enriched for CN individuals and another more heterogeneous group with MCI/AD cases. This aligns with the hypothesis of distinguishing a "healthy" cluster from a "disease-prone" cluster. The decision was also based on the Silhouette Score and Davies-Bouldin Index calculated for k. Furthermore, using k>2 would fragment the dataset into smaller clusters, reducing the power for downstream longitudinal modeling. As for comparison with other benchmark models, the algorithm was compared with Agglomerative, Spectral Clustering and DBScan. Overall, the clustering quality was modest for all methods, reflecting the known overlap between diagnostic categories in Alzheimer's disease. However, the results were consistent across algorithms, with ARI values ranging between 0.14-0.21.



(a) Silhouette Score over k

(b) Davies-Bouldin Index over k

Figure 6.10: Comparison of clustering evaluation metrics over different values of k.



(a) k = 2

(b) k = 3

(c) k = 4

(d) k = 5

Figure 6.11: Comparison of k-Means clustering results for different k values.

Before analyzing the clusters in terms of specific ROIs, it is important to examine their overall composition with respect to clinical diagnosis. Table **??** summarizes the distribution of labels across the two clusters, as well as the overall proportion of cognitively normal (CN) versus non-CN individuals. The results indicate the cluster 0 contains the majority of CN individuals (445 out of 449), but it also shows substantial overlap with MCI (556 individuals) and a number of AD cases (92). Cluster 1 appears to be representative of the MCI/AD group, as its almost exclusively composed of MCI and AD participants, with nearly equal represenation of both. Thus, cluster 0 can be viewed as a ”healthy” cluster with some heterogeneity, while cluster 1 appears to capture a more impaired population. This separation is consistent with expectations, as the MLP embeddings were designed to emphasize differences between CN and non-CN groups. However, the fact that Cluster 0 still includes many MCI individuals highlights the unseparability of the data.

Table 6.10: Diagnosis distribution per cluster

Cluster	Total	CN	MCI	AD
0	1093	445 (40.7%)	556 (50.9%)	92 (8.4%)
1	370	4 (1.1%)	184 (49.7%)	182 (49.2%)

6.3.3 Subgroup Specific Biomarkers

ANOVA was used for feature selection, along with the Kruskal-Wallis test. Notably, hippocampal volume, the amygdala and the inferior lateral ventricles emerged among the most discriminative, consistent with established AD biomarkers. Several volumetric measures (e.g. hippocampus, amygdala, lateral ventricle, entorhinal area) appear among the top-ranked features under both ANOVA and Kruskal–Wallis, suggesting that these regions of interest may still carry signal relevant to group separation.

Table 6.11: ANOVA results (top significant features by p-values)

Feature	p-value (FDR)
Left Hippocampus	5.74×10^{-141}
Left Ent Entorhinal Area	1.86×10^{-126}
Left Amygdala	9.87×10^{-126}
Right Hippocampus	1.26×10^{-119}
Left Inf Lateral Ventricle	1.08×10^{-118}
Right Amygdala	2.87×10^{-107}
Right Inf Lateral Ventricle	1.46×10^{-105}
Left PHG Parahippocampal Gyrus	1.14×10^{-104}
Right Ent Entorhinal Area	8.07×10^{-104}
Right PHG Parahippocampal Gyrus	2.52×10^{-94}

Table 6.12: Kruskal–Wallis results (top significant features by p-values)

Feature	**p-value (FDR)**
Left Hippocampus	4.90×10^{-110}
Left Ent Entorhinal Area	3.94×10^{-97}
Left Amygdala	8.32×10^{-97}
Right Hippocampus	2.36×10^{-95}
Left Inf Lateral Ventricle	1.60×10^{-94}
Left PHG Parahippocampal Gyrus	6.31×10^{-89}
Right Amygdala	2.20×10^{-86}
Right Inf Lateral Ventricle	3.38×10^{-86}
Right Ent Entorhinal Area	3.05×10^{-85}
Right PHG Parahippocampal Gyrus	3.93×10^{-81}

In the high-risk subgroup (Cluster 1), the most elevated features were the lateral ventricles, with the right and left inferior horn of the lateral ventricles showing the strongest increases. Ventricular enlargement is a well-established marker of neurodegeneration in Alzheimer's disease. The lateral ventricles, which are fluid-filled spaces in the brain, tend to enlarge as the surrounding brain tissue shrinks. This ventricular enlargement is a common finding on brain scans like MRI and is a useful marker of disease progression. It's not the ventricles themselves that cause cognitive decline, but their enlargement is closely linked to the atrophy of other brain regions, such as the hippocampus and cortex, which are critical for memory and other cognitive functions. Ventricular enlargement in individuals with mild cognitive impairment (MCI) is also associated with thinner gray matter in the frontal, temporal, and parietal lobes, supporting its role as an indicator of disease progression. Interestingly, the APOE-related SNP *rs429358* also appeared among the top elevated features in Cluster 1. This genetic variant is strongly associated with AD risk and suggests that the high-risk subgroup is chatacterized not only by structural markers of neurodegeneration but also by genetic susceptibility.

The most decreased features in the high-risk subgroup (Cluster 1) include well-established AD biomarkers such as the hippocampus, amygdala and entorhinal areas (both left and right), alongside additional medial temporal lobe structures such as the parahippocampal gyrus and middle temporal gyrus. These findings are consistent with the characteristic pattern of medial temporal atrophy in Alzheimer's disease. In contrast, the control-enriched subgroup (Cluster 0) showed the largest decreases in ventricular volumes, particularly the inferior lateral and lateral ventricles, as well as the third ventricle, similar to the elevated features of the high-risk subgroup. Interestingly, the APOE-related SNP *rs429358* also emerged among the top features, highlighting the impact of its presence on Alzheimer's disease.

Figure 6.12: Top 10 elevated features in each subgroup

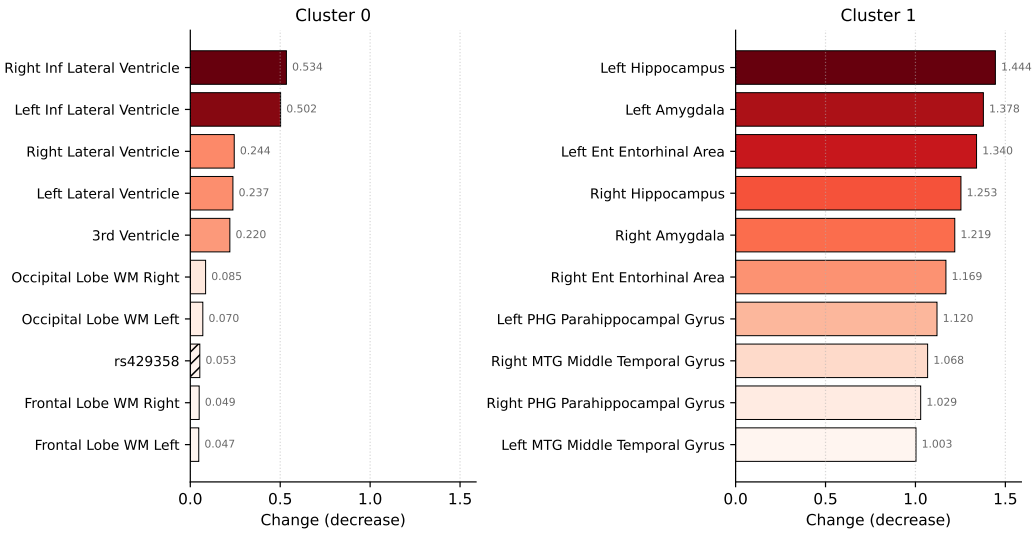


Figure 6.13: Decreased features in each subgroup

The results reveal clear volumetric differences between clusters. Individuals in Cluster 0 display relatively preserved brain volumes, with values close to zero after z-scoring against the CN reference, whereas Cluster 1 exhibits structural deviations: for example, the left hippocampus and right entorhinal area show strong negative shifts, while the left inferior lateral ventricle is notably enlarged. This pattern suggests atrophy in specific cortical and subcortical regions alongside compensatory enlargement elsewhere, which is consistent with known structural alterations in Alzheimer's disease.

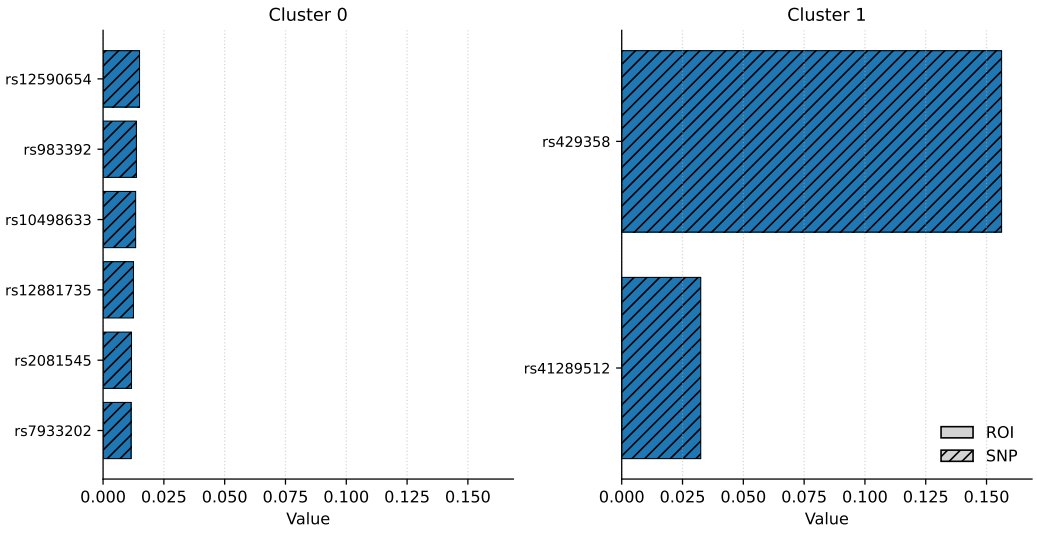


Figure 6.14: Elevated SNPs in each subgroup

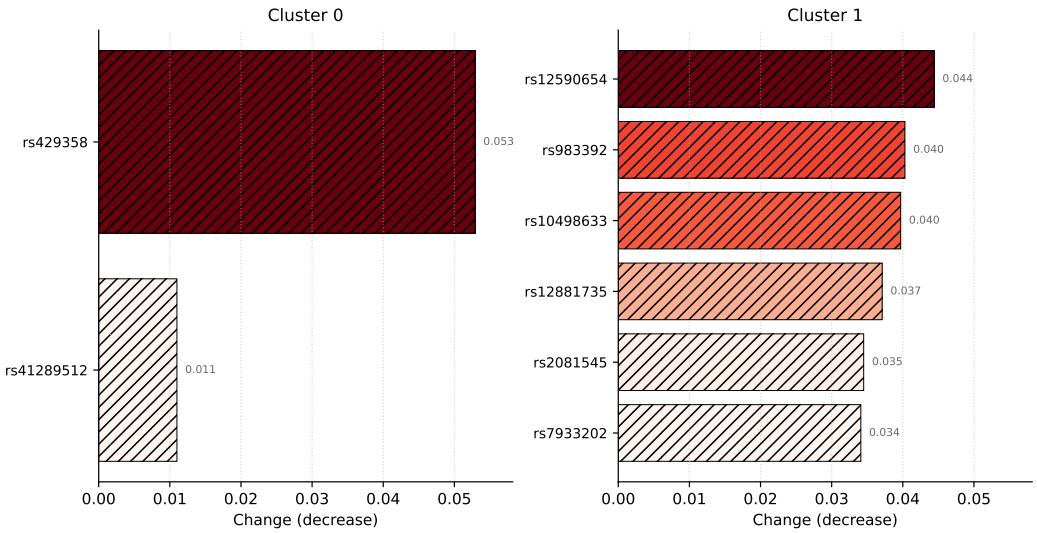


Figure 6.15: Decreased SNPs in each subgroup

Concerning each subgroup's genetic profile, the SNP rs429358 is seen highly elevated in the high-risk subgroup, and especially decreased in the low-risk one. This SNP, located in the fourth exon of the ApoE gene, affects the amino acid at position 130 of the resulting protein. The more common rs429358 allele is (T). If the allele is (C) and the same chromosome also harbors the rs7412 (C) allele, the combination is known as an *APOE-ϵ4* allele. The presence of the *APOE-ϵ4* allele, which arises from the rs429358(C) variant in combination with rs7412(C), is

the strongest known common genetic risk factor for late-onset Alzheimer's disease. Carriers of one or more copies of *APOE-ϵ4* exhibit significantly increased risk and earlier onset of the disease compared to non-carriers [51]. Recent work has further clarified that this allele contributes not only to amyloid-β pathology but also to tau aggregation and neuroinflammation, suggesting a multifactorial role in disease progression [52]. In the context of our subgroup analysis, the enrichment of rs429358 in the high-risk cluster is consistent with these findings, as it reflects the well-established link between *APOE-ϵ4* status and heightened vulnerability to Alzheimer's disease.

The SNP rs41289512 (located in or near the APOE locus on chromosome 19) has also been implicated in Alzheimer's disease risk via large-scale meta-analyses. In the Jansen et al. GWAS meta-analysis combining diagnosed AD cases and AD-by-proxy phenotypes, rs41289512 stood out with an extremely strong statistical association as one of the lead SNPs in the APOE region [53]. This suggests that rs41289512 is tightly linked (in linkage disequilibrium) with APOE alleles that confer high AD risk. In our subgroup genetic profiles, any enrichment of rs41289512 in the high-risk cluster may therefore reflect the same underlying APOE-driven risk architecture. While the functional consequence of rs41289512 itself is not well established, its strong statistical signal reinforces the genetic importance of its surrounding APOE locus in AD susceptibility.

When stratifying by diagnostic label (not-CN), a similar trend emerges. Compared to CN, non-CN individuals present reduced hippocampal and entorhinal volumes, paired with increased ventricular values. This correspondence strengthens the interpretation that Cluster 1 represents an MCI/AD-enriched population subgroup characterized by pronounced structular abnormalities, such as atrophy and verticular enlargement, while Cluster 0 reflects healthier anatomy with milder variation.

Table 6.13: Mean feature values per cluster

Cluster	**Left Hippocampus**	**Left Ent Entorhinal Area**	**Left Inf Lateral Ventricle**
0	-0.399557	-0.183944	0.153328
1	-2.333063	-1.977353	2.139030

Table 6.14: Mean feature values per Not_CN group

Not_CN	**Left Hippocampus**	**Left Ent Entorhinal Area**	**Left Inf Lateral Ventricle**
0	1.94×10^{-17}	-3.26×10^{-17}	3.44×10^{-17}
1	-1.282001	-0.919794	0.945788

To assess and showcase the impact of MCI cases on clustering separability, we reran the clustering pipeline using only CN and AD individuals, excluding all

MCI participants. The presence of MCI subjects introduces overlap that lowers the silhouette score and reduces cluster separability. Once MCI individuals were removed, the resulting clusters became more distinct and well-defined.

Table 6.15: Clustering results for AD vs CN

Scenario	Silhouette score	Cluster sizes
AD vs CN	0.591392	{1: 553, 0: 170}

When MCI individuals are exluded, the silhouette score improves, confirming that this group drives much of the separability. As shown in Table 6.15, the exclusion results in more separable clusters. This finding indicates that the overlap observed in the full dataset is largely driven by the hereogeneous nature of the MCI class. Nevertheless, excluding MCI would undermine the clinical revelance of the analysis, since MCI represents the critical transitional stage in Alzheimer's disease. Instead, these results highlight the intrinsic challenge posed by MCI, which must be addressed through longitudinal modeling rather than exclusion.

Diagnosis Prediction per Subgroup

AD vs MCI/CN Finally, the performance and feature importance results obtained from the AD vs MCI/CN classification task are presented for each subgroup.

Table 6.16: AUCs for AD vs MCI/CN Diagnosis Prediction across Subgroups

Cluster 0		Cluster 1	
AUC (ROC)	**AUC (PR)**	**AUC (ROC)**	**AUC (PR)**
0.846 ± 0.036	0.488 ± 0.106	0.655 ± 0.167	0.678 ± 0.101

Table 6.17: Top Regions for Subgroups 0 and 1

Subgroup 0		Subgroup 1	
Region	**Appearances**	**Region**	**Appearances**
Left Hippocampus	35	Left Hippocampus	35
Right Hippocampus	35	Right Hippocampus	35
Right Ent Entorhinal Area	31	Left Amygdala	34
Right Amygdala	27	Left Ent Entorhinal Area	33
Left Amygdala	24	Right Amygdala	31
Left Ent Entorhinal Area	23	Left Inf Lateral Ventricle	30
Right Inf Lateral Ventricle	21	Right Inf Lateral Ventricle	29
Left PHG Parahippocampal Gyrus	20	Left PHG Parahippocampal Gyrus	27
Left Inf Lateral Ventricle	18	Right Ent Entorhinal Area	26
Left ITG Inferior Temporal Gyrus	15	Left ITG Inferior Temporal Gyrus	26

Although the two subgroups differ in composition - with Subgroup 0 including CN, MCI, and AD individuals and Subgroup 1 consisting primarily of MCI and AD - the results in Table 6.17 show a high degree of overlap between the most frequently selected regions across the two subgroups. In both cases, the hippocampus, entorhinal cortex, amygdala, and inferior lateral ventricles consistently emerge as dominant contributors. As seen before, these regions reflect the core structural changes most strongly associated with Alzheimer's disease and are well-established markers of Alzheimer's disease progression: hippocampal and entorhinal atrophy are among the earliest structural changes, while ventricular enlargement reflects global brain atrophy. The recurrence of these areas strengthens the robustness of our interpretability analysis, confirming that the model has captured biologically meaningful patterns regardless of whether lower-risk individuals (CN) are included in the cohort or not.

Minor variations can still be observed, such as a somewhat higher frequency of middle temporal gyrus contributions in Subgroup 0, likely reflecting the greater heterogeneity of disease stage in the mixed group. However, these differences are modest when compared to the consistency of the medial temporal and ventricular regions across both analyses. Taken together, the findings emphasize that the predictive patterns learned by the model remain robust across cohorts with different risk profiles, reinforcing the central role of medial temporal atrophy and ventricular enlargement as universal hallmarks of disease progression.

AD vs CN AD vs CN is only applicable on the mixed subgroup 0, which mostly contains CN and MCI individuals.

Table 6.18: Per-visit AUC metrics (AD vs CN)

Subgroup 0		
Visit	**AUC (ROC)**	**AUC (PR)**
0	0.898	0.708
1	0.884	0.662
2	0.860	0.694
3	0.922	0.764
4	0.896	0.708
5	0.912	0.750
6	0.866	0.706
Mean	**0.891**	**0.713**

Table 6.19: Top Regions for Subgroup 0 (AD vs CN)

Region	Mean |IG|	Appearances
Left Hippocampus	0.0373	33
Right Hippocampus	0.0311	33
Left Inf Lateral Ventricle	0.0268	21
Left Amygdala	0.0290	21
Right Inf Lateral Ventricle	0.0250	21
Left Ent Entorhinal Area	0.0293	20
Right Thalamus Proper	0.0240	20
Right Ent Entorhinal Area	0.0261	20
Right Amygdala	0.0317	19
Left Thalamus Proper	0.0276	19

For subgroup 0 (lower-risk), the attribution results for AD vs CN diagnosis closely align with our findings in the diagnosis cohort. The hippocampus and amygdala dominate the rankings, with both hemispheres represented among the most influential features. The entorhinal cortex also appears, as well as ventricular expansion, reflecting the structural consequences of adjacent tissue atrophy. Beyond these classical regions, additional cortical contributions emerge, such as the thalamus proper.

Conversion Prediction per Subgroup

In the next step, we applied the Transformer models for prediction to subjects in the test set, stratified by the two derived clusters. For conversion prediction within Cluster 1, representing the high-risk patients (MCI/AD-dominant), only the first four time steps were retained due to limited sample size across visits. These results should therefore be viewed as exploratory, serving as a basis for biologically plausible interpretations.

Table 6.20: AUCs for Conversion Prediction across Subgroups

Cluster 0		Cluster 1	
AUC (ROC)	**AUC (PR)**	**AUC (ROC)**	**AUC (PR)**
0.742 ± 0.029	0.347 ± 0.025	0.671 ± 0.014	0.536 ± 0.014

Table 6.21: Top Regions in Conversion Prediction within Subgroup 0

Region	Mean |IG|	Appearances
Left PCgG Posterior Cingulate Gyrus	0.0071	15
Left Hippocampus	0.0048	15
Right Hippocampus	0.0060	13
Right FuG Fusiform Gyrus	0.0067	13
Right Plns Posterior Insula	0.0060	12
Left PP Planum Polare	0.0078	11
Right IOG Inferior Occipital Gyrus	0.0057	11
Left Plns Posterior Insula	0.0058	10
Right MOG Middle Occipital Gyrus	0.0060	10
Left FRP Frontal Pole	0.0059	10

Table 6.22: Top Regions in Conversion Prediction within Subgroup 1

Region	Mean |IG|	Appearances
Left Amygdala	0.0079	23
Left Hippocampus	0.0063	21
Right Inf Lateral Ventricle	0.0051	21
Left Inf Lateral Ventricle	0.0085	20
Right Amygdala	0.0054	19
Right Hippocampus	0.0078	17
Left Plns Posterior Insula	0.0070	16
Left Alns Anterior Insula	0.0060	13
Left FuG Fusiform Gyrus	0.0087	13
Right Plns Posterior Insula	0.0065	12

In the subgroup-specific conversion analysis, the attribution patterns again show overlap with established Alzheimer's disease biomarkers, but with distinct regional emphasis across clusters.

In the high-risk subgroup (Cluster 1), the hippocampus and amygdala dominate the top-ranked features, alongside strong contributions from the inferior lateral ventricles. Both hemispheres of the inferior lateral ventricles show high attribution, consistent with neurodegenerative ventricular enlargement reflecting tissue loss in surrounding medial temporal regions. The co-occurrence of hippocampal and amygdala regions underscores the central role of medial temporal lobe atrophy in driving conversion. In addition, the posterior and anterior insula, as well as the fusiform gyrus, also exhibit high attribution scores, suggesting that cortical–subcortical interactions extend beyond the classical hippocampal–entorhinal axis. Overall, these findings indicate that Cluster 1 captures a population with pronounced medial temporal vulnerability and distributed cortical involvement, where conversion is mostly influenced by hippocampal and amygdala decline.

By contrast, Cluster 0 presents a more heterogeneous attribution profile. While hippocampal and ventricular regions remain present, they are less dominant, and occipital and frontal regions emerge as relatively more important. The posterior cingulate gyrus and posterior insula appear consistently as key predictors,

together with the fusiform gyrus and frontal pole. The prominence of the posterior cingulate is particularly noteworthy, as this region is a well-established hub of early Alzheimer's pathology and was also highlighted in the global conversion interpretability analysis. This suggests that Cluster 0 represents a subgroup characterized by more distributed or atypical cortical involvement rather than purely medial temporal degeneration.

Chapter 7

Discussion and Future Work

7.1 Summary

The current thesis studies the classification of cognitively normal (CN) individuals and patients with Mild Cognitive Impairment (MCI) or Alzheimer's disease (AD) using longitudinal structural MRI from ADNI. A transformer-based model was used that approaches each patient's follow-up diagnosis as a sequence of visits, adding temporal information. To make the approach clinically realistic, the transformer is trained and evaluated with a causal attention mask - each time step can attend only to the current and past visits - together with explicit padding masks so that variable-length histories can be handled safely. After empirical analysis of the cohort's visit counts and class composition per visit index k, the main experimental horizon was fixed to $T = 7$ visits, which balances sample size and label stability while still capturing meaningful longitudinal change. On the other hand, clustering algorithms were applied to group patients with static and genetic features available. The two groups provided were used to test and interpret the longitudinal model into each cluster separately.

7.2 Future Work

Future research could further refine and extend the findings of this study along several directions. First, the conversion prediction task can be expanded by varying the forecasting horizon — that is, predicting conversion not only at the next visit but also across multiple future time steps. This would allow the model to capture longer-term disease trajectories and to evaluate how predictive information evolves over time. Additionally, the diagnosis prediction task could be extended to study the MCI vs CN or MCI vs AD binary subproblems, along with additional data sources and modalities.

A second line of work concerns the representation of per-visit features before the Transformer encoder. Two approaches were experimentally tested in this

study, even though the latter was omitted: mean pooling across brain regions and an attention-based pooling mechanism that learns a soft weighting over regions. Although the attention pooling variant showed promise in capturing region-specific importance, it also introduced model overfitting due to the increased capacity and parameterization ($T \times F$). Further work could revisit this approach with stronger regularization or sparse attention strategies to achieve a better balance between expressivity and generalization.

Finally, future research could focus on the diagnosis prediction model, extending the analysis to assess model robustness across genetic subgroups rather than the overall combined SNP+ROI baseline profiles. Such stratified evaluation could reveal subgroup-specific model behaviors and potential biomarker differences, contributing to a more personalized understanding of Alzheimer's disease progression.

Appendix A

Additional Tables and Results

The appendix provides the results on the experiments using the raw data, without the linear covariates adjustment.

AD vs MCI/AD Results on Raw Data.

Table A.1: Per visit AUC metrics

Visit	AUC (ROC)	AUC (PR)
0	0.866	0.612
1	0.866	0.660
2	0.836	0.564
3	0.860	0.678
4	0.844	0.564
5	0.812	0.544
6	0.868	0.684
Mean	**0.850**	**0.615**

Table A.2: Per time step metrics (AD vs Rest) (Raw Data)

Visit	Confusion Matrix	Acc.	Bal. Acc.	Prec.	Recall	Spec.	F1	MCC	AUC (ROC)	AUC (PR)
0	[[385, 13], [57, 24]]	0.85	0.64	0.64	0.30	0.96	0.41	0.37	0.87	0.61
1	[[319, 7], [47, 23]]	0.87	0.65	0.76	0.33	0.98	0.46	0.44	0.87	0.66
2	[[250, 12], [43, 21]]	0.83	0.64	0.63	0.33	0.95	0.43	0.36	0.84	0.56
3	[[215, 9], [33, 25]]	0.85	0.69	0.74	0.43	0.96	0.51	0.47	0.86	0.68
4	[[172, 7], [31, 12]]	0.83	0.62	0.64	0.28	0.96	0.39	0.33	0.84	0.57
5	[[128, 5], [19, 8]]	0.84	0.62	0.62	0.29	0.96	0.38	0.34	0.82	0.55
6	[[85, 4], [16, 10]]	0.83	0.68	0.75	0.38	0.96	0.52	0.46	0.87	0.68

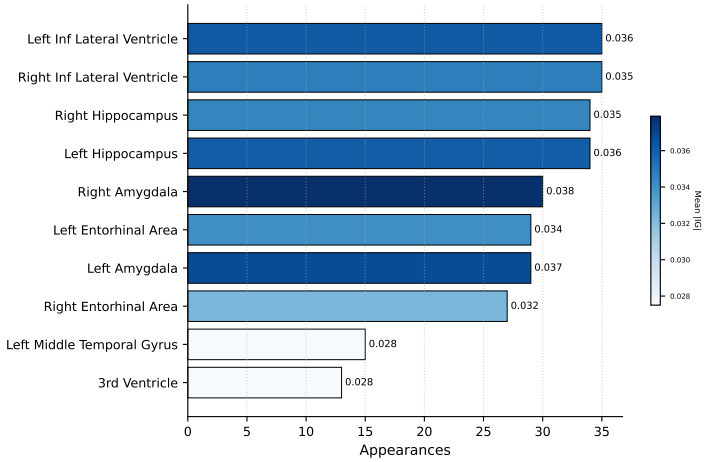


Figure A.1: AD vs non-AD most frequent ROIs (Raw Data)

AD vs CN Results on Raw Data.

Table A.3: Per time step AUC metrics (AD vs CN)

Visit	**AUC (ROC)**	**AUC (PR)**
0	0.930	0.896
1	0.930	0.918
2	0.918	0.910
3	0.914	0.904
4	0.916	0.890
5	0.868	0.836
6	0.928	0.920
Mean	**0.915**	**0.896**

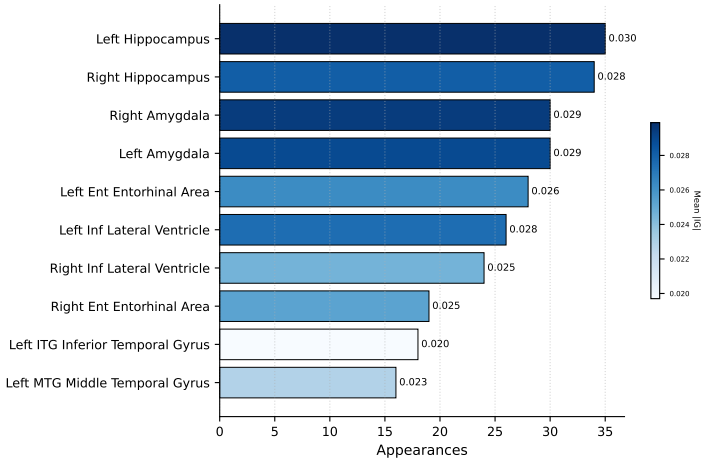


Figure A.2: AD vs CN most frequent ROIs (Raw Data)

Table A.4: AUCs for Subgroup 0 (AD vs CN) (Raw Data)

Cluster 0	
AUC (ROC)	**AUC (PR)**
0.842 ± 0.030	0.667 ± 0.061

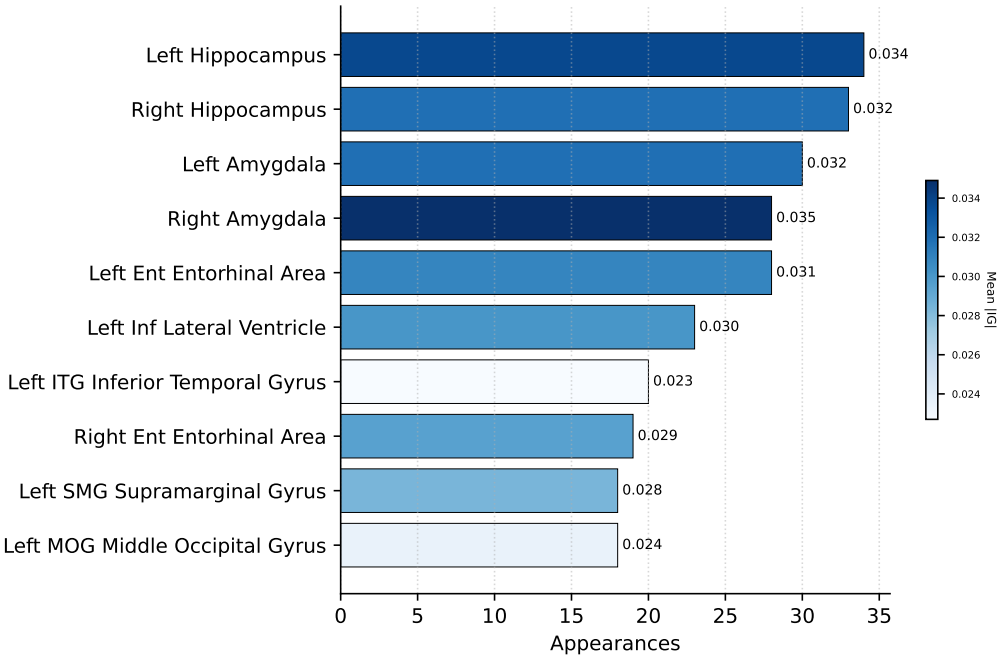


Figure A.3: Top Regions for Subgroup 0 (AD vs CN) (Raw Data)

MCI to AD Conversion Prediction Results on Raw Data.

Table A.5: Per time step AUC metrics (MCI → AD)

Visit	AUC (ROC)	AUC (PR)
0	0.758	0.244
1	0.758	0.208
2	0.552	0.212
3	0.772	0.392
4	0.618	0.276
5	0.592	0.332
6	0.854	0.686
Mean	**0.701**	**0.336**

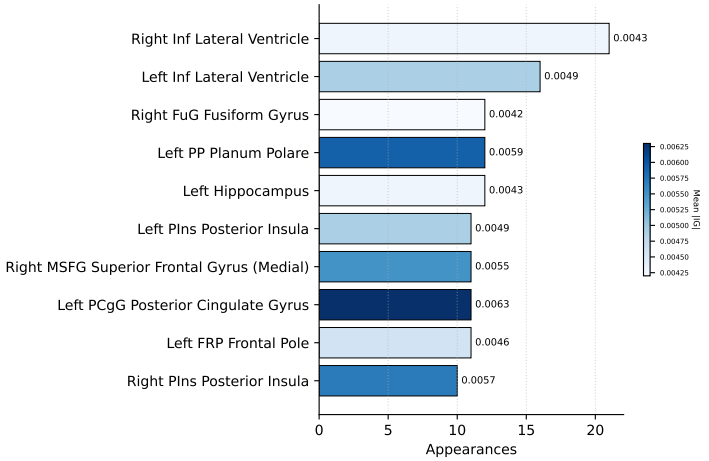


Figure A.4: MCI to AD most frequent ROIs (Raw Data)

Table A.6: AUCs for Conversion Prediction across Subgroups (Raw Data)

Cluster 0		**Cluster 1**	
AUC (ROC)	**AUC (PR)**	**AUC (ROC)**	**AUC (PR)**
0.711 ± 0.012	0.351 ± 0.021	0.729 ± 0.023	0.597 ± 0.019

Table A.7: Top Regions in Conversion Prediction within Subgroup 1 (high-risk) (Raw Data)

Region	**Mean $\lvert IG \rvert$**	**Appearances**
Right Inf Lateral Ventricle	0.0056	25
Left Inf Lateral Ventricle	0.0063	25
Right AIns anterior insula	0.0055	16
Left AIns anterior insula	0.0065	15
Right Lateral Ventricle	0.0039	13
Occipital Lobe WM left	0.0053	11
Right IOG Inferior Occipital Gyrus	0.0059	11
Left Amygdala	0.0041	10
Right PO Parietal Operculum	0.0065	10
Left PCgG Posterior Cingulate Gyrus	0.0073	10

Table A.8: Top Regions in Conversion Prediction within Subgroup 0 (Raw Data)

Region	**Mean |IG|**	**Appearances**
Right Inf Lateral Ventricle	0.0043	17
Left PCgG Posterior Cingulate Gyrus	0.0061	15
Left PIns Posterior Insula	0.0054	12
Left FRP Frontal Pole	0.0053	12
Left Hippocampus	0.0049	12
Left PP Planum Polare	0.0069	11
Left MFC Medial Frontal Cortex	0.0045	11
Right FuG Fusiform Gyrus	0.0045	11
Left Inf Lateral Ventricle	0.0042	10
Left FO Frontal Operculum	0.0055	10

Bibliography

[1] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017, pp. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[2] M. B. Carpenter, *Human Neuroanatomy*, 9th. Baltimore: Williams & Wilkins, 1991, ISBN: 9780683035063.

[3] R. C. Petersen et al., "Mild cognitive impairment: Clinical characterization and outcome," *The Lancet Neurology*, vol. 5, no. 9, pp. 736–746, 2006. DOI: 10.1016/S0140-6736(06)68542-5

[4] R. C. Petersen, "Mild cognitive impairment," *Continuum (Minneapolis, Minn.)*, vol. 22, no. 2, pp. 404–418, 2016. DOI: 10.1212/CON.0000000000000313

[5] D. S. Knopman and R. C. Petersen, "Practical diagnosis and management of dementia," *JAMA*, vol. 321, no. 16, pp. 1589–1599, 2019. DOI: 10.1001/jama.2019.4782

[6] K. Shailaja, B. Seetharamulu, and M. A. Jabbar, "Machine learning in healthcare: A review," in *2018 2nd International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2018, pp. 910–914. DOI: 10.1109/ICECA.2018.8474918

[7] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[8] G. Biau and E. Scornet, "A random forest guided tour," *TEST*, vol. 25, no. 2, pp. 197–227, 2016. DOI: 10.1007/s11749-016-0481-7 [Online]. Available: https://doi.org/10.1007/s11749-016-0481-7

[9] J. Liu, "Logistic regression analysis and its application in medicine," *Journal of Data Analysis and Information Processing*, vol. 7, no. 4, pp. 167–173, 2019. DOI: 10.4236/jdaip.2019.74010 [Online]. Available: https://www.scirp.org/journal/paperinformation?paperid=95655

[10] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, "Applied logistic regression," 2013.

[11] G. Camps-Valls, L. Bruzzone, J. L. Rojo-Álvarez, and M. Martinez-Ramon, "A survey on support vector machine-based methods for classification problems in neuroimaging," *Neurocomputing*, vol. 408, pp. 231–245, 2020. DOI: 10.1016/j.neucom.2019.10.118

[12] J. Peters, H. Brenner, and K.-R. Müller, "Automatic feature extraction for classification of cognitive workload using machine learning," *Frontiers in Neurorobotics*, vol. 7, p. 21, 2013. DOI: 10.3389/fnbot.2013.00021

[13] T. Chen et al., "A review on xgboost algorithm," *International Journal of Database Management Systems*, vol. 11, no. 1, pp. 39–52, 2019. DOI: 10.5121/ijdms.2019.11101 [Online]. Available: https://doi.org/10.5121/ijdms.2019.11101

[14] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785

[15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[16] D. Berrar, "Performance measures for classification," *arXiv preprint arXiv:2008.05756*, 2020. [Online]. Available: https://arxiv.org/abs/2008.05756

[17] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997. DOI: 10.1016/S0031-3203(96)00142-2

[18] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006. DOI: 10.1016/j.patrec.2005.10.010

[19] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982. DOI: 10.1148/radiology.143.1.7063747

[20] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group, 1984, ISBN: 978-0412048418.

[21] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, ACM, 2006, pp. 233–240. DOI: 10.1145/1143844.1143874

[22] J. Blömer, C. Lammersen, M. Schmidt, and C. Sohler, *Theoretical analysis of the k-means algorithm – a survey*, arXiv preprint, arXiv:1602.08254, Feb. 2016. [Online]. Available: https://arxiv.org/abs/1602.08254

[23] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," pp. 1027–1035, 2007.

[24] L. Rokach and O. Maimon, "Clustering methods," in *Data Mining and Knowledge Discovery Handbook*, Springer, 2005, pp. 321–352. DOI: 10.1007/0-387-25465-X_15

[25] F. Murtagh and P. Legendre, "Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion?" *Journal of Classification*, vol. 31, pp. 274–295, 2014. DOI: 10.1007/s00357-014-9161-z

[26] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*, vol. 14, 2002, pp. 849–856.

[27] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007. DOI: 10.1007/s11222-007-9033-z

[28] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, AAAI Press, 1996, pp. 226–231.

[29] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[30] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[31] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017, pp. 3319–3328.

[32] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4765–4774.

[33] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 3145–3153, 2017.

[34] C. Xue et al., "Ai-based differential diagnosis of dementia etiologies on multimodal data," *Nature Medicine*, vol. 30, no. 10, pp. 2977–2989, 2024. DOI: 10.1038/s41591-024-03118-z

[35] V. H. Jasodanand et al., "Ai-driven fusion of multimodal data for alzheimer's disease biomarker assessment," *Nature Communications*, vol. 16, no. 1, p. 7407, 2025. DOI: 10.1038/s41467-025-62590-4

[36] Z. Yao et al., "It: An interpretable transformer model for alzheimer's disease prediction based on pet/mr images," *NeuroImage*, vol. 311, p. 121 210, 2025. DOI: 10.1016/j.neuroimage.2025.121210

[37] M. E. Vlontzou, M. Athanasiou, K. V. Dalakleidi, and I. Skampardoni, "A comprehensive interpretable machine learning framework for mild cognitive impairment and alzheimer's disease diagnosis," *Scientific Reports*, vol. 15, no. 1, p. 8410, 2025. DOI: 10.1038/s41598-025-92577-6

[38] D. Agostinho, M. Simões, M. Castelo-Branco, and the Alzheimer's Disease Neuroimaging Initiative, "Predicting conversion from mild cognitive impairment to alzheimer's disease: A multimodal approach," *Brain Communications*, vol. 6, no. 4, fcae208, 2024. DOI: 10.1093/braincomms/fcae208

[39] C. Xue, S. S. Kowshik, D. Lteif, and S. Puducheri, "Ai-based differential diagnosis of dementia etiologies on multimodal data," *Nature Medicine*, vol. 30, no. 10, pp. 2977–2989, 2024. DOI: 10.1038/s41591-024-03118-z

[40] M. Al Olaimat, J. Martinez, F. Saeed, S. Bozdag, and the Alzheimer's Disease Neuroimaging Initiative, "Ppad: A deep learning architecture to predict progression of alzheimer's disease," *Bioinformatics*, vol. 39, no. Supplement_1, pp. i149–i157, 2023. DOI: 10.1093/bioinformatics/btad249

[41] H. Xue et al., "Artificial intelligence for differential diagnosis of dementia: A multicentre analysis of brain imaging and clinical data from 52,631 individuals," *Nature Medicine*, vol. 30, no. 3, pp. 754–765, 2024. DOI: 10.1038/s41591-024-03118-z

[42] Y. Liu, J. Liu, Y. Li, Z. Wang, H. Jiang, and Y. Zhang, "A transformer-based framework for alzheimer's disease progression prediction using longitudinal multi-modal data," *Bioinformatics*, vol. 39, no. Supplement 1, pp. i149–i159, 2023. DOI: 10.1093/bioinformatics/btad258

[43] Z. Zhu, Y. Liu, C. Wang, Z. Zhang, H. Li, and F. Xu, "A deep learning model for early diagnosis of alzheimer's disease combined with 3d cnn and video swin transformer," *Scientific Reports*, vol. 15, no. 1, p. 4556, 2025. DOI: 10.1038/s41598-025-05568-y [Online]. Available: https://www.nature.com/articles/s41598-025-05568-y

[44] Z. Yang et al., "Disentangling brain heterogeneity via semi-supervised deep learning and mri: Dimensional representations of alzheimer's disease," 2021, Manuscript in preparation or preprint.

[45] D. Bounias et al., "Interactive machine learning-based multi-label segmentation of solid tumors and organs," *Applied Sciences*, vol. 11, no. 16, p. 7488, 2021. DOI: 10.3390/app11167488

[46] Y. L. Rao, B. Ganaraja, B. V. Murlimanju, T. Joy, A. Krishnamurthy, and A. Agrawal, "Hippocampus and its involvement in alzheimer's disease: A review," *3 Biotech*, vol. 12, no. 2, p. 55, 2022. DOI: 10.1007/s13205-022-03123-4

[47] S. P. Poulin, R. Dautoff, J. C. Morris, L. F. Barrett, B. C. Dickerson, et al., "Amygdala atrophy is prominent in early alzheimer's disease and relates to symptom severity," *Psychiatry Research*, vol. 194, no. 1, pp. 7–13, 2011. DOI: 10.1016/j.pscychresns.2011.06.014

[48] K. M. Igarashi, "Entorhinal cortex dysfunction in alzheimer's disease," *Trends in Neurosciences*, vol. 46, no. 2, pp. 124–136, 2023. DOI: 10.1016/j.tins.2022.11.006

[49] S. Chen et al., "Spatially resolved transcriptomics reveals genes associated with the vulnerability of middle temporal gyrus in alzheimer's disease," *Acta Neuropathologica Communications*, vol. 10, p. 188, 2022, PMID: 36544231. DOI: 10.1186/s40478-022-01494-6

[50] J. Hwang et al., "Clinical implications of amyloid-beta accumulation in occipital lobes in alzheimer's continuum," *Brain Sciences*, vol. 11, no. 9, p. 1232, 2021, PMID: 34573252. DOI: 10.3390/brainsci11091232

[51] Y. Yamazaki, N. Zhao, T. R. Caulfield, C.-C. Liu, and G. Bu, "Apolipoprotein e and alzheimer disease: Pathobiology and targeting strategies," *Nature Reviews Neurology*, vol. 16, no. 12, pp. 707–720, 2020. DOI: 10.1038/s41582-020-00430-3 [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC7085286/

[52] Y. He, M. Xu, F. Li, Y. Zhang, X. Chen, and J. Wang, "Apoe ε4 differentially impacts tau aggregation and neuroinflammation in alzheimer's disease," *The Lancet Healthy Longevity*, vol. 6, no. 2, e126–e138, 2025. DOI: 10.1016/S2950-5887(25)00060-6 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2950588725000606

[53] I. E. Jansen et al., "Genome-wide meta-analysis identifies new loci and functional pathways influencing alzheimer's disease risk," *Nature Genetics*, vol. 51, no. 3, pp. 404–413, 2019. DOI: 10.1038/s41588-018-0311-9 [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC6836675/