

ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Εργαστήριο Βιοπληροφορικής και Ανθρώπινης
Ηλεκτροφυσιολογίας



ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

«Ανάπτυξη μοντέλου και εφαρμογής για την πρόγνωση
της γνωστικής ανεπάρκειας»

Γεράσιμος Γραμμένος

Επιβλέπων: Θεμιστοκλής Έξαρχος, Αναπληρωτής Καθηγητής

Περίληψη

Η Ήπια Γνωστική Διαταραχή (ΗΓΔ) είναι μια γνωστική κατάσταση που εμφανίζεται συχνά σε ηλικιωμένους και χαρακτηρίζεται από αξιοσημείωτες αλλαγές στη μνήμη, τη σκέψη και τις ικανότητες συλλογισμού που δεν εμπίπτουν στην φυσιολογική γνωστική έκπτωση. Υπολογίζεται πως περίπου το 10% -15% των ατόμων που πάσχουν από ΗΓΔ θα παρουσιάσουν Alzheimer σε βάθος χρόνου κάνοντας έτσι την ΗΓΔ στάδιο του Alzheimer. Με την παρούσα διατριβή προτείνεται μια μέθοδος με την χρήση eXtreme Gradient Boosting για την πρόγνωση της νόσου Alzheimer όσο ο ασθενής βρίσκεται στο στάδιο της ΗΓΔ. Συγκεκριμένα, αξιοποιώντας δεδομένα από το ADNI αναπτύχθηκε ένα μοντέλο που χρησιμοποιεί διαχρονικά δεδομένα από την baseline επίσκεψη και την επίσκεψη των 12 μηνών από το baseline για να υπολογιστεί σε βάθος 36 μηνών η πιθανότητα μετάβασης σε Alzheimer από ΗΓΔ επιτυγχάνοντας ακρίβεια της τάξεως του 70%. Επιπλέον εφαρμόζεται η επιλογή χαρακτηριστικών με την μέθοδο αναδρομικής εξάλειψης καθώς επίσης και η μέθοδος Shapley για την επεξήγηση των αποφάσεων του μοντέλου. Τέλος αναπτύχθηκε μια διαδικτυακή εφαρμογή προσομοιώνοντας ένα περιβάλλον όπου ο ειδικός θα μπορεί να εισάγει τιμές από τις εξετάσεις του ασθενούς και με την χρήση του μοντέλου να λάβει πρόβλεψη για την πιθανότητα μετάβασης του ασθενούς σε Alzheimer από ΗΓΔ.

Abstract

Mild Cognitive Impairment (MCI) is a cognitive condition frequently observed in older adults, characterized by significant alterations in memory, thinking, and reasoning abilities that extend beyond typical cognitive decline. It is worth noting that around 10%-15% of individuals with MCI are projected to develop Alzheimer's disease, effectively positioning MCI as an early stage of Alzheimer's. In this thesis, a novel approach is presented involving the utilization of eXtreme Gradient Boosting to predict the onset of Alzheimer's disease during the MCI stage. The methodology entails harnessing data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Through the analysis of longitudinal data spanning from the baseline visit to the 12-month follow-up, a predictive model was constructed. This model calculates, over a 36-month period, the likelihood of progression from MCI to Alzheimer's disease, achieving accuracy rate of 70%. To further enhance the precision of the model, the study implements feature selection using the recursive feature elimination technique. Additionally, Shapley method is employed to provide insights into the model's decision-making process, thereby augmenting the transparency and interpretability of the predictions. As a practical application of the developed model, an interactive web application was crafted. This application simulates an environment in which experts can input test values pertaining to a patient. Subsequently, the model processes these inputs and furnishes a prediction regarding the patient's probability of transitioning from MCI to Alzheimer's disease.

Περιεχόμενα

Περίληψη	1
Abstract	2
Λίστα εικόνων	5
Λίστα πινάκων	6
Λίστα συντομογραφιών	6
1. Εισαγωγή	7
1.1 Η νόσος Alzheimer.....	7
1.2 Στάδια της Νόσου.....	7
1.3 Ήπια Γνωστική Διαταραχή.....	8
1.4 Ο ρόλος της Μηχανικής Μάθησης στην πρόγνωση της νόσου	9
1.5 Σκοπός της παρούσας διατριβής	10
2. Βιβλιογραφική ανασκόπηση	11
3. Alzheimer’s Disease Neuroimaging Initiative (ADNI)	18
3.1 Περιγραφή της μελέτης.....	18
3.2 Βιοδείκτες.....	19
3.3 Φάσεις της μελέτης	20
3.3.1 ADNI 1.....	21
3.3.2 ADNI GO.....	22
3.3.3 ADNI 2.....	23
3.3.4 ADNI 3.....	24
4. Διερευνητική Ανάλυση Δεδομένων – Exploratory Data Analysis.....	26
4.1 Δημογραφικά χαρακτηριστικά	26
4.2 Διαγνώσεις	27
4.3 Ελλείπουσες τιμές - Missing values.....	28
4.4 Mini-Mental State Examination (MMSE).....	30
5. Πρόγνωση του Alzheimer με την χρήση Μηχανικής Μάθησης.....	32
5.1 Ορισμός του προβλήματος.....	32
5.2 Βιβλιοθήκες που χρησιμοποιήθηκαν	32
5.2.1 Pandas	32
5.2.2 NumPy	32
5.2.3 Matplotlib.....	32

5.2.4	XGBoost	33
5.2.5	Scikit-learn.....	33
5.3	Προ-επεξεργασία των δεδομένων	33
5.4	eXtream Gradient Boosting - XGBoost	36
5.5	Επιλογή των χαρακτηριστικών.....	38
5.6	Εκπαίδευση του μοντέλου.....	40
5.7	Αποτελέσματα και συζήτηση	41
5.8	Ερμηνεία του μοντέλου και των αποφάσεων του	42
6.	Διαδικτυακή Εφαρμογή για την πρόγνωση της Νόσου	45
6.1	Βιβλιοθήκες και τεχνολογίες που χρησιμοποιήθηκαν	45
6.1.1	ReactJS – Frontend	45
6.1.2	Flask – Backend.....	45
6.1.3	Docker.....	45
6.2	Αρχιτεκτονική εφαρμογής.....	46
7.	Συμπεράσματα, περιορισμοί και μελλοντικές προεκτάσεις.	49
	Αναφορές	50
	Παράρτημα Α.....	54
	Παράρτημα Β.....	57
	Κώδικας Exploratory Data Analysis.....	57
	Κώδικας προ-επεξεργασίας του συνόλου δεδομένων.....	60
	Κώδικας επιλογής χαρακτηριστικών	63
	Κώδικας εκπαίδευσης του μοντέλου	64
	Κώδικας γραφημάτων Shapley	65
	Κώδικας διαδικτυακής εφαρμογής.	68

Λίστα εικόνων

Εικόνα 1:(Αριστερά) φυσιολογική κατάσταση.(Κέντρο) ΗΓΔ.(Δεξιά) Alzheimer	9
Εικόνα 2 Το γράφημα απεικονίζει τους βιοδείκτες ως δείκτες της ΝΑ. Οι καμπύλες δείχνουν τις αλλαγές σε πέντε βιοδείκτες από φυσιολογικούς σε μη φυσιολογικούς κατά την πορεία της ΝΑ (φυσιολογική νόση έως άνοια).	19
Εικόνα 3: Σύνοψη φάσεων ADNI.....	20
Εικόνα 4: Πρόγραμμα κλινικών δοκιμών του ADNI	21
Εικόνα 5: Χρονολόγιο ADNI-1	21
Εικόνα 6: Χρονολόγιο ADNI-GO και ADNI-2	22
Εικόνα 7: Χρονολόγιο ADNI-3	24
Εικόνα 8: Συγκεντρωτικό γράφημα δημογραφικών στοιχείων του ADNIMERGE.....	26
Εικόνα 9: Αριθμός δειγμάτων κάθε διάγνωσης(κατηγορίας)	27
Εικόνα 10: Αριθμός δειγμάτων διαγνώσεων στο baseline στάδιο.....	27
Εικόνα 11:Αριθμός διαγνώσεων συναρτήσει του χρόνου	28
Εικόνα 12: Ελλείπουσες τιμές στο baseline στάδιο.....	29
Εικόνα 13: Ελλείπουσες τιμές στις δωδεκάμηνες επισκέψεις	30
Εικόνα 14: Ελλείπουσες τιμές στις εικοσιτετράμηνες επισκέψεις	30
Εικόνα 15: Τιμές του MMSE για κάθε διάγνωση συναρτήσει του χρόνου	31
Εικόνα 16:Πλήθος ασθενών που παραμένουν σταθεροί σε ΗΓΔ μέχρι την έλευση των 48 μηνών (αριστερά). Πλήθος ασθενών που παρουσιάσουν Alzheimer μέχρι και την έλευση των 48 μηνών(δεξιά).....	35
Εικόνα 17:Πλήθος ασθενών που παραμένουν σταθεροί σε ΗΓΔ μέχρι την έλευση των 48 μηνών (πάνω). Πλήθος ασθενών που παρουσιάσουν Alzheimer μέχρι και την έλευση των 48 μηνών(κάτω)	36
Εικόνα 18: Διαδικασία εκπαίδευσης του μοντέλου	40
Εικόνα 19: Επιρροή του κάθε χαρακτηριστικού στην τελική απόφαση του μοντέλου	42
Εικόνα 20: Επιρροή του κάθε χαρακτηριστικού στην τελική απόφαση του μοντέλου	43
Εικόνα 21: Διάγραμμα λήψης απόφασης με βάση τις τιμές Shapley για μια τυχαία πρόβλεψη από το σύνολο δεδομένων. Η ανάγνωση του διαγράμματος γίνεται από κάτω προς τα πάνω. Στον άξονα Υ απεικονίζονται τα χαρακτηριστικά και ενώ στον άξονα Χ βρίσκονται οι τιμές Shapley. Στις παρενθέσεις βρίσκονται οι τιμές των χαρακτηριστικών για το συγκεκριμένο δείγμα.	44
Εικόνα 22: Αρχιτεκτονική διαδικτυακής εφαρμογής.....	46
Εικόνα 23: Οθόνη προβολής αποτελέσματος πρόβλεψης του μοντέλου	47
Εικόνα 24: Κεντρική οθόνη εφαρμογής	47
Εικόνα 25: Κώδικας διαχείρισης του αιτήματος POST της εφαρμογής προς τον server.	48

Λίστα πινάκων

Πίνακας 1: Συμμετέχοντες ADNI-1.....	22
Πίνακας 2: Συμμετέχοντες ADNI-GO.....	23
Πίνακας 3: Συμμετέχοντες ADNI-2.....	24
Πίνακας 4: Συμμετέχοντες ADNI-3.....	25
Πίνακας 5: Αναπαράσταση κλάσεων (συμβολοσειρών) με ακέραιους αριθμούς.....	33
Πίνακας 6: Δείγμα του χαρακτηριστικού "SEX" πριν την κωδικοποίηση.....	34
Πίνακας 7: Μετά την κωδικοποίηση του χαρακτηριστικού SEX έχουν δημιουργηθεί 2 "νέα" χαρακτηριστικά.....	34
Πίνακας 8: Τελικές παράμετροι για την εκπαίδευση του μοντέλου	37
Πίνακας 9: Χαρακτηριστικά συνόλου δεδομένων μετά την εφαρμογή της μεθόδου αναδρομικής εξάλειψης χαρακτηριστικών	38
Πίνακας 10: Αποτελέσματα εκπαίδευσης του μοντέλου	41
Πίνακας 11: Χαρακτηριστικά συνόλου δεδομένων πριν την προ-επεξεργασία.	54

Λίστα συντομογραφιών

- AD: Alzheimer
- NA: Νόσος Alzheimer
- MCIC: Mild Cognitive Impairment Converters
- MCInc: Mild Cognitive Impairment Non-Converters
- NC: Normal Cognition
- CN: Cognitive Normal
- ΗΓΔ: Ήπια Γνωστική Διαταραχή

1. Εισαγωγή

1.1 Η νόσος Alzheimer

Η νόσος Alzheimer (AD) είναι μια νόσος που προσβάλλει των ανθρώπινο εγκέφαλο στο σύνολο της δομής του προκαλώντας δυσλειτουργίες σε μέρη του εγκεφάλου που σχετίζονται με την μνήμη, την γνωστική ικανότητα και στην συμπεριφορά. Ανακαλύφθηκε από τον Γερμανό νευρολόγο Alois Alzheimer στα τέλη της δεκαετίας του 1890 όταν ο ίδιος παρακολουθούσε μια ασθενή του που παρουσίαζε συμπτώματα απώλειας μνήμης και διαταραχής της συμπεριφοράς. Αφού η ασθενής απεβίωσε ο ίδιος μελέτησε τον εγκέφαλό της και παρατήρησε δομικές διαφορές - δυσλειτουργίες σε σχέση με έναν υγιές εγκέφαλο καταλήγοντας στο συμπέρασμα πως η ίδια έπασχε από μια σπάνια μορφή άνοιας[1].

Η νόσος ανήκει στην κατηγορία των νευροεκφυλιστικών νόσων προκαλώντας μη αναστρέψιμα προβλήματα στην μνήμη την αντίληψη και την συμπεριφορά. Στα πρώτα στάδια της νόσου, ο ασθενής μπορεί να λειτουργήσει χωρίς φροντίδα από κάποιον τρίτο. Ωστόσο, ενδεχομένως να νιώθει αδυναμία στην ανάκληση ονομάτων κοντινών προσώπων του, τοποθεσιών η ακόμα και λέξεων κατά την διάρκεια της ομιλίας. Όσο η νόσος αναπτύσσεται τα συμπτώματα επιμένουν και γίνονται εντονότερα. Στο τελικό στάδιο ο ασθενής αδυνατεί να αυτοεξυπηρετηθεί και χρειάζεται την βοήθεια κάποιου τρίτου για τις δραστηριότητες της καθημερινότητας. Συγκεκριμένα στα τελευταία στάδια, η επικοινωνία είναι σχεδόν αδύνατη, η αντίληψη και η γνωστική ικανότητα έχουν εξασθενήσει σημαντικά, η συμπεριφορά του ατόμου έχει αλλάξει ενώ εξασθενούν και κινητικές λειτουργίες όπως το περπάτημα και η κατάποση[2].

Θα πρέπει να σημειωθεί πως το Alzheimer δεν είναι το ίδιο με την άνοια. Σύμφωνα με την Διεθνή Ένωση για το Alzheimer η άνοια ορίζεται τυπικά ως ένα σύνολο παθήσεων και καταστάσεων όπου παρατηρείται γενική έκπτωση της γνωστικής ικανότητας του ατόμου και αλλαγή της συμπεριφοράς του στην καθημερινότητα. Επιπλέον, υπάρχουν διάφορα είδη άνοιας που η πρόκληση τους εξαρτάται από διάφορες συνθήκες[3].

Αντίθετα το Alzheimer είναι ο πιο συνηθισμένος τύπος άνοιας και αποτελεί το 50-60% των περιπτώσεων της. Όπως και η άνοια, η νόσος δεν συγκαταλέγεται στις σύνηθες παθήσεις που προκαλούνται λόγω της γήρανσης του ανθρώπου [3]. Στην πιο συχνή μορφή του προκαλείται από την συσσώρευση της β αμυλοειδούς πρωτεΐνης και πρωτεΐνης Ταυ γύρω από τα νευρικά κύτταρα. Αυτές οι πρωτεΐνες σχηματίζουν τις β αμυλοειδείς πλάκες και νευροϊνιδιακές δεσμίδες δημιουργώντας μόνιμη βλάβη στα νευρικά κύτταρα. Προκαλεί παρόμοια συμπτώματα με την άνοια ωστόσο το πιο χαρακτηριστικό σύμπτωμα είναι η εξασθένηση της αντίληψης του ατόμου. Για την διάγνωση του ο πιθανός ασθενής υποβάλλεται σε μια σειρά νευρολογικών, ψυχικών, εργαστηριακών και απεικονιστικών εξετάσεων[4].

1.2 Στάδια της Νόσου

Η εξέλιξη της νόσου ταξινομείται σε 4 βασικά στάδια, το ασυμπτωματικό, το πρώιμο, το μεσαίο και το προχωρημένο στάδιο. Το ασυμπτωματικό η προ-κλινικό στάδιο διαρκεί για χρόνια και το άτομο δεν παρουσιάζει κάποιο κλινικό σύμπτωμα που να δηλώνει Alzheimer. Ο μόνος τρόπος διάγνωσης του σταδίου είναι η ύπαρξη βιοδεικτών που να υποδηλώνουν εμφάνιση της νόσου στο μέλλον[5]. Στο πρώιμο στάδιο παρατηρείται ήπια απώλεια μνήμης με μερική επίδραση στην

καθημερινότητα του ατόμου. Σε αυτό το στάδιο παρατηρείται δυσκολία στην συγκέντρωση, απώλεια στην αίσθηση του χρόνου και ενδεχομένως εμφάνιση κατάθλιψης. Το στάδιο αυτό είναι γνωστό και ως «Ήπια Γνωστική Διαταραχή - ΗΓΔ».

Στο μεσαίο στάδιο η νόσος επηρεάζει και άλλες περιοχές του εγκεφάλου προκαλώντας αυξημένη απώλεια μνήμης , δυσκολία στην ομιλία και ανάγνωση ,στην γραφή καθώς και αλλαγή συμπεριφοράς του ατόμου.

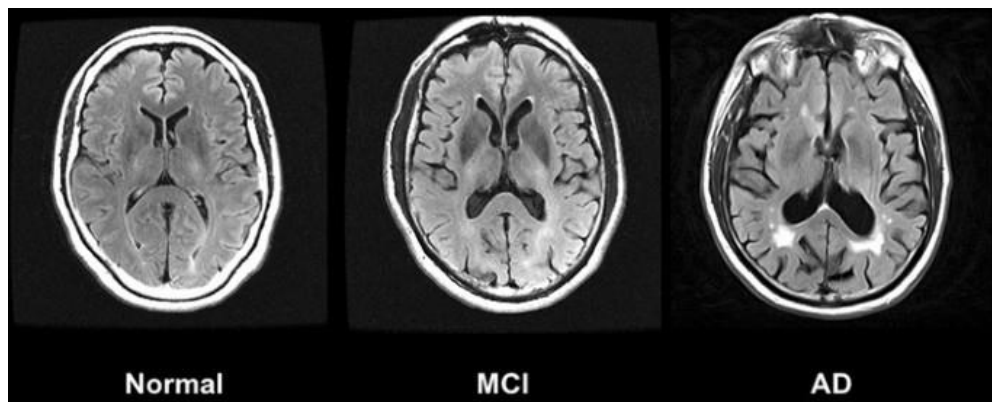
Στο προχωρημένο στάδιο η νόσος έχει βλάψει τον εγκεφαλο σε μεγάλο βαθμό. Στο στάδιο αυτό η μνήμη του ατόμου έχει εξασθενήσει χρίζοντας το ανίκανο να αναγνωρίσει κοντινά πρόσωπα και το άτομο δεν μπορεί να ανταπεξέλθει στην καθημερινότητα του έχοντας επιπλέον προβλήματα στην κατάποση, στην κίνηση, στη αντίληψη και στην επικοινωνία με του με τους γύρω του. Η βλάβη πλέον έχει επηρεάσει πολλές περιοχές του εγκεφάλου καταλήγοντας στον θάνατο του ασθενούς[4], [6].

1.3 Ήπια Γνωστική Διαταραχή

Η Ήπια Γνωστική Διαταραχή (Mild Cognitive Impairment – MCI) είναι μια κατάσταση κατά την οποία το άτομο παρουσιάζει έκπτωση της γνωστικής τους ικανότητας και της εξασθένηση της μνήμης. Έχει αποδειχθεί πώς αν συνάπτει με την υπάρξει βιοδείκτη για το Alzheimer αποτελεί πρώιμο στάδιο αυτού[7]. Συγκεκριμένα περίπου 10%-15% των πασχόντων παρουσιάζει μετάβαση σε Alzheimer κάθε χρόνο. Η ΗΓΔ έχει οριστεί ως το μεταβατικό στάδιο από την κανονική γνωστική λειτουργία του εγκεφάλου στην ανάπτυξη Alzheimer.

Η ΗΓΔ ταξινομείται σε 2 υποκατηγορίες ανάλογα με τα συμπτώματα που παρουσιάζονται. Στην περίπτωση που ΗΓΔ διαταράζει περισσότερο λειτουργίες της μνήμης ονομάζεται αμνησιακή ΗΓΔ .Στην περίπτωση που επηρεάζονται περισσότερο οι γνωστικές ικανότητες του ατόμου τότε ονομάζεται μη αμνησιακή ΗΓΔ. Και στις δύο περιπτώσεις η έκπτωση της γνωστικής ικανότητας και της μνήμης είναι ήπιας μορφής και το άτομο μπορεί να λειτουργήσει κανονικά στην καθημερινότητα του χωρίς την υποστήριξη κάποιου τρίτου.[7], [8]

Η έγκαιρη διάγνωση της ΗΓΔ είναι ύψιστης σημασίας και μπορεί να αποτελέσει μεγάλη πρόκληση. Συγκεκριμένα το σύνδρομο, ως έκφραση νευροεκφυλιστικής ασθένειας, είναι εξαιρετικά περίπλοκο και μπορεί να συνυπάρχει με άλλες νευρολογικές οι ψυχιατρικές παθήσεις, οι οποίες μπορεί να προκαλούν γνωστική εξασθένηση. Επιπλέον μελέτες έχουν δείξει πώς κάποιοι ασθενείς με ΗΓΔ αναπτύσσουν Alzheimer, κάποιοι παραμένουν σταθεροί και κάποιοι επανέρχονται προσωρινά στην φυσιολογική γνωστική κατάσταση αλλά στο μέλλον ενδεχομένως να αναπτύξουν άνοια [9]



Εικόνα 1: (Αριστερά) φυσιολογική κατάσταση, (Κέντρο) ΗΓΔ, (Δεξιά) Alzheimer

1.4 Ο ρόλος της Μηχανικής Μάθησης στην πρόγνωση της νόσου

Η Μηχανική Μάθηση (Machine Learning – ML), που είναι κλάδος της Τεχνητής Νοημοσύνης, μπορεί να παίξει καθοριστικό παράγοντα στην εύρεση τέτοιων μεθόδων. Τα τελευταία χρόνια το πεδίο της μηχανικής μάθησης έχει κεντρίσει το ενδιαφέρον της ερευνητικής κοινότητας στον κλάδο της υγείας λόγω των υψηλών αποδόσεων της στην πρόγνωση και διάγνωση ασθενειών. Στην περίπτωση του Alzheimer η διάγνωση του σε πρώιμο στάδιο είναι μια εξαιρετικά πολύπλοκη διαδικασία που απαιτεί πληθώρα εξετάσεων καθώς την συνδρομή ενός έμπειρου ειδικού. Η μηχανική μάθηση μπορεί να βελτιώσει σημαντικά αυτή την διαδικασία καθώς μπορεί να συμβάλει στην ανάπτυξη συστημάτων υποστήριξης αποφάσεων για την πρόγνωση ή διάγνωση της νόσου ελαττώνοντας σημαντικά το ανθρώπινο σφάλμα [2]. Με την επεξεργαστική δύναμη που είναι διαθέσιμη σήμερα, οι αλγόριθμοι μηχανικής μάθησης μπορούν να επεξεργαστούν μεγάλου όγκου ετερογενή δεδομένα και να εξάγουν κανόνες συσχέτισης, να διακρίνουν σημαντικούς βιοδείκτες, να εντοπίσουν μοτίβα και ανωμαλίες που ο άνθρωπος αδυνατεί να διακρίνει. Τέτοιου είδους δεδομένα είναι τα εργαστηριακά, τα απεικονιστικά, τα βιολογικά, αλλά και δεδομένα σήματος όπως είναι το εγκεφαλογράφημα [10].

Η χρήση της μηχανικής μάθησης για εντοπισμό βιοδεικτών που να υποδηλώνουν την πιθανή εμφάνιση της νόσου είναι υψίστης σημασίας, δεδομένου της μη αναστρεψιμότητας της, ώστε ο ασθενής να μπορεί να λάβει έγκαιρα την κατάλληλη αγωγή με σκοπό την καθυστέρηση της εξάπλωσης της νόσου. Επιπλέον, μια τέτοια εφαρμογή ανοίγει τον δρόμο για ευκολότερες μη παρεμβατικές διαδικασίες με γρηγορότερους ρυθμούς διάγνωσης, λιγότερη ταλαιπωρία για τον ασθενή αλλά και μειωμένο συνολικό κόστος [11].

Η γνώση της ύπαρξης της ΗΓΔ είναι πολύ σημαντική πληροφορία για την πρόγνωση του Alzheimer. Αξιοποιώντας αυτή την πληροφορία, μπορούμε να εκπαιδεύσουμε αλγόριθμους μηχανικής μάθησης για να προβλέψουμε αν η κατάσταση του ασθενούς θα παραμείνει σταθερή στο μέλλον ή αν πρόκειται η ΗΓΔ να εξελιχθεί σε Alzheimer. Αυτό επιτυγχάνεται εκπαιδεύοντας αλγόριθμους ταξινόμησης με δεδομένα προηγούμενων περιπτώσεων όπου μέρος των ασθενών είχε ΗΓΔ που εξελίχθηκε Alzheimer και μέρος τους παρέμεινε σταθερό. Παράλληλα μπορούμε να

αντλήσουμε πληροφορίες για τους βιοδείκτες κάθε ασθενούς και να τους αξιολογήσουμε ποιοτικά ώστε να διακρίνουμε την συμβολή και τον ρόλο του κάθε βιοδείκτη στην εξέλιξη της νόσου [12].

1.5 Σκοπός της παρούσας διατριβής

Η παρούσα διατριβή αποσκοπεί στην ανάπτυξη ενός κλινικού συστήματος υποστήριξη απόφασης σχετικά με την έγκαιρη πρόγνωση της νόσου Alzheimer σε διάστημα 48 μηνών από την πρώτη καταγεγραμμένη εξέταση από ειδικό, όσο ο ασθενής βρίσκεται στο στάδιο της «Ηπιας γνωστικής διαταραχής (Mild Cognitive Impairment – MCI)». Παρόλο που η Ηπία γνωστική διαταραχή είναι δύσκολη στην διάγνωση αποτελεί σημάδι για πιθανή εξέλιξη σε Alzheimer, συνεπώς η ανάπτυξη ενός συστήματος που θα προβλέπει την πιθανή εμφάνιση της νόσου θα αποφέρει τεράστια οφέλη τόσο σε ερευνητικό επίπεδο για τους λόγους που αναφέρθηκαν παραπάνω όσο και σε πρακτικό επίπεδο καθώς μπορεί να συνεισφέρει στις προσπάθειες για έγκαιρη πρόγνωση και πρόληψη της νόσου

Για την ανάπτυξη και εκπαίδευση του αλγόριθμου μηχανικής μάθησης χρησιμοποιήθηκαν δεδομένα από το Alzheimer's Disease Neuroimaging Initiative (ADNI Dataset). Πρόκειται για ένα από τα μεγαλύτερα προγράμματα συλλογής για δημογραφικών, κλινικών, βιολογικών και απεικονιστικών δεδομένων, που καταγράφονται ανά τακτά χρονικά διαστήματα, από ασθενείς διάφορων ομάδων που εγείρουν κίνδυνο να εμφανίσουν Alzheimer.

Στην συνέχεια θα παρουσιαστεί μια διαδικτυακή εφαρμογή που έχει αναπτυχθεί για την εισαγωγή των δεδομένων στο μοντέλο και την λήψη των προβλέψεων από αυτό. Ο γιατρός ή ειδικός θεωρείται ο τελικός χρήστης της εφαρμογής όπου μέσα από αυτήν θα μπορεί να εισάγει τα αποτελέσματα των εξετάσεων του ασθενή και να λαμβάνει την πρόβλεψη του μοντέλου για πιθανή εμφάνιση Alzheimer σε 48 μήνες.

Τα βασικά ερευνητικά ερωτήματα διατυπώνονται ως εξής:

- «Είναι δυνατό να αναπτυχθεί μοντέλο Μηχανικής Μάθησης που να μπορεί να προβλέψει με μεγάλη ακρίβεια την πιθανή εξέλιξη της ΗΓΔ σε νόσο του Alzheimer;»
- «Είναι δυνατό να εξηγήσουμε τον τρόπο που το μοντέλο λαμβάνει αποφάσεις ώστε να θεωρηθούν αξιόπιστες οι προβλέψεις του;»

2. Βιβλιογραφική ανασκόπηση

Στον τομέα της πρόγνωσης και διάγνωσης της νόσου έχουν γίνει οι παρακάτω εργασίες.

Οι A.Hadeer et.al (2022) προτείνουν ένα πλαίσιο που περιλαμβάνει την πολλαπλή ταξινόμηση τεσσάρων σταδίων του φάσματος Alzheimer (AD) και την εκτέλεση ξεχωριστών δυαδικών ταξινομήσεων ιατρικών εικόνων μεταξύ κάθε κατηγορίας δύο ζευγών σταδίων AD. Χρησιμοποιούν δύο μεθόδους για την ταξινόμηση ιατρικών εικόνων και την ανίχνευση της νόσου. Η πρώτη μέθοδος χρησιμοποιεί απλές αρχιτεκτονικές CNN που επεξεργάζονται 2D και 3D MRIs από το σύνολο δεδομένων Alzheimer's Disease Neuroimaging Initiative (ADNI) χρησιμοποιώντας 2D και 3D convolutions. Η δεύτερη μέθοδος εφαρμόζει τις αρχές του transfer learning, χρησιμοποιώντας προ-εκπαιδευμένα μοντέλα όπως το VGG19.

Λόγω της πανδημίας COVID-19, η εργασία προτείνει μια διαδικτυακή εφαρμογή ελέγχου της νόσου Alzheimer που επιτρέπει την εξ αποστάσεως αξιολόγηση της νόσου. Η εφαρμογή αυτή χρησιμοποιεί τις εξειδικευμένες αρχιτεκτονικές CNN και βοηθά τους γιατρούς και τους ασθενείς στον εξ αποστάσεως έλεγχο της ΝΑ και στον προσδιορισμό του σταδίου της ΝΑ του ασθενούς, παρέχοντας τις κατάλληλες συμβουλές με βάση το φάσμα της ΝΑ.

Εννέα μετρικές απόδοσης χρησιμοποιούνται για την αξιολόγηση και τη σύγκριση μεταξύ των δύο μεθόδων. Τα αποτελέσματα υποδεικνύουν διάφορα ευρήματα:

Για την πρώτη μέθοδο (χρήση απλών αρχιτεκτονικών CNN), οι δομές CNN παρουσιάζουν απλότητα, μειώνοντας την υπολογιστική πολυπλοκότητα, τις απαιτήσεις μνήμης και τα προβλήματα overfitting. Οι ακρίβειες είναι 93,61% και 95,17% για ταξινομήσεις 2D και 3D πολλαπλών κατηγοριών σταδίων AD, αντίστοιχα.

Για τη δεύτερη μέθοδο (χρησιμοποιώντας το εκ των προτέρων εκπαιδευμένο μοντέλο VGG19), επιτυγχάνεται ακρίβεια 97% για ταξινομήσεις σταδίου AD πολλαπλών κλάσεων.[13]

Αντίθετα η εργασία των S.Danso et.al (2021) αντιμετωπίζει την πρόκληση της έγκαιρης ανίχνευσης και πρόβλεψης του κινδύνου της νόσου Αλτσχάιμερ (ΝΑ) προτείνοντας ένα πλαίσιο που χρησιμοποιεί συνδυασμό transfer learning και ensemble μεθόδων. Οι ερευνητές στοχεύουν στην ανάπτυξη εξηγήσιμων εξατομικευμένων μοντέλων πρόβλεψης κινδύνου για την άνοια, ξεπερνώντας τους περιορισμούς των υφιστάμενων μοντέλων πρόβλεψης που πάσχουν από φτωχή γενίκευση, έλλειψη επικύρωσης και περιορισμένη κλινική χρησιμότητα.

Το προτεινόμενο πλαίσιο αποτελείται από δύο τύπους μοντέλων πρόβλεψης: μοντέλα πηγής και μοντέλα στόχου. Αρχικά, τα μοντέλα πηγής εκπαιδεύονται και δοκιμάζονται χρησιμοποιώντας ένα δημόσιο διαθέσιμο σύνολο δεδομένων με μεγάλο μέγεθος δείγματος ($n = 84.856$, μέση ηλικία = 69 έτη) και δείγματα παρακολούθησης 14 ετών για την πρόβλεψη του ατομικού κινδύνου άνοιας. Για τη βελτίωση της γενίκευσης, τα όρια απόφασης του καλύτερου μοντέλου πηγής ενημερώνονται χρησιμοποιώντας ένα εναλλακτικό σύνολο δεδομένων από διαφορετικό και νεότερο πληθυσμό ($n = 473$, μέση ηλικία = 52 έτη) για τη δημιουργία του μοντέλου στόχου. Για να εξασφαλιστεί η επεξηγηματικότητα των μοντέλων πρόβλεψης, εφαρμόζεται ο αλγόριθμος SHapely Additive exPlanation (SHAP) για την απεικόνιση των παραγόντων κινδύνου που ευθύνονται για την πρόβλεψη τόσο σε επίπεδο πληθυσμού όσο και σε ατομικό επίπεδο.

Η αξιολόγηση της προτεινόμενης προσέγγισης δείχνει τα εξής αποτελέσματα:

Το καλύτερο μοντέλο πηγής επιτυγχάνει γεωμετρική ακρίβεια 87%, specificity 99% και sensitivity 76% στην πρόβλεψη του κινδύνου άνοιας.

Το μοντέλο-στόχος, το οποίο εκπαιδεύεται με την χρήση transfer learning, ξεπερνά το μοντέλο πηγής σε διάφορες μετρήσεις απόδοσης, με αύξηση της γεωμετρικής ακρίβειας κατά 16,9%, της ειδικότητας κατά 2,7% και της ευαισθησίας κατά 19,1%. Η περιοχή κάτω από τη καμπύλη ROC (Receiver Operating Characteristics -AUC ROC) βελτιώνεται κατά 11% και το ποσοστό αποτελεσματικότητας του transfer learning φτάνει το 20,6%. [14]

Μια άλλη μελέτη των T.Zhang et al (2021) επικεντρώνεται στη χρήση της μηχανικής μάθησης για την ταξινόμηση των διαφόρων σταδίων της νόσου του Alzheimer (AD) συνδυάζοντας μετρήσεις της δομικής μαγνητικής τομογραφίας και της λειτουργικής μαγνητικής τομογραφίας resting state (rs-fMRI). Οι ερευνητές χρησιμοποίησαν χαρακτηριστικά πάχους φλοιού του εγκεφάλου, χαρακτηριστικά δομικού εγκεφαλικού δικτύου και χαρακτηριστικά λειτουργικού εγκεφαλικού δικτύου υποσυχνοτήτων για να ταξινομήσουν τα άτομα εντός της ίδιας κλάσης με σταθερή ΗΓΔ ή ΗΓΔ που θα εξελιχθεί σε Alzheimer (MCInc vs. MCIC) και επίσης μεταξύ διαφορετικών κλάσεων (MCIC vs. AD).

Χρησιμοποιήθηκαν τρεις μέθοδοι επιλογής χαρακτηριστικών, δηλαδή ο αλγόριθμος επιλογής χαρακτηριστικών τυχαίου υποσυνόλου (RSFS), ο αλγόριθμος επιλογής χαρακτηριστικών ελάχιστου πλεονασμού - μέγιστης συνάφειας (MRMR) και ο αλγόριθμος επιλογής χαρακτηριστικών αραιής γραμμικής παλινδρόμησης με βάση τη σταθερή επιλογή (SS-LR), για την επιλογή διακριτικών χαρακτηριστικών στους επαναληπτικούς συνδυασμούς των μετρήσεων MRI και δικτύου. Για την ταξινόμηση χρησιμοποιήθηκε ο Support Vector Machine (SVM) με cross validation

Τα αποτελέσματα της μελέτης δείχνουν ότι ο προτεινόμενος αλγόριθμος RSFS πέτυχε τις καλύτερες ακρίβειες στις ταξινομήσεις MCIC έναντι MCInc και MCIC έναντι AD (84,71% και 89,80%, αντίστοιχα) σε σύγκριση με άλλους αλγορίθμους. Οι περιοχές του εγκεφάλου υψηλής ευαισθησίας που βρέθηκαν στις δύο ομάδες ταξινόμησης ήταν ασυνεπείς. Στην ταξινόμηση MCIC vs. MCInc, οι περιοχές του εγκεφάλου υψηλής ευαισθησίας που σχετίζονταν τόσο με δομικά όσο και με λειτουργικά χαρακτηριστικά περιλάμβαναν μετωπιαία, κροταφική, ουραία, εντερορινική, παραϊπποκάμπειο, και ασβεστολιθική σχισμή και τον περιβάλλοντα φλοιό. Από την άλλη πλευρά, στην ταξινόμηση MCIC vs. AD, οι περιοχές του εγκεφάλου υψηλής ευαισθησίας που σχετίζονταν μόνο με λειτουργικά χαρακτηριστικά περιλάμβαναν τις μετωπιαίες, κροταφικές, θαλαμικές, οσφρητικές και γωνιακές περιοχές.

Τα συμπεράσματα της μελέτης υποδηλώνουν ότι η προτεινόμενη μέθοδος θα μπορούσε να προβλέψει αποτελεσματικά τη μετατροπή από MCI σε AD και η ασυνέπεια συγκεκριμένων εγκεφαλικών περιοχών παρέχει μια νέα εικόνα για την κλινική διάγνωση του Alzheimer. [15]

Η εργασία των A.Singh et.al (2022) παρουσιάζει μια ολοκληρωμένη μελέτη σχετικά με τη χρήση ραδιομορφολογικών χαρακτηριστικών και αλγορίθμων μηχανικής μάθησης για την πρόβλεψη της

νόσου Alzheimer. Οι συγγραφείς εξήγαγαν χαρακτηριστικά από ιατρικές εικόνες με τη χρήση της ραδιομικής. Χρησιμοποίησαν 105 radiomics χαρακτηριστικά για την πρόβλεψη της νόσου Alzheimer και συνέκριναν την απόδοση διαφόρων αλγορίθμων μηχανικής μάθησης, συμπεριλαμβανομένων των Support Vector Machine, K-Nearest Neighbour, Gaussian Naïve Bayes, eXtreme Gradient Boosting (XGBoost) και Random Forest.

Η προτεινόμενη λύση βασίζεται στον αλγόριθμο Random Forest με ραδιομετρικά χαρακτηριστικά όπου πέτυχε ακρίβεια 85% στην πρόβλεψη της νόσου Alzheimer.

Η μέθοδος κατέγραψε τα εξής αποτελέσματα: 88% ακρίβεια, 88% recall, 88% precision και 87% F1-score στην διάκριση μεταξύ AD και CN, 72% ακρίβεια, 73% recall, 72% precision και 71% F1-score για το AD έναντι MCI και 69% ακρίβεια, 69% recall, 68% precision και 69% F1-σκορ για MCI έναντι CN. [16]

Οι Salvatore et al. ανέπτυξαν μια μέθοδο με την χρήση Μηχανικής μάθησης για την έγκαιρη διάγνωση της νόσου χρησιμοποιώντας δείκτες που εξήγαγαν από εικόνες μαγνητικής τομογραφίας. Στο προτεινόμενο μοντέλο ML, χρησιμοποίησαν PCA για την εξαγωγή χαρακτηριστικών από τις εικόνες και SVM για την ταξινόμηση της άνοιας. Είχαν επιτύχει ακρίβεια ταξινόμησης 76% χρησιμοποιώντας ένα σύστημα 20-cross-validation [17]

Η εργασία των T.Sheng et.al (2018) επικεντρώνεται στην πρόβλεψη της πιθανότητας μετάβασης των ατόμων από την Ήπια Γνωστική Διαταραχή (MCI) στη Νόσο Αλτσχάιμερ (AD) εντός ενός έτους. Χρησιμοποιεί τεχνικές βαθιάς μάθησης για την εξαγωγή πολύτιμων χαρακτηριστικών από σαρώσεις μαγνητικής τομογραφίας και χρησιμοποιεί έναν συνδυασμό συνελκτικών νευρωνικών δικτύων (CNN) για την εξαγωγή χαρακτηριστικών και μηχανής διανυσμάτων υποστήριξης (SVM) για την ταξινόμηση. Η έρευνα αναλύει 165 δείγματα από τη βάση δεδομένων του Alzheimer's Disease Neuroimaging Initiative (ADNI), τα οποία αφορούν όλους τους ασθενείς με MCI που τελικά εξελίχθηκαν σε AD σε διαφορετικές χρονικές περιόδους. Το προτεινόμενο μοντέλο επιτυγχάνει ακρίβειες ταξινόμησης 91,0%, 90,0% και 92,3% χρησιμοποιώντας γραμμικές, πολυωνμικές και ακτινικές μεθόδους πυρήνων (RBF), αντίστοιχα. Κατά συνέπεια, η μελέτη υποδηλώνει ότι το μοντέλο που αναπτύχθηκε για την υποστήριξη αποφάσεων υπόσχεται την πρόβλεψη της πιθανότητας μετατροπής από MCI σε-AD εντός χρονικού διαστήματος ενός έτους.[18]

Η εργασία των K.Skolariki et.al (2020) επικεντρώθηκε στην εξεύρεση αποτελεσματικών και αξιόπιστων διαγνωστικών βιοδεικτών για τη νόσο του Alzheimer (AD) με τη χρήση (ημι)αυτοματοποιημένων μεθόδων επεξεργασίας δομικής μαγνητικής τομογραφίας. Η έγκαιρη διάγνωση είναι ζωτικής σημασίας για την αποτελεσματική θεραπεία. Η έρευνα χρησιμοποίησε τεχνικές μηχανικής μάθησης με επίβλεψη για τη διάκριση των μετατρεπόμενων σε MCI (MCIc) από τους μη μετατρεπόμενους (MCInc) και την πρόβλεψη των ποσοστών μετατροπής χρησιμοποιώντας δεδομένα βασικής μαγνητικής τομογραφίας, συγκεκριμένα το πάχος του φλοιού (CTH) και τον όγκο του ιππόκαμπου (HCV).

Η μελέτη περιελάμβανε 803 συμμετέχοντες από το ADNI, που αποτελούνταν από ασθενείς με AD, MCIc, MCInc και υγιείς άτομα. Χρησιμοποιήθηκαν τρεις διαφορετικοί ταξινομητές WEKA (μηχανή διανυσμάτων υποστήριξης, δέντρα απόφασης, Naive Bayes) για τη δημιουργία έξι

μοντέλων ταξινόμησης - τρία με βάση το CTH και τρία με βάση τον HCV (CTH-SVM, CTH-J48, CTH-NB, HCV-SVM, HCV-J48, HCV-NB). Τα πειράματα ταξινόμησης πέτυχαν ευαισθησία έως 71% και ειδικότητα έως 56%. Η ακρίβεια πρόβλεψης της μετατροπής έφτασε το 84%. Ορισμένα πολυμεταβλητά μοντέλα που προέκυψαν από πειράματα ταξινόμησης αναγνώρισαν αποτελεσματικά το MCIc.[19]

Οι Diogo et.al (2022) αξιοποιώντας τις βάσεις δεδομένων AD Neuroimaging Initiative και Open Access Series of Imaging Studies, προτείνουν μια πολυδιαγνωστική προσέγγιση, συνδυάζοντας ταξινομητές μέσω της διαδικασίας ψηφοφορίας (Voting classifiers). Οι ταξινομητές αξιολογούνται αυστηρά ως προς τη γενικευσιμότητά τους σε διαφορετικά σύνολα δεδομένων και πρωτόκολλα απεικόνισης (IR-SPGR και MPRAGE). Ειδικότερα, ο ταξινομητής για το πρόβλημα “Υγιείς (HC) έναντι Alzheimer (AD)” επιτυγχάνει ακρίβεια (BAC – Balanced accuracy) 90,6% και συντελεστή συσχέτισης Matthew (MCC) 0,811. Ο ταξινομητής “HC vs. MCI vs. AD”, που εκπαιδεύτηκε και δοκιμάστηκε σε δεδομένα του ADNI, επιτυγχάνει BAC 62,1% και MCC 0,438. Τα χαρακτηριστικά του ιππόκαμπου επηρεάζουν σημαντικά την ταξινόμηση (25-45%), μαζί με τις κροταφικές, φλοιώδεις και μετωπιαίες περιοχές (περίπου 13% η καθεμία) [20].

Οι El-Sappaght et al (2022) προτείνουν μια μέθοδο που αντιμετωπίζει την πρόκληση της έγκαιρης πρόβλεψης και του ακριβούς χρονοδιαγράμματος της εξέλιξης της νόσου του Alzheimer (NA) από την ήπια γνωστική εξασθένηση (MCI). Η έρευνα προτείνει ένα νέο πλαίσιο βαθιάς μάθησης δύο σταδίων που συγχωνεύει διαχρονικά πολυμεταβλητά δεδομένα, συμπεριλαμβανομένων της νευροαπεικόνισης, των γνωστικών αποτελεσμάτων, των βιοδεικτών και των δημογραφικών στοιχείων. Το πρώτο στάδιο περιλαμβάνει ταξινόμηση πολλαπλών κατηγοριών για τη διάγνωση των ασθενών (φυσιολογικό, MCI, AD), ενώ το δεύτερο στάδιο χρησιμοποιεί παλινδρόμηση για την πρόβλεψη του χρόνου μετατροπής του MCI σε AD. Χρησιμοποιώντας δεδομένα από 1.371 άτομα από την Πρωτοβουλία Νευροαπεικόνισης της Νόσου Alzheimer, συγκρίνουν το προτεινόμενο μοντέλο μακράς-βραχυπρόθεσμης μνήμης (LSTM) με διάφορες προσεγγίσεις μηχανικής μάθησης. Το μοντέλο LSTM υπερέρχει, επιτυγχάνοντας 93,87% ακρίβεια, 94,07% ακρίβεια, 94,07% ανάκληση και 94,07% F1-score στην ταξινόμηση. Για την παλινδρόμηση, το μοντέλο LSTM υπερτερεί σημαντικά έναντι των κανονικών παλινδρομητών με μέσο απόλυτο σφάλμα 0,1375, παρέχοντας μεγαλύτερη σταθερότητα και κλινική βιωσιμότητα. Αυτό το καινοτόμο πλαίσιο υπόσχεται ένα προγνωστικό εργαλείο για την ανίχνευση της εξέλιξης της AD, προσδιορίζοντας με ακρίβεια τον χρόνο μετατροπής σε MCI μέσω της ενσωμάτωσης διαχρονικών δεδομένων των ασθενών.[21]

Οι Grassi et.al (2019) προτείνουν έναν αλγόριθμο μηχανικής μάθησης για τον εντοπισμό ατόμων με MCI που διατρέχουν κίνδυνο εξέλιξης σε AD. Ο αλγόριθμος αναπτύχθηκε χρησιμοποιώντας δεδομένα από την Πρωτοβουλία Νευροαπεικόνισης της Νόσου Alzheimer (ADNI), εστιάζοντας σε 550 άτομα με MCI που παρακολούθηθηκαν για τουλάχιστον 3 έτη. Ο αλγόριθμος επέδειξε ισχυρή απόδοση με AUC ROC 0,88, sensitivity 77,7% και specificity 79,9% σ. Η μελέτη καταλήγει στο συμπέρασμα ότι ο αλγόριθμος που αναπτύχθηκε, αξιοποιώντας εύκολα προσβάσιμες κλινικές πληροφορίες, προσφέρει ένα πολύτιμο εργαλείο για την πρόβλεψη της μετατροπής της AD, ενισχύοντας ενδεχομένως την πρόσληψη σε κλινικές δοκιμές και καθοδηγώντας έγκαιρες παρεμβάσεις για ασθενείς με AD υψηλού κινδύνου [22].

Η μελέτη των E.Ocasio και .Tim Duong παρουσιάζει έναν καινοτόμο αλγόριθμο βαθιάς μάθησης για την πρόβλεψη της εξέλιξης από την ήπια γνωστική εξασθένηση (MCI) στην νόσο Alzheimer (AD) τρία χρόνια μετά τη διάγνωση, χρησιμοποιώντας διαχρονικά και δεδομένα τρισδιάστατης μαγνητικής τομογραφίας ολόκληρου του εγκεφάλου. Η έρευνα περιλαμβάνει 320 άτομα με φυσιολογική κατάσταση(NC), 554 με MCI και 237 με AD. Ο αλγόριθμος χρησιμοποιεί σαρώσεις 3D MRI T1-weighted που συλλέχθηκαν κατά τη διάγνωση του MCI. Αξιοποιεί όγκους μαγνητικής τομογραφίας ολόκληρου του εγκεφάλου χωρίς προηγούμενη περιφερειακή τμηματοποίηση, εκπαιδεύοντας ένα μοντέλο βαθιάς μάθησης με χρήση τεχνικών μάθησης μεταφοράς για την πρόβλεψη της μετατροπής από MCI σε AD. Συγκρίνονται τρεις αρχιτεκτονικές νευρωνικών δικτύων με συνελκτικό τρόπο. Το καλύτερο μοντέλο CNN επιτυγχάνει ισορροπημένη ακρίβεια 0,793 στην πρόβλεψη της μετατροπής από MCI σε AD τρία χρόνια μετά τη διάγνωση. Οι θερμικοί χάρτες αναδεικνύουν τις σχετικές περιοχές του εγκεφάλου. Αυτή η νέα προσέγγιση, χρησιμοποιώντας τρισδιάστατες μαγνητικές τομογραφίες ολόκληρου του εγκεφάλου και βαθιά μάθηση, παρουσιάζει μια πολλά υποσχόμενη μέθοδο για την έγκαιρη πρόβλεψη της εξέλιξης της AD, διευκολύνοντας τη βελτίωση της διαχείρισης της νόσου[23]

Οι E.Papadaki et.al (2022) προτείνουν ένα υβριδικό μοντέλο βαθιάς μάθησης, το CNN-SVM, το οποίο συνδυάζει τα Νευρωνικά Δίκτυα Συνελκτικής Μάθησης (CNN) και τον αλγόριθμο Support Vector Machine (SVM), για την πρόβλεψη των πρώιμων σταδίων της ΝΑ με τη χρήση δεδομένων μαγνητικής τομογραφίας (MRI). Η σύγκριση με μεθόδους που βασίζονται μόνο σε SVM δείχνει ανώτερη αποτελεσματικότητα του CNN-SVM. Οι μετρικές αξιολόγησης καταδεικνύουν την υψηλότερη ακρίβεια του CNN-SVM σε όλες τις φάσεις εκπαίδευσης, επικύρωσης και δοκιμής σε σύγκριση με τις παραδοσιακές αρχιτεκτονικές CNN. Ειδικότερα, το CNN-SVM επιτυγχάνει ακρίβεια 94,37%, ξεπερνώντας τα παραδοσιακά μοντέλα CNN (87,57%) και CNN-LSTM (57,86%).[24]

Οι Fuliang et.al (2023) προτείνουν ένα επεξηγματικό μοντέλο, με χρήση extreme gradient boosting-Shapley additive explanations (XGBoost-SHAP), για τη βελτίωση της ταξινόμησης πολλαπλών κατηγοριών της φυσιολογικής νόησης (NC), της ήπιας γνωστικής εξασθένησης (MCI) και της Alzheimer (AD) χρησιμοποιώντας τη βάση δεδομένων της Πρωτοβουλίας για τη Νευροαπεικόνιση της Νόσου Alzheimer (ADNI). Ενσωματώνοντας κλινικά δεδομένα, αποτελέσματα νευροψυχολογικών δοκιμών, απεικονιστικούς βιοδείκτες και γενετικές πληροφορίες, εφαρμόζει τον αλγόριθμο XGBoost, προσαρμόζοντας την ανισορροπία των κλάσεων. Χρησιμοποιώντας το SHAP, γίνεται ερμηνεία της κάθε απόφασης του μοντέλου καθώς και αξιολόγηση χρήσιμων χαρακτηριστικών κατά την εκπαίδευση. Η μελέτη επιτυγχάνει ανώτερες επιδόσεις ταξινόμησης σε σύγκριση με το Random Forest (RF), το Bagging, το AdaBoost, το naive Bayes (NB) και το μη σταθμισμένο XGBoost, με βελτιώσεις στην ακρίβεια έως και 13,18%. Υψηλό accuracy, specificty, sensitivity και ROC AUC Score καταδεικνύονται τόσο στα σύνολα δεδομένων ADNI όσο και στα σύνολα δεδομένων του Εθνικού Συντονιστικού Κέντρου για τη νόσο Alzheimer (NACC). Τα 10 κορυφαία χαρακτηριστικά που εντοπίζονται μέσω των τιμών SHAP παρουσιάζουν ποικίλες συσχετίσεις με τον κίνδυνο εμφάνισης της νόσου [25].

Η εργασία των Nanni et.al (2020) εξετάζει τις δυνατότητες της βαθιάς μάθησης και της μάθησης μεταφοράς (transfer learning) για την έγκαιρη διάγνωση και πρόγνωση της ΝΑ, συγκρίνοντας την

αποτελεσματικότητά τους με τις συμβατικές μεθόδους Machine Learning (ML). Εξετάστηκαν πάνω από 600 άτομα από το αποθετήριο ADNI, χρησιμοποιώντας εγκεφαλικές σαρώσεις T1-weighted-MRI. Η μελέτη αξιολόγησε ένα σύνολο πέντε αρχιτεκτονικών μάθησης μεταφοράς, ένα τρισδιάστατο συνεπτυγμένο νευρωνικό δίκτυο (CNN) και μια συγχώνευση συμβατικών ταξινομητών ML. Οι εργασίες διάκρισης περιλάμβαναν AD-vs-CN, MCIC-vs-CN και MCIC-vs-MCInc. Το ensemble transfer-learning πέτυχε ισχυρά αποτελέσματα στη διάκριση AD από CN (90,2% AUC), MCIC από CN (83,2% AUC) και MCIC από MCInc (70,6% AUC), παρόμοια ή ελαφρώς χαμηλότερα από τη συγχώνευση των συμβατικών συστημάτων-ML (AD από CN: 93,1% AUC, MCIC από CN: 89,6% AUC, MCIC από MCInc: 69,1-73,3% AUC). Το δίκτυο βαθιάς μάθησης που εκπαιδεύτηκε από το μηδέν είχε λιγότερο αποτελεσματική απόδοση λόγω των περιορισμένων δεδομένων εκπαίδευσης. Αυτά τα ευρήματα υποδηλώνουν την υπόσχεση της μάθησης μεταφοράς και της νευροαπεικόνισης για την αυτοματοποιημένη έγκαιρη διάγνωση και πρόγνωση της AD, ακόμη και όταν προ-εκπαιδεύονται σε γενικές εικόνες.[26]

Οι Zhentao et.al (2023) προτείνουν μια πρωτοποριακή προσέγγιση γνωστή ως VGG-TSwinformer, η οποία συγχωνεύει τα Συνελκτικά Νευρωνικά Δίκτυα (CNN) και τους Transformers για να αντιμετωπίσει την πρόκληση της πρόβλεψης της εξέλιξης της Ήπιας Γνωστικής Διαταραχής (MCI). Αξιοποιώντας τις ικανότητες χωρικής κατανόησης των CNNs, το VGG-TSwinformer εξάγει χαρακτηριστικά χαμηλού επιπέδου από διαδοχικές εικόνες μαγνητικής τομογραφίας εγκεφάλου, τα οποία στη συνέχεια μετατρέπει σε αναπαραστάσεις υψηλού επιπέδου. Αυτή η μεθοδολογία χρησιμοποιεί την μέθοδο συρόμενου παραθύρου (sliding window), συγχωνεύοντας σχολαστικά γειτονικές αναπαραστάσεις χαρακτηριστικών για να συλλάβει λεπτές χωρικές αλλαγές κατά τη διάρκεια του χρονοδιαγράμματος εξέλιξης. Η αξιολόγηση του μοντέλου στο σύνολο δεδομένων ADNI, επιτυγχάνει ακρίβεια ταξινόμησης (77,2%), sensitivity (79,97%), specificity (71,59%) και ROC AUC (0,8153%) κατά τη διάκριση μεταξύ διαφορετικών σταδίων MCI. [27].

Οι Helaly et.al 2022 προτείνουν την ανάπτυξη ενός ολοκληρωμένου πλαισίου για την έγκαιρη ανίχνευση της νόσου του Alzheimer (AD) και την ταξινόμηση ιατρικών εικόνων με τη χρήση βαθιάς μάθησης, και συγκεκριμένα με τη χρήση συνελκτικών νευρωνικών δικτύων (CNN). Η έμφαση δίνεται στην ταξινόμηση των τεσσάρων σταδίων του φάσματος της ΝΑ και στην εκτέλεση δυαδικών ταξινομήσεων μεταξύ γειτονικών κατηγοριών σταδίων της νόσου. Χρησιμοποιούνται δύο μέθοδοι: η πρώτη χρησιμοποιεί απλές αρχιτεκτονικές CNN για 2D και 3D δομικές εγκεφαλικές σαρώσεις από το σύνολο δεδομένων Alzheimer's Disease Neuroimaging Initiative (ADNI), ενώ η δεύτερη χρησιμοποιεί μάθηση μεταφοράς με το προ-εκπαιδευμένο μοντέλο VGG19. Η μελέτη εισάγει επίσης μια διαδικτυακή εφαρμογή ελέγχου της νόσου Αλτσχάιμερ για την εξ αποστάσεως αξιολόγηση της, και την παροχή καθοδήγησης ανάλογα με το στάδιο. Η αξιολόγηση δείχνει ότι οι αρχιτεκτονικές CNN επιτυγχάνουν ακρίβειες 93,61% και 95,17% για ταξινομήσεις σταδίων AD 2D και 3D πολλαπλών κατηγοριών. Το μοντέλο VGG19 επιτυγχάνει ακρίβεια 97% για ταξινόμηση σταδίων AD [28].

Οι Pirrone et.al 2022 παρουσιάζουν μια νέα, απλή και αποτελεσματική μέθοδο για την εξαγωγή χαρακτηριστικών που χρησιμοποιεί φίλτρα πεπερασμένης απόκρισης (FIR) στο διπλό πεδίο του

χρόνου από σήματα EEG. Ο πρωταρχικός στόχος είναι η διάκριση μεταξύ ατόμων που πάσχουν από AD, MCI και υγιών ελέγχων (HC).

Η προτεινόμενη μέθοδος υπολογίζει την ένταση ισχύος τόσο για τις ζώνες υψηλής όσο και για τις ζώνες χαμηλής συχνότητας, χρησιμοποιώντας τις απόλυτες διαφορές μεταξύ αυτών των τιμών για τη διαφοροποίηση μεταξύ των τριών κατηγοριών ατόμων. Η διαφοροποίηση αυτή επιτυγχάνεται μέσω της εφαρμογής διαφόρων τεχνικών μηχανικής μάθησης. Η έρευνα χρησιμοποιεί καταγραφές HEG από μια ομάδα 105 ατόμων που υποβλήθηκαν σε αξιολόγηση της άνοιας στο IRCCS Centro Neurolesi "Bonino-Pulejo" στη Messina της Ιταλίας.

Τα ποσοστά ακρίβειας ανέρχονται σε 97%, 95% και 83% για δυαδικές ταξινομήσεις (HC vs. AD, HC vs. MCI και MCI vs. AD) και στο 75% όταν πρόκειται για τις τρεις κατηγορίες (HC vs. AD vs. MCI).[29]

3. Alzheimer's Disease Neuroimaging Initiative (ADNI)

3.1 Περιγραφή της μελέτης

Το ADNI (Alzheimer's Disease Neuroimaging Initiative) είναι μια παγκόσμια ερευνητική μελέτη με στόχο τη διερεύνηση και την ανάπτυξη θεραπειών για την επιβράδυνση ή την ανακοπή της εξέλιξης της νόσου του Alzheimer (AD). Πρόκειται για μια διαχρονική μελέτη που διεξάγεται σε 63 τοποθεσίες στις Ηνωμένες Πολιτείες και τον Καναδά και παρακολουθεί την εξέλιξη της νόσου Alzheimer στον ανθρώπινο εγκέφαλο από τη φυσιολογική γήρανση έως τα πρώιμα και τα όψιμα στάδια της ήπιας γνωστικής διαταραχής (MCI) και της άνοιας. Ο πρωταρχικός στόχος του ADNI είναι η διερεύνηση βιοδεικτών για χρήση σε κλινικές δοκιμές θεραπείας για της νόσου.

Η μελέτη έχει σημαντικό παγκόσμιο αντίκτυπο μέσω της ανάπτυξης τυποποιημένων πρωτοκόλλων που επιτρέπουν τη σύγκριση των αποτελεσμάτων σε πολλά ερευνητικά κέντρα. Επιπλέον, η πολιτική ανταλλαγής δεδομένων διαδραματίζει καθοριστικό ρόλο, παρέχοντας όλα τα δεδομένα του ADNI χωρίς καμία καθυστέρηση σε ειδικευμένους ερευνητές σε όλο τον κόσμο. Η προσέγγιση αυτή έχει οδηγήσει σε περισσότερες από 1000 επιστημονικές δημοσιεύσεις που έχουν χρησιμοποιήσει δεδομένα ADNI.

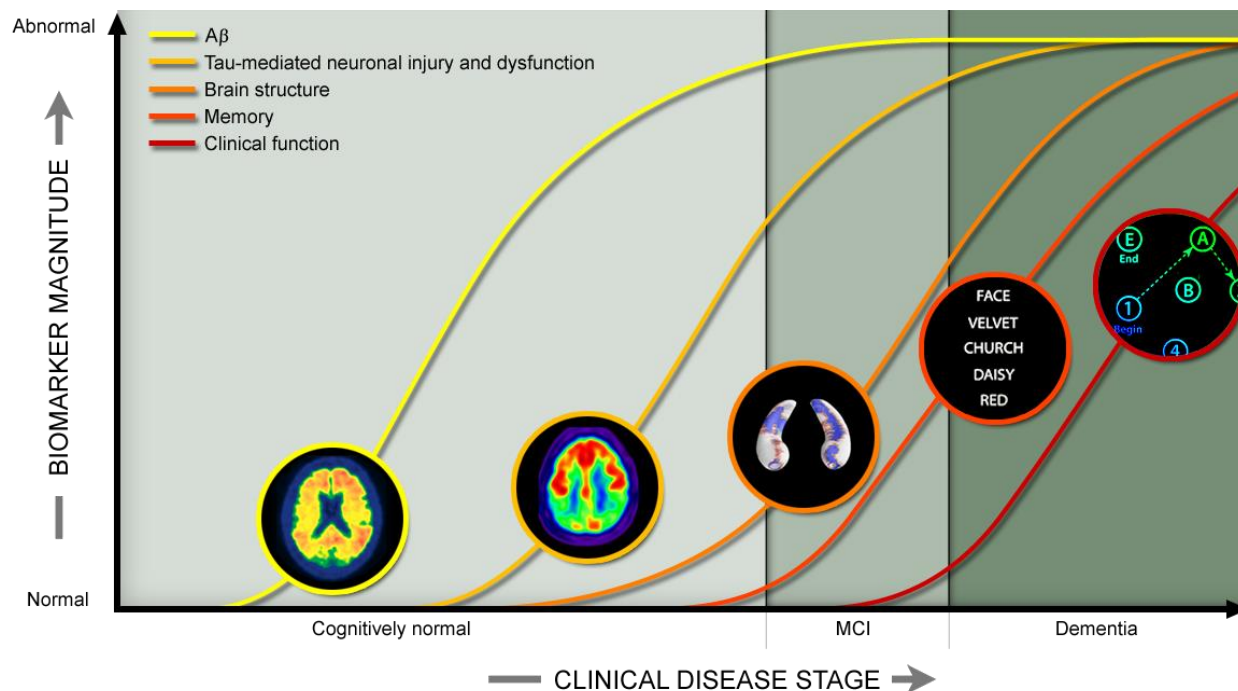
Τα δεδομένα που παράγονται από τους ερευνητές της μελέτης ADNI αποθηκεύονται σε ένα αποθετήριο δεδομένων που ονομάζεται LONI Image & Data Archive (IDA), το οποίο φιλοξενείται στο Εργαστήριο Νευροαπεικόνισης (LONI) του Πανεπιστημίου της Νότιας Καλιφόρνιας. Οι ειδικευμένοι ερευνητές μπορούν να υποβάλλουν ηλεκτρονικά αιτήματα πρόσβασης σε δεδομένα και συνήθως αποκτούν πρόσβαση σε δεδομένα ADNI, συμπεριλαμβανομένων των γνωστικών, απεικονιστικών, γενετικών και βιορευστών, εντός λίγων ημερών από την υποβολή του αιτήματός τους.

Εκτός από τον αντίκτυπο στην έρευνα για τη νόσο Αλτσχάιμερ, το ADNI έχει χρησιμεύσει ως πρότυπο για άλλες πρωτοβουλίες που σχετίζονται με τη νόσο και άλλες ασθένειες. Έχει συνεισφέρει δεδομένα σε διάφορες κοινοπραξίες και έργα μεγάλων δεδομένων, όπως η κοινοπραξία Enhancing Neuro Imaging Genetics through Meta Analysis (ENIGMA) και η Dialogue on Reverse Engineering Assessment and Methods (DREAM) Alzheimer's Disease Big Data Challenge. Αυτές οι συνεργασίες έχουν τη δυνατότητα να αποκαλύψουν πολύτιμες γνώσεις για τις νευρολογικές ασθένειες και να ανακαλύψουν νέους προγνωστικούς βιοδείκτες για το Alzheimer.[30]

Τρεις πρωταρχικοί στόχοι της μελέτης είναι οι εξής:

1. Η ανίχνευση της νόσου στο νωρίτερο δυνατό στάδιο (προ-άνοια) και ο εντοπισμός τρόπων παρακολούθησης της εξέλιξης της με βιοδείκτες.
2. Να υποστηριχθούν οι πρόοδοι στην παρέμβαση, την πρόληψη και τη θεραπεία της νόσου Alzheimer μέσω της εφαρμογής νέων διαγνωστικών μεθόδων στα πρώτα δυνατά στάδια
3. Να διαχειρίζεται συνεχώς την καινοτόμο πολιτική πρόσβασης στα δεδομένα του ADNI, η οποία παρέχει όλα τα δεδομένα χωρίς εμπόριο σε όλους τους επιστήμονες στον κόσμο

3.2 Βιοδείκτες



Εικόνα 2 Το γράφημα απεικονίζει τους βιοδείκτες ως δείκτες της ΝΑ. Οι καμπύλες δείχνουν τις αλλαγές σε πέντε βιοδείκτες από φυσιολογικούς σε μη φυσιολογικούς κατά την πορεία της ΝΑ (φυσιολογική νόσηση έως άνοια).

Η Εικόνα 2 απεικονίζει τους βιοδείκτες ως δείκτες του Alzheimer. Οι καμπύλες δείχνουν τις αλλαγές σε πέντε βιοδείκτες από φυσιολογικούς σε μη φυσιολογικούς κατά την πορεία της νόσου (φυσιολογική λειτουργία έως και άνοια).

1. Το β-αμυλοειδές (b-amyloid) που μετριέται στο εγκεφαλονωτιαίο υγρό ή με απεικόνιση PET αμυλοειδούς
2. Ο Νευροεκφυλισμός που υποδεικνύεται από την πρωτεΐνη tau, μετρούμενη στο εγκεφαλονωτιαίο υγρό, ή από συναπτική δυσλειτουργία, μετρούμενη με FDG-PET
3. Ατροφία του εγκεφάλου, κυρίως στον μέσο κροταφικό λοβό, μετρούμενη με δομική μαγνητική τομογραφία
4. Απώλεια μνήμης, μετρούμενη με γνωστικές δοκιμασίες
5. Κλινική λειτουργία, που υποδεικνύεται από γενική γνωστική έκπτωση, μετρούμενη με γνωστικές δοκιμασίες.

Οι μεταβολές 1-3 υποδεικνύονται από βιοδείκτες που μπορούν να παρατηρηθούν πριν από τη διάγνωση της άνοιας, ενώ οι μεταβολές 4-5 είναι οι κλασικοί δείκτες διάγνωσης της άνοιας.

3.3 Φάσεις της μελέτης

Το ADNI ως διαχρονική μελέτη που βρίσκεται σε εξέλιξη έως, και την ημέρα συγγραφής της παρούσας εργασίας, χωρίζεται σε 4 φάσεις (στάδια). Κάθε φάση του ADNI βασίζεται στην προηγούμενη, επεκτείνοντας το πεδίο συλλογής δεδομένων και βελτιώνοντας τα ερευνητικά πρωτόκολλα. Μέσω αυτών των φάσεων, το ADNI συγκέντρωσε πληθώρα διαχρονικών δεδομένων, επιτρέποντας στους ερευνητές να παρακολουθούν την εξέλιξη της νόσου, να εντοπίζουν βιοδείκτες και να διερευνούν πιθανές θεραπευτικές οδούς.

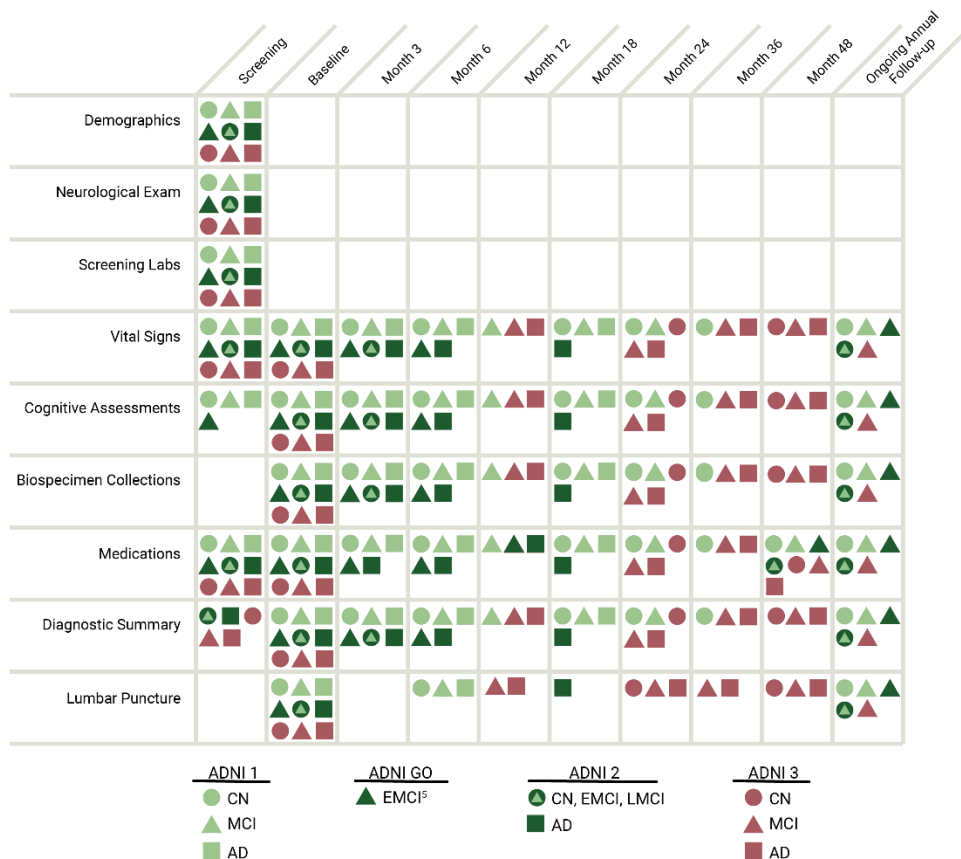
Στην Εικόνα 3 απεικονίζονται συνοψισμένα οι στόχοι κάθε φάσης του ADNI καθώς και ο αριθμός των συμμετεχόντων.

STUDY CHARACTERISTICS	ADNI-1	ADNI-GO (Grand Opportunities)	ADNI-2	ADNI-3
Primary goal	Develop biomarkers as outcome measures for clinical trials	Examine biomarkers in earlier stages of disease	Develop biomarkers as predictors of cognitive decline, and as outcome measures	Study the use of tau PET and functional imaging techniques in clinical trials
Funding	\$40 million federal (NIA), \$27 million industry and foundation	\$24 million American Recovery Act funds	\$40 million federal (NIA), \$27 million industry and foundation	\$ 40 million federal (NIA), up to \$20 million industry and foundation
Duration/start date	5 years/October 2004	2 years/September 2009	5 years/September 2011	5 years/September 2016
Cohort	200 elderly controls 400 MCI 200 AD	Existing ADNI-1 + 200 early MCI	Existing ADNI-1 and ADNI-GO + 150 elderly controls 150 late MCI 150 AD	Existing ADNI-1, ADNI-GO, ADNI-2 + 133 elderly controls 151 MCI 87 AD

Εικόνα 3: Σύνοψη φάσεων ADNI

Κατά τη διάρκεια των τεσσάρων φάσεων του ADNI, όπου ήταν δυνατό, οι συμμετέχοντες μεταφέρθηκαν από τις προηγούμενες φάσεις για συνέχισή της παρακολούθησής τους, ενώ σε κάθε φάση προστέθηκαν νέοι συμμετέχοντες για την περαιτέρω διερεύνηση της εξέλιξης της νόσου Alzheimer

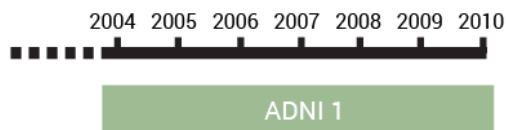
Στην Εικόνα 4 απεικονίζεται συγκεντρωτικά το πρόγραμμα των κλινικών δοκιμών για όλες τις φάσεις του ADNI



Εικόνα 4: Πρόγραμμα κλινικών δοκιμών του ADNI

3.3.1 ADNI 1

Η αρχική φάση του ADNI ξεκίνησε τον Οκτώβριο του 2004, με χρηματοδότηση ύψους 67 εκατομμυρίων δολαρίων από το Εθνικό Ινστιτούτο Γήρανσης, 13 φαρμακευτικές εταιρείες και δύο ιδρύματα. Πρωταρχικός στόχος της ήταν ο εντοπισμός ακριβέστερων και πιο ευαίσθητων βιοδεικτών για την έγκαιρη ανίχνευση και παρακολούθηση του Alzheimer. Η μελέτη ADNI1 συνέλεξε και ανέλυσε τεράστιο όγκο δεδομένων, συμπεριλαμβανομένων χιλιάδων εγκεφαλικών σαρώσεων, γενετικών προφίλ και βιοδεικτών αίματος και εγκεφαλονωτιαίου υγρού.



Εικόνα 5: Χρονολόγιο ADNI-1

Η μελέτη περιλάμβανε συνολικά 800 συμμετέχοντες, αποτελούμενη από 400 άτομα που είχαν διαγνωστεί με ήπια γνωστική εξασθένηση (MCI), 200 άτομα στα αρχικά στάδια της AD και 200 ηλικιωμένα άτομα που χρησίμευαν ως άτομα ελέγχου. Κατά τη διάρκεια της ADNI1 και των επόμενων φάσεων, οι ερευνητές χρησιμοποίησαν διάφορες τεχνικές απεικόνισης του εγκεφάλου,

όπως δομική μαγνητική τομογραφία και τομογραφία εκπομπής ποζιτρονίων (PET). Οι σαρώσεις PET περιελάμβαναν τη χρήση FDG-PET, η οποία μετρά τον μεταβολισμό της γλυκόζης στον εγκέφαλο, και μια πιλοτική μελέτη που χρησιμοποιούσε PET αμυλοειδούς με μια ραδιενεργό ένωση που ονομάζεται "Pittsburgh Compound B" για την αξιολόγηση της συσσώρευσης αμυλοειδούς στον εγκέφαλο.

Πίνακας 1: Συμμετέχοντες ADNI-1

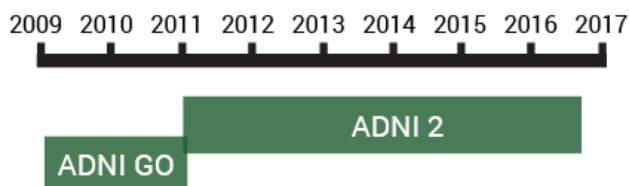
Ανάλυση συμμετεχόντων	
Cognitive Normal	200
MCI	400
AD	200

Στόχοι της 1^η φάσης είναι:

1. Η ανάπτυξη βελτιωμένων μεθόδων που δημιουργούν ομοιόμορφα πρότυπα για την απόκτηση διαχρονικών, πολυτοπικών δεδομένων MRI και PET σε ασθενείς με νόσο Alzheimer (AD), ήπια γνωστική εξασθένηση (MCI) και ηλικιωμένους μάρτυρες.
2. Η απόκτηση ενός γενικά προσβάσιμου αποθετηρίου δεδομένων που περιγράφει τις διαχρονικές αλλαγές στη δομή και το μεταβολισμό του εγκεφάλου, καθώς και κλινικά/γνωστικά δεδομένα και δεδομένα βιοδεικτών για την επικύρωση των υποκατάστατων απεικόνισης.
3. Η ανάπτυξη μεθόδων που θα παρέχουν τη μέγιστη δυνατή ισχύ για τον προσδιορισμό των αποτελεσμάτων της θεραπείας σε δοκιμές που αφορούν αυτούς τους ασθενείς.
4. Η δοκιμή μιας σειράς υποθέσεων με βάση τα κλινικά δεδομένα και τα δεδομένα βιοδεικτών.

3.3.2 ADNI GO

Το ADNI επεκτάθηκε το 2009 κατά τη διάρκεια της φάσης ADNI GO, η οποία αξιολόγησε την υπάρχουσα ομάδα ADNI 1 μαζί με 200 νέους συμμετέχοντες με πρώιμη ήπια γνωστική εξασθένηση (EMCI). Στόχος αυτής της φάσης ήταν να εξεταστούν βιοδείκτες σε ένα πρώιμο στάδιο της νόσου. Τα πρωτόκολλα μαγνητικής τομογραφίας προσαρμόστηκαν επίσης στη φάση ADNI GO.



Εικόνα 6: Χρονολόγιο ADNI-GO και ADNI-2

Πίνακας 2: Συμμετέχοντες ADNI-GO

Ανάλυση συμμετεχόντων	
Cognitive Normal and MCI (συνέχεια από ADNI1))	500
EMCI (new)	200

3.3.3 ADNI 2

Το 2011 ξεκίνησε η δεύτερη του ADNI2 με πρόσθετη χρηματοδότηση ύψους 67 εκατομμυρίων δολαρίων. Αυτή η φάση διέυρυνε τη δεξαμενή συμμετεχόντων, συμπεριλαμβάνοντας άτομα από τις φάσεις ADNI1 και ADNI GO, καθώς και εισάγοντας νέες ομάδες συμμετεχόντων. Αυτές οι νέες ομάδες αποτελούνταν από 150 ηλικιωμένα άτομα ελέγχου, 100 άτομα με πρώιμη ήπια γνωστική εξασθένηση (EMCI), 150 άτομα με όψιμη ήπια γνωστική εξασθένηση (LMCI) και 150 άτομα με διάγνωση ήπιας νόσου Alzheimer (AD).

Το ADNI2 εισήγαγε επίσης μια νέα ομάδα που ονομάστηκε Significant Memory Concern (SMC) για να γεφυρώσει το χάσμα μεταξύ των υγιών ατόμων και των ατόμων με ήπια γνωστική εξασθένηση. Τα κριτήρια ένταξης για τους συμμετέχοντες στην κατηγορία SMC απαιτούσαν την αυτοαναφορά σημαντικών ανησυχιών για τη μνήμη. Συνολικά 107 άτομα συμπεριλήφθηκαν σε αυτή την ομάδα.

Μια σημαντική προσθήκη στο ADNI2 ήταν η ενσωμάτωση σαρώσεων PET αμυλοειδούς με τη χρήση Florbetapir σε όλες τις περιοχές του ADNI2 και για όλους τους συμμετέχοντες στο ADNI2 και το ADNI GO. Αυτή η τεχνική απεικόνισης επέτρεψε την αξιολόγηση της συσσώρευσης αμυλοειδούς στον εγκέφαλο, παρέχοντας πολύτιμες πληροφορίες για την παθολογία της νόσου του Alzheimer.

Στόχοι της 2^η φάσης είναι:

1. Ο καθορισμός των σχέσεων μεταξύ των κλινικών, απεικονιστικών, γενετικών και βιοχημικών χαρακτηριστικών βιοδεικτών ολόκληρου του φάσματος της νόσου του Alzheimer (NA) καθώς εξελίσσεται η παθολογία.
2. Η ενημέρωση της κοινότητας της νευροεπιστήμης της νόσου, προσδιορισμός διαγνωστικών και προγνωστικών δεικτών και μέτρων έκβασης που μπορούν να χρησιμοποιηθούν σε κλινικές δοκιμές και βοήθεια στην ανάπτυξη των πιο αποτελεσματικών σεναρίων κλινικών δοκιμών.
3. Η ανάπτυξη ομοιόμορφων προτύπων για την απόκτηση διαχρονικών πολυτοπικών δεδομένων MRI και PET σε ασθενείς με AD, MCI και ηλικιωμένους μάρτυρες.
4. Η εκτέλεση διαχρονικών κλινικών, γνωστικών, MRI, PET (18F-Florbetapir και FDG), καθώς και μελετών βιοδεικτών αίματος και ENY σε 550 νεοεισερχόμενα άτομα, επιπλέον της συνέχισης αυτών των μελετών για περίπου 700 άτομα από τα ADNI1 και ADNI GO για επιπλέον 5 έτη.
5. Η συλλογή δειγμάτων αίματος για την εξαγωγή DNA και RNA. Τα νεοεισαχθέντα άτομα σε αυτή την φάση θα έχουν επίσης δείγματα που θα συλλεχθούν για κυτταρική αθανασία και γονοτυπία APOE.

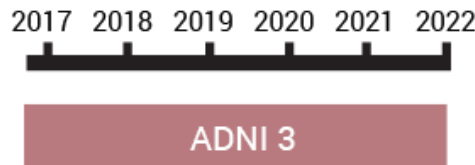
6. Η επικύρωση των κλινικών διαγνώσεων και των υποκατάστατων απεικόνισης και βιοδεικτών μέσω της νευροπαθολογικής εξέτασης των συμμετεχόντων ADNI1, ADNI GO και ADNI2 που προσέρχονται για αυτοψία.
- 7.

Πίνακας 3: Συμμετέχοντες ADNI-2

Ανάλυση συμμετεχόντων	
Cognitive Normal	150
Cognitive Normal and MCI rollover from ADNI 1	450-500
EMCI	150
EMCI (ADNI GO rollover)	200
LMCI	150
Mild AD	200

3.3.4 ADNI 3

Το 2016 ξεκίνησε η τρίτη φάση του ADNI με διευρυμένο στόχο τη διερεύνηση των αλληλεπιδράσεων μεταξύ των κλινικών, γνωστικών, απεικονιστικών, γενετικών και βιοχημικών χαρακτηριστικών βιοδεικτών σε όλο το φάσμα της νόσου Alzheimer (NA). Το ADNI3 εισήγαγε μια σημαντική προσθήκη στα πρωτόκολλα απεικόνισης με την ενσωμάτωση εγκεφαλικών σαρώσεων που μπορούν να ανιχνεύσουν τις συσσωματώσεις της πρωτεΐνης tau, γνωστές ως tau PET. Οι εν λόγω δεσμίδες χρησιμεύουν ως κρίσιμος δείκτης της εξέλιξης και της παθολογίας της νόσου.



Εικόνα 7: Χρονολόγιο ADNI-3

Επιπλέον, το ADNI3 επικεντρώνεται στη συνεχή ανακάλυψη, βελτιστοποίηση, τυποποίηση και επικύρωση μέτρων κλινικών δοκιμών και βιοδεικτών που χρησιμοποιούνται στην έρευνα για τη νόσο Alzheimer.

Το ADNI3 περιλαμβάνει τη συνεργασία επιστημόνων από 59 ερευνητικά κέντρα στις Ηνωμένες Πολιτείες και τον Καναδά. Στόχος της φάσης είναι να συμπεριλάβει εκατοντάδες νέους συμμετέχοντες με ήπια γνωστική εξασθένηση (MCI), ήπια AD και ηλικιωμένα άτομα που θα χρησιμεύσουν ως μάρτυρες. Αυτή η επέκταση του αριθμού των συμμετεχόντων θα προσφέρει μια ευρύτερη και πιο ολοκληρωμένη κατανόηση της νόσου Αλτσχάιμερ σε όλα τα διάφορα στάδιά της.

Στόχοι της 3^η φάσης του ADNI είναι:

1. Ο προσδιορισμός και καθορισμός των μετρήσεων της γνώσης και της λειτουργίας, συμπεριλαμβανομένων των σύνθετων μετρήσεων, και των μετρήσεων βιοδεικτών, οι οποίες καταγράφουν τη διαχρονική αλλαγή με τη μεγαλύτερη στατιστική ισχύ για την ανίχνευση των επιδράσεων της θεραπείας σε κλινικές δοκιμές. Η διαχρονική μεταβολή του εγκεφαλικού tau που μετράται με 18F-AV-1451 PET (AV-1451) θα συσχετιστεί/συγκριθεί με άλλες μετρήσεις.
2. Ο καθορισμός των κλινικών, γνωστικών και βιοδεικτικών μέτρων που προβλέπουν καλύτερα την έκπτωση της γνώσης σε συμμετέχοντες με CN, MCI και AD. Επιπλέον, προσδιορίστε ποιες μεταβολές βιοδεικτών συσχετίζονται με τη γνωστική έκπτωση, με έμφαση στο AV-1451 PET.
3. Η επικύρωση των μετρήσεων βιοδεικτών που λαμβάνονται κατά την έναρξη και διαχρονικά με συσχέτιση των αποτελεσμάτων με τις κλινικές μετρήσεις "χρυσού προτύπου" και την παθολογία.
4. Ο καθορισμός των βέλτιστων μέτρων έκβασης με προσοχή στη γνωστική έκπτωση και το AV-1451 PET (tau PET) , των προγνωστικών παραγόντων της γνωστικής έκπτωσης και των κριτηρίων ένταξης/αποκλεισμού για κλινικές δοκιμές με γνωστικά φυσιολογικούς συμμετέχοντες (για δευτερογενείς προκλινικές δοκιμές AD), ασθενείς με MCI (για δοκιμές προδρομικής AD) και συμμετέχοντες με πρόωμη άνοια λόγω AD.
5. Ο προσδιορισμός των επιδράσεων άλλων γνωστών πρωτεϊνών της νόσου που βρίσκονται σε εγκεφάλους και γονίδια της AD, καθώς και νεοανακαλυφθέντων γονιδίων, πρωτεϊνών και αναλυτών που παρέχουν χρήσιμες πληροφορίες σχετικά με την παθογένεια/διάγνωση της AD.

Πίνακας 4: Συμμετέχοντες ADNI-3

Ανάλυση συμμετεχόντων	
Cognitive Normal (new)	135-500
Cognitive Normal rollover from ADNI 2	295-330
MCI (new)	150-515
MCI Rollover from ADNI 2	275-320
Mild AD (new)	85-185
Mild AD rollover from ADNI 2	130-150

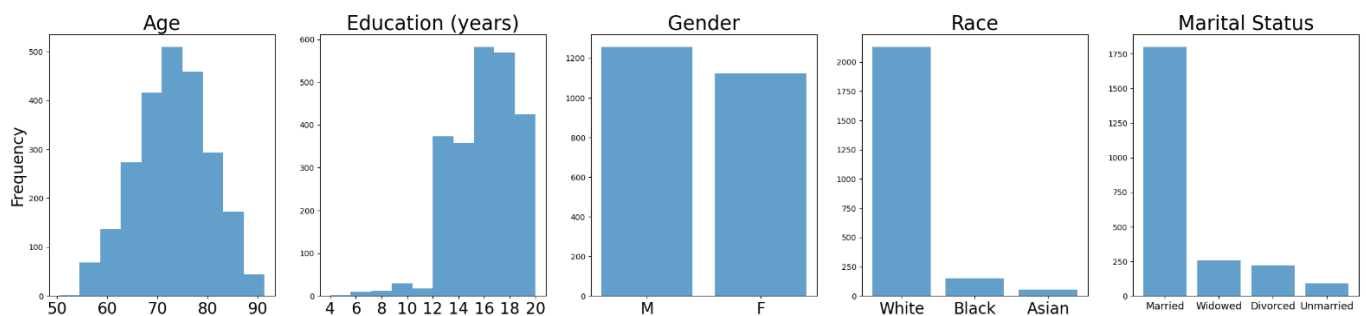
4. Διερευνητική Ανάλυση Δεδομένων – Exploratory Data Analysis

Για τους σκοπούς της παρούσας διατριβής πραγματοποιήθηκε μια βασική διερευνητική ανάλυση των δεδομένων στο ADNIMERGE dataset με σκοπό την καλύτερη κατανόηση του συνόλου δεδομένων και την εύρεση μοτίβων ή σχέσεων που ενδεχομένως να παρατηρείται μέσα σε αυτό. Επιπλέον στόχος της ανάλυσης ήταν η αξιολόγηση της ποιότητας των δεδομένων, ο εντοπισμός τυχόν ανωμαλιών ή ζητημάτων που ενδέχεται να επηρεάσουν την εγκυρότητα της μετέπειτα χρήσης και επεξεργασίας και η βαθύτερη κατανόηση της υποκείμενης δομής των δεδομένων.

Το σύνολο δεδομένων αποτελείται από 2378 μοναδικούς ασθενείς και καταγράφει για αυτούς πολλαπλές επισκέψεις σε βάθος χρόνου. Συνολικά στην αρχική του μορφή περιέχει πάνω από δέκα χιλιάδες καταγραφές για όλες τις φάσεις του ADNI.

Επιπλέον έχει 116 χαρακτηριστικά εκ των οποίων τα 50 είναι τα χαρακτηριστικά της πρώτης επίσκεψης (baseline) τα οποία παραμένουν σταθερά για κάθε επίσκεψη.

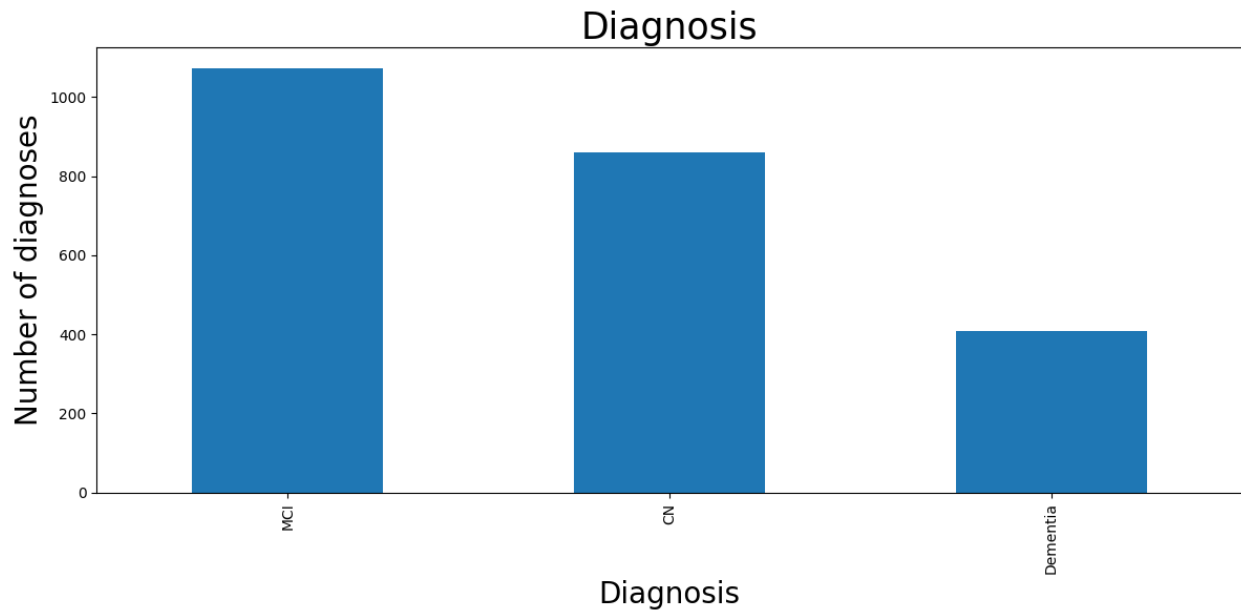
4.1 Δημογραφικά χαρακτηριστικά



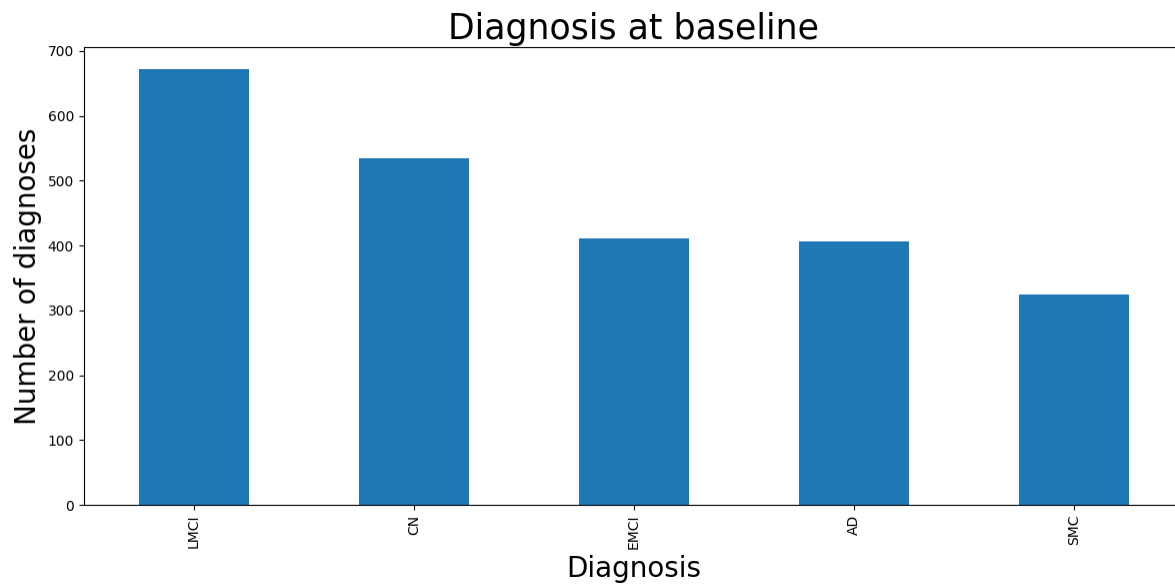
Εικόνα 8: Συγκεντρωτικό γράφημα δημογραφικών στοιχείων του ADNIMERGE

Όπως παρατηρούμε στην Εικόνα 8 η πλειοψηφία των συμμετεχόντων είναι ηλικιακά μεταξύ περίπου των 67-85 ετών χωρίς δείγματα ηλικίας κάτω των 55 και άνω των 90, ενώ υπερτερούν ελάχιστα οι άντρες στο δείγμα σε σχέση με τις γυναίκες. Επιπλέον οι περισσότεροι συμμετέχοντες έχουν πάνω από 14 χρόνια εκπαίδευσης ενώ ταυτόχρονα είναι παντρεμένοι.

4.2 Διαγνώσεις



Εικόνα 9: Αριθμός δειγμάτων κάθε διάγνωσης(κατηγορίας)

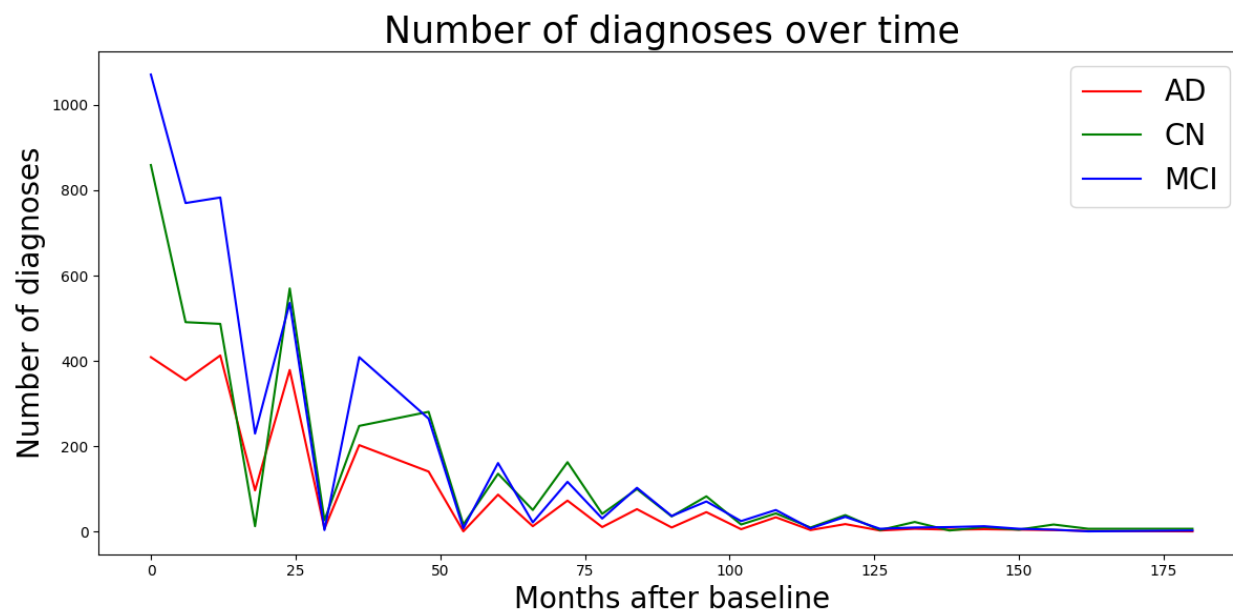


Εικόνα 10: Αριθμός δειγμάτων διαγνώσεων στο baseline στάδιο

Όπως παρατηρούμε στην Εικόνα 9 οι πλειοψηφία των συμμετέχων έχει ήδη διαγνωστεί με ΗΓΔ επομένως είχαν ήδη συμπτώματα της νόσου όταν ξεκίνησα να συμμετέχουν στις κλινικές μελέτες. Ακολουθεί δεύτερο πληθυσμιακά το Control group, δηλαδή οι συμμετέχοντες που δεν είχαν κάποιο σύμπτωμα που να υποδηλώνει ΗΓΔ η Alzheimer. Τέλος ακολουθούν οι συμμετέχοντες με Alzheimer που είναι λογικό καθώς σκοπός της μελέτης είναι η παρακολούθηση της εξέλιξης της

νόσου πριν τα τελικά στάδια της ,συνεπώς είναι αναμενόμενο να είναι χαμηλό το πλήθος συμμετεχόντων που έχουν ήδη Alzheimer .

Για την στήλη DX_bl χρησιμοποιείται διαφορετικός τρόπος καταγραφής της ΗΓΔ σε επιμέρους στάδια. Όπως φαίνεται στην Εικόνα 10 η ΗΓΔ(MCI) έχει διαχωριστεί σε 3 στάδια : EMCI, SMC και LMCI. Για τους σκοπούς της παρούσας διατριβής θα χρησιμοποιηθεί η πρώτη μέθοδος καταγραφής καθώς επικρατεί και στις εγγραφές των μετέπειτα επισκέψεων.

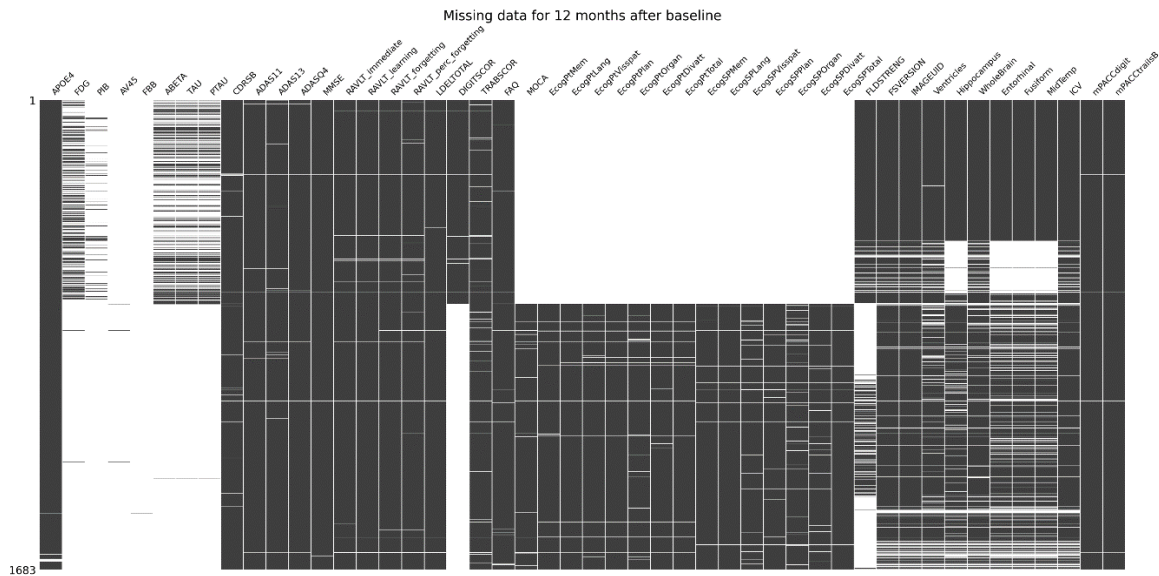


Εικόνα 11:Αρίθμος διαγνώσεων συναρτήσει του χρόνου

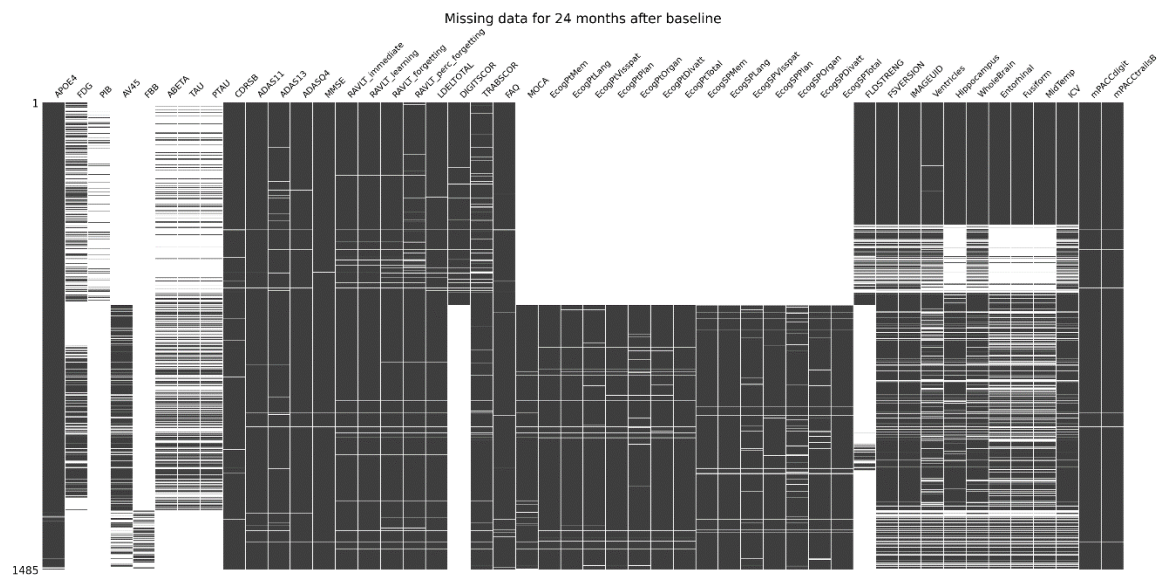
Στην συνέχεια ακολουθεί μια επισκόπηση των διαγνώσεων σε βάθος χρόνου. Όπως παρατηρείται στην Εικόνα 11, η πλειονότητα των διαγνώσεων και επισκέψεων παρατηρούνται κατά τους πρώτους μήνες από την αρχική επίσκεψη. Καθώς περνά ο χρόνος, παρατηρείται μείωση στον αριθμό των επισκέψεων. Μια πιθανή εξήγηση γι' αυτό το φαινόμενο είναι ότι στην αρχή της κλινικής μελέτης, οι συμμετέχοντες επέδειξαν μεγαλύτερο ενδιαφέρον για τη συμμόρφωση προς τη μελέτη σε σχέση με αργότερα, όταν έχαναν το ενδιαφέρον τους. Επιπλέον, αξίζει να σημειωθεί ότι, παρά τη μείωση της συχνότητας των επισκέψεων στη διάρκεια του χρόνου, παρατηρείται μεγαλύτερη συνέπεια στις ετήσιες επισκέψεις, δηλαδή σε διαστήματα πολλαπλάσια των 12 μηνών (24 μηνές, 36 μηνές κλπ.).

4.3 Ελλείπουσες τιμές - Missing values

Υπάρχουν ελλείπουσες τιμές σε όλη τη έκταση του συνόλου δεδομένων στην Εικόνα 12 απεικονίζονται οι ελλείπουσες τιμές συγκεντρωτικά από όλους τους ασθενείς σε όλες της φάσεις του της μελέτης.



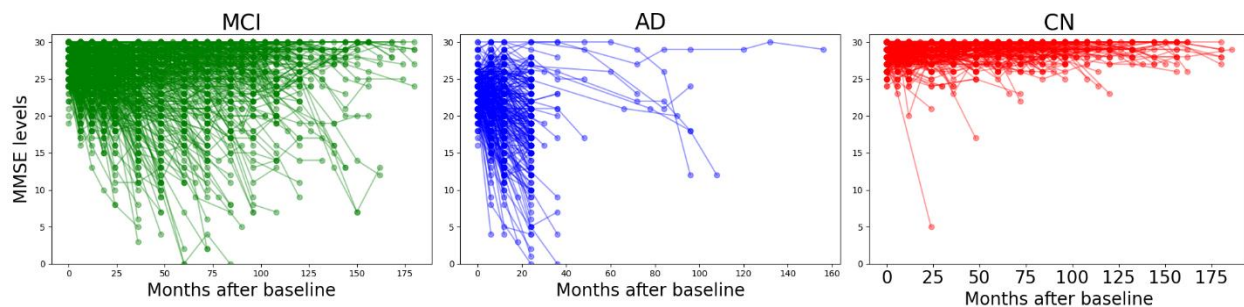
Εικόνα 13: Ελλείπουσες τιμές στις δωδεκάμηνες επισκέψεις



Εικόνα 14: Ελλείπουσες τιμές στις εικοσιτετράμηνες επισκέψεις

4.4 Mini-Mental State Examination (MMSE)

Η εξέταση Mini-Mental State Examination (MMSE) είναι ένα ευρέως χρησιμοποιούμενο εργαλείο διαλογής για τη γνωστική εξασθένηση, ιδίως στον ηλικιωμένο πληθυσμό. Αξιολογεί διάφορους γνωστικούς τομείς, συμπεριλαμβανομένων του προσανατολισμού, της μνήμης, της προσοχής, της γλώσσας και των οπτικών-χωρικών δεξιοτήτων. Το MMSE έχει διαπιστωθεί ότι έχει υψηλή αξιοπιστία και εγκυρότητα στην ανίχνευση της γνωστικής εξασθένησης, με βαθμολογία 24 ή χαμηλότερη να υποδηλώνει πιθανή γνωστική εξασθένηση[31]. Παρακάτω παρουσιάζεται η πρόοδος του MMSE score για κάθε κατηγορία διάγνωσης



Εικόνα 15: Τιμές του MMSE για κάθε διάγνωση συναρτήσει του χρόνου

Όπως παρατηρούμε στην Εικόνα 15 οι ασθενείς που διαγνώστηκαν αρχικά με ΗΓΔ(MCI) παρουσιάζουν πτώση στις γνωστικές τους δυνατότητες με την πάροδο του χρόνου ενώ άλλοι παραμένουν σταθεροί. Αυτό υποδεικνύει την σημαντικότητα του δείκτη καθώς οι ασθενείς που παρουσιάζουν πτώση έχουν αυξημένη πιθανότητα να παρουσιάσουν Alzheimer ενώ όσοι δεν παρουσιάζουν πτώση δείχνουν να σταθεροποιούνται στην ΗΓΔ. Επίσης αξίζει να σημειωθεί πως όσοι συμμετέχοντες διαγνώστηκαν στην αρχή της μελέτης με Alzheimer δεν παρουσιάζουν καταγραφές μετά τα 2 χρόνια πιθανώς λόγω αποχής από την μελέτη ή θανάτου από τις επιπλοκές της νόσου.

5. Πρόγνωση του Alzheimer με την χρήση Μηχανικής Μάθησης.

5.1 Ορισμός του προβλήματος

Σκοπός της παρούσας εργασίας είναι η ανάπτυξη ενός μοντέλου πρόγνωσης του Alzheimer για ασθενείς που έχουν ήδη παρουσιάσει Ήπια Γνωστική Διαταραχή. Ως διάστημα πρόγνωσης ορίστηκαν οι 48 μήνες από την 2^η ετήσια καταγεγραμμένη επίσκεψη του ασθενούς (baseline visit). Το ανώτατο όριο των 48 μηνών προκύπτει τόσο από την βιβλιογραφία, αν αναλογιστεί κανείς ότι το 10%-15% των ασθενών με ΗΓΔ παρουσιάζει Alzheimer ετησίως, όσο και από την ανάλυση των δεδομένων από το ADNIMERGE (βλ. Κεφάλαιο 4). Επιπλέον, πέραν της ανάπτυξης του μοντέλου στόχος είναι η ερμηνεία και η κατανόηση του αποτελέσματος ώστε να μπορεί να θεωρηθεί έγκαιρη και εμπεριστατωμένη η πρόβλεψη.

5.2 Βιβλιοθήκες που χρησιμοποιήθηκαν

Για την ανάπτυξη του μοντέλου χρησιμοποιήθηκαν οι εξής βιβλιοθήκες:

5.2.1 Pandas

Η βιβλιοθήκη Pandas είναι μια βιβλιοθήκη ανάλυσης και διαχείρισης συνόλων δεδομένων ανοικτού κώδικα για τη γλώσσα προγραμματισμού Python. Σύμφωνα με την επίσημη τεκμηρίωση, έχει ως στόχο να παρέχει "γρήγορες, ευέλικτες και εκφραστικές δομές δεδομένων σχεδιασμένες να κάνουν την εργασία με "σχεσιακά" ή "επισημασμένα" δεδομένα εύκολη και διαισθητική".

Η Pandas προσφέρει ένα ευρύ φάσμα λειτουργιών για τον καθαρισμό, την προεπεξεργασία, το φιλτράρισμα, τη συγκέντρωση και την οπτικοποίηση δεδομένων. Παρέχει επίσης πολλές ενσωματωμένες λειτουργίες για στατιστική ανάλυση και ενσωμάτωση με άλλες βιβλιοθήκες Python για μηχανική μάθηση, επιστημονικούς υπολογισμούς και οπτικοποίηση δεδομένων[32].

5.2.2 NumPy

Η NumPy είναι μια βιβλιοθήκη αριθμητικών υπολογισμών ανοικτού κώδικα για τη γλώσσα προγραμματισμού Python. Παρέχει γρήγορους και αποδοτικούς πολυδιάστατους πίνακες, οι οποίοι είναι απαραίτητοι για την εκτέλεση αριθμητικών υπολογισμών στην Python. Τα κύρια χαρακτηριστικά της NumPy περιλαμβάνουν διαχείριση N-διάστατων πινάκων, εξελιγμένες συναρτήσεις (εκπομπής), εργαλεία για την ενσωμάτωση κώδικα C/C++ και Fortran, χρήσιμες δυνατότητες γραμμικής άλγεβρας, μετασχηματισμού Fourier και τυχαίων αριθμών κ.α.

5.2.3 Matplotlib

Η Matplotlib είναι μια βιβλιοθήκη γραφικών παραστάσεων για τη γλώσσα προγραμματισμού Python. Παρέχει ένα ευρύ φάσμα συναρτήσεων για τη δημιουργία υψηλής ποιότητας 2D και 3D γραφικών παραστάσεων, διαγραμμάτων και γραφικών παραστάσεων. Τα κύρια χαρακτηριστικά της Matplotlib περιλαμβάνουν "σχήματα ποιότητας δημοσίευσης σε πολλές μορφές, διαδραστικά σχήματα και στοιχειώδη υποστήριξη κινούμενων σχεδίων". Η Matplotlib χρησιμοποιείται ευρέως στην επιστήμη δεδομένων, στην επιστημονική πληροφορική και σε άλλους τομείς όπου η οπτικοποίηση δεδομένων είναι σημαντική.

5.2.4 XGBoost

Ο XGBoost είναι μια βιβλιοθήκη μηχανικής μάθησης ανοικτού κώδικα για τη γλώσσα Python. Έχει σχεδιαστεί για να εκτελεί gradient boosting, μια δημοφιλής τεχνική μηχανικής μάθησης για εργασίες παλινδρόμησης και ταξινόμησης. Ο XGBoost είναι βελτιστοποιημένος για ταχύτητα και αποδοτικότητα και χρησιμοποιείται ευρέως σε εφαρμογές επιστήμης δεδομένων όπου η ακρίβεια και η απόδοση είναι κρίσιμες. Τα κύρια χαρακτηριστικά του XGBoost περιλαμβάνουν παραλληλοποιημένη ενίσχυση δέντρων, βελτιστοποιημένο κατανεμημένο υπολογισμό και υποστήριξη για προσαρμοσμένες συναρτήσεις και μετρικές αξιολόγησης.

5.2.5 Scikit-learn

Η Scikit-learn είναι μια βιβλιοθήκη μηχανικής μάθησης ανοικτού κώδικα για τη γλώσσα Python. Παρέχει ένα ευρύ φάσμα συναρτήσεων για ταξινόμηση, παλινδρόμηση, ομαδοποίηση, μείωση διαστάσεων και άλλες εργασίες μηχανικής μάθησης. Η Scikit-learn έχει σχεδιαστεί για να είναι εύκολη στη χρήση όσο και στην ενσωμάτωση με άλλες βιβλιοθήκες Python. Χρησιμοποιείται ευρέως στην επιστήμη των δεδομένων και άλλους τομείς όπου η μηχανική μάθηση είναι σημαντική.

5.3 Προ-επεξεργασία των δεδομένων

Αρχικά, φορτώνουμε το σύνολο δεδομένων ADNI, το οποίο περιέχει δεδομένα από πολλαπλές χρονικές στιγμές και πρωτόκολλα. Διαγράφουμε τις καταγραφές που δεν έχουν διάγνωση (DX), καθώς δεν είναι χρήσιμες για την ανάλυσή μας. Στη συνέχεια, επιλέγουμε τις επισκέψεις από τη χρονική στιγμή των 12 μηνών (m12) για κάθε υποκείμενο που έχει πραγματοποιήσει όλες τις προγραμματισμένες επισκέψεις μέχρι και τους 48 μήνες από το baseline. Δημιουργούμε μια νέα στήλη «Transition» που να υποδηλώνει την μετάβαση σε Alzheimer από ΗΓΔ και την συμπληρώνουμε με μονάδα «1» όπου υπάρχει μετάβαση στο διάστημα από 12 μήνες έως και 48 και μηδέν «0» όπου η διάγνωση παραμένει σταθερή για κάθε ασθενή. Έτσι έχουμε ανάγει το πρόβλημα μας σε δυαδική ταξινομήση. Τις καταγραφές που έχουν φυσιολογική διάγνωση τις απαλείφουμε καθώς δεν είναι αντικείμενο της παρούσας μελέτης. Επίσης απαλείφουμε και τις καταγραφές στο διάστημα μετά τους 12 μήνες και έως τους 48 καθώς στόχος είναι να προβλέψουμε την μετάβαση γνωρίζοντας την ιατρική κατάσταση του ασθενούς στους έως τους 12 μήνες. Στην συνέχεια αντικαταστήσουμε στις στήλες DX και DX_bl τις διαγνώσεις που είναι σε μορφή συμβολοσειράς με αριθμούς όπως φαίνεται στον Πίνακα 5. Επίσης στην στήλη DX_bl αντικαταστήσουμε τα στάδια της ΗΓΔ με την τιμή που αναπαριστά την ΗΓΔ.

Πίνακας 5: Αναπαράσταση κλάσεων (συμβολοσειρών) με ακέραιους αριθμούς

Κλάση	Ακεραία τιμή αναπαράστασης
CN	0
MCI	1
AD	2

Στην συνέχεια κωδικοποιούμε τις διακριτές τιμές σε διανύσματα τοποθετώντας μονάδα στην θέση της τιμής και μηδέν στις υπόλοιπες θέσεις. Παράδειγμα της κωδικοποίησης φαίνεται στους Πίνακες 6 και 7

Πίνακας 6: Δείγμα του χαρακτηριστικού "SEX" πριν την κωδικοποίηση

SEX
Male
Female

Πίνακας 7: Μετά την κωδικοποίηση του χαρακτηριστικού SEX έχουν δημιουργηθεί 2 "νέα" χαρακτηριστικά

Male	Female
1	0
0	1

Στην συνέχεια υπολογίζουμε τις ελλείπουσες τιμές καθώς όπως εξηγήθηκε στο 4^ο κεφάλαιο δεν θεωρούμε πως σκοπίμως υπάρχουν κενές τιμές με εξαίρεση τους δείκτες Every Day Cognition και MOCA που δεν καταγράφηκαν στην πρώτη φάση του ADNI.

Για τον υπολογισμό των ελλειπουσών τιμών χρησιμοποιήθηκε η μέθοδος k-κοντινότερων γειτόνων. Η μέθοδος αυτή αντικαθιστά τις ελλείπουσες τιμές υπολογίζοντας το μέσο όρο ή τη διάμεσο των τιμών των K πλησιέστερων γειτόνων για κάθε ελλείπουσα τιμή. Ο αριθμός των γειτόνων, K, καθορίζεται ως παράμετρος εισόδου. Στην παρούσα περίπτωση, ορίζουμε το K σε 5, που σημαίνει ότι οι αναμενόμενες τιμές εκτιμώνται με βάση τις τιμές των πέντε πλησιέστερων γειτόνων. Για να εξασφαλιστεί ο ακριβής υπολογισμός, χωρίζουμε πρώτα το σύνολο δεδομένων σε δύο υποσύνολα το ADNI-1 και το υπόλοιπο ADNI (ADNI-GO και ADNI-2,ADNI-3) επειδή τα υποσύνολα ADNI GO, ADNI-2 και ADNI-3 περιέχουν τιμές που δεν συγκεντρώθηκαν στην πρώτη φάση του ADNI σε σύγκριση με τις μελλοντικές φάσεις. Αντιμετωπίζοντάς τα ξεχωριστά, μπορούμε να προσαρμόσουμε τη διαδικασία υπολογισμού στα χαρακτηριστικά κάθε υποσυνόλου και να υπολογίσουμε λιγότερο μεροληπτικές τιμές.

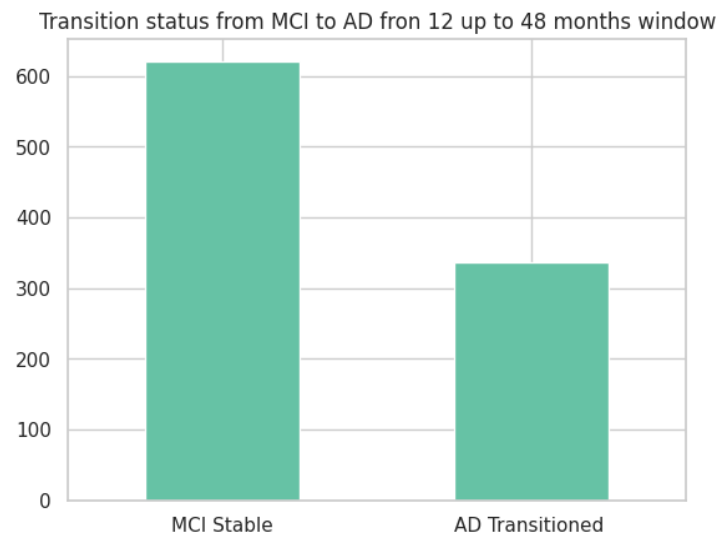
Για το υποσύνολο ADNI-1, δημιουργούμε ένα νέο DataFrame, εφαρμόζοντας τη διαδικασία υπολογισμού μόνο σε αυτό το υποσύνολο. Αποκλείουμε όσες στήλες δημιουργήθηκαν κατά την διάρκεια επεξεργασίας, το αναγνωριστικό συμμετέχοντα (RID), το πρωτόκολλο μελέτης (COLPROT) για να αποφύγουμε τον υπολογισμό περιττών και μεροληπτικών πληροφοριών. Επιπλέον αποκλείουμε τους δείκτες «Every Day Cognition Scale» και «MOCA» καθώς δεν συλλέχθηκαν στο ADNI-1. Περνάμε τις υπόλοιπες στήλες στον KNNImputer (αντικείμενο της βιβλιοθήκης scikit-learn για εφαρμογή του KNN αλγόριθμου), μαζί με τη μεταβλητή-στόχο "DX", για να υπολογίσουμε τις ελλείπουσες τιμές βάσει της διάγνωσης του κάθε ασθενούς.

Ομοίως, για τα υπόλοιπα υποσύνολα ADNI (ADNI-GO ,ADNI-2 και ADNI-3) εφαρμόζουμε την ίδια μέθοδο εξαιρώντας στο το ίδιο σύνολο στηλών αλλά συμπεριλαμβάνουμε τους δείκτες που

δεν μετρήθηκαν στο ADNI-1 και υπολογίζουμε τις ελλείπουσες τιμές με βάση την μεταβλητή-στόχο την διάγνωση.

Τέλος, συνδυάζουμε το υποσύνολο ADNI-1 με το υπόλοιπο σύνολο δεδομένων καταλήγοντας σε ένα σύνολο δεδομένων όπου μόνο οι δείκτες που δεν συλλέχθηκαν παραμένουν κενοί.

Το τελικό σύνολο δεδομένων περιλαμβάνει 957 ασθενείς. Από αυτούς, 777 έχουν διαγνωστεί με ΗΓΔ από την αρχική επίσκεψη, ενώ 336 έχουν ήδη διαγνωσθεί με Αλτσχάιμερ (AD). Από τους συνολικούς 957 ασθενείς, οι 666 παραμένουν σταθεροί σε ΗΓΔ από την επίσκεψη των 12 μηνών έως τους 48 μήνες, ενώ οι 336 μεταβαίνουν σε Alzheimer (AD). (Εικόνες 16 κ 17)



Εικόνα 16: Πλήθος ασθενών που παραμένουν σταθεροί σε ΗΓΔ μέχρι την έλευση των 48 μηνών (αριστερά). Πλήθος ασθενών που παρουσιάζουν Alzheimer μέχρι και την έλευση των 48 μηνών (δεξιά)

Στο νέο προ-επεξεργασμένο σύνολο δεδομένων προκύπτει πως 621 ασθενείς παραμένουν σταθεροί σε ΗΓΔ ενώ 333 παρουσιάζουν σε Alzheimer στο διάστημα που μεσολαβεί από την επίσκεψη των 12 μηνών έως τους 48 μήνες

Patients converting from MCI to AD



Εικόνα 17: Πλήθος ασθενών που παραμένουν σταθεροί σε ΗΓΔ μέχρι την έλευση των 48 μηνών (πάνω). Πλήθος ασθενών που παρουσιάσουν Alzheimer μέχρι και την έλευση των 48 μηνών (κάτω)

5.4 eXtream Gradient Boosting - XGBoost

Για την υλοποίηση της παρούσας εργασίας επιλέχθηκε ο αλγόριθμος eXtream Gradient Boosting (XGBoost) που ανήκει στην κατηγορία των αλγορίθμων ενισχύσεις κλίσης (gradient boosting). Οι αλγόριθμοι ενίσχυσης κλίσης αποτελούν μια μέθοδο που υποθέτει ότι ο επαναληπτικός συνδιασμός πολλαπλών αδύναμων ταξινομητών οδηγεί στη δημιουργία ενός ισχυρού ταξινομητή. Ο συγκεκριμένος αλγόριθμος, γνωστός ως Gradient boosting, ικανοποιεί αυτήν την υπόθεση εκπαιδεύοντας τον πρώτο ταξινομητή για να μάθει την ανεξάρτητη μεταβλητή και τους επόμενους ταξινομητές για να μάθουν τις κλίσεις των προηγούμενων ταξινομητών.

Ως κλίσεις $l(y_i, \hat{y}_i)$ ορίζουμε την απόκλιση μεταξύ της προσθετικής ταξινόμησης ($y_i^{(t-1)}$) από την προηγούμενη επανάληψη $t - 1$ και της πραγματικής ταξινόμησης y_i για την παρατήρηση i . Η συνάρτηση κόστους $L(t)$, η οποία χρησιμοποιείται στην επανάληψη t με n παρατηρήσεις, περιγράφεται από την Εξίσωση 1. Εδώ, η συνάρτηση f_t αναπαριστά τον αδύναμο ταξινομητή στην επανάληψη t , ενώ η $\Omega(f_t)$ αναφέρεται σε έναν συνάρτηση ρύθμισης (regularization) που ελέγχει την πολυπλοκότητα του ταξινομητή [33], [34]

$$L^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \Omega(f_t) \quad (1)$$

Το άθροισμα των συνδυασμών όλων των αδύναμων ταξινομητών f_k ορίζει την τελική ταξινόμηση $\hat{y}_i$ για την παρατήρηση i (Εξίσωση 2)

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i) \quad (2)$$

Ο eXtreme Gradient Boosting (XGBoost) αποτελεί μια βιβλιοθήκη λογισμικού ανοικτού κώδικα που υλοποιεί την μέθοδο gradient boosting με έμφαση στην επεκτασιμότητα, τον παραλληλισμό και την κατανεμημένη εκτέλεση. Ο XGBoost χρησιμοποιεί δέντρα ταξινόμησης και παλινδρόμησης (CARTs) ως τα αδύναμους ταξινομητές. Η παράμετρος $nrounds$ (n) καθορίζει τον αριθμό των επαναλήψεων κατά τη διάρκεια της διαδικασίας εκπαίδευσης. Ο ρυθμός μάθησης eta (η) ελέγχει την επίδραση κάθε αδύναμου ταξινομητή στο τελικό μοντέλο και

αποτρέπει το υπερ-ταίριασμα των δεδομένων(overfitting). Η παράμετρος γάμμα (γ) καθορίζει το ελάχιστο ποσοστό μείωσης των απωλειών που απαιτείται για την εξειδίκευση των φύλλων των δέντρων. Υψηλές τιμές του γ οδηγούν σε μοντέλα με μεγαλύτερη συντηρητικότητα. Η παράμετρος `max_depth` (d_{max}) καθορίζει το μέγιστο βάθος που μπορεί να έχει ένα δέντρο. Τα βαθύτερα δέντρα είναι πιο πολύπλοκα και ενδέχεται να υποφέρουν από υπερπροσαρμογή. Η παράμετρος `min_child_weight` (w_{min}) ορίζει τον ελάχιστο αριθμό παρατηρήσεων που απαιτούνται για να δημιουργηθεί ένας κόμβος-παιδί. Η παράμετρος `subsample` (s) καθορίζει το ποσοστό των παρατηρήσεων που επιλέγονται τυχαία σε κάθε επανάληψη κατά την διάρκεια της εκπαίδευσης. Μικρές τιμές του `subsample` αποτρέπουν το υπερ-ταίριασμα. Τέλος, η παράμετρος `colsample_bytree` (c) αναφέρεται στο ποσοστό των τυχαία επιλεγμένων χαρακτηριστικών σε κάθε επανάληψη. Μικρές τιμές του c οδηγούν σε πιο αξιόπιστα μοντέλα, ενώ τιμές που πλησιάζουν το μηδέν μπορεί να οδηγήσουν σε χαμηλή ποιότητα αποτελεσμάτων.[34]

Οι παράμετροι για την εκπαίδευση του αλγορίθμου απεικονίζονται στον Πίνακα 8

Πίνακας 8: Τελικές παράμετροι για την εκπαίδευση του μοντέλου

Παράμετρος	Τιμή
<code>n_rounds</code>	600
<code>max_depth</code>	5
<code>Eta</code>	0.01
<code>scale_pos_weigh</code>	2
<code>tree_method</code>	<code>gpu_hist</code>
<code>subsample</code>	0.8
<code>col_sample</code>	0.8

5.5 Επιλογή των χαρακτηριστικών

Η σωστή επιλογή των χαρακτηριστικών (feature selection) από το σύνολο δεδομένων είναι καίριας σημασίας για στη μηχανική μάθηση καθώς συμβάλει στην βελτίωση της απόδοσης του μοντέλου, τη μείωση της πολυπλοκότητας και την ενίσχυση της ερμηνείας του αποτελέσματος. Για τους σκοπούς της παρούσης επιλέχθηκε η μέθοδος της αναδρομικής εξάλειψης χαρακτηριστικών (Recursive Feature Elimination) όπως υλοποιείται στην βιβλιοθήκη scikit-learn.

Η μέθοδος αυτή υλοποιείται ως εξής:

- Σε πρώτη φάση γίνεται εκπαίδευση του μοντέλου XGBoost χρησιμοποιώντας ολόκληρο το σύνολο χαρακτηριστικών και εφαρμόζοντας 5-fold cross validation. Με το πέρας της εκπαίδευσης το μοντέλο βαθμολογεί κάθε χαρακτηριστικό ως προς την σημαντικότητα του στην πρόβλεψη. Η βαθμολογία αυτή προκύπτει από την συχνότητα χρησιμοποιήθηκε το κάθε χαρακτηριστικό για ως ρίζα για την δημιουργία των επί μέρους δένδρων του αλγορίθμου[33].
- Τα χαρακτηριστικά με χαμηλότερες βαθμολογίες σημαντικότητας θεωρούνται λιγότερο σημαντικά και υποψήφια για απαλοιφή. Παράλληλα υπολογίζεται μια μετρική αξιολόγησης του το μοντέλου για την ορθότητα των προβλέψεων
- Στην συνέχεια απαλείφεται ένας προκαθορισμένος από τον χρήστη αριθμός χαρακτηριστικών (βήμα) από το σύνολο δεδομένων και εκτελείται ξανά η εκπαίδευση με 5-fold cross validation.
- Αυτή η διαδικασία εκτελείται επαναληπτικά έως ότου είτε επιτευχθεί ο επιθυμητός αριθμός χαρακτηριστικών είτε η απόδοση του μοντέλου σταματήσει να βελτιώνεται.

Τα χαρακτηριστικά που προέκυψαν μετά την παραπάνω διαδικασία παρουσιάζονται στον Πίνακα 9

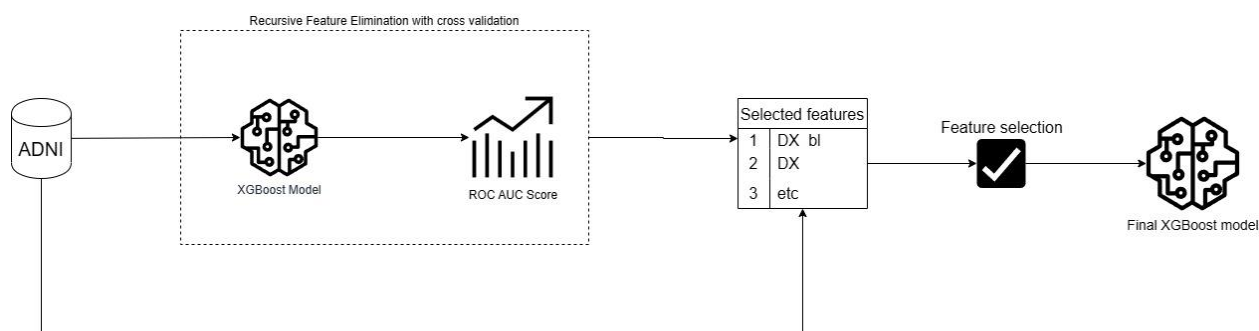
Πίνακας 9: Χαρακτηριστικά συνόλου δεδομένων μετά την εφαρμογή της μεθόδου αναδρομικής εξάλειψης χαρακτηριστικών

Χαρακτηριστικό
LDELTOTAL
TRABSCOR
FAQ
Hippocampus
Entorhinal
MidTemp
ICV
mPACCdigit
mPACCtrailsB
CDRSB_bl
ADAS13_bl
ADAS11_bl
MMSE_bl

RAVLT_learning_bl
TRABSCOR_bl
FAQ_bl
mPACCdigit_bl
Entorhinal_bl
mPACCtrailsB_bl
Fusiform_bl
MidTemp_bl
ICV_bl
PTMARRY_Widowed
PTETHCAT_Not_Hisp/Latino
DX
EcogPtMem
EcogPtPlan
EcogSPPlan
EcogSPOrgan
EcogSPDivatt
EcogSPTotal
MOCA_bl
EcogPtMem_bl
EcogSPLang_bl
EcogSPTotal_bl
RAVLT_immediate
ADASQ4
ADAS13
CDRSB
ADAS11
APOE4
DX_bl

Η λίστα με όλα τα χαρακτηριστικά του συνόλου δεδομένων πριν την προ-επεξεργασία είναι διαθέσιμη στο **Παράρτημα Α**

5.6 Εκπαίδευση του μοντέλου



Εικόνα 18: Διαδικασία εκπαίδευσης του μοντέλου

Η διαδικασία που ακολουθήθηκε για την δημιουργία του μοντέλου περιγράφεται στην Εικόνα 18.

Αρχικά, αφού εγίνε προ-επεξεργασία του dataset εκπαιδύτηκε μια εκδοχή του μοντέλου XGboost για την εφαρμογή της μεθόδου επιλογής χαρακτηριστικών «Recursive Feature Elimination» όπως αυτή περιγράφεται στην ενότητα 5.5. Ως μέτρο αξιολόγησης της ποιότητας της επιλογής χαρακτηριστικών επιλέχθηκε το ROC AUC Score (Receiver Operating Characteristic Area Under the Curve Score - Εξίσωση 1) το οποίο αποτελεί ένα σημαντικό μέτρο απόδοσης για τα μοντέλα ταξινόμησης και μετράει την ικανότητά τους να διακρίνουν τις διάφορες κλάσεις του προβλήματος. Μετά την επιλογή των σημαντικότερων χαρακτηριστικών, πραγματοποιήθηκε εκ νέου εκπαίδευση του μοντέλου με τη χρήση μόνο των επιλεγμένων χαρακτηριστικών και παράλληλη εφαρμογή 5-fold cross validation για την ορθή αξιολόγηση της απόδοσης του. Η τεχνική αυτή διασφαλίζει την αξιόπιστη εκτίμηση της γενικής ικανότητας του μοντέλου και επιτρέπει την αξιολόγηση της απόδοσής του σε διάφορα υποσύνολα δεδομένων. Τα αποτελέσματα του 5-fold cross validation αναγράφονται στον πίνακα 10.

Για την αξιολόγηση της απόδοσης, χρησιμοποιήθηκαν διάφορα μέτρα. Το Recall (Εξίσωση 3) πού αναφέρεται στην ικανότητα του μοντέλου να εντοπίζει σωστά τα θετικά δείγματα σε σχέση με το συνολικό αριθμό των θετικών δειγμάτων. Το Precision (Εξίσωση 4) που αναφέρεται στην ικανότητα του μοντέλου να προβλέπει σωστά τα θετικά δείγματα σε σχέση με το συνολικό αριθμό των προβλεπόμενων θετικών δειγμάτων. Η ακρίβεια αναφέρεται στο ποσοστό των σωστών προβλέψεων του μοντέλου συνολικά. Τέλος, το ROC AUC Score αναφέρεται στο συνολικό εμβαδόν της περιοχής κάτω από την καμπύλη ROC (Receiver Operating Characteristic), το οποίο μετράει την ικανότητα του μοντέλου να διακρίνει τις διαφορές μεταξύ των κλάσεων.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (3)$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (4)$$

5.7 Αποτελέσματα και συζήτηση

Πίνακας 10: Αποτελέσματα εκπαίδευσης του μοντέλου

Fold	Class	Precision	Recall	Accuracy	ROC AUC
1	Transition	0.69	0.73	0.79	0.778
	No Transition	0.85	0.82		
2	Transition	0.80	0.71	0.83	0.805
	No Transition	0.85	0.90		
3	Transition	0.79	0.75	0.84	0.821
	No Transition	0.87	0.90		
4	Transition	0.75	0.78	0.83	0.820
	No Transition	0.88	0.86		
5	Transition	0.74	0.81	0.83	0.826
	No Transition	0.89	0.85		
Average	Transition	0.754	0.756	0.824	0.81

Συνολικά, το μοντέλο επέδειξε ικανοποιητικά αποτελέσματα στην πρόγνωση της μετάβασης από ΗΓΔ σε Alzheimer. Το precision για την κατηγορία «Transition» ήταν 0,754, γεγονός που υποδηλώνει ότι όταν το μοντέλο προέβλεπε τα δείγματα, ήταν σωστό περίπου στο 75,4% των περιπτώσεων. Η μέση ευαισθησία (Recall) κυμαίνεται στο 0,756 γεγονός που υποδηλώνει ότι το μοντέλο αναγνώρισε με επιτυχία περίπου το 75,6% των πραγματικών περιπτώσεων της κλάσης «Transition». Η μέση ακρίβεια του μοντέλου ήταν 0,824, που σημαίνει ότι έκανε σωστές προβλέψεις για περίπου 82,4% όλων των περιπτώσεων (Transition – No transition). Επιπλέον, το μέσος ROC AUC score είναι 0,81 που σημαίνει ότι το μοντέλο πέτυχε μια αρκετά καλή διακριτική ικανότητα μεταξύ των κλάσεων μετάβασης και μη μετάβασης.

Όσον αφορά την κλάση μετάβασης, οι τιμές του precision κυμαίνονται σε όλες τις αναδιπλώσεις από 0,69 έως 0,80, υποδεικνύοντας ένα αρκετά ικανοποιητικό επίπεδο ακρίβειας στην πρόβλεψη περιπτώσεων μετάβασης σε Alzheimer. Η ευαισθησία για την κλάση «Transition» κυμαίνονταν από 0,71 έως 0,81, υποδηλώνοντας μια σχετικά υψηλή ικανότητα του μοντέλου να εντοπίζει ένα σημαντικό ποσοστό πραγματικών περιπτώσεων μετάβασης.

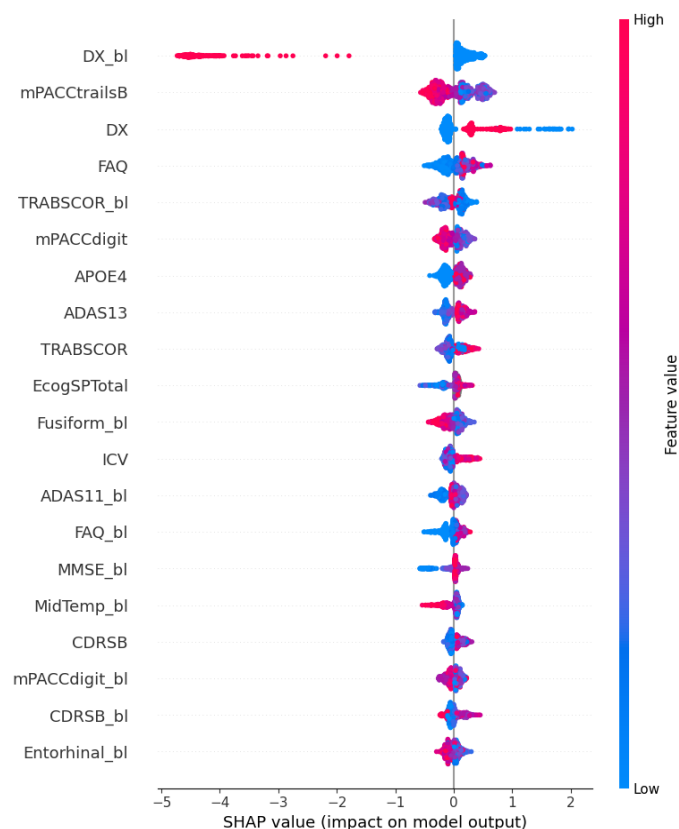
5.8 Ερμηνεία του μοντέλου και των αποφάσεων του

Η ερμηνεία των αποφάσεων του μοντέλου επιτρέπει την κατανόηση της διαδικασίας λήψης αποφάσεων που υιοθετεί το μοντέλο και την ανάλυση των λόγων που κρύβονται πίσω από μια συγκεκριμένη απόφαση. Μέσω της εν λόγω ερμηνείας, μπορούμε να αναδείξουμε τις συσχετίσεις, τις επιρροές και τις συνεισφορές των διαφόρων χαρακτηριστικών προς την εκάστοτε πρόβλεψη, παρέχοντας έτσι σημαντική κατανόηση και πληροφορίες που ενισχύουν την επιστημονική γνώση και συνεισφέρουν στη βελτίωση της ακρίβειας και της αξιοπιστίας του μοντέλου. Για την ερμηνεία των αποτελεσμάτων χρησιμοποιήθηκαν οι τιμές Shapley από την θεωρία παιγνίων.[35]. Η υλοποίηση σε κώδικα έγινε με την βιβλιοθήκη shap. Στην Εικόνα 19 απεικονίζεται η επίδραση μερικών χαρακτηριστικών στην ικανότητα πρόβλεψης του μοντέλου.

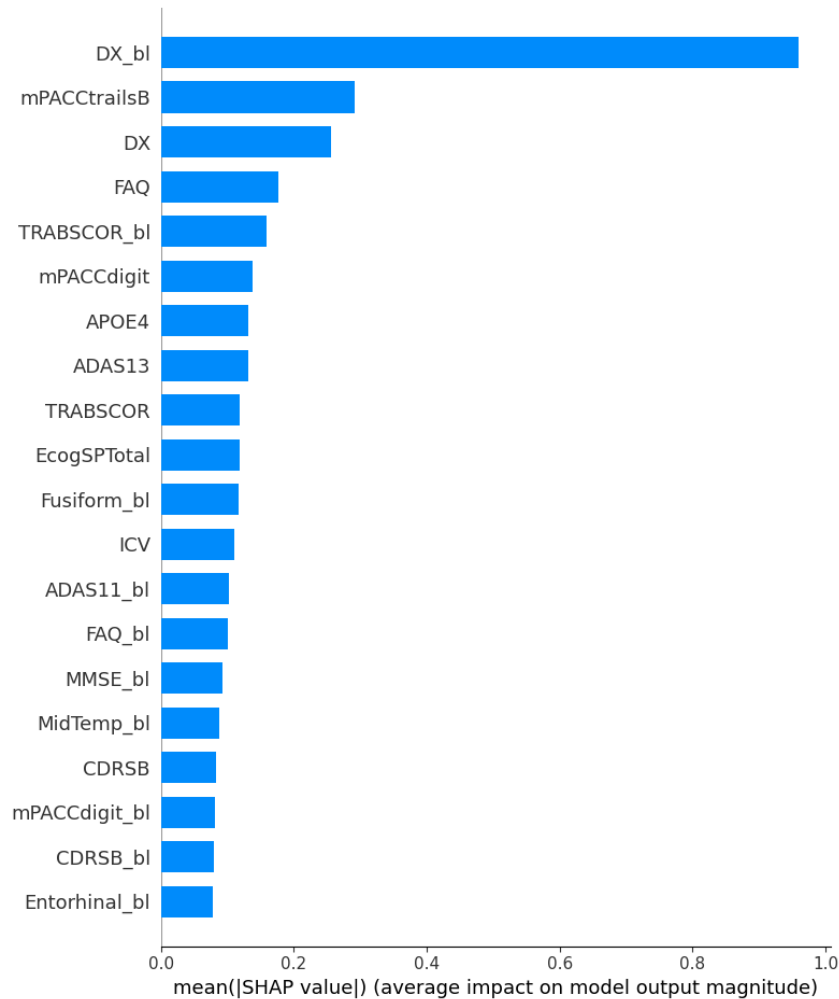
Το χρώμα της μπάρας στα δεξιά αντιστοιχεί στην τιμή του χαρακτηριστικού: κόκκινο για υψηλές τιμές και μπλε για χαμηλές τιμές. Αυτό βοηθά στη διάκριση της θετικής και της αρνητικής συνεισφοράς των χαρακτηριστικών. Το διάγραμμα περιλαμβάνει επίσης μια κατακόρυφη γραμμή αναφοράς στο μηδέν, η οποία αντιπροσωπεύει τη μέση απόδοση του μοντέλου. Τα χαρακτηριστικά στα αριστερά αυτής της γραμμής έχουν αρνητικό αντίκτυπο στις προβλέψεις, ενώ τα χαρακτηριστικά στα δεξιά έχουν θετικό αντίκτυπο. Έτσι, το διάγραμμα μας επιτρέπει να παρατηρήσουμε τόσο την κατεύθυνση όσο και το μέγεθος της συμβολής κάθε χαρακτηριστικού.

Όπως παρατηρούμε, η διάγνωση κατά την επίσκεψη στο δωδεκάμηνο έχει μεγάλη επιρροή στην απόφαση του μοντέλου. Το εύρος της επίδρασης ποικίλει καθώς παρατηρούμε πως ορισμένες χαμηλές τιμές (μπλέ) έχουν υψηλή επίδραση στην θετική πρόγνωση σε σχέση με τις υπόλοιπες αρνητικές -ως προς τον οριζόντιο άξονα – χαμηλές τιμές. Μία λογική εξήγηση για αυτό πως πιθανότατα τα δείγματα που έχουν αρνητικό αντίκτυπο στην πρόβλεψη είναι πιθανοί ασθενείς που δεν παρουσιάζουν μετάβαση σε Alzheimer, ενώ τα δείγματα που έχουν θετικό αντίκτυπο πληρούν τους κανόνες του μοντέλου που να υποδηλώνει πιθανή μετάβαση σε Alzheimer.

Στην Εικόνα 20 απεικονίζονται τα χαρακτηριστικά του dataset ταξινομημένα ως προς την επιρροή της τελικής απόφασης.



Εικόνα 19: Επιρροή του κάθε χαρακτηριστικού στην τελική απόφαση του μοντέλου

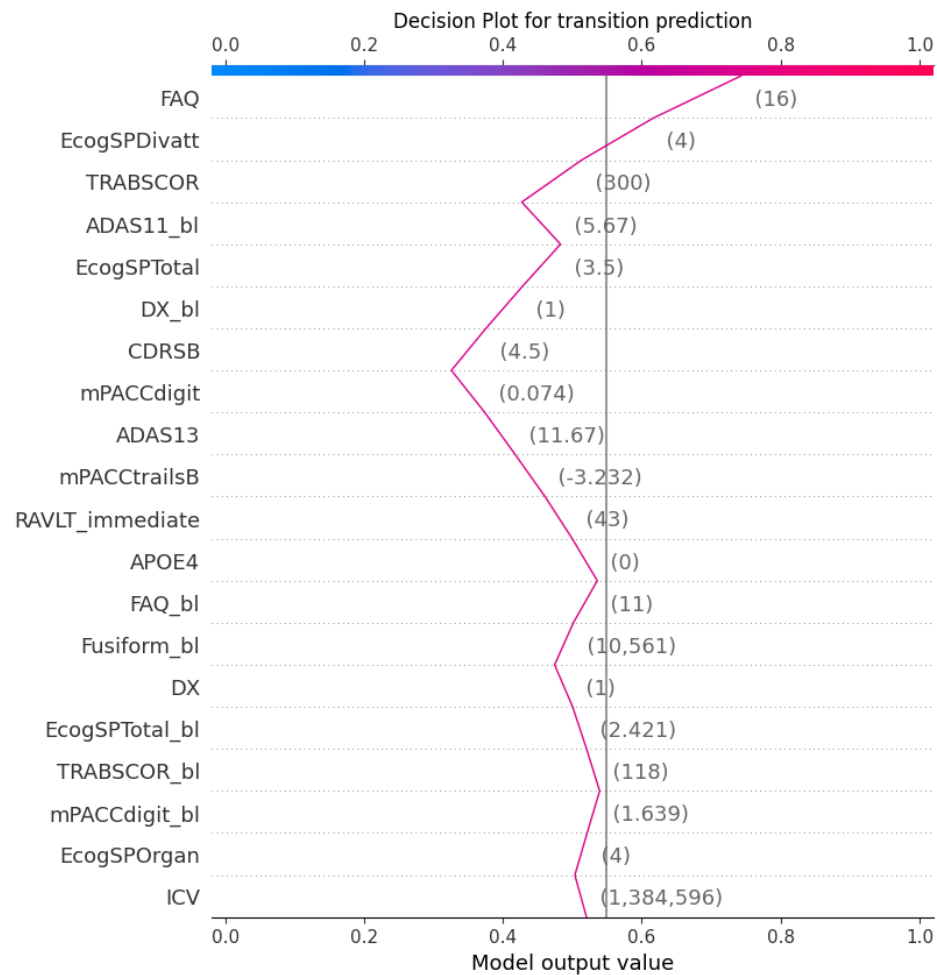


Εικόνα 20: Επιρροή του κάθε χαρακτηριστικού στην τελική απόφαση του μοντέλου

Μια από τις θεμελιώδεις ιδιότητες των τιμών Shapley είναι ότι αθροίζουν πάντα τη διαφορά μεταξύ του αποτελέσματος του παιγνίου όταν είναι παρόντες όλοι οι παίκτες (δείγματα) και του αποτελέσματος του παιγνίου όταν δεν υπάρχουν παίκτες. Για τα μοντέλα μηχανικής μάθησης αυτό σημαίνει ότι οι τιμές SHAP όλων των χαρακτηριστικών εισόδου θα αθροίζουν πάντα στη διαφορά μεταξύ της βασικής (αναμενόμενης) εξόδου του μοντέλου και της τρέχουσας εξόδου του μοντέλου για την εξηγούμενη πρόβλεψη [35]. Λαμβάνοντας αυτό υπόψιν μπορούμε να απεικονίσουμε ένα διάγραμμα λήψης απόφασης (Εικόνα 21) που να περιγράφει τον τρόπο που έφτασε το μοντέλο σε μια πρόβλεψη για ένα ή περισσότερα δείγματα.

Το διάγραμμα είναι κεντραρισμένο στον άξονα x στην αναμενόμενη τιμή. Όλες οι τιμές SHAP είναι σχετικές με την αναμενόμενη τιμή του μοντέλου. * Ο άξονας y παραθέτει τα χαρακτηριστικά του μοντέλου. Τα χαρακτηριστικά είναι ταξινομημένα κατά φθίνουσα ως προς την σπουδαιότητα. Η σπουδαιότητα υπολογίζεται από την συχνότητα εμφάνισης στο σύνολο των προβλέψεων. Και συνήθως διαφέρει από την σπουδαιότητα για ολόκληρο το σύνολο δεδομένων. Η πρόβλεψη κάθε παρατήρησης αναπαρίσταται με μια χρωματιστή γραμμή. Στην κορυφή του διαγράμματος, κάθε γραμμή τέμνει τον άξονα x στην προβλεπόμενη τιμή της αντίστοιχης παρατήρησης. Αυτή η τιμή

καθορίζει το χρώμα της γραμμής σε ένα φάσμα. * Προχωρώντας από το κάτω μέρος του διαγράμματος προς την κορυφή, οι τιμές SHAP για κάθε χαρακτηριστικό προστίθενται στη βασική τιμή του μοντέλου. Αυτό δείχνει πώς συμβάλλει κάθε χαρακτηριστικό στη συνολική πρόβλεψη.



Εικόνα 21: Διάγραμμα λήψης απόφασης με βάση τις τιμές Shapley για μια τυχαία πρόβλεψη από το σύνολο δεδομένων. Η ανάγνωση του διαγράμματος γίνεται από κάτω προς τα πάνω. Στον άξονα Y απεικονίζονται τα χαρακτηριστικά και ενώ στον άξονα X βρίσκονται οι τιμές Shapley. Στις παρενθέσεις βρίσκονται οι τιμές των χαρακτηριστικών για το συγκεκριμένο δείγμα.

6. Διαδικτυακή Εφαρμογή για την πρόγνωση της Νόσου

Στα πλαίσια της παρούσας διατριβής δημιουργήθηκε μια διαδικτυακή εφαρμογή για την χρήση του μοντέλου που αναπτύχθηκε στο 5^ο Κεφάλαιο. Σκοπός της εφαρμογής είναι να διευκολύνει την διεπαφή του τελικού μη τεχνικού χρήστη με το μοντέλο.

6.1 Βιβλιοθήκες και τεχνολογίες που χρησιμοποιήθηκαν

6.1.1 ReactJS – Frontend

Η ReactJS είναι μια δημοφιλής βιβλιοθήκη JavaScript για την κατασκευή web περιβαλλόντων. Παρέχει μια προσέγγιση βασισμένη σε στοιχεία που επιτρέπει στους προγραμματιστές να δημιουργούν επαναχρησιμοποιήσιμα κομμάτια του UI και να διαχειρίζονται αποτελεσματικά την κατάσταση και τη ροή δεδομένων εντός της εφαρμογής. Η React προσφέρει ένα εικονικό DOM, επιτρέποντας την αποτελεσματική ενημέρωση και απόδοση των στοιχείων. Με τη δηλωτική σύνταξή της, απλοποιεί τη διαδικασία δημιουργίας σύνθετων εφαρμογών, αναλύοντάς τις σε μικρότερα, διαχειρίσιμα στοιχεία. Υποστηρίζει ένα τεράστιο οικοσύστημα βιβλιοθηκών και εργαλείων, καθιστώντας το μια εξαιρετική επιλογή για τη δημιουργία δυναμικών και διαδραστικών front-end εφαρμογών.

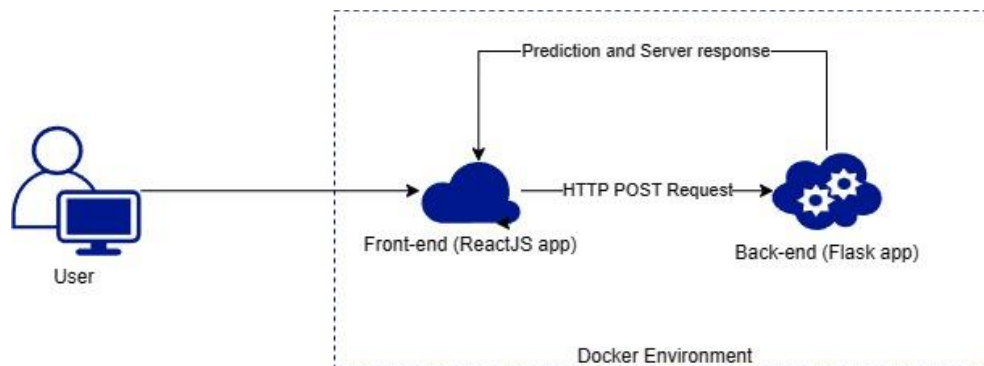
6.1.2 Flask – Backend

Το Flask είναι ένα ελαφρύ και ευέλικτο web framework της Python που παρέχει έναν απλό και αποδοτικό τρόπο για τη δημιουργία web εφαρμογών. Ακολουθεί έναν ελαφρύ σχεδιασμό, επιτρέποντας στους προγραμματιστές να επιλέξουν τα εργαλεία και τις βιβλιοθήκες που χρειάζονται από όλο το οικοσύστημα της python, διατηρώντας τον πυρήνα του framework ελαφρύ. Το Flask είναι κατάλληλο για μικρά έως μεσαία έργα και την κατασκευή RESTful APIs. Παρέχει δυνατότητες δρομολόγησης, απεικόνισης σελίδων και εύκολη πρόσβαση σε βάσεις δεδομένων.

6.1.3 Docker

Το Docker (Ντόκερ) είναι μια πλατφόρμα λογισμικού ανοιχτού κώδικα που υλοποιεί εικονικοποίηση σε επίπεδο λειτουργικού συστήματος. Βασικά, το Docker παρέχει αυτοματοποιημένες διαδικασίες για την ανάπτυξη εφαρμογών σε απομονωμένους χώρους εντός του συστήματος που ονομάζονται containers. Το Docker χρησιμοποιεί τεχνολογίες του πυρήνα του Linux, όπως τα cgroups και τα kernel namespaces, προκειμένου να επιτρέψει την εκτέλεση ανεξάρτητων λογισμικών containers στο ίδιο λειτουργικό σύστημα. Με αυτόν τον τρόπο, αποφεύγεται η ανάγκη για επιπλέον υπολογιστικούς πόρους που απαιτούνται από μια εικονική μηχανή (virtual machine).

6.2 Αρχιτεκτονική εφαρμογής



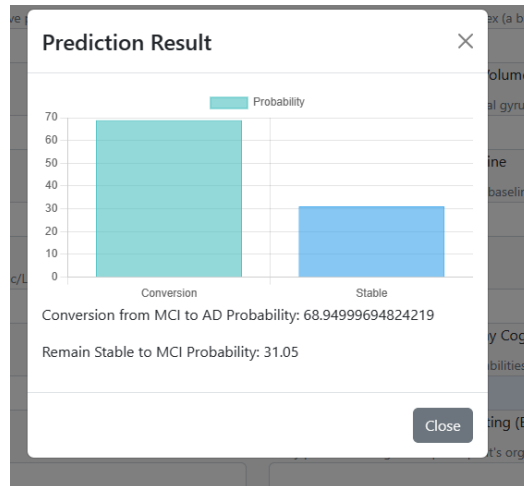
Εικόνα 22: Αρχιτεκτονική διαδικτυακής εφαρμογής

Για τον σχεδιασμό της εφαρμογής επιλέχθηκε η αρκετά απλή αρχιτεκτονική του frontend – backend με την χρήση Docker containers για καλύτερη διαχείριση και ανάπτυξη του συστήματος (Εικόνα 22). Για το front-end μέρος αναπτύχθηκε μια web εφαρμογή με την χρήση της ReactJS όπου εκεί ο τελικός χρήστης εισάγει τις τιμές από τις διάφορες μετρήσεις των εξετάσεων του ασθενούς (Επιλεγμένα χαρακτηριστικά – Κεφάλαιο 5). Στην συνέχεια ο χρήστης πατάει το κουμπί «Submit» και η εφαρμογή στέλνει ένα αίτημα HTTP POST στο backend όπου γίνεται ο υπολογισμός της πιθανότητας μετάβασης σε Alzheimer από MCI. Στη συνέχεια το backend επιστρέφει την πιθανότητα.

Για το backend σχεδιάστηκε ένα απλό RESTful API με την χρήση του Flask web framework για Python. Το backend λαμβάνει σε JSON μορφή (από το frontend) τις τιμές του κάθε χαρακτηριστικού και στην συνέχεια «καλεί» το αποθηκευμένο μοντέλο και εισάγει τις τιμές σε αυτό. Το μοντέλο υπολογίζει τη πιθανότητα μετάβασης σε Alzheimer από MCI και με βάση αυτή την τιμή υπολογίζεται και η πιθανότητα να μείνει σταθερός ο ασθενής. Στην συνέχεια επιστρέφει ως απάντηση στο POST αίτημα του frontend τις υπολογισθείσες τιμές.

Στην Εικόνα 23 απεικονίζεται το περιβάλλον της εφαρμογής και στην Εικόνα 24 η οθόνη παρουσίασης του αποτελέσματος

Στο Παράρτημα Β βρίσκεται ολόκληρος ο κώδικας της εφαρμογής



Εικόνα 23: Οθόνη προβολής αποτελέσματος πρόβλεψης του μοντέλου

Cognitive Impairment Prediction using eXtreme Gradient Boosting

Normal MCI AD

Features Input

Diagnosis at Baseline
Diagnosis at the baseline visit

Clinical Dementia Rating - Sum of Boxes
Clinical Dementia Rating-Sum of Boxes (assessment of dementia severity)

Alzheimer's Disease Assessment Scale - Cognitive Subscale 13
Alzheimer's Disease Assessment Scale - Cognitive Subscale 13 (assessment of cognitive functioning)

Rey Auditory Verbal Learning Test - Immediate Recall
Rey Auditory Verbal Learning Test - Immediate Recall (assessment of verbal memory)

Trail Making Test Part B - Time to Completion
Trail Making Test Part B - Time to Completion (assessment of executive function and mental flexibility)

Hippocampus Volume

Presence of APOE4 Allele
Presence of APOE4 allele (a genetic risk factor for Alzheimer's disease)

Alzheimer's Disease Assessment Scale - Cognitive Subscale 11
Alzheimer's Disease Assessment Scale - Cognitive Subscale 11 (assessment of cognitive functioning)

Alzheimer's Disease Assessment Scale - Cognitive Subscale 4
Alzheimer's Disease Assessment Scale - Cognitive Subscale 4 (assessment of cognitive functioning)

Logical Memory IIa from Wechsler Memory Scale - Revised
Logical Memory IIa from the Wechsler Memory Scale - Revised (assessment of verbal memory)

Functional Activities Questionnaire
Functional Activities Questionnaire (assessment of activities of daily living)

Entorhinal Cortex Volume

Εικόνα 24: Κεντρική οθόνη εφαρμογής

```

@app.route('/predict', methods=['POST'])
def predict():
    # Load xgboost model

    # Get JSON data from the request
    json_data = request.get_json()

    for key, value in json_data.items():
        if value == "":
            json_data[key] = None

    if 'DX' in json_data:
        if json_data['DX'] == 'MCI':
            json_data['DX'] = 1
        elif json_data['DX'] == 'AD':
            json_data['DX'] = 2
    if 'DX_bl' in json_data:
        if json_data['DX_bl'] == 'MCI':
            json_data['DX_bl'] = 1
        elif json_data['DX_bl'] == 'AD':
            json_data['DX_bl'] = 2

    # Perform prediction using the loaded model
    data = pd.DataFrame.from_dict(json_data, orient='index').transpose()
    # convert all data to float    data = data.astype(float)

    data = xgb.DMatrix(data)
    prediction = model.predict(data)
    prediction = prediction * 100

    return np.round(prediction, 2).tolist()

```

Εικόνα 25: Κώδικας διαχείρισης του αιτήματος POST της εφαρμογής προς τον server.

7. Συμπεράσματα, περιορισμοί και μελλοντικές προεκτάσεις.

Με την πάροδο του χρόνου, παρατηρείται σημαντική αύξηση στον αριθμό των ατόμων που πάσχουν από Alzheimer . Οι εκτιμήσεις δείχνουν ότι τα περιστατικά αναμένεται να τριπλασιαστούν κατά τη διάρκεια των επόμενων τριών δεκαετιών, δημιουργώντας επιτακτική ανάγκη για την εύρεση αποτελεσματικών μεθόδων πρόγνωσης και έγκαιρης διάγνωσης και κατανόησης της νόσου. Η ανεύρεση τέτοιων μεθόδων έχει κρίσιμη σημασία για την επίτευξη έγκαιρης αντιμετώπισης της νόσου, καθώς επιτρέπει την έναρξη της αγωγής από τα πρώτα στάδια της. Αυτό διαδραματίζει καθοριστικό ρόλο στην αύξηση του προσδόκιμου ζωής και τη βελτίωση της ποιότητας ζωής των ασθενών. Επιπλέον αυτές οι μέθοδοι συμβάλουν στην καλύτερη κατανόηση της νόσου και των παραγόντων που την προκαλούν από την ερευνητική κοινότητα.

Υπό αυτό το πρίσμα, η παρούσα μελέτη κατέληξε σε μια ολοκληρωμένη εφαρμογή για την πρόγνωση του Alzheimer για άτομα που πάσχουν από Ήπια Γνωστική Διαταραχή. Με την χρήση της μηχανικής μάθησης αναπτύχθηκε ένα μοντέλο που να μπορεί να προβλέψει με ακρίβεια της τάξεως του 70% την πιθανή μετάβαση του ασθενούς από ΗΓΔ σε Alzheimer σε διάστημα 48 μηνών από την πρώτη επίσκεψη. Σε συνδυασμό με το μοντέλο, παρουσιάστηκε μια μέθοδος επεξήγησης και ερμηνείας του αποτελέσματος της πρόβλεψης με την χρήση τιμών Shapley ώστε ο ειδικός να μπορεί να κατανοήσει την διαδικασία λήψης της απόφασης του μοντέλου. Επιπλέον επιλέχθηκαν τα χαρακτηριστικά, από το σύνολο δεδομένων, που έχουν την μεγαλύτερη επίδραση στην τελική πρόβλεψη του μοντέλου, μειώνοντας την διάσταση των δεδομένων σημαντικά.

Πάρα τα θετικά αποτελέσματα της μελέτης πρέπει να ληφθεί υπόψιν το σχετικά μικρό μέγεθος του συνόλου δεδομένων, στο οποίο εκπαιδευτικό το μοντέλο, όπως αυτό προέκυψε μετρά την επεξεργασία. Επίσης, είναι σημαντικό να ληφθούν υπόψιν τυχόν περιορισμοί που τίθενται από τα πρωτόκολλα των κλινικών δοκιμών του ADNI. Τα πρωτόκολλα αυτά καθορίζουν συγκεκριμένες μεθόδους συλλογής δεδομένων, τα κριτήρια επιλογής των συμμετεχόντων, τις διαδικασίες διαγνωστικής αξιολόγησης και άλλες παράμετρους της μελέτης. Αυτοί οι περιορισμοί μπορεί να επηρεάσουν την αναπαραγωγιμότητα των αποτελεσμάτων σε διαφορετικά περιβάλλοντα και να περιορίσουν τη γενίκευση των ευρημάτων σε πληθυσμούς που απέχουν από τις ομάδες συμμετεχόντων στην μελέτη του ADNI.

Παράλληλα, υπάρχει ανάγκη για περεταίρω έρευνα και διερεύνηση των μεθόδων πρόγνωσης της νόσου. Τέτοιο παράδειγμα είναι η επέκταση της παρούσας μελέτης σε μεγαλύτερο δείγμα πληθυσμού και ενδεχομένως σε δεδομένα που δεν συμπίπτουν με τα πρωτόκολλά του ADNI ώστε το μοντέλο που θα προκύψει να ανταποκρίνεται καλύτερα σε πραγματικές συνθήκες. Επιπλέον, θα μπορούσαν να αξιοποιηθούν και μέθοδοι έκτος της κλασσικής μηχανικής μάθησης όπως το Deep Reinforcement Learning ώστε να διερευνηθούν τυχόν τρόποι πρόγνωσης της νόσου όπου δεν μπορούν να υποστηριχθούν μαθηματικά από την κλασσική μηχανική μάθηση.

Αναφορές

- [1] J. Ramirez-Bermudez, “Alzheimer’s Disease: Critical Notes on the History of a Medical Concept,” *Arch Med Res*, vol. 43, no. 8, pp. 595–599, Nov. 2012, doi: 10.1016/J.ARCMED.2012.11.008.
- [2] C. Kavitha, V. Mani, S. R. Srividhya, O. I. Khalaf, and C. A. Tavera Romero, “Early-Stage Alzheimer’s Disease Prediction Using Machine Learning Models,” *Front Public Health*, vol. 10, p. 240, Mar. 2022, doi: 10.3389/FPUBH.2022.853294/BIBTEX.
- [3] “What is Dementia? Symptoms, Causes & Treatment | alz.org.” <https://www.alz.org/alzheimers-dementia/what-is-dementia> (accessed May 03, 2023).
- [4] Z. Breijyeh and R. Karaman, “Comprehensive Review on Alzheimer’s Disease: Causes and Treatment,” *Molecules*, vol. 25, no. 24. MDPI, Dec. 01, 2020. doi: 10.3390/MOLECULES25245789.
- [5] B. Dubois *et al.*, “Preclinical Alzheimer’s disease: Definition, natural history, and diagnostic criteria,” *Alzheimers Dement*, vol. 12, no. 3, p. 292, Mar. 2016, doi: 10.1016/J.JALZ.2016.02.002.
- [6] B. Dubois *et al.*, “Preclinical Alzheimer’s disease: Definition, natural history, and diagnostic criteria,” *Alzheimer’s & Dementia*, vol. 12, no. 3, pp. 292–323, Mar. 2016, doi: 10.1016/J.JALZ.2016.02.002.
- [7] “Alzheimers facts and figures report 2022,” 2022. Accessed: May 06, 2023. [Online]. Available: <https://www.alz.org/media/Documents/alzheimers-facts-and-figures-special-report-2022.pdf>
- [8] “Mild Cognitive Impairment (MCI) | Symptoms & Treatments | alz.org.” https://www.alz.org/alzheimers-dementia/what-is-dementia/related_conditions/mild-cognitive-impairment (accessed May 06, 2023).
- [9] O. L. Lopez, “Mild Cognitive Impairment,” *Continuum : Lifelong Learning in Neurology*, vol. 19, no. 2 Dementia, p. 411, Apr. 2013, doi: 10.1212/01.CON.0000429175.29601.97.
- [10] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, “Big data in healthcare: management, analysis and future prospects,” *J Big Data*, vol. 6, no. 1, pp. 1–25, Dec. 2019, doi: 10.1186/S40537-019-0217-0/FIGURES/6.
- [11] A. G. Vrahatis, K. Skolariki, M. G. Krokidis, K. Lazaros, T. P. Exarchos, and P. Vlamos, “Revolutionizing the Early Detection of Alzheimer’s Disease through Non-Invasive Biomarkers: The Role of Artificial Intelligence and Deep Learning,” *Sensors*, vol. 23, no. 9, p. 4184, Apr. 2023, doi: 10.3390/s23094184.

- [12] M. Ansart *et al.*, “Predicting the progression of mild cognitive impairment using machine learning: A systematic, quantitative and critical review,” *Med Image Anal*, vol. 67, p. 101848, Jan. 2021, doi: 10.1016/J.MEDIA.2020.101848.
- [13] H. A. Helaly, M. Badawy, and A. Y. Haikal, “Deep Learning Approach for Early Detection of Alzheimer’s Disease,” *Cognit Comput*, vol. 14, no. 5, pp. 1711–1727, Sep. 2022, doi: 10.1007/S12559-021-09946-2/FIGURES/15.
- [14] S. O. Danso, Z. Zeng, G. Muniz-Terrera, and C. W. Ritchie, “Developing an Explainable Machine Learning-Based Personalised Dementia Risk Prediction Model: A Transfer Learning Approach With Ensemble Learning Algorithms,” *Front Big Data*, vol. 4, p. 613047, May 2021, doi: 10.3389/FDATA.2021.613047/BIBTEX.
- [15] T. Zhang *et al.*, “Predicting MCI to AD Conversion Using Integrated sMRI and rs-fMRI: Machine Learning and Graph Theory Approach,” *Front Aging Neurosci*, vol. 13, p. 688926, Jul. 2021, doi: 10.3389/FNAGI.2021.688926/BIBTEX.
- [16] A. Singh, R. Kumar, and A. K. Tiwari, “Prediction of Alzheimer’s Using Random Forest with Radiomic Features,” *Computer Systems Science and Engineering*, vol. 45, no. 1, pp. 513–530, Aug. 2022, doi: 10.32604/CSSE.2023.029608.
- [17] C. Salvatore, A. Cerasa, P. Battista, M. C. Gilardi, A. Quattrone, and I. Castiglioni, “Magnetic resonance imaging biomarkers for the early diagnosis of Alzheimer’s disease: A machine learning approach,” *Front Neurosci*, vol. 9, no. SEP, 2015, doi: 10.3389/fnins.2015.00307.
- [18] T. Shen, J. Jiang, Y. Li, P. Wu, C. Zuo, and Z. Yan, “Decision Supporting Model for One-year Conversion Probability from MCI to AD using CNN and SVM,” *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2018-July, pp. 738–741, Oct. 2018, doi: 10.1109/EMBC.2018.8512398.
- [19] K. Skolariki, G. M. Terrera, and S. Danso, “Multivariate Data Analysis and Machine Learning for Prediction of MCI-to-AD Conversion,” *Adv Exp Med Biol*, vol. 1194, pp. 81–103, 2020, doi: 10.1007/978-3-030-32622-7_8/TABLES/7.
- [20] V. S. Diogo, H. A. Ferreira, and D. Prata, “Early diagnosis of Alzheimer’s disease using machine learning: a multi-diagnostic, generalizable approach,” *Alzheimers Res Ther*, vol. 14, no. 1, pp. 1–21, Dec. 2022, doi: 10.1186/S13195-022-01047-Y/FIGURES/4.
- [21] S. El-Sappagh, H. Saleh, F. Ali, E. Amer, and T. Abuhmed, “Two-stage deep learning model for Alzheimer’s disease detection and prediction of the mild cognitive impairment time,” *Neural Comput Appl*, vol. 34, no. 17, pp. 14487–14509, Sep. 2022, doi: 10.1007/S00521-022-07263-9/FIGURES/8.
- [22] M. Grassi *et al.*, “A novel ensemble-based machine learning algorithm to predict the conversion from mild cognitive impairment to Alzheimer’s disease using socio-demographic characteristics, clinical information, and neuropsychological measures,”

- Front Neurol*, vol. 10, no. JUL, p. 463970, Jul. 2019, doi: 10.3389/FNEUR.2019.00756/BIBTEX.
- [23] E. Ocasio and T. Q. Duong, “Deep learning prediction of mild cognitive impairment conversion to Alzheimer’s disease at 3 years after diagnosis using longitudinal and wholebrain 3D MRI,” *PeerJ Comput Sci*, vol. 7, pp. 1–21, May 2021, doi: 10.7717/PEERJ-CS.560/SUPP-2.
 - [24] E. Papadaki, T. Exarchos, P. Vlamos, and A. Vrahatis, “A Hybrid Deep Learning model for predicting the early Alzheimer’s Disease stages using MRI,” in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Sep. 2022. doi: 10.1145/3549737.3549779.
 - [25] F. Yi *et al.*, “XGBoost-SHAP-based interpretable diagnostic framework for alzheimer’s disease,” *BMC Med Inform Decis Mak*, vol. 23, no. 1, pp. 1–14, Dec. 2023, doi: 10.1186/S12911-023-02238-9/FIGURES/6.
 - [26] L. Nanni *et al.*, “Comparison of Transfer Learning and Conventional Machine Learning Applied to Structural Brain MRI for the Early Diagnosis and Prognosis of Alzheimer’s Disease,” *Front Neurol*, vol. 11, p. 576194, Nov. 2020, doi: 10.3389/FNEUR.2020.576194/BIBTEX.
 - [27] Z. Hu, Z. Wang, Y. Jin, and W. Hou, “VGG-TSwinformer: Transformer-based deep learning model for early Alzheimer’s disease prediction,” *Comput Methods Programs Biomed*, vol. 229, p. 107291, Feb. 2023, doi: 10.1016/J.CMPB.2022.107291.
 - [28] H. A. Helaly, M. Badawy, and A. Y. Haikal, “Deep Learning Approach for Early Detection of Alzheimer’s Disease,” *Cognit Comput*, vol. 14, no. 5, pp. 1711–1727, Sep. 2022, doi: 10.1007/S12559-021-09946-2/FIGURES/15.
 - [29] D. Pirrone, E. Weitschek, P. Di Paolo, S. De Salvo, and M. C. De Cola, “EEG Signal Processing and Supervised Machine Learning to Early Diagnose Alzheimer’s Disease,” *Applied Sciences 2022, Vol. 12, Page 5413*, vol. 12, no. 11, p. 5413, May 2022, doi: 10.3390/APP12115413.
 - [30] “ADNI | About.” <https://adni.loni.usc.edu/about/> (accessed Jun. 29, 2023).
 - [31] I. Arevalo-Rodriguez *et al.*, “Mini-Mental State Examination (MMSE) for the detection of Alzheimer’s disease and other dementias in people with mild cognitive impairment (MCI),” *Cochrane Database Syst Rev*, vol. 2015, no. 3, Mar. 2015, doi: 10.1002/14651858.CD010783.PUB2.
 - [32] “pandas documentation — pandas 2.0.1 documentation.” <https://pandas.pydata.org/docs/> (accessed May 14, 2023).
 - [33] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System”, Accessed: May 19, 2023. [Online]. Available: <https://github.com/dmlc/xgboost>

- [34] L. Bloch and C. M. Friedrich, “Using Bayesian Optimization to Effectively Tune Random Forest and XGBoost Hyperparameters for Early Alzheimer’s Disease Diagnosis,” *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, vol. 362 LNICST, pp. 285–299, 2021, doi: 10.1007/978-3-030-70569-5_18/TABLES/7.
- [35] S. M. Lundberg, P. G. Allen, and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions”, Accessed: Jun. 18, 2023. [Online]. Available: <https://github.com/slundberg/shap>

Παράρτημα Α

Πίνακας 11: Χαρακτηριστικά συνόλου δεδομένων πριν την προ-επεξεργασία.

RID
COLPROT
ORIGPROT
PTID
SITE
VISCODE
EXAMDATE
DX_b1
AGE
PTGENDER
PTEDUCAT
PTETHCAT
PTRACCAT
PTMARRY
APOE4
FDG
PIB
AV45
FBB
ABETA
TAU
PTAU
CDRSB
ADAS11
ADAS13
ADASQ4
MMSE
RAVLT_immediate
RAVLT_learning
RAVLT_forgetting
RAVLT_perc_forgetting
LDELTOTAL
DIGITSCOR
TRABSCOR
FAQ
MOCA
EcogPtMem
EcogPtLang
EcogPtVisspat
EcogPtPlan

EcogPtOrgan
EcogPtDivatt
EcogPtTotal
EcogSPMem
EcogSPLang
EcogSPVisspat
EcogSPPlan
EcogSPOrgan
EcogSPDivatt
EcogSPTotal
FLDSTRENG
FSVERSION
IMAGEUID
Ventricles
Hippocampus
WholeBrain
Entorhinal
Fusiform
MidTemp
ICV
DX
mPACCdigit
mPACCtrailsB
EXAMDATE_bl
CDRSB_bl
ADAS11_bl
ADAS13_bl
ADASQ4_bl
MMSE_bl
RAVLT_immediate_bl
RAVLT_learning_bl
RAVLT_forgetting_bl
RAVLT_perc_forgetting_bl
LDELTOTAL_BL
DIGITSCOR_bl
TRABSCOR_bl
FAQ_bl
mPACCdigit_bl
mPACCtrailsB_bl
FLDSTRENG_bl
FSVERSION_bl
IMAGEUID_bl
Ventricles_bl
Hippocampus_bl
WholeBrain_bl

Entorhinal_bl
Fusiform_bl
MidTemp_bl
ICV_bl
MOCA_bl
EcogPtMem_bl
EcogPtLang_bl
EcogPtVisspat_bl
EcogPtPlan_bl
EcogPtOrgan_bl
EcogPtDivatt_bl
EcogPtTotal_bl
EcogSPMem_bl
EcogSPLang_bl
EcogSPVisspat_bl
EcogSPPlan_bl
EcogSPOrgan_bl
EcogSPDivatt_bl
EcogSPTotal_bl
ABETA_bl
TAU_bl
PTAU_bl
FDG_bl
PIB_bl
AV45_bl
FBB_bl
Years_bl
Month_bl
Month
M
update_stamp

Παράρτημα Β

Κώδικας Exploratory Data Analysis

```
fig, ax = plt.subplots(nrows=1, ncols=5, figsize=(30,6))
axes = ax.ravel()
fields = ['AGE', 'PTEDUCAT', 'PTGENDER', 'PTETHCAT', 'PTMARRY']

# AGE
visits = unique_patients['AGE'].dropna(axis=0)
axes[0].hist(visits, bins=10, alpha=0.7)
ax[0].set_title('Age', fontsize=25)
ax[0].set_ylabel('Frequency', fontsize=20)
for tick in ax[0].xaxis.get_major_ticks():
    tick.label.set_fontsize(20)

# EDUCATION
visits = unique_patients['PTEDUCAT'].dropna(axis=0)
axes[1].hist(visits, bins=10, alpha=0.7)
ax[1].set_title('Education (years)', fontsize=25)
for tick in ax[1].xaxis.get_major_ticks():
    tick.label.set_fontsize(20)

# GENDER
visits = unique_patients['PTGENDER'].dropna(axis=0)
names = [val[0] for val in visits.value_counts().index.values]
heights = visits.value_counts().values
ax[2].bar(names, heights, alpha=0.7)
ax[2].set_title('Gender', fontsize=25)
for tick in ax[2].xaxis.get_major_ticks():
    tick.label.set_fontsize(20)

# RACE
visits = unique_patients['PTRACCAT'].dropna(axis=0)
names = visits.value_counts().index.values[0:3]
heights = visits.value_counts().values[0:3]
ax[3].bar(names, heights, alpha=0.7)
ax[3].set_title('Race', fontsize=25)
for tick in ax[3].xaxis.get_major_ticks():
    tick.label.set_fontsize(20)

# MARITAL STATUS
visits = unique_patients['PTMARRY'].dropna(axis=0)
names = ['Married', 'Widowed', 'Divorced', 'Unmarried']
heights = visits.value_counts().values[0:4]
ax[4].bar(names, heights, alpha=0.7)
ax[4].set_title('Marital Status', fontsize=25)
for tick in ax[4].xaxis.get_major_ticks():
    tick.label.set_fontsize(13)
```

```

data['VISCODE'] = data['VISCODE'].apply(keep_numbers)
[9]

data['VISCODE'] = np.where(data['VISCODE'] == 'bl', 0, data['VISCODE'])
[10]

data['VISCODE'] = data['VISCODE'].astype(int)
[11]

```

```

kde_df = data[['RID', 'M', 'DX']]
dicts = []
name_dict = {'Dementia': 'AD', 'MCI': 'MCI', 'CN': 'CN'}
kde_df['DX'] = kde_df['DX'].map(name_dict)
for m in sorted(kde_df['M'].unique()):
    df = kde_df[kde_df['M'] == m]
    dict_c = {'M': m}
    if len(df['DX'].value_counts()) > 2:
        dict_c.update(dict(df['DX'].value_counts()))
        dicts.append(dict_c)
table_df = pd.DataFrame(dicts)

fig, ax = plt.subplots(figsize=(12,6))
ax.plot(table_df['M'], table_df['AD'], c='red', label='AD')
ax.plot(table_df['M'], table_df['CN'], c='green', label='CN')
ax.plot(table_df['M'], table_df['MCI'], c='blue', label='MCI')
ax.set_title('Number of diagnoses over time', fontsize=25)
ax.set_ylabel('Number of diagnoses', fontsize=20)
ax.set_xlabel('Months after baseline', fontsize=20)
ax.legend(fontsize=20)
fig.tight_layout()
[13]

```

```

demographics = data[demographics_cols]
# plot
def plot_bar(df, col, title, xlabel, ylabel):
    fig, ax = plt.subplots(figsize=(12,6))
    df[col].value_counts().plot(kind='bar', ax=ax)
    ax.set_title(title, fontsize=25)
    ax.set_xlabel(xlabel, fontsize=20)
    ax.set_ylabel(ylabel, fontsize=20)
    fig.tight_layout()
plot_bar(unique_patients, 'DX_bl', 'Diagnosis at baseline', 'Diagnosis', 'Number of diagnoses')
plot_bar(unique_patients, 'DX', 'Diagnosis', 'Diagnosis', 'Number of diagnoses')

```

```
def plot_missing_data(df,viscode=0,colprot=None,title=None):
    adni_df= df[[col for col in data.columns if col[-3:] != "_bl"]]

    # plot missing values for baseline observations, sorted by protocol (i.e. ADNI stage)
    keep_cols = list(adni_df.columns[adni_df.isna().sum() > 10]) + ['VISCODE','COLPROT']
    if colprot is not None:
        adni_anymissing = adni_df.loc[(adni_df.VISCODE == viscode) & (adni_df.COLPROT == colprot),keep_cols].sort_values(['VISCODE','COLPROT'])
    else:
        adni_anymissing = adni_df.loc[(adni_df.VISCODE == viscode),keep_cols].sort_values(['VISCODE','COLPROT'])
    if len(adni_anymissing) == 0:
        print('No data for VISCODE {} and COLPROT {}'.format(viscode, colprot))
        return

    fig, ax = plt.subplots(figsize=(30,15))
    msno.matrix(adni_anymissing, ax=ax)
    ax.set_title(title, fontsize=25)
    fig.tight_layout()
    msno.matrix(adni_anymissing.loc[:, (adni_anymissing.columns != 'VISCODE') & (adni_anymissing.columns != 'COLPROT')],ax=ax)
```

```
color_dict = {'CN': 'red', 'AD': 'blue', 'Dementia': 'blue', 'LMCI': 'green', 'MCI': 'green', 'EMCI': 'orange',
              'SMC': 'purple'}
name_dict = {'CN': 'CN', 'Dementia': 'AD', 'MCI': 'MCI'}
traj_df = data[['RID', 'M', 'MMSE', 'DX_bl', 'DX']].dropna(axis=0)
traj_df = traj_df.groupby('RID').aggregate({'M': lambda x: list(x), 'MMSE': lambda x: list(x), 'DX_bl': 'max',
                                           'DX': 'max'})
diags = traj_df['DX'].unique()

fig, ax = plt.subplots(ncols=len(diags), nrows=1, figsize=(20,5))
axes = ax.ravel()

for i,dx in enumerate(diags):
    if i == 0:
        axes[i].set_ylabel('MMSE levels', fontsize=20)
    for diag, vals, mths in traj_df[traj_df['DX'] == dx][['DX', 'MMSE', 'M']].values:
        sort_mths = [mth for mth,val in sorted(zip(mths,vals))]
        sort_vals = [val for mth,val in sorted(zip(mths,vals))]
        axes[i].plot(sort_mths, sort_vals, 'o-', alpha=0.4, color=color_dict[dx])
    axes[i].set_title(name_dict[dx], fontsize=25)
    axes[i].set_xlabel('Months after baseline', fontsize=20)
    axes[i].set_ylim((0,31))
    for tick in axes[i].xaxis.get_major_ticks():
        tick.label.set_fontsize(20)
fig.tight_layout()
```

Patients across all adni phases

```
adni_groups_obs = data.groupby(['COLPROT','VISCODE']).size().reset_index(name='count')
adni_groups_obs = adni_groups_obs.pivot(index='VISCODE', columns='COLPROT', values='count')

adni_groups_obs.plot(kind='bar', figsize=(15,5), title='Number of observations per protocol and visit')

<Axes: title={'center': 'Number of observations per protocol and visit'}, xlabel='VISCODE'>
```

```

baseline_dx.groupby(['DX_bl', 'COLPROT']).size().reset_index(name='count')
baseline_dx.groupby(['DX_bl', 'COLPROT']).size().\
reset_index(name='count').pivot(index='DX_bl', columns='COLPROT', values='count').plot(kind='bar', figsize=(15,5), title='Number of patients per diagnosis at baseline and protocol')

#count transitions per viscode
new = new[new['Transition'] != 0]
new.groupby(['VISCODE', 'Transition']).size().reset_index(name='count').pivot(index='VISCODE',
columns='Transition', values='count').plot(kind='bar',
figsize=(15,5), title='Number of transitions per visit')

```

Κώδικας προ-επεξεργασίας του συνόλου δεδομένων.

```

data.dropna(subset=['DX'], inplace=True)

m12 = data[data['Month'] == 12]

dx_48 = data[(data['Month'] <= 48) & (data['Month'] > 12)][['DX', 'RID', 'Month']]
dx_48.sort_values(by=['RID', 'Month'], inplace=True)
dx_48.drop(columns=['Month'], inplace=True)
dx_48.drop_duplicates(subset=['RID'], keep='last', inplace=True)
dx_48.rename(columns={'DX': 'DX_48'}, inplace=True)

df = pd.merge(m12, dx_48, on='RID')

df['DX_bl'].replace({'EMCI': 'MCI', 'LMCI': 'MCI'}, inplace=True)

df = df[df['DX_bl'] != 'SMC']

df['DX_bl'] = df['DX_bl'].map({'CN': 0, 'MCI': 1, 'AD': 2})

df['DX'] = df['DX'].map({'CN': 0, 'MCI': 1, 'Dementia': 2})

```

```

test = df['DX_bl']!=df['DX']
test.value_counts()

False    1202
True      134
dtype: int64

df['DX_48'] = df['DX_48'].map({'CN':0, 'MCI':1, 'Dementia':2})

df = df[df['DX_bl']!=0]

df['Transition'] = df['DX_bl'] != df['DX_48']

df['Transition'] = df['Transition'].map({True:1, False:0})

df['Transition'].value_counts()

0      621
1      336
Name: Transition, dtype: int64

```

```

df = df.drop(['ORIGPROT', 'SITE', 'EXAMDATE', 'update_stamp', 'FLDSTRENG', 'PTID',
             'FLDSTRENG_bl', 'EXAMDATE_bl',
             'FSVERSION',
             'FSVERSION_bl', 'IMAGEUID', 'IMAGEUID_bl'], axis=1)

genes = ['TAU_bl', 'ABETA_bl', 'PTAU_bl', 'TAU', 'ABETA', 'PTAU']

for val in genes:
    all_vals = df[val].values
    df = df.drop(val, axis=1)
    df[val] = [float(str(x).replace('<', '').replace('>', ''))) for x in all_vals]

df.drop(['VISCODE', 'Month'], axis=1, inplace=True)

cat_cols = ['PTGENDER', 'PTRACCAT', 'PTMARRY', 'PTETHCAT']

df = pd.get_dummies(df, columns=cat_cols, drop_first=False)

```

```

from sklearn.impute import KNNImputer

}

imputer = KNNImputer(n_neighbors=5)

}

adni1_missing_cols = ['MOCA',
    'EcogPtMem',
    'EcogPtLang',
    'EcogPtVisspat',
    'EcogPtPlan',
    'EcogPtOrgan',
    'EcogPtDivatt',
    'EcogPtTotal',
    'EcogSPMem',
    'EcogSPLang',
    'EcogSPVisspat',
    'EcogSPPlan',
    'EcogSPOrgan',
    'EcogSPDivatt',
    'EcogSPTotal',]
adni1_missing_cols = [col + '_bl' for col in adni1_missing_cols ] + adni1_missing_cols

}

}

adni1.drop(['FBB', 'FBB_bl', 'PIB', 'PIB_bl', 'DIGITSCOR', 'PTAU', 'ABETA', 'AV45_bl', 'AV45', 'DIGITSCOR_bl'], axis=1, inplace=True)

}

```

```

def impute_missing(df, cols_to_exclude, target):
    df = df.copy()
    imputer = KNNImputer(n_neighbors=5)
    imputed = imputer.fit_transform(df.drop(cols_to_exclude, axis=1), df[target])
    imputed = pd.DataFrame(imputed, columns=df.drop(cols_to_exclude, axis=1).columns)
    return imputed

adni1 = df[df['COLPROT'] == 'ADNI1']
imputed_adni1 = impute_missing(adni1, adni1_missing_cols+['DX_48', 'DX', 'RID', 'COLPROT'], 'DX')
imputed_adni1['RID'] = adni1['RID'].values
imputed_adni1['DX'] = adni1['DX'].values
imputed_adni1['DX_48'] = adni1['DX_48'].values
imputed_adni1['COLPROT'] = adni1['COLPROT'].values

imputed_adni1.reset_index(inplace=True, drop=True)
adni1.reset_index(inplace=True, drop=True)

rest_of_adni = df[df['COLPROT'] != 'ADNI1']
rest_of_adni.reset_index(inplace=True, drop=True)

imputed = impute_missing(rest_of_adni, ['RID', 'COLPROT', 'DX_48', 'DX', 'Transition'], 'DX')

```

```
imputed['RID'] = rest_of_adni['RID']
imputed[['DX', 'DX_48', 'Transition', 'COLPROT']] = rest_of_adni[['DX', 'DX_48', 'Transition', 'COLPROT']]
```

```
pd.concat([imputed_adni1, imputed]).to_csv('preprocessed1.csv', index=False)
```

```
477     ADNIGO
Name: COLPROT, dtype: object
```

```
imputed = pd.merge(imputed, df[['RID', 'DX_48', 'DX', 'Transition']], on='RID')
```

```
imputed.to_csv('preprocessed.csv', index=False)
```

Κώδικας επιλογής χαρακτηριστικών

```
from sklearn.feature_selection import RFECV
```

```
model_rfe = xgb.XGBClassifier(n_estimators=600, learning_rate=0.01, max_depth=5, scale_pos_weight=2, enable_categorical=True, tree_method='gpu_hist', subsample=0.8, colsample_bytree=0.8)
```

```
rfe = RFECV(model_rfe, step=1, cv=5, scoring='roc_auc', verbose=1, n_jobs=-1)
```

```
rfe.fit(X, Y)
```

```
print(X.columns[rfe.get_support()])
```

```
Index(['DX_bl', 'APOE4', 'CDRSB', 'ADAS11', 'ADAS13', 'ADASQ4',
      'RAVLT_immediate', 'LDELTOTAL', 'TRABSCOR', 'FAQ', 'Hippocampus',
      'Entorhinal', 'MidTemp', 'ICV', 'mPACCdigit', 'mPACCtrailsB',
      'CDRSB_bl', 'ADAS11_bl', 'ADAS13_bl', 'MMSE_bl', 'RAVLT_learning_bl',
      'TRABSCOR_bl', 'FAQ_bl', 'mPACCdigit_bl', 'mPACCtrailsB_bl',
      'Entorhinal_bl', 'Fusiform_bl', 'MidTemp_bl', 'ICV_bl',
      'PTMARRY_Widowed', 'PTETHCAT_Not_Hisp/Latino', 'DX', 'EcogPtMem',
      'EcogPtPlan', 'EcogSPPlan', 'EcogSPOrgan', 'EcogSPDivatt',
      'EcogSPTotal', 'MOCA_bl', 'EcogPtMem_bl', 'EcogSPLang_bl',
      'EcogSPTotal_bl'],
      dtype='object')
```

```
data_selected = data[X.columns[rfe.get_support()]]
```

Κώδικας εκπαίδευσης του μοντέλου

```
import numpy as np
import pandas as pd
import xgboost as xgb
from sklearn.model_selection import StratifiedKFold
from sklearn.metrics import roc_auc_score, classification_report, confusion_matrix, ConfusionMatrixDisplay, recall_score, precision_score, accuracy_score
import mlflow

data = pd.read_csv('data/selected_features.csv')

X = data.drop(['Transition'], axis=1)
Y = data['Transition']

kfold = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)

mlflow.start_run()

i = 0
recall = []
precision = []
accuracy = []
roc_auc = []
for train_index, test_index in kfold.split(X, Y):
    i += 1
    X_train, X_test = X.iloc[train_index], X.iloc[test_index]
    Y_train, Y_test = Y.iloc[train_index], Y.iloc[test_index]

    train = xgb.DMatrix(X_train, label=Y_train)
    test = xgb.DMatrix(X_test, label=Y_test)
    param = {'max_depth': 5, 'eta': 0.01, 'objective': 'binary:logistic', 'scale_pos_weight': 2, 'tree_method': 'gpu_hist', 'subsample': 0.8, 'colsample_bytree': 0.8}
    model = xgb.train(param, train, num_boost_round=600)
    pred = model.predict(test)
    pred = np.where(pred > 0.5, 1, 0)

    print("Classification Report: Fold: ", i)
    print(classification_report(Y_test, pred, target_names=['No', 'Yes']))
    ConfusionMatrixDisplay(confusion_matrix(Y_test, pred), display_labels=['No', 'Yes']).plot()
    print("ROC AUC SCORE: ", np.round(roc_auc_score(Y_test, pred), 3))

    recall.append(np.round(recall_score(Y_test, pred), 3))
    precision.append(np.round(precision_score(Y_test, pred), 3))
    accuracy.append(np.round(accuracy_score(Y_test, pred), 3))
    roc_auc.append(np.round(roc_auc_score(Y_test, pred), 3))

# Log average scores per class
mlflow.log_metric('average_recall', np.mean(recall))
mlflow.log_metric('average_precision', np.mean(precision))
mlflow.log_metric('average_accuracy', np.mean(accuracy))
mlflow.log_metric('average_roc_auc', np.mean(roc_auc))

final_train = input("Do you want to train the model on the whole dataset? (y/n): ")

model = xgb.XGBClassifier(n_estimators=600, learning_rate=0.01, max_depth=5, scale_pos_weight=2, enable_categorical=True, tree_method='gpu_hist', subsample=0.8,
                           colsample_bytree=0.8)

if final_train == 'y':
    param = {'max_depth': 5, 'eta': 0.01, 'objective': 'binary:logistic', 'scale_pos_weight': 2, 'tree_method': 'gpu_hist', 'subsample': 0.8, 'colsample_bytree': 0.8}
    train = xgb.DMatrix(X, label=Y)
    model = xgb.train(param, train, num_boost_round=600)
    pred = model.predict(test)
    pred = np.where(pred > 0.5, 1, 0)

    print("Classification Report: Fold: ", i)
    print(classification_report(Y_test, pred, target_names=['No', 'Yes']))
    ConfusionMatrixDisplay(confusion_matrix(Y_test, pred), display_labels=['No', 'Yes']).plot()
    print("ROC AUC SCORE: ", np.round(roc_auc_score(Y_test, pred), 3))

    mlflow.log_artifact('model_api.bst')
else:
    print("Model not saved")

mlflow.end_run()
```

Κώδικας γραφημάτων Shapley

```
import shap
shap.initjs()
```



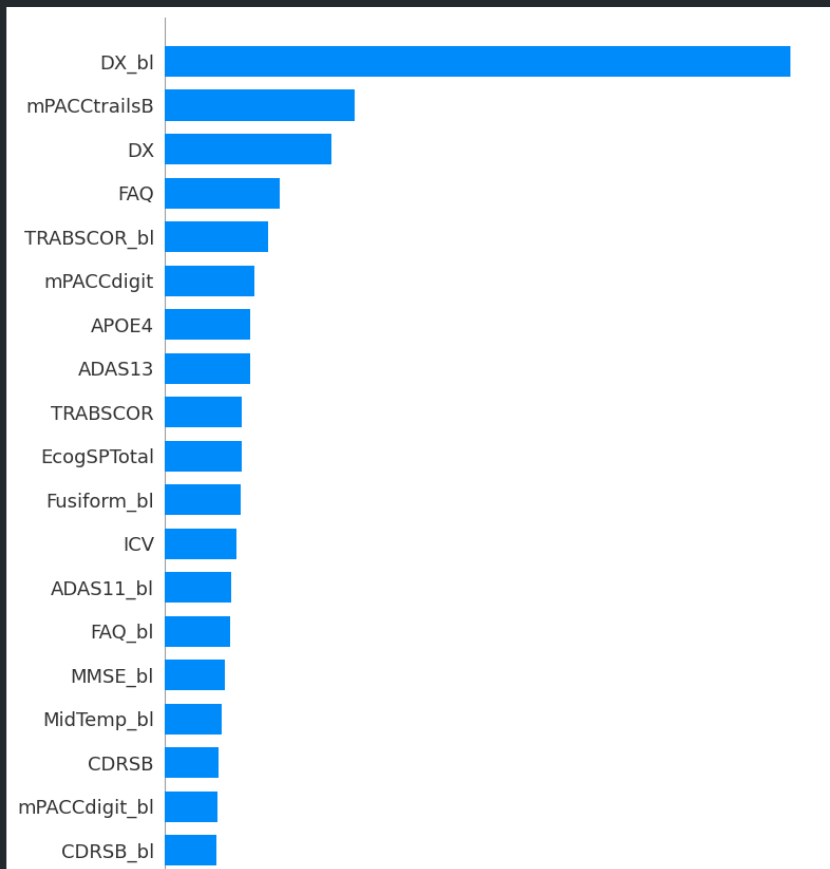
```
explainer = shap.TreeExplainer(final_model)
```

```
shap_values = explainer.shap_values(X)
```

```
shap.force_plot(explainer.expected_value, shap_values[0,:], X.iloc[0,:])
```

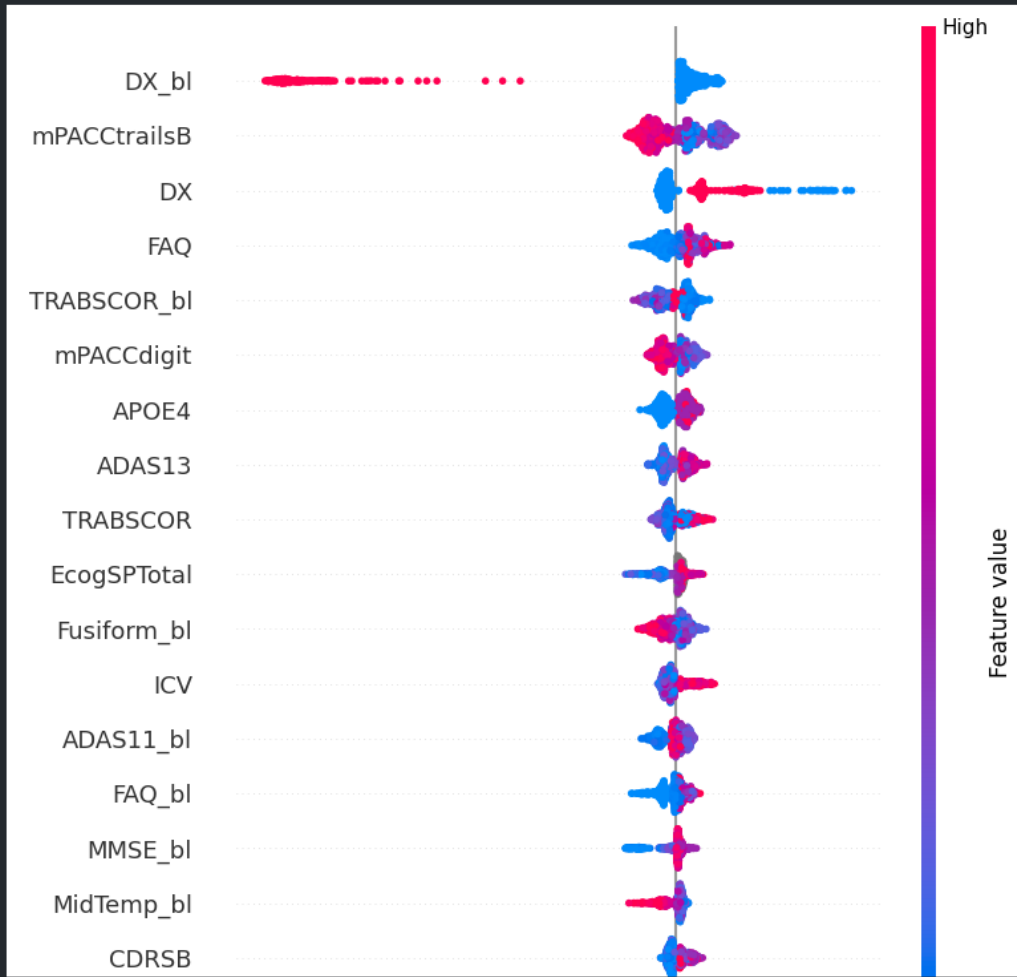
higher ⇌ lower

```
shap.summary_plot(shap_values, X, plot_type='bar')
```

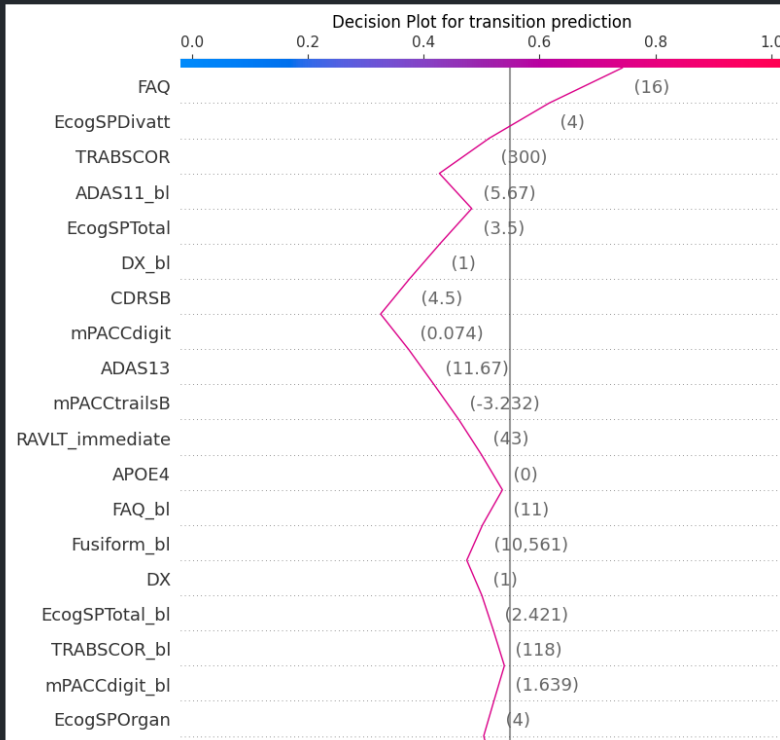


```
shap.summary_plot(shap_values,X)
```

No data for colormapping provided via 'c'. Parameters 'vmin', 'vmax' will be ignored



```
#decision plot for 1 prediction
shap.decision_plot(explainer.expected_value,shap_values[956:957,:],X.iloc[956:957,:],link='logit',title='Decision Plot for transition prediction')
```



Κώδικας διαδικτυακής εφαρμογής.

- Home.js

```
import React from 'react';
import { Container, Row, Col, Button, Image } from 'react-bootstrap';
import './HomeScreen.css';
import Form from '../components/Form'

const Home = () => {
  return (
    <Container fluid className="home-screen">
      <Row className="justify-content-center align-items-center">
        <Col md={6} className="text-center">
          <Image src={require('./images/alzheimer.png')} alt="Machine Learning Illustration" className="ml-illustration" />
        </Col>
        <Col md={6}>
          <div className="home-content">
            <h2>Cognitive Impairment Prediction using eXtreme Gradient Boosting</h2>
            You, 1 second ago • Uncommitted changes
          </div>
        </Col>
      </Row>
      <br/>
      <Row>
        <Form />
      </Row>
    </Container>
  );
};

export default Home;
```

- Form.js

```
You, 2 months ago | 2 authors (You and others)
import React, { useState, useEffect } from 'react';
import { Container, Form, Button, Row, Col, Modal } from 'react-bootstrap';
import axios from 'axios';
import { Bar } from 'react-chartjs-2';
import { Chart, registerables } from 'chart.js';
Chart.register(...registerables);

function PredictionForm() {
  const [features, setFeatures] = useState([]);
  const [formData, setFormData] = useState({});
  const [showModal, setShowModal] = useState(false);
  const [predictionResult, setPredictionResult] = useState(null);
  const [chartData, setChartData] = useState(null);

  useEffect(() => {
    loadFeatures();
  }, []);

  const loadFeatures = () => {
    try {
      const featuresData = require('./features.json');
      const nullFormData = {};
      featuresData.forEach((feature) => {
        if (feature.id === 'DX' || feature.id === 'DX_bl') {
          feature.type = 'dropdown';
          feature.options = ['', 'MCI', 'AD'];
        }
        nullFormData[feature.id] = null;
      });
      setFormData(nullFormData);
      setFeatures(featuresData);
    } catch (error) {
      console.error('Error loading features:', error);
    }
  }
}
```

```

const handleInputChange = (e) => {
  setFormData({
    ...formData,
    [e.target.name]: e.target.value,
  });
};

const handleSubmit = async (e) => {
  e.preventDefault();

  try {
    const response = await axios.post(
      'http://localhost:5000/predict',
      JSON.stringify(formData),
      {
        headers: {
          'Content-Type': 'application/json',
        },
      }
    );

    const predictionResult = response.data;
    console.log('Prediction result:', predictionResult);
    setPredictionResult(predictionResult);

    const chartData = {
      labels: ['Conversion', 'Stable'],
      datasets: [
        {
          label: 'Probability',
          data: [predictionResult.conversion_prob, predictionResult.stable_prob],
          backgroundColor: ['rgba(75, 192, 192, 0.6)', 'rgba(54, 162, 235, 0.6)'],
          borderColor: ['rgba(75, 192, 192, 1)', 'rgba(54, 162, 235, 1)'],
          borderWidth: 1,
        },
      ],
    };

    setChartData(chartData);
    setShowModal(true);
  }
};

```

```

    setChartData(chartData);
    setShowModal(true);
  } catch (error) {
    console.error('Error occurred during prediction:', error);
  }
};

const handleCloseModal = () => {
  setShowModal(false);
  setPredictionResult(null);
  setChartData(null);
};

return (
  <Container className="prediction-form">
    <h2 className="text-center">Features Input</h2>
    {features.length === 0 ? (
      <Button variant="primary" onClick={loadFeatures}>
        Load Features
      </Button>
    ) : (
      <Form onSubmit={handleSubmit}>
        <Row>
          {features.map((feature) => (
            <Col md={6} key={feature.id}>
              <Form.Group controlId={feature.id}>
                <Form.Label className="form-label">{feature.label}</Form.Label><br/>
                <Form.Text className="desc">{feature.description}</Form.Text>
                {feature.type === 'dropdown' ? (
                  <Form.Control
                    as="select"
                    name={feature.id}
                    className="form-control"
                    onChange={handleInputChange}
                    value={formData[feature.id] || ''}
                    required
                  >
                    {feature.options.map((option) => (
                      <option key={option} value={option}>
                        {option}
                      </option>
                    ))}
                  </Form.Control>
                ) : (
                  <Form.Control
                    type={feature.type}
                    name={feature.id}
                    className="form-control"
                    onChange={handleInputChange}
                    value={formData[feature.id] || ''}
                    required

```

```

        value={formData[feature.id] || ''}
        required
      )
    )
  }
  </Form.Group>
</Col>
))
</Row>
<Button variant="primary" className="submit-button" onClick={handleSubmit}>
  Submit
</Button>
</Form>
)}

<Modal show={showModal} onHide={handleCloseModal}>
  <Modal.Header closeButton>
    <Modal.Title>Prediction Result</Modal.Title>
  </Modal.Header>
  <Modal.Body>
    {predictionResult && (
      <
        <Bar data={chartData} />
        <p>Conversion from MCI to AD Probability: {predictionResult.conversion_prob}</p>
        <p>Remain Stable to MCI Probability: {predictionResult.stable_prob}</p>
      </
    )}
  </Modal.Body>
  <Modal.Footer>
    <Button variant="secondary" onClick={handleCloseModal}>
      Close
    </Button>
  </Modal.Footer>
</Modal>
</Container>
);
}

export default PredictionForm;

```

- Backend – App.py

```

import json
from flask import Flask, request, jsonify
import xgboost as xgb
import pandas as pd
import numpy as np
from flask_cors import CORS

app = Flask(__name__)
CORS(app)
model = xgb.Booster()
model.load_model('model.bst')

@app.route('/')
def hello():
    return 'Cognitive Impairment Prediction using eXtreme Gradient Boosting API'

@app.route('/predict', methods=['POST'])
def predict():
    # Load xgboost model

    # Get JSON data from the request
    json_data = request.get_json()

    for key, value in json_data.items():
        if value == "":
            json_data[key] = None

    if 'DX' in json_data:
        if json_data['DX'] == 'MCI':
            json_data['DX'] = 1
        elif json_data['DX'] == 'AD':
            json_data['DX'] = 2
    if 'DX_bl' in json_data:
        if json_data['DX_bl'] == 'MCI':
            json_data['DX_bl'] = 1
        elif json_data['DX_bl'] == 'AD':
            json_data['DX_bl'] = 2

    # Perform prediction using the loaded model
    data = pd.DataFrame.from_dict(json_data, orient='index').transpose()
    # convert all data to float
    data = data.astype(float)

```

```

data = pd.DataFrame.from_dict(json_data, orient='index').transpose()
# convert all data to float
data = data.astype(float)
# data['DX'] = data['DX'].astype(int) if data['DX'] is not None else None
# change type of data

data = xgb.DMatrix(data)
prediction = model.predict(data)
prediction = prediction * 100

conversion_prob = {
    "conversion_prob": np.round(prediction[0], 2).item(),
    "stable_prob": np.round(100 - np.round(prediction[0], 2), 2).item()
}

return jsonify(conversion_prob)

if __name__ == '__main__':
    app.run(host='0.0.0.0', port=5000, debug=True)

```