Δράση «Εμβληματικές δράσεις σε διαθεματικές επιστημονικές περιοχές με ειδικό ενδιαφέρον για την σύνδεση με τον παραγωγικό ιστό» ID 16618

Εθνικό δίκτυο έρευνας για την ανάδειξη της γενετικής βάσης των νευροεκφυλιστικών νόσων Alzheimer και Parkinson, την ανίχνευση αξιόπιστων βιοδεικτών και την ανάπτυξη καινοτόμων υπολογιστικών τεχνολογιών και θεραπευτικών στρατηγικών στη βάση της ιατρικής ακριβείας (BRAIN PRECISION, TAEDR-0535850)

ΤΙΤΛΟΣ ΠΑΡΑΔΟΤΕΟΥ: Υπολογιστική ανάλυση και πρόβλεψη σε ετερογενή αλληλοεπιδρώντα δεδομένα

ΕΝΟΤΗΤΑ ΕΡΓΑΣΙΑΣ 5: Ανάπτυξη νέων υπολογιστικών μοντέλων και τεχνολογιών για την έγκαιρη διάγνωση των νευροεκφυλιστικών νόσων Alzheimer και Parkinson και των πρόδρομων μορφών τους.

ΥΠΕΥΘΥΝΗ ΕΡΕΥΝΗΤΙΚΗ ΟΜΑΔΑ (ΦΟΡΕΑΣ): ΠΑΝΑΓΙΩΤΗΣ ΒΛΑΜΟΣ (ΙΠ)

Υπολογιστική ανάλυση και πρόβλεψη σε ετερογενή αλληλοεπιδρώντα δεδομένα

1. Εισαγωγή

Η συλλογή δεδομένων ασθενών υπό την μορφή Ηλεκτρονικού Φακέλου Υγείας (EHR) έχει εφαρμοστεί σε πάνω από το 90% των νοσοκομείων και κλινικών στις Ηνωμένες Πολιτείες, δημιουργώντας ένα τεράστιο αποθετήριο δεδομένων ασθενών που χρησιμεύει ως πολύτιμη πηγή πραγματικών δεδομένων (real-word data) για την έρευνα (Office of the National Coordinator for Health Information Technology, 2022; Kim et al., 2023). Σε αντίθεση με τα παραδοσιακά μητρώα ασθενών ή τις βάσεις δεδομένων ασφαλιστικών εταιρειών, τα οποία συχνά παρουσιάζουν σημαντικές χρονικές καθυστερήσεις, τα συστήματα EHR καταγράφουν συνεχώς ενημερωμένα, διαχρονικά δεδομένα που παράγονται κατά τη διάρκεια της κλινικής περίθαλψης. Αυτά τα αρχεία περιλαμβάνουν τόσο δομημένα δεδομένα — όπως κωδικοποιημένες διαγνώσεις, εργαστηριακά αποτελέσματα, συνταγογραφούμενα φάρμακα και δημογραφικές πληροφορίες — όσο και μη δομημένα δεδομένα συμπεριλαμβανομένων των σημειώσεων των γιατρών, των περιλήψεων εξιτηρίου και των αναφορών του ιστορικού των ασθενών. Η αμεσότητα και το βάθος των δεδομένων που προέρχονται από EHR παρέχουν σημαντικές πληροφορίες για την ανάπτυξη προγνωστικών μοντέλων και τη δημιουργία πραγματικών στοιχείων στον τομέα της ψηφιακής υγείας. Ένα σημαντικό πλεονέκτημά της εκπαίδευσης μοντέλων σε πραγματικά κλινικά δεδομένα είναι βελτιωμένη γενικευσιμότητα (Bakouny και Patt, 2021; Rashidisabet et al.,2023; Amrollahi et al., 2022). Μαθαίνοντας από διαφορετικούς και ετερογενείς πληθυσμούς ασθενών, αυτά τα μοντέλα αντικατοπτρίζουν καλύτερα τη μεταβλητότητα που συναντάται στην κλινική πράξη, αυξάνοντας την ανθεκτικότητα και την αξιοπιστία όταν εφαρμόζονται σε νέα περιβάλλοντα. Ωστόσο, τα RWD χρειάζονται ιδιαίτερη προσοχή ως προς τις πιθανές μεροληψίες, αφού αυτές μπορεί να εμφανιστούν σε πολλά σημεία της διαδικασίας, από τη δημιουργία και την εξαγωγή των δεδομένων μέχρι και τη μοντελοποίησή τους.

Η πλήρης αξιοποίηση του δυναμικού των RWD απαιτεί την υπέρβαση αρκετών κρίσιμων προκλήσεων (Bastarache et al., 2022; Collins καιTabak, 2014). Τα δεδομένα EHR είναι εγγενώς ετερογενή, συχνά αδόμητα και συχνά ελλιπή, απαιτώντας προηγμένες τεχνικές προ επεξεργασίας, κανονικοποίησης (Kim και Min, 2025), ολοκλήρωσης για αποτελεσματική μάθηση. Προκλήσεις όπως οι ελλείπουσες τιμές (Ren et al., 2024), η ακανόνιστη δειγματοληψία δεδομένων στο χρόνο (Chauhan et al., 2024) και οι συστηματικές μεροληψίες στη συλλογή

δεδομένων (Al-Sahab et al., 2024) μπορούν να επηρεάσουν σημαντικά την απόδοση του μοντέλου, εάν δεν αντιμετωπιστούν σωστά. Πρόσφατες έρευνες έχουν δείξει πολλά υποσχόμενα αποτελέσματα στην αναγνώριση της νόσου του Αλτσχάιμερ και συναφών μορφών άνοιας σε πρώιμαστάδια, χρησιμοποιώντας μεθόδους μηχανικής μάθησης σε δεδομένα EHR. Πολλές μελέτες έχουν δείξει ότι τόσο τα δομημένα όσο και τα αδόμητα κλινικά δεδομένα, όπως το ιστορικό φαρμακευτικής αγωγής, οι κλινικές περιγραφές κατά συμπεριφορικά πρότυπα, μπορούν να αξιοποιηθούν για την ανίχνευση δεικτών γνωστικής έκπτωσης (Ford et al., 2019; Jammeh et al., 2018).

Ορισμένες προσεγγίσεις έχουν προτείνει τη χρήση παθητικών ψηφιακών υπογραφών, οι οποίες εξάγονται από χρονικά δεδομένα Ηλεκτρονικών Φακέλων Υγείας (EHR), για τον προσδιορισμό του κινδύνου άνοιας χρόνια πριν από την εμφάνιση των συμπτωμάτων (Boustani et al., 2020), ενώ άλλες έχουν βελτιώσει την ακρίβεια πρόβλεψης χρησιμοποιώντας δείκτες πολυγονιδιακού κινδύνου (polygenic risk scores), συμπεριφορικά συμπτώματα και κοινωνικοοικονομικούς παράγοντες εντός συνόλων δεδομένων πληθυσμού μεγάλης κλίμακας (Gao et al., 2023; Li et al., 2023). Μια άλλη μελέτη εφάρμοσε τη μάθηση ετικετών (label learning) σε δεδομένα αποζημιώσεων και Ηλεκτρονικών Φακέλων Υγείας (EHR) μεγάλης κλίμακας, επιτυγχάνοντας ισχυρή προγνωστική απόδοση για την εμφάνιση της νόσου Αλτσχάιμερ εντός 2 ετών, αντιμετωπίζοντας τη διαγνωστική αβεβαιότητα που ενυπάρχει στα διοικητικά σύνολα δεδομένων (Nori et al., 2019). Περαιτέρω έρευνες έχουν καταδείξει τη σημασία των πολυτροπικών μοντέλων, τα οποία συνδυάζουν δεδομένα EHR με περιβαλλοντικούς και κοινωνικούς παράγοντες, για την αποτύπωση της ετερογένειας της νόσου (Tang et al., 2024). Οι μελέτες αυτές υπογραμμίζουν επίσης τη σημασία της χρονικής δυναμικής και της λειτουργικής έκπτωσης στη βελτίωση της πρόβλεψης σε χρονικά παράθυρα 1 έως 5 ετών, καθώς και την αυξανόμενη χρήση μεθόδων εξηγήσιμης μηχανικής μάθησης για τον εντοπισμό βασικών προγνωστικών παραγόντων, όπως η υπνική άπνοια, ο αποπροσανατολισμός, τα καταθλιπτικά συμπτώματα και οι συνοσηρότητες (Akter et al., 2025).

2. Μεθοδολογία ανάλυσης ετερογενών αλληλοεπιδρώντων δεδομένων της νόσου Αλτσχάιμερ

2.1 Δεδομένα

Η μελέτη βασίστηκε στην ανάλυση δεδομένων πραγματικού κόσμου (RWD) από το σύστημα ηλεκτρονικών φακέλων υγείας (EHR – Epic) του Johns Hopkins Health System. Αρχικά, η βάση δεδομένων περιείχε 685.765 εγγραφές. Ωστόσο, μετά την εφαρμογή συγκεκριμένων κριτηρίων επιλογής που αφορούσαν το προφίλ των ασθενών, τα χαρακτηριστικά των επισκέψεων σε δομές πρωτοβάθμιας φροντίδας και μνήμης, καθώς και την πληρότητα των στοιχείων, το τελικό δείγμα της μελέτης περιορίστηκε σε 197.481 ασθενείς. Το υποσύνολο αυτό

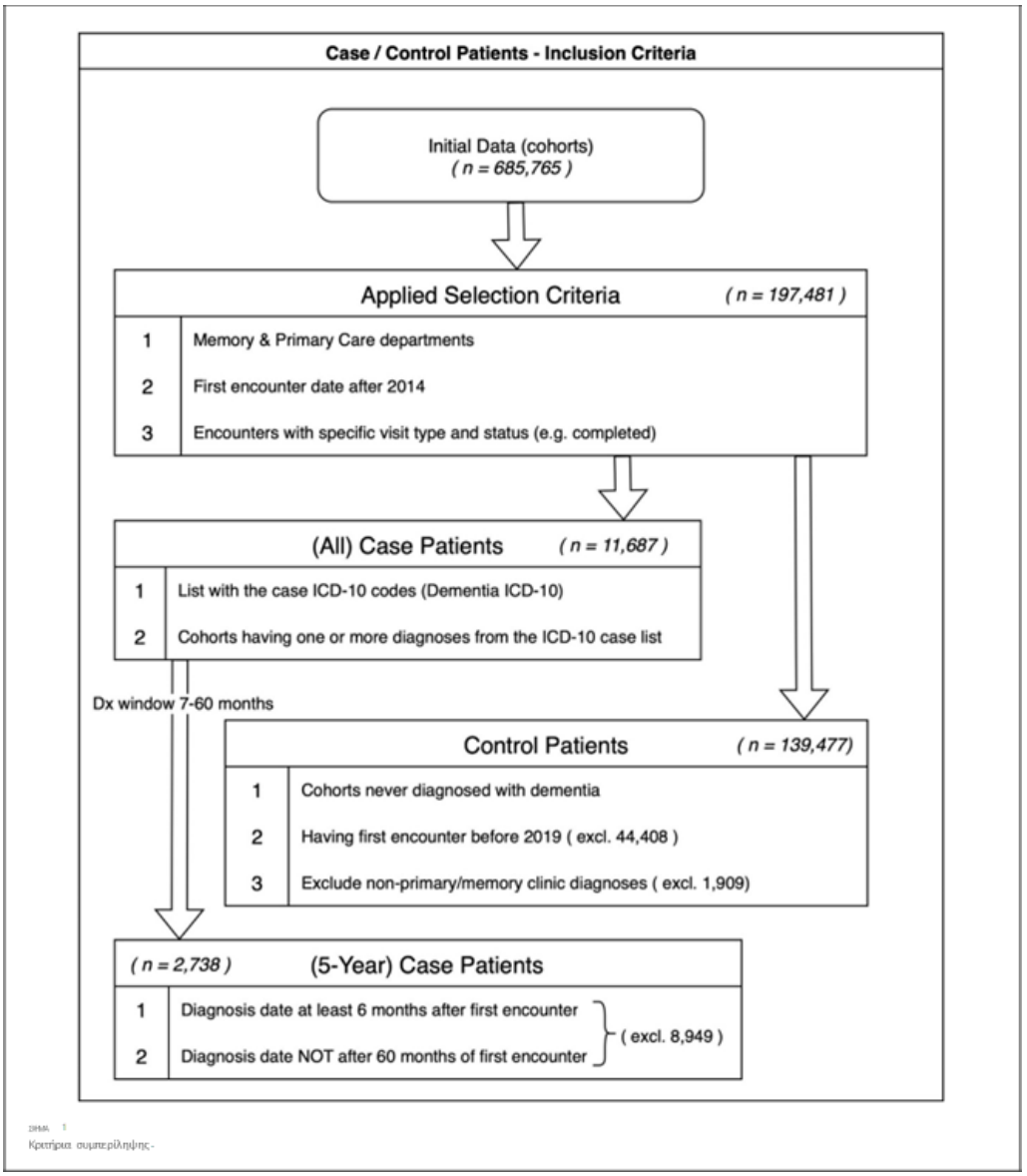
περιλαμβάνει άτομα για τα οποία υπάρχουν διαθέσιμα δεδομένα δεκαετίας, καλύπτοντας την περίοδο από 1η Ιανουαρίου 2014 έως 31 Δεκεμβρίου 2023. Τα δεδομένα προέρχονται τόσο από μονάδες πρωτοβάθμιας Φροντίδας Υγείας όσο και από εξειδικευμένες Κλινικές Μνήμης του συστήματος Johns Hopkins (Εικόνα 1). Συγκεκριμένα, η κατηγορία των κλινικών μνήμης αφορά εξωτερικούς ασθενείς του Johns Hopkins Memory and Alzheimer's Treatment Center (JHMATC) στη Βαλτιμόρη, οι οποίοι επισκέφθηκαν το κέντρο τουλάχιστον μία φορά μέσα στην περίοδο αναφοράς. Αντίστοιχα, τα δεδομένα πρωτοβάθμιας φροντίδας περιλαμβάνουν εξωτερικούς ασθενείς με τουλάχιστον μία επίσκεψη στο δίκτυο κλινικών του Johns Hopkins στην ευρύτερη περιοχή του Μέριλαντ και της Ουάσιγκτον (DC). Ο Πίνακας 1 παρουσιάζει αναλυτικά τα δημογραφικά και κλινικά χαρακτηριστικά του πληθυσμού. Μετά την ταξινόμηση των ασθενών με βάση το στάδιο Γνωστικής Έκπτωσης (CI), το τελικό δείγμα διαμορφώθηκε σε 142.175 άτομα. Η συντριπτική πλειονότητα (139.437 ασθενείς ή 98,1%) εντάχθηκε στην ομάδα ελέγχου, ενώ 2.738 ασθενείς (1,9%) κατατάχθηκαν στην ομάδα της άνοιας. Παρατηρήθηκε σαφής ηλικιακή διαφοροποίηση, με τη μέση ηλικία να αυξάνεται από τα 58,5 έτη στην ομάδα ελέγχου στα 74,7 έτη στην ομάδα της άνοιας. Είναι σημαντικό να τονιστεί μια ιδιαιτερότητα του συνόλου δεδομένων: συνδυάζει έναν γενικό πληθυσμό πρωτοβάθμιας φροντίδας, που καλύπτει ευρύ ηλικιακό φάσμα, με έναν πιο εξειδικευμένο πληθυσμό από κλινικές μνήμης, ο οποίος αποτελείται κυρίως από άτομα μεγαλύτερης ηλικίας. Αυτή η ανομοιογένεια δυσκολεύει την άμεση σύγκριση των δύο ομάδων, καθώς η ηλικία είναι καθοριστικός παράγοντας κινδύνου για την άνοια. Συνεπώς, κάποιες από τις διαφορές που παρατηρούνται ενδέχεται να οφείλονται απλώς στην ηλικία και όχι στη νόσο αυτή καθαυτή. Παρότι η ηλικία δεν χρησιμοποιήθηκε ως μεταβλητή στο μοντέλο μας, ο περιορισμός αυτός πρέπει να λαμβάνεται σοβαρά υπόψη κατά την ερμηνεία των ευρημάτων. Μελλοντικές έρευνες θα ήταν σκόπιμο να χρησιμοποιήσουν μάρτυρες αντίστοιχης ηλικίας (age-matched controls) ή να εφαρμόσουν στατιστικές μεθόδους που εξαλείφουν την επίδραση του ηλικιακού παράγοντα.

Η ανάλυσή μας βασίστηκε στην κωδικοποίηση ICD-10 για τον προσδιορισμό των κλινικών διαγνώσεων και των συννοσηροτήτων. Δεδομένου ότι το πρότυπο ICD-10 τέθηκε σε εφαρμογή την 1η Οκτωβρίου 2015, ένας περιορισμένος αριθμός διαγνώσεων δεν κατέστη δυνατό να συμπεριληφθεί, λόγω της απουσίας συνεπούς κωδικοποίησης πριν από την ημερομηνία υιοθέτησής του. Για να διαχωρίσουμε τους συμμετέχοντες σε ασθενείς και φυσιολογικούς (case-control), ορίσαμε έναν στόχο πρόβλεψης δυαδικής ταξινόμησης σε επίπεδο ασθενούς. Ουσιαστικά, ελέγξαμε εάν ο κάθε ασθενής είχε διαγνωστεί με κάποια μορφή άνοιας κατά τη διάρκεια της μελέτης, χρησιμοποιώντας τους σχετικούς κωδικούς ICD-10 που φαίνονται στον Πίνακα 2. Στο πλαίσιο της μελέτης, οι κωδικοί ICD-10 του δείγματος αντιστοιχίστηκαν σε κατηγορίες νόσων βάσει του συστηματος ICD-10), το οποίο αποτελεί το

παγκοσμίως αναγνωρισμένο πρότυπο κωδικοποίησης παθήσεων του Παγκόσμιου Οργανισμού Υγείας (WHO, 1992).

2.2 Διαδικασία συλλογής δεδομένων

Πραγματοποιήθηκε μετατροπή των εγγραφών του EHR, οι οποίες φέρουν χρονική σήμανση, σε μια δομημένη μορφή πίνακα. Σε αυτή τη μορφή, κάθε ασθενής αντιπροσωπεύεται από ένα σύνολο χαρακτηριστικών (προγνωστικοί δείκτες) και ετικετών ταξινόμησης που ανατέθηκαν βάσει κριτηρίων συμπερίληψης, υποδεικνύοντας εάν η έκβαση σε άνοια εκδηλώνεται εντός της καθορισμένης περιόδου (Shickel et al., 2018; Ferrao et al., 2017). Η διαδικασία συλλογής της υποσυνόλου, όπως απεικονίζεται στο Σχήμα 1, συμπεριέλαβε 197.481 ασθενείς από τις κλινικές JHMATC και JHCP. Ως σημείο αναφοράς ορίστηκε η πρώτη τους επίσκεψη που καταγράφηκε στο EHR μεταξύ 1ης Ιανουαρίου 2014 και 31ης Δεκεμβρίου 2023. Προϋπόθεση για τη συμπερίληψη ήταν οι επισκέψεις να είναι έγκυρες και ολοκληρωμένες, αφορώντας τύπους όπως επίσκεψη σε ιατρείο, κλινική υποστήριξη, βιντεο-επίσκεψη ή επίσκεψη παρακολούθησης (follow-up). Η διαδικασία κατηγοριοποίησης των συμμετεχόντων σε ομάδες ασθενών (cases) και μη νοσούντων (controls — δηλαδή των μη νοσούντων) περιγράφεται στη συνέχεια. Από αυτή τη διαδικασία προέκυψαν αρχικά 11.687 ασθενείς και 139.477 μη νοσούντες.



Εικόνα 1: Κριτήρια επιλογής ασθενών

Καθώς η οριστική διάγνωση της άνοιας απαιτεί συχνά πρόσθετες αξιολογήσεις σε μεταγενέστερες επισκέψεις, συμπεριλάβαμε περιστατικά (cases) μόνο εφόσον η διάγνωσή τους τεκμηριώθηκε σε διάστημα τουλάχιστον 6 μηνών και το πολύ 5 ετών από την πρώτη τους επίσκεψη. Μετά την εφαρμογή αυτών των χρονικών κριτηρίων ένταξης, από τους αρχικούς 11.687 ασθενείς που είχαν εντοπιστεί, παρέμειναν στο τελικό σύνολο δεδομένων 2.738 επιβεβαιωμένα περιστατικά (Πίνακας 3).Αντίστοιχα, χαρακτηρίσαμε ως «μάρτυρες» (controls) τους ασθενείς από τον αρχικό πληθυσμό (197.481) που δεν είχαν λάβει ποτέ διάγνωση άνοιας (δηλαδή δεν περιλαμβάνονταν στη λίστα των περιστατικών). Από την ομάδα των μαρτύρων αποκλείστηκαν 1.909 άτομα, καθώς είχαν διαγνωστεί με άνοια σε δομές εκτός των κλινικών JHMATC ή JHCP. Προκειμένου να αυξήσουμε την πιθανότητα οι μάρτυρες να παραμείνουν στην ομάδα ελέγχου για τουλάχιστον μία πενταετία, θέσαμε ως επιπρόσθετη προϋπόθεση η πρώτη

τους επίσκεψη να έχει πραγματοποιηθεί πριν από την 1η Ιανουαρίου 2019. Μολονότι η άνοια εξελίσσεται συχνά σε βάθος χρόνου, προηγούμενες μελέτες έχουν δείξει ότι το χρονικό παράθυρο των 5 ετών επαρκεί για την καταγραφή σημαντικού ποσοστού περιστατικών και αποτελεί συνήθη πρακτική στα μοντέλα πρόγνωσης και πρόβλεψης. Η επέκταση του χρονικού πλαισίου σε 7-10 έτη θα περιόριζε σημαντικά τις επιλέξιμες περιπτώσεις στην κοόρτη μας, καθώς η περίοδος παρατήρησης εκτείνεται από το 2014 έως το τέλος του 2023.

Πίνακας 1. Χαρακτηριστικά δειγμάτων

Χαρακτηριστικά	Σύνολο *N* = 142.[175a]	Στάδιο γνωστικής δυσλειτουργίας	
		Έλεγχος *N* = 139.437 (98,1%)	Άνοια *N* = 2.738 (1,9%)
Ηλικία[b] (Μέσος όρος και SD)	58,8 (11,1)	58,5 (10,9)	74,7 (10,0)
<65	101.283 (71,2%)	100.908 (72,4%)	375 (13,7%)
65–74	26.838 (18,9%)	25.963 (18,6%)	875 (32,0%)
75–84	10.891 (7,7%)	9.820 (7,0%)	1.071 (39,1%)
85	3.163 (2,2%)	2.746 (2,0%)	417 (15,2%)
Φύλο			
Γυναίκα	81.861 (57,6%)	80.125 (57,5%)	1.736 (63,4%)
Άνδρες	60.314 (42,4%)	59.312 (42,5%)	1.002 (36,6%)
Φυλή			
Λευκό	91.072 (64,1%)	89.284 (64,0%)	1.788 (65,3%)
Μαύροι	35.223 (24,8%)	34.511 (24,8%)	712 (26,0%)
Ασιάτες	2.101 (1,5%)	2.053 (1,5%)	48 (1,8%)
Άλλοι	5.331 (3,7%)	5.242 (3,8%)	89 (3,3%)
Άγνωστο	8.448 (5,9%)	8.347 (6,0%)	101 (3,7%)
Εθνικότητα			
Ισπανόφωνοι	2.280 (1,6%)	2.236 (1,6%)	44 (1,6%)
Άλλοι	139.821 (98,3%)	137.127 (98,3%)	2.694 (98,4%)
Άγνωστο	74 (0,1%)	74 (0,1%)	—

Πίνακας 2: Κωδικοί ICD-10 που χρησιμοποιήθηκαν στην διαλογή ασθενών

Κατηγορία	Περιγραφή	Ασθενείς (*N* = 3.688)[c]
G30.*x*[a]	Νόσος Αλτσχάιμερ (περιλαμβάνει πρώιμη/όψιμη εκδήλωση, άτυπη, μη καθορισμένη)	892
G31.84	Ήπια γνωστική δυσλειτουργία, όπως δηλώθηκε (στάδιο προ-άνοιας)	1.025
G31.83	Νευρογνωστική διαταραχή με σωμάτια Lewy	5
G31.0	Μετωποκροταφική άνοια	—
F01.*x*	Αγγειακή άνοια	334
F02.*x*[b]	Άνοια σε άλλες ασθένειες (π.χ. Πάρκινσον, Πικ, Χάντινγκτον)	144
F03.*x*	Απροσδιόριστη άνοια (χρησιμοποιείται όταν η αιτία είναι άγνωστη ή δεν έχει τεκμηριωθεί)	1.237

Πίνακας 3: Ανάλυση συνόλου δεδομένων

Επιλογή	
Αρχική επιλογή	685.765
Εφαρμογή κριτηρίων επιλογής	197.481
Έλεγχος — πρώτη επίσκεψη <2019	**139.477**
Έλεγχος — πρώτη επίσκεψη ≥2019	44.408
Εξαιρούνται (έλεγχοι) με διαγνώσεις άνοιας που δεν είναι πρωτοπαθείς/από κλινική μνήμης	1.909
Άνοια (Όλα)	1
Άνοια — πρώτη διάγνωση μεταξύ 7 και 60 μηνών	**2.738**
Άνοια — πρώτη διάγνωση εντός των πρώτων 6 μηνών ή μετά από 60 μήνες	8.949

Η ανάλυση των Mitchell και Shiri-Feshki (2009), η οποία συμπεριέλαβε 41 μελέτες κοόρτης σε άτομα με Ήπια Γνωστική Διαταραχή (MCI), κατέγραψε ετήσιο ρυθμό μετάπτωσης (ACR) σε άνοια της τάξης του 6,7% (ειδικότερα 6,5% για τη νόσο Αλτσχάιμερ και 1,6% για την αγγειακή άνοια). Τα δεδομένα αυτά υποδεικνύουν ότι εντός μιας πενταετίας, ένα σημαντικό ποσοστό ατόμων με MCI (συνήθως 25–35% ή και υψηλότερο) αναμένεται να εξελιχθεί σε άνοια. Το ευρήμα αυτό τεκμηριώνει την επάρκεια του παραθύρου παρατήρησης 5 ετών για τον εντοπισμό ικανού αριθμού νέων περιστατικών, καθιστώντας το ένα καθιερωμένο χρονικό πλαίσιο σε προγνωστικά μοντέλα. Επομένως, παρόλο που η εξέλιξη της νόσου μπορεί να εκτείνεται πέραν της πενταετίας, ο σχεδιασμός μας εξασφαλίζει τη διαγνωστική σταθερότητα τη. Ωστόσο, αναγνωρίζουμε ότι ο πρόσθετος περιορισμός που απαιτεί η πρώτη επίσκεψη των μη νοσούντων να πραγματοποιηθεί πριν από την 1η Ιανουαρίου 2019 δεν εξασφαλίζει πλήρη παρακολούθηση, ιδίως για τους ασθενείς της κλινικής μνήμης, πολλοί από τους οποίους ζουν εκτός της πολιτείας και ενδέχεται να μην καταγράφονται με συνέπεια στο EHR καθ' όλη τη διάρκεια της περιόδου.

2.3 Παραμετροποίηση

Αφού ολοκληρώθηκε η κατηγοριοποίηση των ασθενών με βάση το προφίλ των επισκέψεων και των διαγνώσεών τους, το επόμενο βήμα ήταν η ανάλυση των δεδομένων από τους Ηλεκτρονικούς Φακέλους Υγείας (EHR) για να εξάγουμε τις σχετικές ιατρικές μεταβλητές (covariates). Το τελικό πολυτροπικό σύνολο δεδομένων περιλαμβάνει στοιχεία από κλινικές μετρήσεις, ζωτικά σημεία, αποτελέσματα εργαστηριακών εξετάσεων, καθώς και βαθμολογίες από ερωτηματολόγια αξιολόγησης της γνωστικής λειτουργίας. Οι μετρήσεις που αντλήσαμε, τόσο για τους ασθενείς που νοσούν (cases) όσο και για την ομάδα ελέγχου (controls), προέρχονται από την πρώτη καταγεγραμμένη επίσκεψή τους και καλύπτουν ένα χρονικό "παράθυρο" έξι μηνών μετά από αυτή την αρχική επαφή. Με το εξάμηνο αυτό περιθώριο, εξασφαλίζουμε τη συλλογή ενός ευρέος φάσματος μετρήσεων, ελαχιστοποιώντας έτσι τα κενά στα δεδομένα (missing values) για κάθε χαρακτηριστικό. Σε περιπτώσεις όπου υπάρχουν επαναλαμβανόμενες μετρήσεις για το ίδιο χαρακτηριστικό, θεωρούμε ως πιο ακριβή την πιο πρόσφατη μέτρηση (Figure 2). Αν και τα δεδομένα από τον πραγματικό κόσμο είναι ανεκτίμητα λόγω του ρεαλισμού τους, η επεξεργασία τους κρύβει σημαντικές δυσκολίες, ειδικά όταν πρόκειται να μετατραπούν ακατέργαστα κλινικά δεδομένα σε μια δομημένη μορφή κατάλληλη για εφαρμογή μεθόδων μηχανικής μάθησης (Kim and Min, 2025). Συχνά, τα πεδία μετρήσεων στις βάσεις δεδομένων παρατήρησης αποθηκεύονται ως απλό κείμενο (strings), γεγονός που αυξάνει την πιθανότητα τυπογραφικών λαθών. Επιπλέον, ορισμένες τιμές, όπως η αρτηριακή πίεση, καταγράφονται σε μορφή κειμένου τύπου συστολική/διαστολική. Για να τυποποιήσουμε

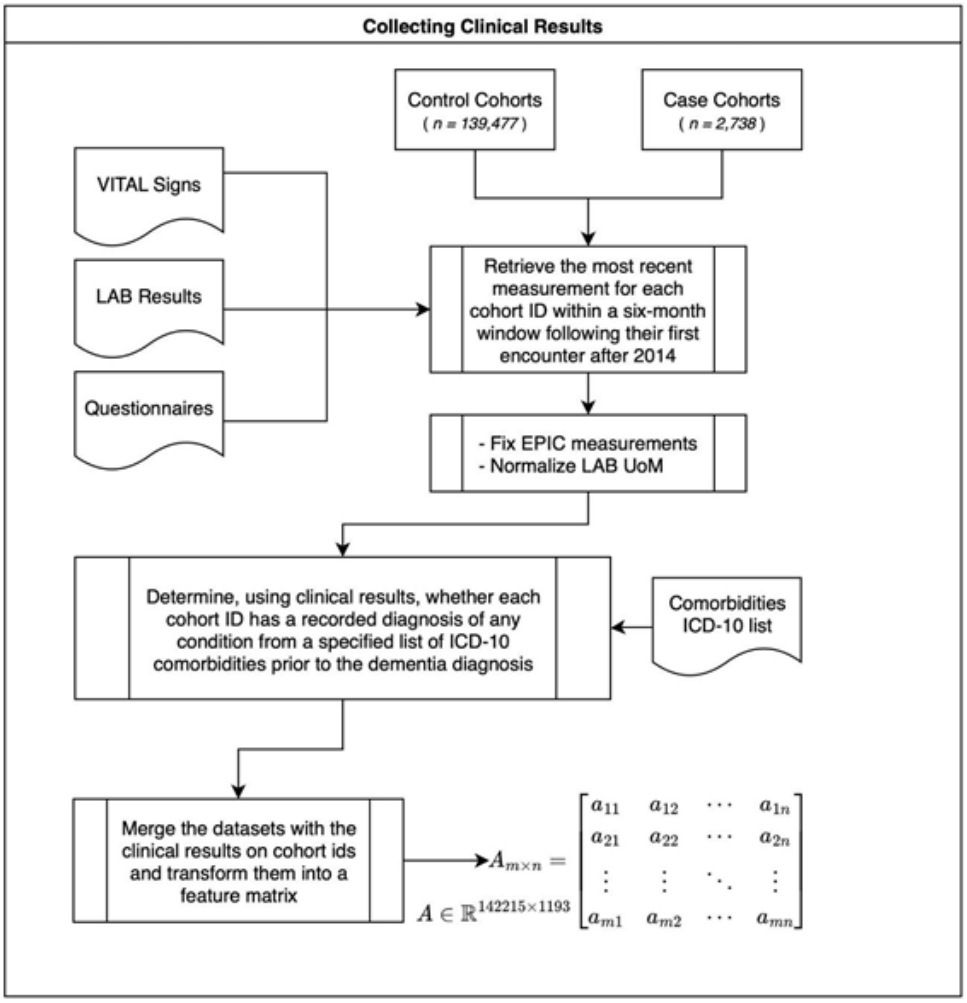
αυτές τις μετρήσεις, μετατρέψαμε τις τιμές πίεσης σε έναν ενιαίο αριθμητικό δείκτη: τη Μέση Αρτηριακή Πίεση (MAP) (DeMers and Wachs, 2025). Για παράδειγμα, μια εγγραφή "130/86" αντικαθίσταται από την τιμή 100.66, που αντιστοιχεί στο MAP της. Αυτή η διαδικασία σημειώνεται ως "Fix EPIC measurements" στο Figure 2. Παρομοίως, τα εργαστηριακά αποτελέσματα στα αρχικά δεδομένα μπορεί να εμφανίζονται με διαφορετικές μονάδες μέτρησης, κάτι που θα απαιτούσε μια εξαντλητική διαδικασία μετατροπών για να ευθυγραμμιστεί κάθε δείκτης με μια πρότυπη μονάδα αναφοράς. Αντ' αυτού, κανονικοποιήσαμε τις μετρήσεις χρησιμοποιώντας το Φυσιολογικό Εύρος Αναφοράς (NR) για κάθε δείκτη, προσαρμόζοντας τα εργαστηριακά αποτελέσματα σε σχέση με τα συγκεκριμένα Ανώτερα και Κατώτερα Φυσιολογικά Όρια (ULN και LLN), όπως ορίζονται για κάθε δείκτη από τα αυτόματα συστήματα των εργαστηρίων. Αυτή η μέθοδος πλεονεκτεί έναντι των παραδοσιακών μετατροπών μονάδων, καθώς εξαλείφει την εξάρτηση από τις μονάδες μέτρησης και μειώνει τον κίνδυνο σφαλμάτων κατά τη μετατροπή. Αντιθέτως, κάθε εργαστηριακή μέτρηση τεκμηριώνεται μαζί με το φυσιολογικό της εύρος, πληροφορία που υπάρχει στο 94% των αρχείων του συστήματος EHR του Johns Hopkins (JH). Επιπλέον, ένα ακόμη βασικό πλεονέκτημα αυτής της μεθόδου είναι ότι βασίζεται σε ελάχιστες υποθέσεις, στηριζόμενη αποκλειστικά στα πρωτογενή δεδομένα (την πραγματική μέτρηση και το εύρος αναφοράς της), ανεξάρτητα από τη μονάδα μέτρησης. Αυτό βελτιώνει τη δυνατότητα σύγκρισης τυποποιημένων μετρήσεων μεταξύ διαφορετικών εργαστηριακών συστημάτων, καθώς και διαφορετικών πηγών δειγμάτων, όπως το αίμα και τα ούρα. Πέρα από τα ζωτικά σημεία, τα εργαστηριακά αποτελέσματα και τα ερωτηματολόγια, ενσωματώσαμε και διαγνώσεις συννοσηροτήτων (Figure 2). Για να εντοπίσουμε την ύπαρξη συννοσηροτήτων, αναζητήσαμε στη βάση δεδομένων EHR κωδικούς ICD-10 που σχετίζονται με συγκεκριμένες κατηγορίες παθήσεων. Για τους ασθενείς-περιστατικά, καταγράψαμε τις συννοσηρότητες που προϋπήρχαν της διάγνωσης άνοιας. Υιοθετήσαμε την παραδοχή ότι, εάν μια διάγνωση δεν είναι καταγεγραμμένη στο σύστημα EHR, τότε δεν υφίσταται. Η λίστα των κωδικών ICD-10 προέρχεται από τον κατάλογο της Διεθνούς Ταξινόμησης Νόσων (World Health Organization, 1992). Επιπλέον, εξαιρέσαμε κατηγορίες που αφορούσαν ατυχήματα ή μη φυσιολογικά ευρήματα εξετάσεων που δεν κατέληξαν σε διάγνωση. Επίσης, για να αποφύγουμε τη διαρροή πληροφορίας προς τον στόχο (target leakage), αφαιρέσαμε από τον κατάλογο συννοσηροτήτων τις ομάδες (χαρακτηριστικά) που περιείχαν διαγνώσεις οι οποίες αλληλεπικαλύπτονται με τμήματα των κωδικών-στόχων ICD-10 [π.χ., DISEASES OF THE NERVOUS SYSTEM (G30–G32)] ή σχετίζονται με τη γνωστική λειτουργία [π.χ., SYMPTOMS AND SIGNS INVOLVING COGNITION PERCEPTION EMOTIONAL STATE AND BEHAVIOR (R40–R46)]. Το τελικό στάδιο στη συλλογή των κλινικών μετρήσεων (Figure 2) ήταν η συγχώνευση όλων των χαρακτηριστικών —

συμπεριλαμβανομένων των κλινικών μετρήσεων (ζωτικά σημεία, εργαστηριακά και σκορ γνωστικής αξιολόγησης) και των διαγνώσεων συννοσηροτήτων— σε έναν ενιαίο πίνακα. Αυτός ο πίνακας μετασχηματίστηκε σε ένα μητρώο που αποτελείται από 1.193 στήλες χαρακτηριστικών και 142.215 σειρές (παρατηρήσεις).

2.4 Προεπεξεργασία των δεδομένων

Στο πλαίσιο της προγνωστικής μοντελοποίησης, η φάση της προεπεξεργασίας εστιάζει στον εντοπισμό, τη διόρθωση ή την εξάλειψη σφαλμάτων, ασυνεπειών και ανακριβειών εντός ενός συνόλου δεδομένων, με σκοπό την αναβάθμιση της ποιότητας και της αξιοπιστίας του για τα μοντέλα μηχανικής μάθησης. Επιπροσθέτως, επιστρατεύονται στρατηγικές διαχείρισης των ανισορροπιών στα δεδομένα, ώστε να διασφαλιστεί η αποτελεσματική λειτουργία του μοντέλου και για τις δύο κλάσεις. Τέτοιες μέθοδοι συμβάλλουν στον περιορισμό της μεροληψίας και ενισχύουν την ικανότητα του μοντέλου να γενικεύει σωστά σε νέα, άγνωστα δεδομένα. Σε γενικές γραμμές, η προεπεξεργασία εγγυάται την πληρότητα, την ακρίβεια και τη συνέπεια των δεδομένων, μειώνοντας παράλληλα τον θόρυβο και τις πιθανές μεροληψίες σε εφαρμογές που βασίζονται σε δεδομένα. Στην περίπτωσή μας, το σύνολο δεδομένων παρουσίαζε σημαντική ανισορροπία, καθώς τα 2.738 περιστατικά (cases) αντιστοιχούσαν μόλις στο 1,9% των 139.477 ατόμων της ομάδας ελέγχου (controls) (Εικόνα 2). Προκειμένου να αντιμετωπιστεί αυτή η ανισορροπία χωρίς να συρρικνωθεί το μέγεθος του δείγματος της ομάδας ελέγχου, εφαρμόσαμε τη μέθοδο Repeated Random Undersampling Cross-Evaluation (He and Ma, 2013). Πιο συγκεκριμένα, η ομάδα ελέγχου διαχωρίστηκε σε 51 ισομεγέθη υποσύνολα, καθένα εκ των οποίων είχε μέγεθος αντίστοιχο με αυτό της μειονοτικής κλάσης. Στη συνέχεια, κάθε υποσύνολο συνδυάστηκε με την ίδια μειονοτική κλάση για τη δημιουργία ενός ισορροπημένου συνόλου εκπαίδευσης. Το μοντέλο εκπαιδεύτηκε ανεξάρτητα σε κάθε υποσύνολο και, στη συνέχεια, οι μετρικές απόδοσης συγκεντρώθηκαν συνολικά.

Η βελτιστοποίηση των υπερπαραμέτρων εκπαίδευσης πραγματοποιήθηκε ξεχωριστά για κάθε υποσύνολο, επιλέγοντας τη διαμόρφωση εκείνη που σημείωσε την υψηλότερη μέση ακρίβεια κατά τη διαδικασία του cross-validation. Στη συνέχεια, και τα δύο μοντέλα τυποποιήθηκαν με βάση το συγκεκριμένο σετ υπερπαραμέτρων που επιλέχθηκε. Η παρούσα μελέτη αξιοποιεί τους αλγορίθμους RandomForest και XGBoost προκειμένου να επικυρώσει τη μεθοδολογία που εφαρμόστηκε και να συγκρίνει την απόδοσή της.»



Εικόνα 2: Κλινικά χαρακτηριστικά του συνόλου δεδομένων

2.5 . Διαχείριση Ελλειπουσών Τιμών

Ο εντοπισμός των ελλειπούσων τιμών και ο καθορισμός της κατάλληλης στρατηγικής διαχείρισής τους, όπως η διαγραφή γραμμών ή στηλών, η συμπλήρωση τιμών (imputation) ή η κωδικοποίηση της ίδιας της απουσίας δεδομένων ως χαρακτηριστικό, αποτελούν θεμελιώδη πτυχή της προεπεξεργασίας (Ren et al., 2023). Δεδομένου ότι το τελικό σύνολο δεδομένων εμφάνιζε ποσοστό ελλιπών τιμών της τάξεως του 78% , γεγονός που οφείλεται κυρίως στη μετατροπή των χρονικά προσδιορισμένων κλινικών δεδομένων σε μορφή πίνακα, εξαιρέθηκαν τα χαρακτηριστικά με ποσοστό έλλειψης άνω του 90%. Αυτό είχε ως αποτέλεσμα τη δραστική μείωση των ενεργών μεταβλητών πρόβλεψης από 1.193 σε 166.

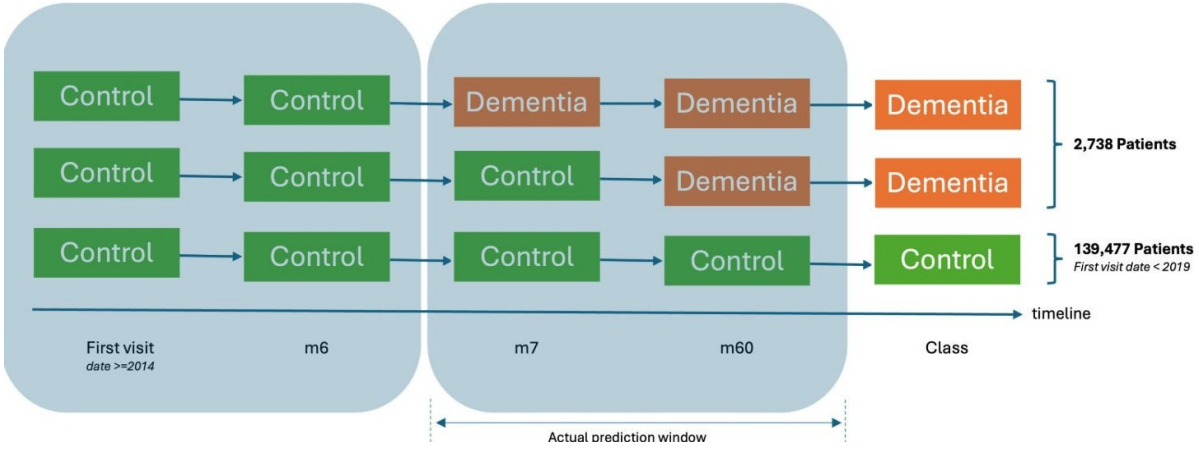
Το συγκεκριμένο κατώφλι επιλέχθηκε με στόχο να μειωθεί η πολυπλοκότητα του μοντέλου και να ελαχιστοποιηθεί ο "θόρυβος" που προέρχεται από μεταβλητές με αραιές παρατηρήσεις. Παρόλο που ορισμένες εργαστηριακές εξετάσεις, οι οποίες ζητούνται σπάνια, ενδέχεται να προσφέρουν σημαντική πληροφορία, η εμπειρική αξιολόγηση μέσω των δεικτών

σημαντικότητας χαρακτηριστικών (feature importance) του XGBoost κατέδειξε ότι τα χαρακτηριστικά με εξαιρετικά υψηλά ποσοστά έλλειψης είχαν ελάχιστη συνεισφορά στην απόδοση του μοντέλου. Για τις εναπομείνασες ελλιπείς τιμές δεν εφαρμόστηκε μέθοδος συμπλήρωσης. Αυτό συμβαίνει διότι στα κλινικά δεδομένα η απουσία πληροφορίας συχνά δεν είναι τυχαία (Missing Not At Random - MNAR) και ενδέχεται να φέρει σημαντικό κλινικό νόημα, με αποτέλεσμα η τεχνητή συμπλήρωση εργαστηριακών μετρήσεων να οδηγεί πιθανώς σε μεροληπτικά αποτελέσματα (Ibrahim et al., 2012; Li et al., 2021).»

2.6 Ανάπτυξη του μοντέλου Μηχανικής Μάθησης

Αξιολογήσαμε την απόδοση δύο ευρέως διαδεδομένων αλγορίθμων συνδυαστικής μάθησης (ensemble learning), του Random Forest (RF) και του XGBoost (XGB), με στόχο να αντιμετωπίσουμε τις προκλήσεις που θέτουν τα πολυδιάστατα και ετερογενή δεδομένα που συναντώνται συνήθως στους Ηλεκτρονικούς Φακέλους Υγείας (EHR). Η επιλογή βασίστηκε στην αποδεδειγμένη ανθεκτικότητα και αποτελεσματικότητά τους στη διαχείριση πολύπλοκων συνόλων δεδομένων (Lebedev et al., 2014; Moore and Bell, 2022). Οι μέθοδοι συνόλου που βασίζονται σε δέντρα, όπως το Random Forest και το XGBoost, είναι ειδικά σχεδιασμένες για να περιορίζουν το φαινόμενο της υπερπροσαρμογής (overfitting) σε σύγκριση με τα μεμονωμένα δέντρα απόφασης. Το Random Forest το επιτυγχάνει αυτό συνδυάζοντας τις προβλέψεις πολλών ασυσχέτιστων δέντρων απόφασης, καθένα από τα οποία εκπαιδεύεται σε δείγματα που έχουν προκύψει από επαναδειγματοληψία (bootstrapped samples) και με τυχαία επιλογή χαρακτηριστικών. Αυτός ο σχεδιασμός μειώνει τη διακύμανση και αποτρέπει το μοντέλο από το να «απομνημονεύει» τον θόρυβο του συνόλου εκπαίδευσης, οδηγώντας σε καλύτερη γενίκευση (Breiman, 2001). Παρομοίως, το XGBoost ενσωματώνει τεχνικές ενίσχυσης (boosting) με μεθόδους κανονικοποίησης (ποινές L1 και L2) και συρρίκνωσης (shrinkage). Οι τεχνικές αυτές εμποδίζουν τον αλγόριθμο να προσαρμόζεται σε θόρυβο ή σε παραπλανητικά μοτίβα στα δεδομένα εκπαίδευσης, μειώνοντας έτσι περαιτέρω τον κίνδυνο υπερπροσαρμογής, διατηρώντας παράλληλα την προγνωστική ακρίβεια (Chen and Guestrin, 2016). Και οι δύο αλγόριθμοι έχουν δείξει σε εφαρμοσμένες μελέτες ότι επιτυγχάνουν υψηλή προγνωστική απόδοση διατηρώντας ανθεκτικότητα απέναντι στην υπερπροσαρμογή, καθιστώντας τους ιδανικούς για εργασίες κλινικής πρόβλεψης. Το Random Forest και το XGBoost επέδειξαν συγκρίσιμη απόδοση όσον αφορά την ακρίβεια και τη γενίκευση. Σκοπός ήταν να εντοπιστεί το μοντέλο που αποτυπώνει αποτελεσματικότερα τα υποκείμενα μοτίβα των δεδομένων και παρουσιάζει ισχυρά χαρακτηριστικά γενίκευσης. Για τον λόγο αυτό, πραγματοποιήσαμε μια ολοκληρωμένη αξιολόγηση χρησιμοποιώντας μια διαδικασία επαναλαμβανόμενης στρωματοποιημένης διασταυρούμενης επικύρωσης (Repeated Stratified

KFold cross-validation) με 5 επαναλήψεις και 5 διπλώσεις (folds) σε όλα τα 51 τμήματα του συνόλου δεδομένων, οδηγώντας σε 1.275 επαναλήψεις συνολικά. Ακολούθησε ο υπολογισμός του μέσου όρου της απόδοσης των προβλέψεων σε άγνωστα δεδομένα (test set), ώστε να εκτιμηθεί η συνολική αποτελεσματικότητα και η ικανότητα γενίκευσης. Οι υπερπαράμετροι εκπαίδευσης βελτιστοποιήθηκαν ξεχωριστά για κάθε μοντέλο και για καθένα από τα 51 υποσύνολα, χρησιμοποιώντας τη μέθοδο Μπεϋζιανής βελτιστοποίησης με διασταυρούμενη επικύρωση (BayesSearchCV) και επιλέγοντας τη διαμόρφωση που απέδωσε την υψηλότερη μέση ακρίβεια στις διπλώσεις της διασταυρούμενης επικύρωσης. Στη συνέχεια, και τα δύο μοντέλα τυποποιήθηκαν με βάση τη βελτιστοποίηση υπερπαραμέτρων. Παρόλο που και το Random Forest και το XGBoost είναι μέθοδοι συνόλου που βασίζονται σε δέντρα, το Random Forest προτιμάται συχνά για την κλινική ερμηνεία λόγω της χρήσης ανεξάρτητων δέντρων απόφασης, γεγονός που διευκολύνει την εξαγωγή και ανάλυση μεμονωμένων δέντρων (Laabs et al., 2024). Επιπλέον, αν και υπολογίζει τον μέσο όρο των προβλέψεων από πολλαπλά δέντρα, κάθε δέντρο παραμένει ερμηνεύσιμο από μόνο του.



Εικόνα 3: Μοντελοποίηση προβλήματος

Η παραμετροποίηση του ταξινομητή περιλαμβάνει 100 εκτιμητές (estimators), πράγμα που σημαίνει ότι κάθε μοντέλο αποτελείται από 100 δέντρα. Για την αξιολόγηση της απόδοσης του μοντέλου, εφαρμόστηκε η μέθοδος επαναλαμβανόμενης στρωματοποιημένης διασταυρούμενης επικύρωσης (Repeated Stratified K-Fold Cross-Validation) με 5 διαχωρισμούς (splits) και 5 επαναλήψεις, δημιουργώντας έτσι 25 μοναδικά ζεύγη συνόλων εκπαίδευσης και ελέγχου. Σε κάθε διαχωρισμό εκπαιδεύεται ένα ξεχωριστό μοντέλο Random Forest, γεγονός που οδηγεί αθροιστικά στη δημιουργία 2.500 δέντρων απόφασης στο σύνολο των επαναλήψεων. Ως βέλτιστος ταξινομητής επιλέγεται το μοντέλο που επιτυγχάνει το υψηλότερο F1 score για την κλάση της νόσου κατά τη διαδικασία της διασταυρούμενης επικύρωσης. Στη συνέχεια, τα 100 δέντρα απόφασης του βέλτιστου αυτού μοντέλου εισάγονται μεμονωμένα σε

μια συνάρτηση, η οποία μετατρέπει κάθε δέντρο σε μορφή πίνακα και τα ενσωματώνει σε ένα dataframe. Η διαδικασία αυτή επιτρέπει την περαιτέρω επιλογή των πιο πληροφοριακών μονοπατιών απόφασης, βάσει συγκεκριμένων προκαθορισμένων κριτηρίων. Στα κριτήρια αυτά περιλαμβάνονταν: κόμβοι-φύλλα (leaf nodes) που περιέχουν περισσότερα από 140 δείγματα, εκ των οποίων τουλάχιστον το 72% να έχει ταξινομηθεί ως περιστατικά νόσου, καθώς και ένα F1 score για την κλάση της νόσου που να υπερβαίνει το 0,70 για το αντίστοιχο δέντρο. Προκειμένου να ενισχύσουμε την κλινική σημασία και την αξιοπιστία των ευρημάτων μας, επιλέχθηκαν 10 μονοπάτια απόφασης από ένα συνολικό σύνολο 5.100 δέντρων, ώστε να αξιολογηθούν από ειδικούς. Αυτά τα μονοπάτια, που βασίζονται σε κανόνες, βελτιώνουν την κλινική ερμηνευσιμότητα και προσφέρουν ένα διαφανές και εύληπτο πλαίσιο, το οποίο υποστηρίζει την ταχύτερη και ακριβέστερη διάγνωση της άνοιας.

Για παράδειγμα, η κλινική ερμηνεία του συνόλου κανόνων υποδεικνύει ότι, σε ένα δείγμα 252 ασθενών οι οποίοι παρουσίαζαν επίπεδα κρεατινίνης εντός του φυσιολογικού εύρους αλλά κοντά στο ανώτατο όριο, τιμές VLDLCALC κοντά στο κατώτατο όριο, σωματικό βάρος κάτω από 3167,60 oz, φυσιολογική αναλογία CHOL/HDL, χαμηλά επίπεδα TSH, φυσιολογική (προς χαμηλή) ολική πρωτεΐνη και χαμηλό αριθμό λευκών αιμοσφαιρίων, το 75% ανέπτυξε άνοια εντός μιας πενταετίας. Σε αντίθεση με τα μοντέλα «μαύρου κουτιού» (black-box), η δομή που βασίζεται σε κανόνες διέπεται από μια σαφή λογική την οποία οι κλινικοί ιατροί μπορούν να εξηγήσουν με ευκολία, ενισχύοντας έτσι την εμπιστοσύνη και τη χρηστικότητα στην πράξη. Τα μοντέλα αυτά αναδεικνύουν τα πλέον σημαντικά κλινικά χαρακτηριστικά και φέρνουν στο φως πολύπλοκες αλληλεπιδράσεις μεταξύ των παραγόντων κινδύνου, οι οποίες ενδεχομένως να μην είναι εμφανείς με τις παραδοσιακές μεθόδους. Το γεγονός αυτό επιτρέπει την πρώιμη ανίχνευση της άνοιας μέσω του εντοπισμού διακριτικών αλλά ουσιαστικών μοτίβων. Παράλληλα, τα μονοπάτια απόφασης καθιστούν δυνατή την ταχεία διαλογή και τη στρωματοποίηση κινδύνου, βοηθώντας τους κλινικούς να κατατάσσουν τους ασθενείς σε κατηγορίες επικινδυνότητας και να ιεραρχούν τις έγκαιρες παρεμβάσεις. Επιπροσθέτως, μπορούν να προσαρμοστούν ώστε να ευθυγραμμίζονται με τις κλινικές κατευθυντήριες οδηγίες, γεγονός που τα καθιστά εξαιρετικά ευέλικτα για πρακτική χρήση. Συνολικά, η μέθοδος της κλινικής ερμηνείας βάσει κανόνων προσφέρει μια βέλτιστη ισορροπία μεταξύ ερμηνευσιμότητας, ταχύτητας και ακρίβειας, στοιχείο ιδιαίτερα πολύτιμο για τη διαχείριση περίπλοκων παθήσεων όπως η άνοια.

3. Αποτελέσματα ανάλυσης και πρόβλεψης

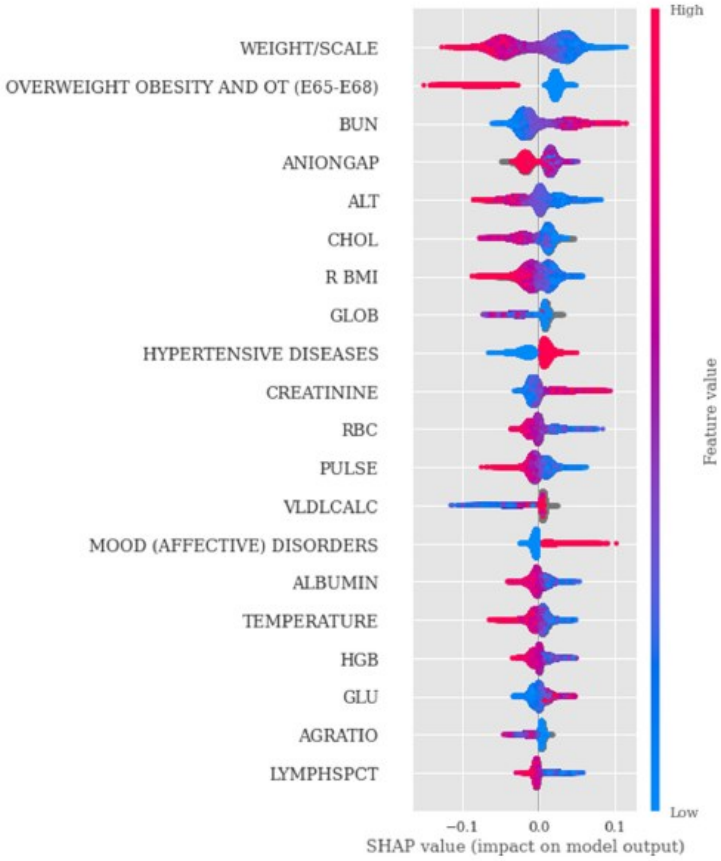
Τα αποτελέσματα της απόδοσης, τόσο για την κλάση των ασθενών όσο και για την ομάδα ελέγχου, κυμαίνονται σε πολύ παρόμοια επίπεδα, γεγονός αναμενόμενο δεδομένης της ισορροπίας του συνόλου δεδομένων. Το μοντέλο XGBoost επέδειξε επιδόσεις εφάμιλλες ή και ελαφρώς ανώτερες από εκείνες του μοντέλου Random Forest, με τη μέση τιμή του F1-score για την κλάση των ασθενών στον πληθυσμό να κυμαίνεται από 72.5% έως 72.6% για τον ταξινομητή XGBoost, και από 71.4% έως 71.5% για το Random Forest (Πίνακας 4). Και τα δύο μοντέλα παρουσιάζουν τιμές AUROC που εμπίπτουν στο εύρος από 0.77 έως 0.79. Τα διαστήματα εμπιστοσύνης 95% (CI) υπολογίστηκαν και για τα δύο μοντέλα με βάση την Student's t-κατανομή παρέχοντας έτσι μια εκτίμηση της ακρίβειας σχετικά με τις μετρικές απόδοσης.

Πίνακας 4: Απόδοση του μοντέλου

Metric	Random Forest[a]		XGBoost[a]	
	Value (Std)	95% CI[b]	Value (Std)	95% CI[b]
Mean AUROC	0.776 (0.014)	0.775–0.777	0.795 (0.014)	0.794–0.795
Mean precision control	0.718 (0.015)	0.718–0.719	0.727 (0.025)	0.726–0.729
Mean precision case	0.698 (0.015)	0.697–0.698	0.718 (0.021)	0.717–0.719
Mean test accuracy	0.707 (0.014)	0.706–0.708	0.722 (0.014)	0.722–0.723
Mean F1-score control	0.699 (0.016)	0.698–0.700	0.719 (0.016)	0.718–0.720
Mean F1-score case	0.715 (0.013)	0.714–0.715	0.725 (0.014)	0.725–0.726

Η ανάλυση των μεταβλητών εισόδου στο μοντέλο (covariate analysis) είναι απαραίτητη για την ερμηνεία του τρόπου με τον οποίο τα χαρακτηριστικά επηρεάζουν τη μεταβλητή-στόχο, συμβάλλοντας τόσο στη βελτίωση της ακρίβειας του μοντέλου όσο και στην εξαγωγή ουσιαστικών συμπερασμάτων από τα δεδομένα. Το γράφημα SHAP (Εικόνα 4) αναδεικνύει τα 20 πιο επιδραστικά χαρακτηριστικά για το μοντέλο Random Forest που χρησιμοποιήθηκε. Δεδομένου ότι υιοθετήσαμε μια προσέγγιση σε επίπεδο ασθενούς, κάθε κουκκίδα στο γράφημα SHAP αντιπροσωπεύει μια μεμονωμένη κλινική μέτρηση (προερχόμενη από εργαστηριακά αποτελέσματα ή ζωτικά σημεία) ή μια διάγνωση συννοσηρότητας για τον εκάστοτε ασθενή. Ο άξονας x απεικονίζει την τιμή SHAP για ένα συγκεκριμένο χαρακτηριστικό, υποδεικνύοντας τον αντίκτυπο που έχει αυτό το χαρακτηριστικό στην πρόβλεψη του μοντέλου για τον συγκεκριμένο ασθενή. Ο άξονας y παραθέτει τα χαρακτηριστικά (π.χ. εργαστηριακά

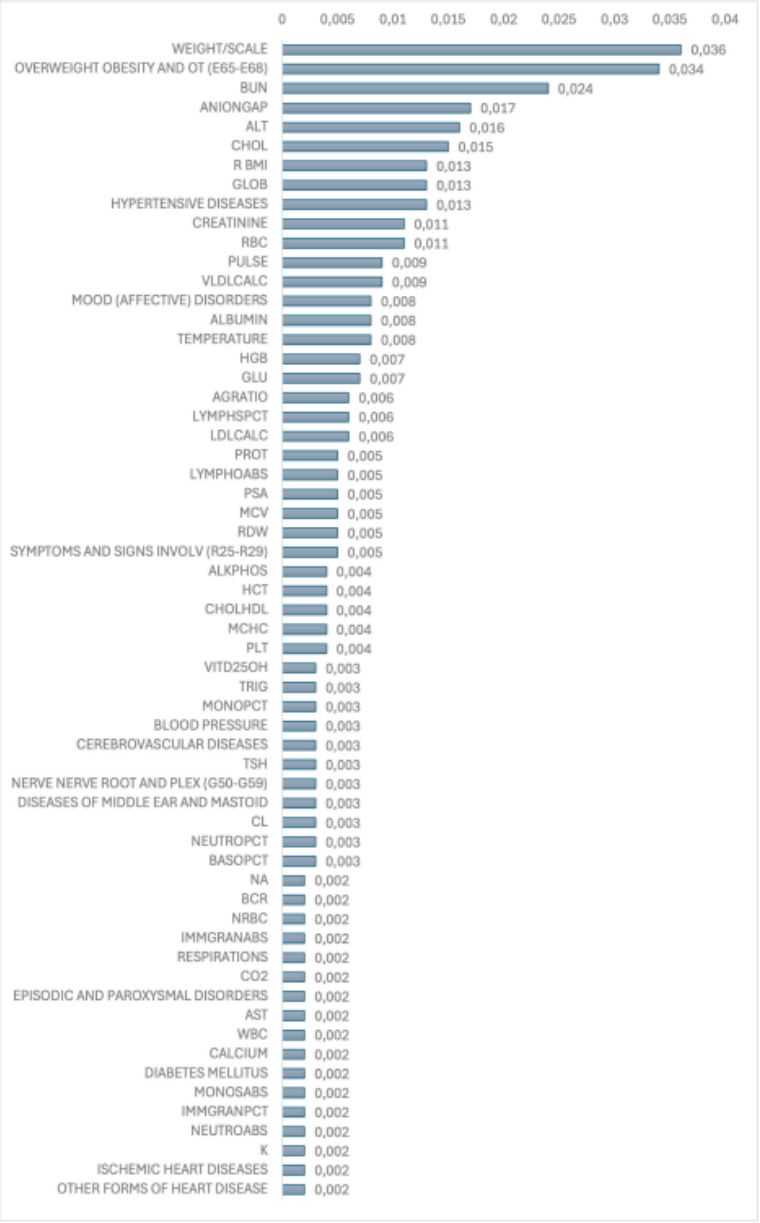
αποτελέσματα ή διαγνώσεις συννοσηροτήτων) ταξινομημένα βάσει της σπουδαιότητάς τους στο μοντέλο, ενώ οι κουκκίδες δείχνουν τις τιμές των μεμονωμένων ασθενών για το χαρακτηριστικό αυτό και την αντίστοιχη τιμή SHAP, η οποία αντικατοπτρίζει τον βαθμό συμβολής του χαρακτηριστικού στην πρόβλεψη. Το χρώμα των κουκκίδων υποδηλώνει την τιμή του χαρακτηριστικού για τον ασθενή, με τη χρωματική διαβάθμιση να αποσκοπεί στην απεικόνιση της αλληλεπίδρασης μεταξύ της τιμής του χαρακτηριστικού και της επίδρασής του στην πρόβλεψη. Στην περίπτωση των συννοσηροτήτων (π.χ. διάγνωση με τιμή 1 για «Αληθές» και 0 για «Ψευδές»), οι τιμές SHAP στον άξονα x κυμαίνονταν συνεχώς στο εύρος ± 0.10, καθώς οι τιμές SHAP αντιπροσωπεύουν τον αντίκτυπο ή τη συνεισφορά κάθε χαρακτηριστικού στο αποτέλεσμα του μοντέλου (πρόβλεψη) για κάθε περίπτωση, και όχι την ίδια τη δυαδική τιμή.



Εικόνα 4: Ερμηνείες SHAP που προέκυψαν από το μοντέλο Random Forest

Συνεπώς, στην περίπτωση των υπερτασικών νόσων, το κόκκινο χρώμα (Dx True) υποδηλώνει ότι η ύπαρξη προηγούμενης διάγνωσης υπερτασικής νόσου συνδέεται με αυξημένο κίνδυνο εμφάνισης άνοιας (θετικές τιμές SHAP προς τα δεξιά). Η σημαντικότητα χαρακτηριστικών SHAP (Εικόνα 5) προσδιορίζεται υπολογίζοντας τον μέσο όρο των απόλυτων τιμών SHAP για την κλάση της νόσου σε όλα τα δείγματα. Με αυτόν τον τρόπο μετράται το μέσο μέγεθος της επίδρασης κάθε χαρακτηριστικού στην πρόβλεψη του μοντέλου, ανεξαρτήτως

κατεύθυνσης (θετικής ή αρνητικής). Οι τιμές που προκύπτουν δείχνουν τη σχετική συνεισφορά κάθε χαρακτηριστικού στην πρόβλεψη για την κλάση της νόσου, όπου οι υψηλότερες τιμές υποδηλώνουν μεγαλύτερη σπουδαιότητα. Τα υπόλοιπα χαρακτηριστικά, μέχρι τον αριθμό 166, εμφανίζουν μηδενική σημαντικότητα SHAP, καθώς δεν έχουν καμία μετρήσιμη επίδραση στις προβλέψεις του μοντέλου.



Εικόνα 5. Γράφημα σημαντικότητας χαρακτηριστικών που καθορίζουν την τελική πρόβλεψη του μοντέλου

Στο πλαίσιο περαιτέρω συζήτησης των ανωτέρω αποτελεσμάτων, στόχος μας ήταν η ανάπτυξη μοντέλων πρόβλεψης άνοιας πενταετίας, αξιοποιώντας δεδομένα από τον πραγματικό κόσμο. Η μελέτη βασίστηκε σε μία ενιαία πληθυσμιακή ομάδα (cohort) που

περιλάμβανε εξωτερικούς ασθενείς από την κλινική μνήμης του Κέντρου Μνήμης και Θεραπείας Αλτσχάιμερ του Johns Hopkins (JHMATC) στη Βαλτιμόρη των ΗΠΑ, καθώς και εξωτερικούς ασθενείς από κλινικές πρωτοβάθμιας περίθαλψης στην ευρύτερη περιοχή του Maryland και της Washington DC. Σε αυτή την κατεύθυνσή, αξιολογήσαμε δύο ευρέως διαδεδομένες μεθόδους συνδυαστικής μάθησης (Random Forest και XGBoost) προκειμένου να αντιμετωπίσουμε τις προκλήσεις που θέτουν τα πολυδιάστατα και ετερογενή δεδομένα των Ηλεκτρονικών Φακέλων Υγείας (EHR). Οι συγκεκριμένοι αλγόριθμοι επιλέχθηκαν για την ικανότητά τους να μοντελοποιούν μη γραμμικές σχέσεις και πολύπλοκες αλληλεπιδράσεις μεταξύ πολλαπλών παραγόντων πρόβλεψης. Σε μελλοντικές εργασίες, σκοπεύουμε να επεκτείνουμε τη σύγκριση συμπεριλαμβάνοντας επιπλέον προσεγγίσεις, όπως το LASSO και το RUSBoost, χρησιμοποιώντας την ίδια διαδικασία αξιολόγησης.

Το γράφημα SHAP και η ανάλυση σημαντικότητας χαρακτηριστικών υποδεικνύουν ότι η διάγνωση ΥΠΕΡΒΑΡΟΥ, ΠΑΧΥΣΑΡΚΙΑΣ ΚΑΙ ΑΛΛΩΝ ΜΟΡΦΩΝ ΥΠΕΡΣΙΤΙΣΜΟΥ (E65-E68) συνδέεται με μειωμένο κίνδυνο άνοιας. Το εύρημα αυτό έρχεται σε αντίθεση με την επικρατούσα υπόθεση ότι η παχυσαρκία στη μέση ηλικία αυξάνει την πιθανότητα εμφάνισης άνοιας αργότερα στη ζωή. Αντιθέτως, τα στοιχεία δείχνουν ότι το χαμηλό σωματικό βάρος (underweight) ενδέχεται να συνδέεται με υψηλότερο κίνδυνο άνοιας. Αυτά τα απροσδόκητα ευρήματα υπογραμμίζουν την ανάγκη για περαιτέρω διερεύνηση των βαθύτερων αιτιών και των πιθανών επιπτώσεών τους στη δημόσια υγεία (Qizilbash et al., 2015). Παρόμοιο μοτίβο παρατηρείται και σε άλλα βασικά ανθρωπομετρικά χαρακτηριστικά, όπως το ΒΑΡΟΣ/ΖΥΓΙΣΗ και ο Δείκτης Μάζας Σώματος (BMI), τα οποία αποτελούν συνεχείς μεταβλητές που προέρχονται από το σύνολο δεδομένων του Epic (ζωτικά σημεία). Άλλα ευρήματα της παρούσας μελέτης υποδηλώνουν ότι χαμηλότερες τιμές αλανινικής αμινοτρανσφεράσης (ALT), χοληστερόλης (CHOL) και ερυθρών αιμοσφαιρίων (RBCs) σχετίζονται με αυξημένο κίνδυνο άνοιας. Η μελέτη Atherosclerosis Risk in Communities (ARIC) διαπίστωσε ότι άτομα με επίπεδα ALT κάτω από το 10ο εκατοστημόριο είχαν 34% υψηλότερο κίνδυνο ανάπτυξης άνοιας σε σύγκριση με άτομα στο δεύτερο πεμπτημόριο. Η συσχέτιση αυτή παρέμεινε σημαντική ακόμη και μετά τη στάθμιση παραγόντων όπως η ηλικία, το φύλο, η φυλή, το επίπεδο εκπαίδευσης και ο γονότυπος APOE4. Η μελέτη προτείνει ότι τα χαμηλά επίπεδα ALT ενδέχεται να υποδηλώνουν μειωμένη ηπατική λειτουργία, η οποία θα μπορούσε να συμβάλει σε εγκεφαλικό υπομεταβολισμό και διαταραχή των νευροδιαβιβαστών, αυξάνοντας έτσι τον κίνδυνο άνοιας (Lu et al., 2021).

Παρομοίως, η μελέτη των Wang et al. (2018) έδειξε ότι, σε παρακολούθηση διάρκειας 7 ετών σε 1.800 ηλικιωμένους, όσοι είχαν συγκριτικά χαμηλά επίπεδα λευκωματίνης (ALBUMIN) στον ορό αντιμετώπιζαν υπερδιπλάσιο κίνδυνο εμφάνισης ήπιας γνωστικής διαταραχής (MCI). Τα ευρήματα αυτά δείχνουν ότι η λευκωματίνη μπορεί να λειτουργεί ως

ανεξάρτητος παράγοντας κινδύνου για MCI στους ηλικιωμένους. Μια άλλη μελέτη, στην οποία συμμετείχαν πάνω από 300.000 άτομα από τη βιοτράπεζα του Ηνωμένου Βασιλείου (UK Biobank), εντόπισε ότι τα χαμηλά επίπεδα αιμοσφαιρίνης (HGB) και το εύρος κατανομής ερυθρών αιμοσφαιρίων (RDW) συνδέονταν με αυξημένο κίνδυνο άνοιας. Συγκεκριμένα, η αναιμία συνδέθηκε με 56% υψηλότερο κίνδυνο εμφάνισης άνοιας. Η μελέτη διαπίστωσε επίσης ότι τα χαμηλά επίπεδα ερυθρών αιμοσφαιρίων και αιμοσφαιρίνης θα μπορούσαν να οδηγήσουν σε μειωμένη ικανότητα μεταφοράς οξυγόνου στο αίμα, συμβάλλοντας στην παθογένεια της άνοιας (Qiang et al., 2023).

Αντιθέτως, υπάρχουν αυξανόμενες ενδείξεις ότι τα αυξημένα επίπεδα αζώτου ουρίας αίματος (BUN) —ένας δείκτης δυσλειτουργίας των νεφρών— συνδέονται με αυξημένο κίνδυνο άνοιας. Μελέτη Μενδελιανής Τυχαιοποίησης χρησιμοποίησε γενετικά δεδομένα για να υποστηρίξει μια αιτιώδη συσχέτιση μεταξύ της διαταραγμένης νεφρικής λειτουργίας (συμπεριλαμβανομένων δεικτών όπως το BUN) και του αυξημένου κινδύνου άνοιας. Αυτό ενισχύει το επιχείρημα ότι η σχέση δεν είναι απλώς συσχετιστική (Huang et al., 2024). Ομοίως, τα αυξημένα επίπεδα κρεατινίνης ορού (CREATININE), που υποδηλώνουν μειωμένη νεφρική λειτουργία, έχουν συσχετιστεί με αυξημένο κίνδυνο γνωστικής έκπτωσης και άνοιας (Xiao et al., 2023). Η διάγνωση της ΥΠΕΡΤΑΣΗΣ (υψηλή αρτηριακή πίεση) συνδέεται ισχυρά με αυξημένο κίνδυνο άνοιας, συμπεριλαμβανομένης της νόσου Αλτσχάιμερ και της αγγειακής άνοιας, με πολυάριθμες μελέτες μεγάλων κοορτών να δείχνουν σταθερά ότι άτομα με υπέρταση, ειδικά στη μέση ηλικία, διατρέχουν σημαντικά υψηλότερο κίνδυνο να αναπτύξουν άνοια αργότερα στη ζωή (Kennelly et al., 2009).

Όταν συγκρίναμε τα αποτελέσματά μας με τις πιο γνωστές προσεγγίσεις που δεν βασίζονται στην Τεχνητή Νοημοσύνη (non-AI) στη βιβλιογραφία κινδύνου άνοιας, το χάσμα στην απόδοση και η προέλευσή του έγιναν σαφέστερα. Τα κλασικά σκορ κινδύνου μέσης ηλικίας, όπως το CAIDE, τα οποία βασίζονται σε έναν αθροιστικό συνδυασμό ηλικίας, αρτηριακής πίεσης, χοληστερόλης, δείκτη μάζας σώματος και εκπαίδευσης, διακρίνουν με συνέπεια τη μελλοντική άνοια με δείκτη AUROC μεταξύ 0.64 και 0.78 σε εξωτερικές επικυρώσεις (Pietilä et al., 2025). Η δική μας προσέγγιση με Random Forest πέτυχε μέσο AUROC 0.776 και ακρίβεια ελέγχου (test accuracy) 0.707, μια μέτρια αλλά σταθερή βελτίωση που εξηγείται πλήρως από την ικανότητα του μοντέλου να συλλαμβάνει πολύπλοκες αλληλεπιδράσεις υψηλής τάξης.

Για παράδειγμα, ένα από τα δέκα μονοπάτια απόφασης που διατήρησε το μοντέλο εντόπισε κίνδυνο 75% για ανάπτυξη άνοιας εντός 5 ετών. Αυτό το μοτίβο υψηλού κινδύνου εμφανίστηκε σε 252 ασθενείς που μοιράζονταν επτά κοινά κλινικά ευρήματα: επίπεδα κρεατινίνης κοντά στο ανώτατο όριο, χαμηλή VLDL χοληστερόλη, βάρος κάτω του μέσου όρου,

φυσιολογική αναλογία χοληστερόλης προς HDL, χαμηλή θυρεοειδοτρόπο ορμόνη (TSH), χαμηλή ολική πρωτεΐνη και μειωμένο αριθμό λευκών αιμοσφαιρίων. Από μόνες τους, αυτές οι τιμές μπορεί να φαίνονται αδιάφορες ή ακόμα και προστατευτικές, ωστόσο συνδυαστικά σηματοδότησαν μια ομάδα σημαντικά αυξημένου κινδύνου. Η παραδοσιακή βηματική παλινδρόμηση (stepwise regression) πιθανότατα θα είχε χάσει αυτό το μοτίβο, καθώς κάθε μεταβλητή από μόνη της δείχνει μόνο ασθενή συσχέτιση με το αποτέλεσμα. Αυτή η παρατηρούμενη αύξηση της απόδοσης δεν πρέπει να ερμηνεύεται αποκλειστικά ως επικρότηση των μεθόδων AI έναντι των παραδοσιακών προσεγγίσεων. Μάλλον, αντικατοπτρίζει την ικανότητα της συνδυαστικής μάθησης να εξερευνά συστηματικά αλληλεπιδράσεις πολλών διαστάσεων και να αποκαλύπτει κλινικά σημαντικά προφίλ κινδύνου που πιθανότατα θα παρέμεναν αθέατα από τα συμβατικά στατιστικά μοντέλα, τα οποία περιορίζονται από προκαθορισμένους όρους και ένα στενό σύνολο όρων αλληλεπίδρασης.

Για να πλαισιώσουμε τα ευρήματά μας στο τρέχον τοπίο της μοντελοποίησης κινδύνου άνοιας, ανασκοπήσαμε αρκετές πρόσφατες, υψηλής ποιότητας μελέτες. Οι Schliep et al. (2024) συνέδεσαν 4.206 συμμετέχοντες από την κοόρτη της Cache County με 163 διαγνωστικές κατηγορίες κωδικών ICD και έξι κοινωνικοδημογραφικές μεταβλητές. Χρησιμοποιώντας ορίζοντα πρόβλεψης ενός έτους, το μοντέλο τους πέτυχε AUROC 0.67, το οποίο αυξήθηκε στο 0.77 όταν η άνοια ορίστηκε απευθείας από τους κωδικούς ICD αντί μέσω κλινικής κρίσης, αναδεικνύοντας τους περιορισμούς στην απόδοση που επιβάλλονται από τα αραιά σύνολα χαρακτηριστικών. Οι Tang et al. (2024) εκπαίδευσαν έναν ταξινομητή Random Forest σε 749 περιπτώσεις Αλτσχάιμερ και 250.545 μάρτυρες, αναφέροντας AUROC που αυξήθηκε από 0.72 επτά χρόνια πριν την έναρξη σε 0.81 κατά την ημερομηνία αναφοράς. Η χρήση γραφήματος βιοϊατρικής γνώσης από πλευράς τους εντόπισε περαιτέρω την υπερλιπιδαιμία και την οστεοπόρωση ως πρώιμους προγνωστικούς δείκτες ειδικούς για κάθε φύλο. Ο αλγόριθμος Emergency Department Dementia Algorithm (EDDA), που αναπτύχθηκε από 759.665 επισκέψεις στα επείγοντα χρησιμοποιώντας μόνο ένα περιορισμένο σύνολο ζωτικών σημείων διαλογής και πεδίων φαρμακευτικής αγωγής, πέτυχε AUROC 0.85 σε ένα ξεχωριστό σύνολο ελέγχου και 0.93 σε εξωτερική επικύρωση, αποδεικνύοντας ότι εξαιρετικά εστιασμένα χαρακτηριστικά πραγματικού χρόνου μπορούν να αποδώσουν ισχυρή βραχυπρόθεσμη προγνωστική απόδοση (Cohen et al., 2025).

Συγκρινόμενη με αυτά τα σημεία αναφοράς, η προσέγγισή μας πέτυχε AUROC 0.776 και ακρίβεια ελέγχου 0.707 (Πίνακας 4). Αξίζει να σημειωθεί ότι η ανάλυση SHAP εντόπισε νεφρικούς δείκτες (BUN, κρεατινίνη) και μεταβλητές σχετικές με τα λιπίδια που ευθυγραμμίζονται με το σήμα υπερλιπιδαιμίας που περιγράφουν οι Tang et al. (2024), ενώ παράλληλα ανέδειξε χαμηλού βαθμού αναιμία και επίπεδα χολερυθρίνης, χαρακτηριστικά που

δεν είναι ανιχνεύσιμα σε προσεγγίσεις που βασίζονται σε απεικόνιση, όπως το Eye-AD. Τα ευρήματα αυτά υποδηλώνουν ότι, αν και διαφορετικές μέθοδοι συλλαμβάνουν ξεχωριστές βιολογικές υπογραφές, ένα ολοκληρωμένο και ευρέως διαθέσιμο σύνολο χαρακτηριστικών EHR μπορεί να επιτύχει απόδοση συγκρίσιμη ή και ανώτερη από εκείνη εξειδικευμένων μοντέλων, παρέχοντας παράλληλα πρόσθετους, κλινικά ερμηνεύσιμους παράγοντες κινδύνου.

Επανεκτιμήσαμε το μοντέλο χρησιμοποιώντας μόνο τα πιο σχετικά χαρακτηριστικά, συγκεκριμένα αυτά με σκορ σημαντικότητας μεγαλύτερο ή ίσο του 0.005, προκειμένου να αξιολογήσουμε την απόδοση του μοντέλου υπό συνθήκες μειωμένης διαστατικότητας. Μετά από αυτή τη μείωση, το μοντέλο επαναξιολογήθηκε με 27 χαρακτηριστικά αντί των αρχικών 166, γεγονός που οδήγησε σε πτώση της απόδοσης AUROC μόλις κατά περίπου 3%. Μια πιθανή πηγή μεροληψίας στο σύνολο δεδομένων μας είναι η χρήση μεταβλητών που προέρχονται από ερωτηματολόγια, λόγω των αυτοαναφερόμενων μετρήσεων που είναι εγγενώς επιρρεπείς σε σφάλματα ανάκλησης (recall bias), σφάλματα απόκρισης (response bias) και υποκειμενική ερμηνεία, τα οποία μπορεί να εισάγουν μεταβλητότητα και να επηρεάσουν την αξιοπιστία των δεδομένων. Για παράδειγμα, οι ασθενείς μπορεί να αναφέρουν ελλιπώς ή υπερβολικά παράγοντες του τρόπου ζωής, οι φροντιστές μπορεί να παρέχουν ασυνεπείς πληροφορίες ανάλογα με τις αντιλήψεις τους και οι απαντήσεις μπορεί να ποικίλλουν ανάλογα με την εκπαίδευση, το πολιτισμικό υπόβαθρο ή τη γνωστική κατάσταση. Αυτά τα ζητήματα μπορούν να περιορίσουν την ακρίβεια και να εισάγουν συστηματικές διαφορές μεταξύ των ομάδων, οι οποίες με τη σειρά τους μπορεί να επηρεάσουν την προγνωστική μοντελοποίηση εάν οι μεταβλητές των ερωτηματολογίων αποτελούν σημαντικό ποσοστό του συνόλου δεδομένων. Ωστόσο, στη μελέτη μας, τα δεδομένα βάσει ερωτηματολογίων αντιπροσώπευαν μόνο ένα πολύ μικρό κλάσμα των χαρακτηριστικών που χρησιμοποιήθηκαν στα μοντέλα μας, σε αντίθεση με το πολύ μεγαλύτερο σύνολο αντικειμενικών μετρήσεων, όπως αποτελέσματα εργαστηριακών εξετάσεων, διαγνωστικοί κωδικοί και συννοσηρότητες. Λόγω αυτής της ανισορροπίας, οι μεταβλητές των ερωτηματολογίων είχαν αμελητέα επιρροή στα αποτελέσματα των μοντέλων και δεν αλλοίωσαν τη σχετική συνεισφορά των χαρακτηριστικών. Η ανάλυση SHAP επιβεβαίωσε ότι οι κύριοι παράγοντες πρόβλεψης προήλθαν από αντικειμενικά κλινικά δεδομένα. Αν και είναι σημαντικό να αναγνωριστεί η πιθανότητα μεροληψίας στις αυτοαναφερόμενες μετρήσεις, στη συγκεκριμένη περίπτωση ο περιορισμένος ρόλος τους μειώνει τον κίνδυνο ουσιαστικής επίδρασης στην εγκυρότητα του μοντέλου.

Συμπερασματικά, η ανωτέρω μεθοδολογία καταδεικνύει την δυνατότητα αξιοποίησης της μηχανικής μάθησης για την πρόβλεψη του κινδύνου εμφάνισης της νόσου Αλτσχάιμερ και συναφών ανοιών, χρησιμοποιώντας δεδομένα πραγματικού κόσμου από Ηλεκτρονικούς Φακέλους Υγείας (EHR). Μέσω του μετασχηματισμού των χρονικά προσδιορισμένων κλινικών

εγγραφών σε δομημένους προγνωστικούς δείκτες, τα μοντέλα σημείωσαν ισχυρές επιδόσεις, με τις τιμές AUROC να κυμαίνονται μεταξύ 0.77 και 0.79. Ένα σημαντικό στοιχείο είναι ότι η προσέγγισή μας υπερβαίνει την εξέταση μεμονωμένων παραγόντων κινδύνου, εντοπίζοντας συνδυασμούς δεικτών που συγκροτούν κλινικά μονοπάτια —ή αλλιώς σύνολα κανόνων— τα οποία σχετίζονται με την εκδήλωση της άνοιας. Το γεγονός αυτό προσφέρει ερμηνεύσιμα δεδομένα σχετικά με την έναρξη της νόσου και υποστηρίζει μια πιο αποτελεσματική στρωματοποίηση κινδύνου.

4.1 Διερεύνηση της συσχέτισης μεταξύ της προφλεγμονής και της έγκαιρης διάγνωσης της νόσου του Alzheimer σε παρειακά κύτταρα με τη χρήση ανοσοκυτταροχημείας και τεχνικών μηχανικής μάθησης

4.1.1 Συλλογή δειγμάτων

Στη μελέτη αυτή αναλύθηκαν δεδομένα ομάδας 162 ατόμων, που περιλάμβανε 140 υγιή άτομα χωρίς συμπτώματα άνοιας ή γνωστικών ελλειμμάτων και 22 ασθενείς με διάγνωση ήπιας γνωστικής δυσλειτουργίας (MCI) ή νόσου Αλτσχάιμερ. Οι συμμετέχοντες, ηλικίας 18 έως 80 ετών, επιλέχθηκαν τυχαία, χωρίς περιορισμούς ως προς το φύλο, το επίπεδο μόρφωσης ή την κοινωνικοοικονομική κατάσταση. Είναι σημαντικό να σημειωθεί ότι κανένας από τους συμμετέχοντες δεν παρουσίαζε σημάδια στοματικής βλάβης κατά τη στιγμή της συλλογής των στοματικών κυττάρων. Το συλλεγμένο υλικό προετοιμάστηκε εν μέρει ως επιχρίσματα, σταθεροποιήθηκε αμέσως σε 95% αλκοόλη και χρωματίστηκε χρησιμοποιώντας τη μέθοδο Papanicolaou για κυτταρομορφολογική ανάλυση.

4.1.2 Ανοσολογική ανάλυση

Για την ανοσολογική ανάλυση, προετοιμάστηκαν επιπλέον αντικειμενοφόρες πλάκες από τα δείγματα στοματικών κυττάρων. Αυτές οι αντικειμενοφόρες πλάκες μεταφέρθηκαν σε 4% φορμαλδεΰδη σε PBS για τουλάχιστον 30 λεπτά, στεγνώθηκαν στον αέρα για 1 ώρα και αποθηκεύτηκαν σε σφραγισμένα κουτιά στους −4 °C μέχρι να πραγματοποιηθούν οι διαδικασίες ανοσοχρώσης. Τα αντισώματα που χρησιμοποιήθηκαν για την ανοσοκυτταροχημεία περιελάμβαναν TNFα (52B83), IL-1β (E7-2-hIL1β) και IL-6Rα (H-7), όλα μονοκλωνικά αντισώματα ποντικού από την Santa Cruz Biotechnology Inc., το καθένα σε αραίωση 1:50. Τα αντισώματα παρουσίασαν καφέ κυτταροπλασματική έκφραση όταν χρωματίστηκαν με DAB Quanto Chromogen. Τα αντικειμενικά γυαλιά μικροσκοπίου που περιείχαν επιχρίσματα στοματικών κυττάρων αξιολογήθηκαν με οπτική βαθμολόγηση. Τα παρειακά κύτταρα ταξινομήθηκαν ως βασικά, ενδιάμεσα ή διαφοροποιημένα με βάση τα

κυτταροπλασματικά και πυρηνικά χαρακτηριστικά και τις αναλογίες τους (χρώση Pap). Επιλέχθηκε η αιματοξυλίνη Gill καθώς είναι η ελαφρύτερη βασοφιλική χρώση και δεν χρωματίζει υπερβολικά το δείγμα, αλλά συμπληρώνει τις άλλες πτυχές του δείγματος, όπως το κυτταρόπλασμα. Αναλύθηκε η συχνότητα και η αναλογία αυτών των διαφορετικών κυτταρικών πληθυσμών. Επιπλέον, η κυτταροπλασματική έκφραση των κυττάρων που χρωματίστηκαν με την ανοσοκυτταροχημική μέθοδο αξιολογήθηκε και βαθμολογήθηκε ως ήπια-αρνητική (<10%), μέτρια (10-50%) και υψηλή (>50%) έκφραση αντισωμάτων σύμφωνα με δύο μεταβλητούς παράγοντες, δηλαδή την καταμέτρηση του αριθμού των θετικά χρωματισμένων κυττάρων και τη βαθμολόγηση της έντασης της χρώσης.

4.1.3 Ανάλυση δεδομένων

Η προεπεξεργασία των δεδομένων πραγματοποιήθηκε χρησιμοποιώντας τη βιβλιοθήκη pandas της Python. Το σύνολο δεδομένων περιλάμβανε 162 περιπτώσεις σε 13 μεταβλητές, με 9 μεταβλητές να διατηρούνται για ανάλυση. Οι κατηγορικές μεταβλητές μετατράπηκαν σε αριθμητικές τιμές. Για παράδειγμα, το φύλο κωδικοποιήθηκε ως 1 για τους άνδρες και 0 για τις γυναίκες. Η ηλικία κατηγοριοποιήθηκε σε έξι ομάδες (0–19, 20–25, 26–35, 36–50, 51–65, 65+) και κωδικοποιήθηκε αριθμητικά από 0 έως 5. Η ετικέτα για ανάλυση μετατράπηκε σε τρεις κατηγορικές τιμές: 0 για υγεία, 1 για νευρολογικά και 2 για άλλα προφίλ ασθενειών. Η υγεία αντιπροσωπεύει μια ομάδα χωρίς συμπτώματα άνοιας ή γνωστικών ελλειμμάτων, η νευροαντιπροσωπεύει ασθενείς που έχουν διαγνωστεί με MCI και AD, και η άλλη υποδηλώνει μια ομάδα με θετική έκφραση TNF-α, IL-1β και IL-6Rα, κάπνισμα και ιστορικό νευροεκφυλισμού.

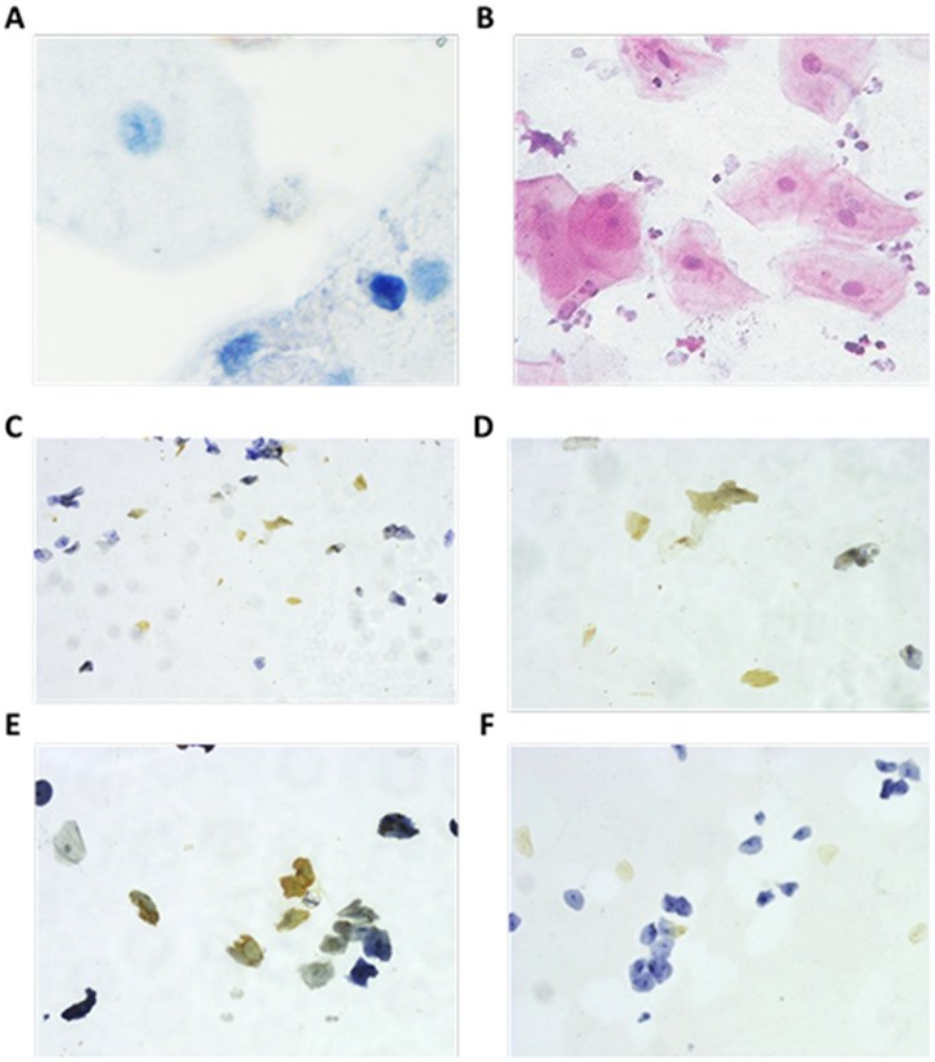
Η οπτικοποίηση των δεδομένων πραγματοποιήθηκε χρησιμοποιώντας τις βιβλιοθήκες matplotlib και seaborn. Δημιουργήθηκε ένας χάρτης θερμότητας συσχέτισης χρησιμοποιώντας τον συντελεστή συσχέτισης του Pearson για να εξεταστούν οι σχέσεις μεταξύ των μεταβλητών και οι συσχετίσεις τους με την ετικέτα (προφίλ) του συνόλου δεδομένων. Η μείωση της διαστατικότητας πραγματοποιήθηκε χρησιμοποιώντας ανάλυση κύριων συνιστωσών (PCA), ομοιόμορφη προσέγγιση και προβολή πολλαπλών διαστάσεων (UMAP) και t-Distributed Stochas tic Neighbor Embedding (t-SNE), με οπτικές αναπαραστάσεις που δημιουργούνται σε μορφές 2D και 3D.

Η σημασία των χαρακτηριστικών αξιολογήθηκε χρησιμοποιώντας μια υβριδική προσέγγιση που συνδυάζει τρεις ταξινομητές gradi ent boosting: XgBoost, CatBoost και LightGBM. Αυτοί οι αλγόριθμοι ταξινόμησαν τα χαρακτηριστικά με βάση τη σημασία τους και οι κατατάξεις συγκεντρώθηκαν χρησιμοποιώντας μια μέθοδο καταμέτρησης βασισμένη στην κατάταξη Borda. Κατασκευάστηκε ένα γράφημα γραμμών/κουκκίδων για να εμφανιστούν οι τιμές

σημασίας των χαρακτηριστικών. Η ταξινόμηση πολλαπλών κατηγοριών πραγματοποιήθηκε χρησιμοποιώντας τον αλγόριθμο τυχαίου δάσους, με μια διαδικασία 10-πλάσιας διασταυρούμενης επικύρωσης για την αξιολόγηση της απόδοσης. Αυτός ο συγκεκριμένος ταξινομητής επιλέχθηκε λόγω της καθιερωμένης ισχυρής απόδοσής του σε σύνθετα βιοϊατρικά σύνολα δεδομένων, σε συνδυασμό με την ερμηνευσιμότητά του. Προσφέρει μια πολύτιμη μέτρηση της σημασίας των χαρακτηριστικών για την αξιολόγηση της σημασίας των μεμονωμένων χαρακτηριστικών. Τα αποτελέσματα συντέθηκαν σε έναν πίνακα σύγχυσης. Επιπλέον, διαμορφώθηκε ένας ταξινομητής δέντρου αποφάσεων με μέγιστο βάθος 10, ελάχιστο 15 δείγματα που απαιτούνται για τη διαίρεση ενός εσωτερικού κόμβου και ελάχιστο 5 δείγματα ανά κόμβο φύλλου για τη διατήρηση της ερμηνευσιμότητας του μοντέλου. Αυτές οι συνδυασμένες βιολογικές και υπολογιστικές μεθοδολογίες εξασφάλισαν μια ολοκληρωμένη ανάλυση, ενσωματώνοντας λεπτομερείς κυτταρικές παρατηρήσεις με προηγμένες τεχνικές επεξεργασίας δεδομένων και μηχανικής μάθησης, προκειμένου να παρέχουν αξιόπιστες πληροφορίες για το σύνολο των δεδομένων. Η μελέτη αυτή στοχεύει να γεφυρώσει το χάσμα μεταξύ κλινικών και υπολογιστικών μεθοδολογιών, προκειμένου να βελτιώσει την κατανόηση και την ανίχνευση της νόσου του Αλτσχάιμερ. Ενσωματώνοντας λεπτομερείς κυτταρικές παρατηρήσεις με προηγμένες τεχνικές επεξεργασίας δεδομένων και μηχανικής μάθησης, η έρευνα επιδιώκει να εντοπίσει σημαντικούς βιοδείκτες και μοτίβα που σχετίζονται με την AD. Ο στόχος είναι να αναπτυχθούν ελάχιστα επεμβατικές, οικονομικά αποδοτικές και αξιόπιστες μέθοδοι για την έγκαιρη διάγνωση και πρόληψη της νόσου του Alzheimer, συμβάλλοντας έτσι στην καλύτερη διαχείριση και ενδεχομένως στην άμβλυνση των επιπτώσεων αυτής της νευροεκφυλιστικής διαταραχής στον γηράσκοντα πληθυσμό.

4.2 Αποτελέσματα

Τα αποτελέσματα αυτής της μελέτης αποκάλυψαν σημαντικά ευρήματα τόσο από τη βιολογική όσο και από την υπολογιστική ανάλυση. Οι αντικειμενοφόρες πλάκες μικροσκοπίου που περιείχαν επιχρίσματα στοματικών κυττάρων αξιολογήθηκαν για την ταξινόμηση των στοματικών κυττάρων με βάση τα κυτταροπλασματικά και πυρηνικά χαρακτηριστικά τους, όπως φαίνεται στη Εικόνα 6. Επιπλέον, αξιολογήθηκε και βαθμολογήθηκε η κυτταροπλασματική έκφραση των στοματικών κυττάρων που χρωματίστηκαν με την ανοσοκυτταροχημική μέθοδο. Η ανάλυση έδειξε ποικίλα επίπεδα έκφρασης των προφλεγμονωδών κυτοκινών TNFα, IL-1β και IL-6, με διακριτά μοτίβα που παρατηρήθηκαν σε άτομα με AD και MCI σε σύγκριση με τους υγιείς μάρτυρες.

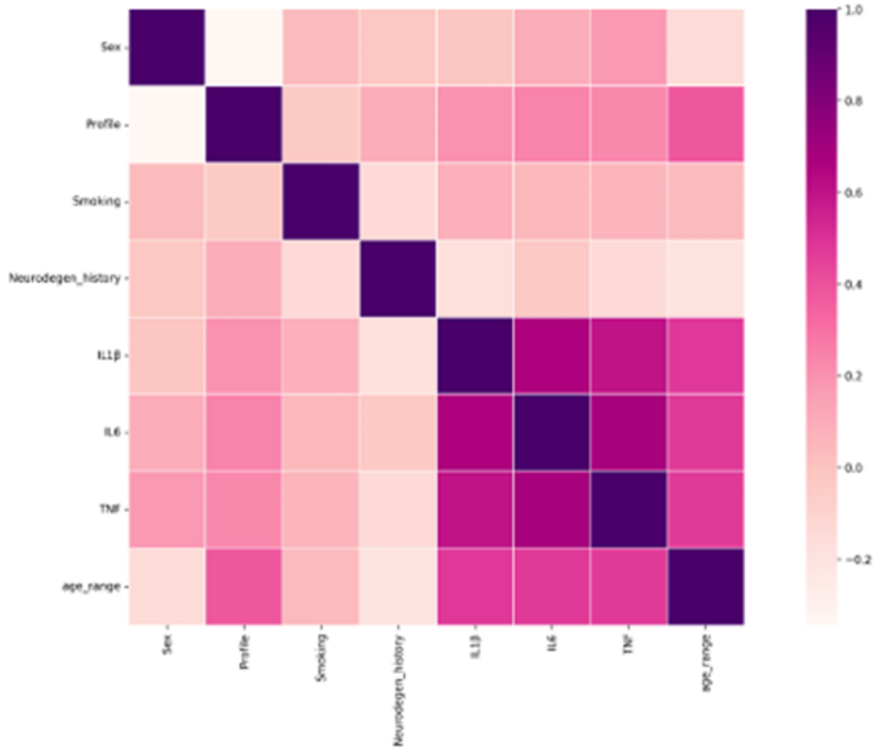


Εικόνα 6. Εικόνες μικροσκοπίου παρειακών κυττάρων χρησιμοποιώντας χρώση Pap σε υγείες εθελοντές (A), σε ασθενείς με AD (B)**,** χρώση Pap που παρατηρήθηκε σε επιφανειακά παρειακά κύτταρα ασθενών με AD, ανοσοϊστοχημική μελέτη της έκφρασης της IL-6 σε ασθενείς. (C x10 και D x20), TNF-a (E), IL-1b (F).

Από υπολογιστική άποψη, ο χάρτης συσχετίσεων έδειξε σημαντικές θετικές συσχετίσεις μεταξύ της μεταβλητής-στόχου (προφίλ) και του κατασκευασμένου ηλικιακού εύρους και, σε μικρότερο βαθμό, με τις μεταβλητές TNF, IL-6 και IL-1β, όπως φαίνεται στην Εικόνα 7. Αυτό υποδηλώνει ότι αυτοί οι παράγοντες συνδέονται στενά με τα προφίλ της νόσου που μελετώνται. Οι απεικονίσεις μείωσης διαστάσεων αποκάλυψαν ότι τα σημεία δεδομένων ομαδοποιήθηκαν ξεκάθαρα, με τις πιο καθορισμένες ομαδοποιήσεις να παρατηρούνται. Προσδιορίστηκαν τέσσερις ξεχωριστές ομάδες, με μία ομάδα να περιλαμβάνει δείγματα που χαρακτηρίζονται τόσο από «νευρολογικά» όσο και από «άλλα» προφίλ, ενώ οι υπόλοιπες ομάδες περιλάμβαναν δείγματα με «νευρολογικά» και/ή «άλλα» προφίλ, όπως φαίνεται στην Εικόνα 3. Αυτό υποδηλώνει τον σαφή διαχωρισμό μεταξύ των διαφορετικών προφίλ ασθενειών με βάση τις αναλυθείσες μεταβλητές. Στο διάγραμμα 2D PCA, καταγράφεται το 74% της

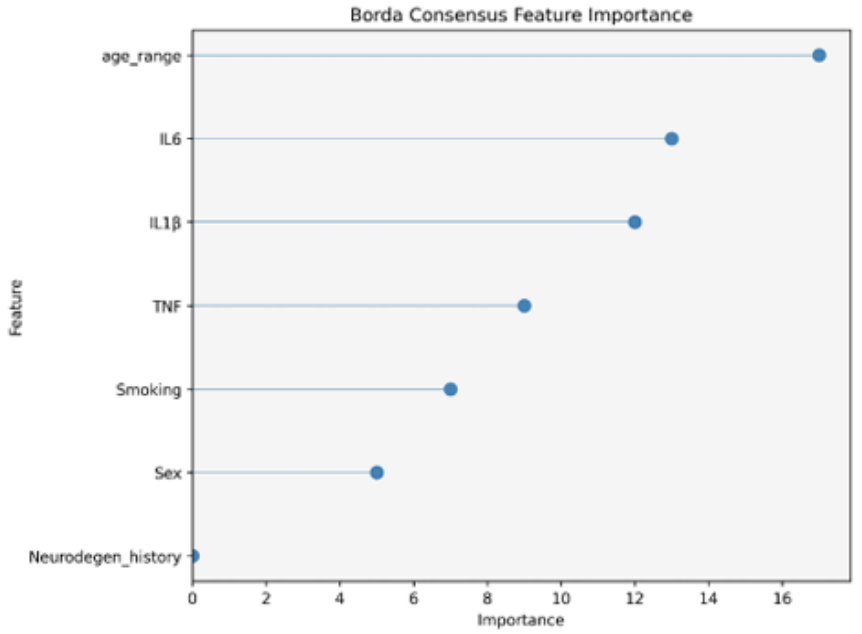
συνολικής διακύμανσης, ενώ το διάγραμμα 3D PCA αντιπροσωπεύει το 81,4% της διακύμανσης. Η μείωση της διαστατικότητας πραγματοποιήθηκε επίσης χρησιμοποιώντας τα τέσσερα κορυφαία χαρακτηριστικά που προσδιορίστηκαν μέσω της μεθόδου συναίνεσης για τη σημασία των χαρακτηριστικών. Σύμφωνα με αυτήν την προσέγγιση, τα δύο πρώτα κύρια συστατικά αντιπροσώπευαν το 87% της συνολικής διακύμανσης, ενώ τα τρία πρώτα κύρια συστατικά εξηγούσαν το 95%. Ωστόσο, είναι σημαντικό να τονιστεί ότι οι τελικές απεικονίσεις PCA, τόσο σε 2D όσο και σε 3D, δεν αποκάλυψαν σαφή μοτίβα διαχωρισμού μεταξύ των δειγμάτων. Ως αποτέλεσμα, τα ευρήματα που σχετίζονται με τα τέσσερα κορυφαία χαρακτηριστικά δεν περιλαμβάνονται σε αυτό το έγγραφο.

Όπως αναμενόταν, όλα τα στοιχεία που βρίσκονται στην κύρια διαγώνιο του διαγράμματος αποδίδονται με βαθύ μωβ χρώμα, υποδηλώνοντας τη συσχέτιση μιας μεταβλητής με τον εαυτό της (Εικόνα 7). Συγκεκριμένα, η μεταβλητή-στόχος του συνόλου δεδομένων (προφίλ) εμφανίζει σημαντική θετική συσχέτιση με το κατασκευασμένο εύρος ηλικιών και, σε μικρότερο βαθμό, με τις μεταβλητές TNF, IL-6 και IL-1β.



Εικόνα 7. Χάρτης θερμότητας συσχετίσεων που περιλαμβάνει τις εννέα μεταβλητές που επιλέχθηκαν για το πεδίο εφαρμογής της παρούσας έρευνας. Κάθε κελί του χάρτη θερμότητας έχει χρωματικό κώδικα που αντιπροσωπεύει την αντίστοιχη τιμή συσχέτισης, η οποία κυμαίνεται από −1 για μεταβλητές που παρουσιάζουν αρνητική συσχέτιση (υποδεικνύεται με λευκό χρώμα) έως 1 για μεταβλητές που παρουσιάζουν θετική συσχέτιση (υποδεικνύεται με μοβ χρώμα).

Η ανάλυση σημασίας χαρακτηριστικών ανέδειξε την ηλικιακή ομάδα ως τον κυρίαρχο παράγοντα διαφοροποίησης μεταξύ των τριών κατηγοριών των μεταβλητών εξόδου (προφίλ), όπως φαίνεται στην Εικόνα 8. Άλλα σημαντικά χαρακτηριστικά ήταν τα IL-6, IL-1β και TNF, υπογραμμίζοντας τη σημασία τους στο πλαίσιο των προφίλ της νόσου. Η απόδοση του αλγορίθμου τυχαίου δάσους έδειξε την υψηλότερη ακρίβεια στην ορθή αναγνώριση δειγμάτων από την κατηγορία «άλλα», με μέτρια αποτελεσματικότητα για τις κατηγορίες «υγιή» και «νευρολογικά», όπως φαίνεται στην Εικόνα 9. Συγκεκριμένα, από τα 55 δείγματα που χαρακτηρίστηκαν ως υγιή, ο αλγόριθμος ταξινόμησε με ακρίβεια τα 37. Για την κατηγορία «νευρολογικά», αναγνώρισε σωστά 16 από τα 21 δείγματα. Στην κατηγορία «άλλα», ο αλγόριθμος κατάφερε να ταξινομήσει με ακρίβεια 60 από τα συνολικά 85 δείγματα, υπογραμμίζοντας την ιδιαίτερη ισχύ του στην αναγνώριση αυτής της κατηγορίας.

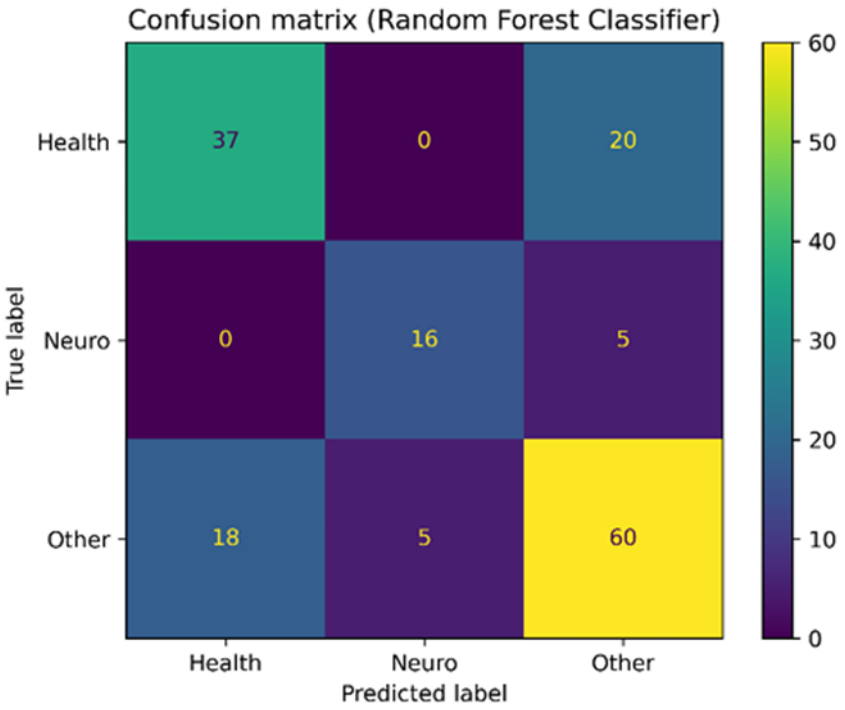


Εικόνα 8. Γράφημα γραμμών/κουκκίδων (line/dot plot) για να εμφανίσει τα αποτελέσματα που προέκυψαν μέσω της μεθοδολογίας συναίνεσης για τη σημασία των χαρακτηριστικών Borda.

Σε αυτό το γράφημα, ο άξονας y απαριθμεί τα ονόματα των χαρακτηριστικών που υπάρχουν στο σύνολο δεδομένων, ενώ ο άξονας x ποσοτικοποιεί τις τιμές σημασίας που προσδιορίστηκαν μέσω της μέτρησης με βάση την κατάταξη Borda. Η εξέταση αυτού του διαγράμματος αποκαλύπτει τη συμφωνία μεταξύ των αποτελεσμάτων του σχήματος συναίνεσης για τη σημασία των χαρακτηριστικών και των μοτίβων που παρατηρήθηκαν στον προηγούμενο χάρτη θερμότητας συσχετίσεων. Συγκεκριμένα, η ηλικιακή ομάδα αναδεικνύεται ως ο κυρίαρχος παράγοντας στη διαφοροποίηση μεταξύ των τριών κατηγοριών της μεταβλητής εξόδου (προφίλ), αντανακλώντας τη θέση της ως το χαρακτηριστικό με τη σημαντικότερη θετική συσχέτιση στον χάρτη θερμότητας. Μια παρόμοια σχέση παρατηρείται για τα χαρακτηριστικά

IL-6, IL-1β και TNF, υπογραμμίζοντας τη συνάφειά τους σύμφωνα τόσο με το σχήμα συναινετικής σημασίας χαρακτηριστικών όσο και με τον χάρτη θερμότητας συσχετίσεων.

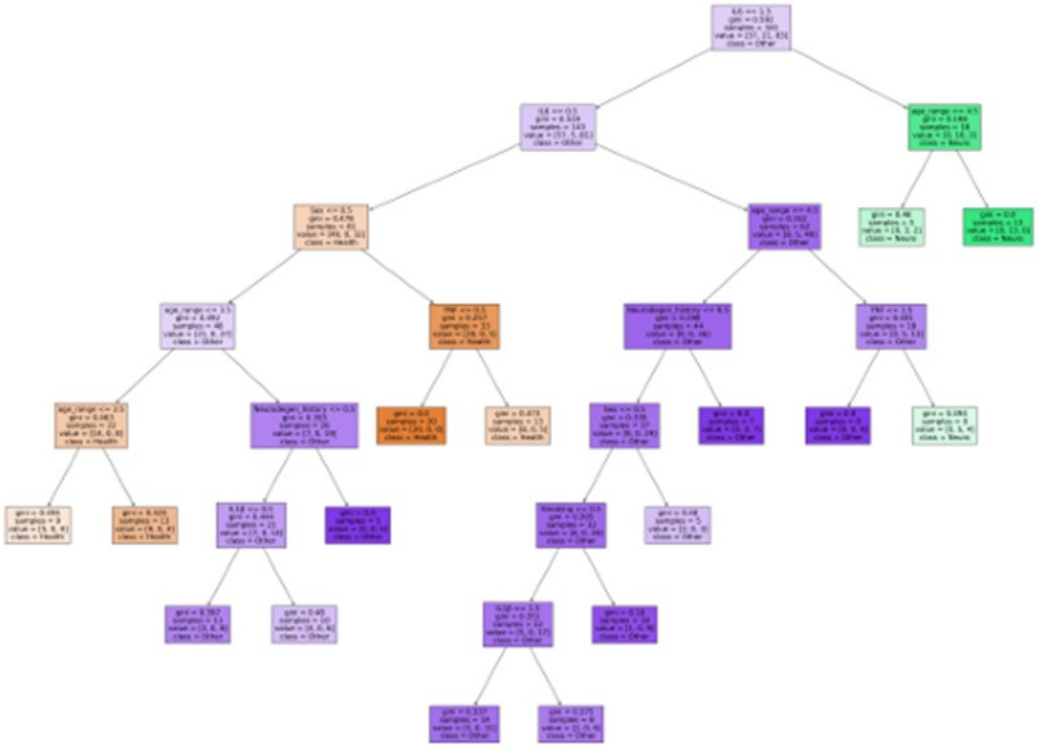
Ο ταξινομητής δέντρου αποφάσεων που επίσης εφαρμόστηκε έχει διαμορφωθεί έτσι ώστε να εξισορροπεί την πολυπλοκότητα και την ερμηνευσιμότητα. Το δέντρο διευκόλυνε την σαφή οπτικοποίηση της τμηματοποίησης του χώρου χαρακτηριστικών, καταγράφοντας βασικά μοτίβα δεδομένων και αποφεύγοντας ψευδείς συσχετίσεις, όπως φαίνεται στην Εικόνα 6. Αυτή η προσέγγιση επέτρεψε τη διαφάνεια και την ερμηνευσιμότητα των αποτελεσμάτων του μοντέλου, συμβάλλοντας στην ολοκληρωμένη κατανόηση των δεδομένων. Η ενσωμάτωση λεπτομερών κυτταρικών παρατηρήσεων με προηγμένες τεχνικές επεξεργασίας δεδομένων και μηχανικής μάθησης παρείχε αξιόπιστες πληροφορίες για το σύνολο δεδομένων. Τα ευρήματα υπογραμμίζουν το δυναμικό του συνδυασμού βιολογικών και υπολογιστικών μεθοδολογιών για την ενίσχυση της κατανόησης και της ανίχνευσης της νόσου του Αλτσχάιμερ, ανοίγοντας το δρόμο για την ανάπτυξη αποτελεσματικών στρατηγικών έγκαιρης διάγνωσης και πρόληψης.



Εικόνα 9. Πίνακας σύγχυσης, που προέρχεται από την ταξινόμηση πολλαπλών κατηγοριών μετά από 10-πλήρη διασταυρούμενη επικύρωση χρησιμοποιώντας τον αλγόριθμο τυχαίου δάσους, απεικονίζει την απόδοση του μοντέλου.

Συγκεκριμένα, το τυχαίο δάσος παρουσιάζει τη μεγαλύτερη ακρίβεια στην ορθή αναγνώριση δειγμάτων από την κατηγορία «άλλα», ενώ παράλληλα επιδεικνύει μέτρια αποτελεσματικότητα για τις δύο υπόλοιπες κατηγορίες εντός της μεταβλητής-στόχου. Συγκεκριμένα, από τα 55 δείγματα που χαρακτηρίστηκαν ως υγιή, ο αλγόριθμος ταξινόμησε με ακρίβεια τα 37. Για την κατηγορία νευρολογικών, αναγνώρισε σωστά 16 από τα 21 δείγματα. Αξίζει να σημειωθεί ότι στην κατηγορία «άλλα», ο αλγόριθμος πέτυχε την ακριβή ταξινόμηση

60 από τα συνολικά 85 δείγματα, υπογραμμίζοντας την ιδιαίτερη ισχύ του στην αναγνώριση αυτής της κατηγορίας.



Εικόνα 10. Οπτικοποίηση δέντρου αποφάσεων με έλεγχο βάθους και περιορισμούς δειγμάτων.

Η Εικόνα 10 παρουσιάζει έναν ταξινομητή δέντρου αποφάσεων που έχει εκπαιδευτεί στο σύνολο δεδομένων. Το δέντρο έχει αρχικοποιηθεί με μέγιστο βάθος 10 για να αποφευχθεί η υπερβολική πολυπλοκότητα, ελάχιστη απαίτηση δειγμάτων 15 για κάθε εσωτερική διαίρεση κόμβου για να αποφευχθεί η υπερβολική προσαρμογή στο θόρυβο και ελάχιστα 5 δείγματα για κάθε κόμβο φύλλου για να εξασφαλιστεί επαρκής υποστήριξη δεδομένων για τελικές αποφάσεις. Κάθε κόμβος αντιπροσωπεύει ένα σημείο απόφασης με βάση την τιμή ενός συγκεκριμένου χαρακτηριστικού, με διακλαδώσεις που οδηγούν σε αποτελέσματα ή περαιτέρω αποφάσεις. Τα φύλλα, που υποδηλώνονται με τα μοναδικά τους χρώματα, αντιστοιχούν στα τελικά αποτελέσματα της ταξινόμησης. Η παράμετρος βάθους διασφαλίζει ότι το δέντρο παραμένει διαχειρίσιμο και ερμηνεύσιμο, ενώ οι περιορισμοί στις διαχωρίσεις και τα δείγματα φύλλων καθοδηγούν το δέντρο να καταγράψει βασικά μοτίβα δεδομένων και να αποφύγει ψευδείς συσχετίσεις. Αυτή η στρατηγική διαμόρφωση διευκολύνει ένα διαφανές και ερμηνεύσιμο μοντέλο, επιτρέποντας την σαφή οπτικοποίηση της τμηματοποίησης του χώρου χαρακτηριστικών σύμφωνα με τη λογική που έχει μάθει το μοντέλο πρόβλεψης.

5. Μελέτη της γενετικής βάσης της νόσου Alzheimer με χρήση ολοκληρωμένων δεδομένων single-nucleus RNA sequencing

Δεδομένης της αυξανόμενης ανάγκης για προηγμένες υπολογιστικές προσεγγίσεις για την ανάλυση δεδομένων αλληλούχισης RNA (scRNA-seq) μεγάλης κλίμακας, η παρούσα μελέτη παρουσιάζει μια προσαρμοσμένη υπολογιστική διαδικασία σχεδιασμένη για την επεξεργασία και ανάλυση δεδομένων γονιδιακής έκφρασης που σχετίζονται με τη νόσο του Αλτσχάιμερ. Η διαδικασία αντιμετωπίζει συγκεκριμένα την πολυπλοκότητα των συνόλων δεδομένων scRNA-seq και snRNA-seq, επιτρέποντας τον εντοπισμό πιθανών βιοδεικτών και μοριακών μηχανισμών που συνδέονται με την εξέλιξη της νόσου. Η ανάλυση βασίζεται σε δεδομένα snRNA-seq από μεταθανάτιο εγκεφαλικό ιστό, τα οποία είναι διαθέσιμα μέσω του Gene Expression Omnibus (GSE147528) [Leng et al 2021]. Η μελέτη αυτή προφίλησε τον ουραίο ενδορινικό φλοιό και τον ανώτερο μετωπιαίο γύρο (SFG) σε όλα τα στάδια Braak της νόσου Αλτσχάιμερ. Για την εργασία αυτή, επιλέχθηκαν μόνο δείγματα από τον SFG, με έμφαση στα στάδια 0 και 2, προκειμένου να εξασφαλιστεί η υπολογιστική σκοπιμότητα και μια ισορροπημένη σύγκριση μεταξύ του ελέγχου και της νόσου σε πρώιμο στάδιο.

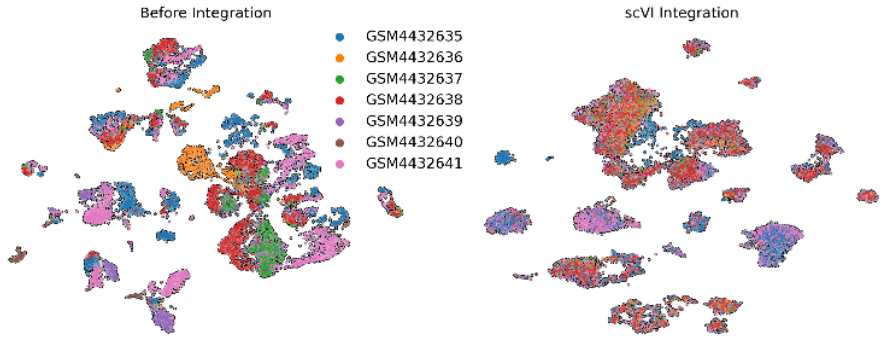
Η αναγνώριση διπλών στοιχείων είναι ένα βασικό στάδιο προεπεξεργασίας στις αναλύσεις scRNA-seq και snRNA-seq, με στόχο την ανίχνευση περιπτώσεων όπου δύο κύτταρα καταγράφονται εσφαλμένα ως ένα, οδηγώντας σε τεχνητά προφίλ γονιδιακής έκφρασης που μπορούν να παραπλανήσουν τις αναλύσεις (Zheng et al., 2017). Σε αυτή τη μελέτη, τα διπλά στοιχεία εντοπίστηκαν και αφαιρέθηκαν χρησιμοποιώντας το SOLO, ένα εργαλείο βασισμένο σε νευρωνικό δίκτυο που ανιχνεύει με ακρίβεια και φιλτράρει τέτοια τεχνητά στοιχεία για να εξασφαλίσει την ποιότητα και την αξιοπιστία του συνόλου δεδομένων [Bernstein et al. 2020].

Η αναφορά του τύπου κυττάρου πραγματοποιήθηκε χρησιμοποιώντας το decoupleR [Badia-i Mompel], μια προσέγγιση βασισμένη σε δείκτες που αξιοποιεί γνωστά γονίδια-δείκτες για τον προσδιορισμό συγκεκριμένων τύπων κυττάρων. Αυτά τα γονίδια-δείκτες επιλέχθηκαν από το CellMarker 2.0, μια ολοκληρωμένη βάση δεδομένων με χειροκίνητα αναφερόμενους δείκτες συγκεκριμένων τύπων κυττάρων. Για τον προσδιορισμό γονιδίων με σημαντικές διαφορές έκφρασης μεταξύ των τύπων κυττάρων, εφαρμόστηκε η δοκιμή κατάταξης Wilcoxon. Από αυτή την ανάλυση, επιλέχθηκαν τα 100 γονίδια με τη μεγαλύτερη διαφοροποίηση στην έκφραση και υποβλήθηκαν σε ανάλυση λειτουργικού εμπλουτισμού χρησιμοποιώντας το εργαλείο Enrichr για να διερευνηθούν οι σχετικές βιολογικές οδοί και διαδικασίες.

5.1 Αποτελέσματα

Μετά την προεπεξεργασία και τη συγχώνευση, το σύνολο δεδομένων περιλάμβανε 12.616 κύτταρα και 19.965 γονίδια/χαρακτηριστικά. Η οπτικοποίηση των δεδομένων

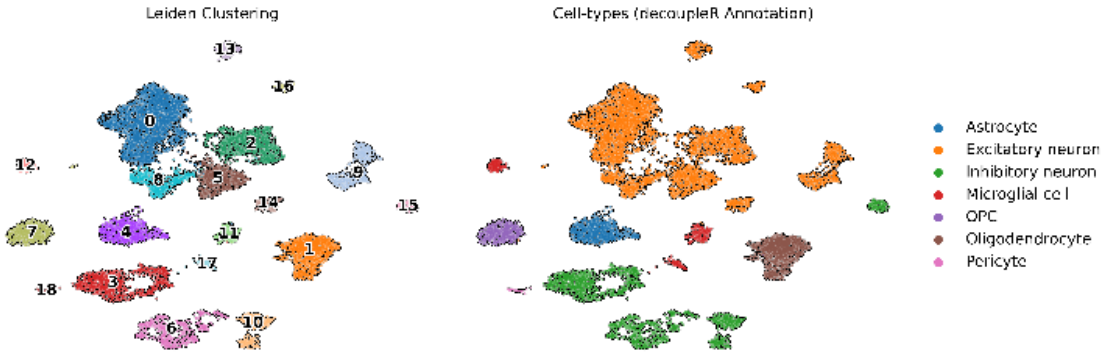
πραγματοποιήθηκε χρησιμοποιώντας το UMAP, μια ευρέως χρησιμοποιούμενη τεχνική μη γραμμικής μείωσης διαστάσεων στην ανάλυση RNA-seq μονής κυψέλης. Σε κάθε γράφημα UMAP, κάθε κουκκίδα αντιπροσωπεύει ένα μεμονωμένο κύτταρο, με τα χρώματα να υποδεικνύουν συγκεκριμένα χαρακτηριστικά μεταδεδομένων. Για παράδειγμα, η Εικόνα 2 εμφανίζει προβολές UMAP πριν και μετά τη διόρθωση του φαινομένου παρτίδας χρησιμοποιώντας scVI. Τα κύτταρα χρωματίζονται ανάλογα με την αντίστοιχη παρτίδα τους. Αξίζει να σημειωθεί ότι, μετά τη διόρθωση, τα κύτταρα από διαφορετικές παρτίδες εμφανίζονται πιο αναμεμειγμένα, υποδηλώνοντας την αποτελεσματική απομάκρυνση της τεχνικής διακύμανσης που είναι χαρακτηριστική της παρτίδας, διατηρώντας παράλληλα τη σημαντική βιολογική δομή των δεδομένων.



Εικόνα 11. Προβολές UMAP του συνόλου δεδομένων πριν και μετά τη διόρθωση του φαινομένου παρτίδας με χρήση scVI. Κάθε σημείο αντιπροσωπεύει ένα μεμονωμένο κύτταρο, χρωματισμένο ανάλογα με την παρτίδα προέλευσής του. Μετά τη διόρθωση, τα κύτταρα από διαφορετικές παρτίδες εμφανίζουν αυξημένη ανάμειξη, υποδηλώνοντας την επιτυχή μείωση της τεχνικής διακύμανσης που σχετίζεται με την παρτίδα, με παράλληλη διατήρηση της υποκείμενης βιολογικής δομής.

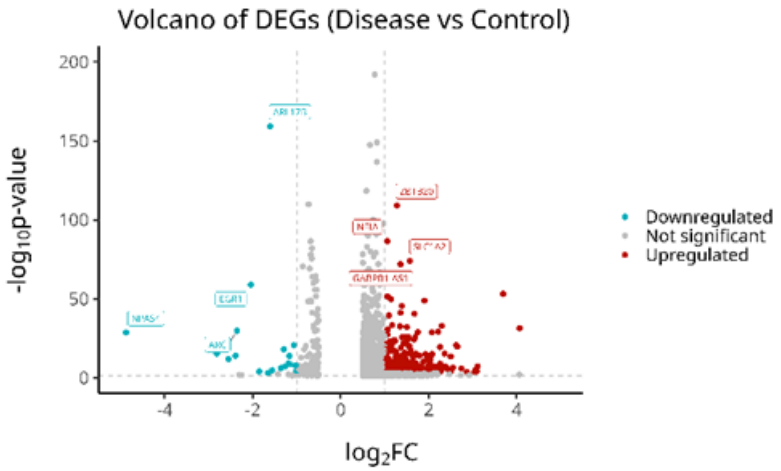
Για την αξιολόγηση της αποτελεσματικότητας της ενοποίησης δεδομένων, χρησιμοποιήθηκε ο δείκτης Local Inverse Simpson’s Index (LISI)]. Η βαθμολογία iLISI ποσοτικοποιεί την ανάμειξη παρτίδων, κυμαινόμενη από 1 (καμία ανάμειξη) έως B (πλήρης ανάμειξη σε B παρτίδες), και κανονικοποιείται από το scIB σε κλίμακα 0–1, με υψηλότερες τιμές να υποδηλώνουν καλύτερη ενσωμάτωση. Μετά την εφαρμογή του scVI, η βαθμολογία LISI αυξήθηκε από 0,057 σε 0,38, αντανακλώντας μια σαφή βελτίωση στην ανάμειξη παρτίδων. Αυτό το ποσοτικό αποτέλεσμα υποστηρίζει τις οπτικοποιήσεις UMAP, επιβεβαιώνοντας ότι το scVI ελαχιστοποίησε αποτελεσματικά τις επιδράσεις παρτίδων, διατηρώντας παράλληλα τη βιολογική δομή των δεδομένων. Η ομαδοποίηση Leiden εφαρμόστηκε στο ενοποιημένο σύνολο δεδομένων με ανάλυση 0,25, με αποτέλεσμα 19 καλά καθορισμένα και, σε ορισμένες περιπτώσεις, διακριτά σύμπλεγματα. Η σχολιασμός του τύπου κυττάρου πραγματοποιήθηκε χρησιμοποιώντας το

decoupleR μαζί με γονίδια δεικτών για τον ανθρώπινο εγκεφαλικό ιστό από το αποθετήριο CellRank 2.0. Με βάση την έκφραση των γονιδίων δεικτών, προσδιορίστηκαν επτά τύποι κυττάρων, με μεγάλο ποσοστό κυττάρων να ταξινομούνται ως διεγερτικά νευρώνες. Η Εικόνα 3 εμφανίζει τα αποτελέσματα της ομαδοποίησης (αριστερά) και τους αντίστοιχους σχολιασμούς του τύπου κυττάρου (δεξιά).



Εικόνα 12. Οπτικοποίηση UMAP των αποτελεσμάτων ομαδοποίησης και σχολιασμού των κυττάρων. Στα αριστερά, εμφανίζονται 19 ομάδες κυττάρων μετά από ομαδοποίηση Leiden με ανάλυση 0,25. Στα δεξιά, οι τύποι κυττάρων προσδιορίστηκαν χρησιμοποιώντας το decoupleR με γονίδια δεικτών του ανθρώπινου εγκεφάλου από το αποθετήριο CellRank 2.0, αποκαλύπτοντας επτά κύριους πληθυσμούς κυττάρων. Ένα μεγάλο μέρος των κυττάρων ταυτοποιήθηκε ως διεγερτικά νευρώνες, με επιπλέον πληθυσμούς που περιλαμβάνουν αστροκύτταρα, ανασταλτικά νευρώνες, μικρογλοιακά κύτταρα, OPC, ολιγοδενδροκύτταρα και περικύτταρα.

Η ανάλυση διαφορικής έκφρασης πραγματοποιήθηκε χρησιμοποιώντας το τεστ Wilcoxon rank-sum, εφαρμόζοντας ένα όριο σημαντικότητας $p < 0,001$ και μια απόλυτη λογαριθμική μεταβολή ($|logFC|$) ≥ 1 για τον προσδιορισμό στατιστικά σημαντικών γονιδίων διαφορικής έκφρασης (DEGs). Τα αποτελέσματα απεικονίζονται σε ένα διάγραμμα ηφαιστείου (Εικόνα 13), όπου κάθε σημείο αντιπροσωπεύει ένα μεμονωμένο γονίδιο. Τα γονίδια που εμφανίζονται με ανοιχτό μπλε χρώμα είναι σημαντικά υποεκφρασμένα στα κύτταρα της νόσου σε σχέση με τα κύτταρα ελέγχου, ενώ αυτά με κόκκινο χρώμα είναι σημαντικά υπερεκφρασμένα. Τα γονίδια με γκρι χρώμα δεν παρουσιάζουν στατιστικά σημαντικές διαφορές έκφρασης μεταξύ των δύο καταστάσεων.



Εικόνα 13. Διάγραμμα ηφαιστείου που απεικονίζει τα αποτελέσματα της ανάλυσης διαφορικής έκφρασης χρησιμοποιώντας το τεστ κατάταξης Wilcoxon. Κάθε σημείο αντιπροσωπεύει ένα γονίδιο, με το ανοιχτό μπλε να υποδηλώνει σημαντική υποέκφραση στα κύτταρα της νόσου σε σύγκριση με τα κύτταρα ελέγχου και το κόκκινο να υποδηλώνει σημαντική υπερέκφραση. Τα γκρι σημεία αντιστοιχούν σε γονίδια χωρίς στατιστικά σημαντικές διαφορές έκφρασης μεταξύ των δύο καταστάσεων. Τα όρια ορίστηκαν σε $p < 0,001$ και $|logFC| \geq 1$.

Η ανάλυση της γονιδιακής οντολογίας (GO) των κυτταρικών συστατικών ανέδειξε τις κυτταρικές συνδέσεις ως την πιο σημαντική κατηγορία. Αυτές οι συνδέσεις είναι απαραίτητες για τη διατήρηση της δομικής ακεραιότητας και της σηματοδότησης μέσω του αιματοεγκεφαλικού φραγμού (BBB). Διαταραχές στις στενές, κενές και προσκολλητικές συνδέσεις που σχηματίζονται από πρωτεΐνες όπως οι κλαουδίνες, η οκλουδίνη, η ZO-1 και η VE-καδερίνη έχουν συνδεθεί με αυξημένη διαπερατότητα του BBB και εξέλιξη της AD. Επιπλέον, η βιβλιοθήκη OMIM Disease αποκάλυψε ότι η ηλικιακή εκφύλιση της ωχράς κηλίδας (AMD) είναι η πιο σημαντική σχετιζόμενη πάθηση. Πρόσφατες μελέτες αναφέρουν ότι τα άτομα με AMD αντιμετωπίζουν υψηλότερο κίνδυνο ανάπτυξης AD, ιδιαίτερα σε νεαρή ηλικία, υποστηρίζοντας έναν πιθανό κοινό παθολογικό μηχανισμό μεταξύ των διαταραχών του αμφιβληστροειδούς και των νευροεκφυλιστικών διαταραχών. Επιπλέον, η ανάλυση GO που σχετίζεται με βιολογικές διεργασίες αποκάλυψε ότι η διαμεμβρανική μεταφορά L-γλουταμινικού ήταν ο πιο σημαντικός όρος. Η δυσλειτουργία των μεταφορέων γλουταμινικού — βασικών πρωτεϊνών που είναι υπεύθυνες για τη διατήρηση της ομοιόστασης του γλουταμινικού στη συναπτική σχισμή — έχει συσχετιστεί με συναπτική δυσλειτουργία και νευροεκφυλισμό στην AD. Η μειωμένη απομάκρυνση της περίσσειας γλουταμινικού μπορεί να οδηγήσει σε διεγερτική τοξικότητα και να συμβάλει στην προοδευτική απώλεια των συνάψεων.

Συμπερασματικά, στην παρούσα μελέτη προχωρήσαμε στην ανάλυση δεδομένων snRNA-seq που προέρχονται από μια συγκεκριμένη περιοχή του εγκεφάλου μεταθανάτιων δειγμάτων, που ονομάζεται ενδορινικός φλοιός. Τα δείγματα προέρχονταν από άτομα που είχαν διαγνωστεί με AD (στάδιο Braak II) ή από υγιείς μάρτυρες. Τα αποτελέσματά μας έδειξαν δυσλειτουργία των μεταφορέων γλουταμινικού και σημαντική ενίσχυση της οδού σηματοδότησης BDNF, η οποία αποτελεί ισχυρή ένδειξη που σχετίζεται με τη συναπτική δυσλειτουργία και τη νευροεκφυλισμό στην AD.

6. Ανάλυση ετερογενών δεδομένων για τον εντοπισμό ισχυρών ενεργοποιητών χρησιμοποιώντας βιβλιοθήκες φαρμάκων

Η λανθασμένη αναδίπλωση πρωτεϊνών αποτελεί χαρακτηριστικό γνώρισμα των νευροεκφυλιστικών διαταραχών (ΝΕΔ), παίζοντας κεντρικό ρόλο στην παθογένεσή τους μέσω της διατάραξης της κυτταρικής πρωτεϊνικής ομοιόστασης και οδηγώντας σε νευρωνική εκφύλιση. Οι μοριακοί σαρωτές (molecular chaperones), όπως οι πρωτεΐνες θερμικού σοκ, είναι κρίσιμοι για τη διατήρηση της πρωτεϊνικής ομοιόστασης, βοηθώντας στη σωστή αναδίπλωση των πρωτεϊνών, αποτρέποντας τη συσσώρευση και διευκολύνοντας την απομάκρυνση των λανθασμένα αναδιπλωμένων πρωτεϊνών.

Έχει σχεδιαστεί ένα πλαίσιο βαθιάς μάθησης που προβλέπει σε ποιο βαθμό μια δεδομένη ένωση μπορεί να ενισχύσει τη δραστηριότητα μιας στοχευόμενης πρωτεΐνης. Το μοντέλο αυτό μπορεί να χρησιμοποιηθεί για σάρωση βιβλιοθηκών φαρμάκων προκειμένου να εντοπιστούν ενώσεις που αυξάνουν τη δραστηριότητα των σαρωτών σχετικών με ΝΕΔ, προσφέροντας ενδεχομένως θεραπευτικά οφέλη, ενώ ταυτόχρονα στοχεύει στην επιτάχυνση της επαναχρησιμοποίησης φαρμάκων για τις ΝΕΔ, συμβάλλοντας σε καλύτερη κατανόηση των θεραπευτικών επιλογών που στοχεύουν την πρωτεϊνική ομοιόσταση.

Το μοντέλο εκπαιδεύτηκε σε 60.591 ζεύγη φάρμακων-πρωτεϊνών, επισημασμένα με το log10 της τιμής δραστηριότητάς τους, που προήλθαν από τη βάση δεδομένων BindingDB και χωρίστηκαν τυχαία σε σύνολα εκπαίδευσης, επικύρωσης και δοκιμής με αναλογίες 70%, 10% και 20%, αντίστοιχα. Τα φάρμακα και οι πρωτεΐνες-στόχοι αναπαραστάθηκαν με SMILES και ακολουθία αμινοξέων αντίστοιχα, ενώ οι τιμές δραστηριότητας εκφράστηκαν με EC50 (η συγκέντρωση ενός φαρμάκου, εκφρασμένη σε νανομόρια (nM), που απαιτείται για την παραγωγή του μισού της μέγιστης βιολογικής απόκρισης της πρωτεΐνης-στόχου). Για να αποδειχθεί η ανθεκτικότητά του, το μοντέλο επικυρώθηκε εξωτερικά και στη βάση δεδομένων ChEMBL, η οποία περιέχει 78.805 δραστηριότητες φάρμακου-στόχου. Οι μετρικές

παλινδρόμησης του μοντέλου ήταν: μέσο τετραγωνικό σφάλμα (MSE) = 0,73, συσχέτιση Pearson = 0,79 και συντελεστής συμφωνίας (concordance index) = 0,8.

7. Ανάλυση ετερογενών αλληλοεπιδρώντων δεδομένων της νόσου Πάρκινσον

7.1 Μεθοδολογία ανάλυση ετερογενών αλληλοεπιδρώντων δεδομένων της νόσου Πάρκινσον

Χρησιμοποιήθηκε ένας ικανός όγκος δεδομένων, ο οποίος αποτελείται από 94 αρχεία Excel που περιλαμβάνουν στοιχεία για περίπου 900 ασθενείς. Πιο συγκεκριμένα, τα αρχεία αυτά περιέχουν απαντήσεις σε ερωτήσεις τις οποίες κλήθηκαν να συμπληρώσουν οι ασθενείς στο πλαίσιο της έρευνας. Τα δεδομένα συλλέχθηκαν από 34 συνεργαζόμενες κλινικές σε διάφορα μέρη του κόσμου, με στόχο τη βαθύτερη κατανόηση των αιτίων που οδηγούν στην επιδείνωση της νόσου Parkinson, καθώς και τη δημιουργία μίας εκτεταμένης βάσης δεδομένων για περαιτέρω μελέτη του φαινομένου. Στην παρούσα έρευνα, ο βασικός στόχος είναι η πρόβλεψη της μετάβασης ενός ασθενούς από μία ηπιότερη κατάσταση της νόσου σε μία επόμενη, πιο σοβαρή κατάσταση. Τα σύνολα δεδομένων που χρησιμοποιήθηκαν περιλαμβάνουν τρεις διαφορετικές καταστάσεις της νόσου. Η πρώτη είναι η φυσιολογική γνωστική κατάσταση (PD-NC), στην οποία ο ασθενής έχει διαγνωστεί με τη νόσο χωρίς εμφανή γνωστική εξασθένηση. Η δεύτερη κατάσταση είναι η ήπια γνωστική εξασθένηση (PD-MCI), η οποία προκύπτει ως αποτέλεσμα της εξέλιξης της νόσου. Η τρίτη και πιο προχωρημένη κατάσταση είναι το στάδιο της άνοιας (PDD), στο οποίο ενδέχεται να καταλήξει ένας ασθενής με νόσο Parkinson.

Τα δεδομένα περιλαμβάνουν ερωτήσεις, απαντήσεις και αποτελέσματα εξετάσεων από διαφορετικές επισκέψεις των ασθενών στον θεράποντα ιατρό τους, όπου καταγράφεται κάθε φορά η γνωστική τους κατάσταση (cognitive state). Η συγκεκριμένη μεταβλητή αποτέλεσε τον βασικό στόχο της πρόβλεψης, με σκοπό την εκτίμηση της μετάβασης από μία κατάσταση της νόσου στην επόμενη, πιο επιβαρυμένη, βάσει των υπόλοιπων διαθέσιμων δεδομένων. Στη συνέχεια, παρουσιάζεται αναλυτικά η διαδικασία που ακολουθήθηκε, προκειμένου να εξαχθούν τα τελικά συμπεράσματα της έρευνας.

7.2 Επεξεργασία Δεδομένων

Προκειμένου να καταστούν τα δεδομένα κατάλληλα για ανάλυση και περαιτέρω αξιοποίηση, απαιτήθηκαν αρκετές προσαρμογές. Αρχικά, τα δεδομένα ήταν διαχωρισμένα σε διαφορετικά σύνολα δεδομένων (datasets), τα οποία έπρεπε να ενοποιηθούν βάσει ενός κοινού χαρακτηριστικού. Για τον σκοπό αυτό χρησιμοποιήθηκε το αναγνωριστικό του ασθενούς (ID), καθώς αποτελεί κοινό στοιχείο μεταξύ των διαφορετικών datasets. Δεδομένου ότι ο στόχος

της έρευνας είναι η πρόβλεψη της επιδείνωσης της κατάστασης του ασθενούς, κρίθηκε απαραίτητη η συσχέτιση των επισκέψεών του με τα υπόλοιπα διαθέσιμα δεδομένα. Για τον λόγο αυτό αξιοποιήθηκαν τα στοιχεία που αφορούν την πρώτη επίσκεψη του ασθενούς (Baseline), τα οποία συσχετίστηκαν με τη στήλη-στόχο που αποτυπώνει τη γνωστική κατάσταση (cognitive state). Ωστόσο, για την επίτευξη της πρόβλεψης της γνωστικής κατάστασης στην επόμενη επίσκεψη, χρησιμοποιήθηκαν οι επόμενες επισκέψεις του ασθενούς, εξαιρουμένης της αρχικής επίσκεψης (Baseline). Με τον τρόπο αυτό κατέστη δυνατή η εκτίμηση της εξέλιξης της γνωστικής κατάστασης με βάση τα δεδομένα της πρώτης επίσκεψης.

Η συγκεκριμένη διαδικασία αποδείχθηκε ιδιαίτερα χρονοβόρα, καθώς απαιτήθηκε η ενοποίηση όλων των επιμέρους αρχείων, λαμβάνοντας υπόψη τις προαναφερθείσες παραμέτρους. Στη συνέχεια, πραγματοποιήθηκε απομάκρυνση μεγάλου αριθμού περιττών στηλών, οι οποίες δεν παρείχαν χρήσιμη πληροφορία για την προβλεπτική διαδικασία. Στόχος ήταν η μείωση των χαρακτηριστικών σε έναν εύλογο και διαχειρίσιμο αριθμό, κατάλληλο για την ανάπτυξη του μοντέλου πρόβλεψης. Ο αρχικός αριθμός χαρακτηριστικών που προέκυψε μετά την ένωση των διαφορετικών αρχείων ανερχόταν σε 382.

7.3 Ανάλυση δεδομένων

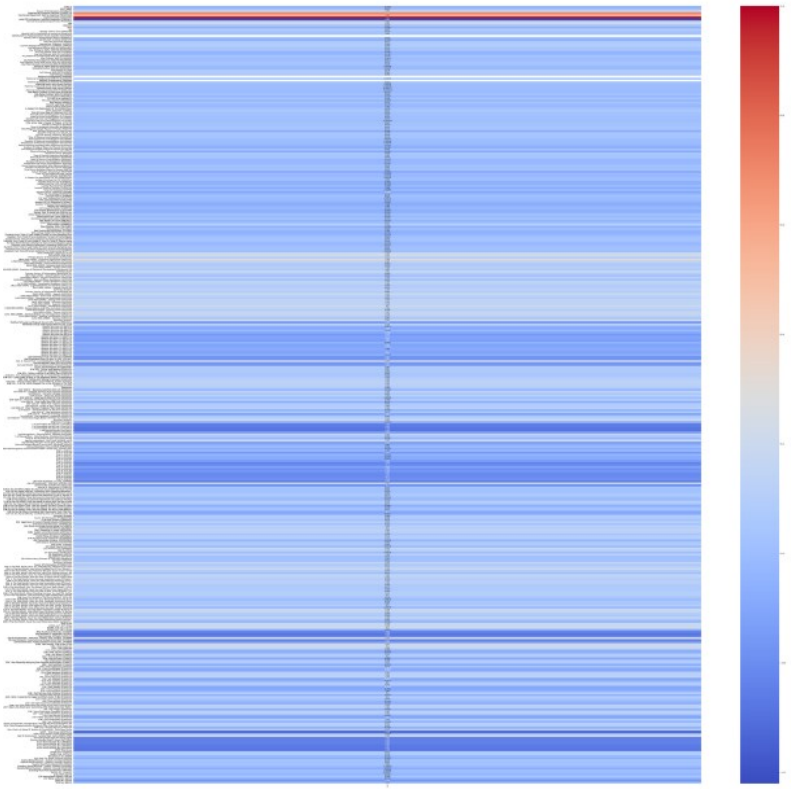
Προκειμένου να προχωρήσει στο στάδιο της ανάλυσης των δεδομένων με στόχο την εξαγωγή αξιόπιστων συμπερασμάτων, κρίθηκε απαραίτητος ο καθαρισμός των δεδομένων. Αφού τα δεδομένα μετασχηματίστηκαν σε κατάλληλη μορφή, όπως παρουσιάστηκε προηγουμένως, ακολούθησε το στάδιο της προεπεξεργασίας, το οποίο περιλάμβανε τον καθαρισμό τους.

Αρχικά, τα δεδομένα περιείχαν σημαντικό αριθμό μηδενικών και μη συμπληρωμένων τιμών, γεγονός που έπρεπε να αντιμετωπιστεί. Για τον λόγο αυτό, τα κενά δεδομένα αντικαταστάθηκαν με τις τιμές των προηγούμενων διαθέσιμων καταγραφών. Στη συνέχεια, μετά τη διαμόρφωση του τελικού συνόλου δεδομένων, διαπιστώθηκε ότι εξακολουθούσε να υπάρχει μεγάλος αριθμός χαρακτηριστικών, παρά τον αρχικό καθαρισμό που είχε πραγματοποιηθεί.

Προκειμένου να επιτευχθεί ένας πιο διαχειρίσιμος αριθμός χαρακτηριστικών, πραγματοποιήθηκε η δημιουργία πίνακα συσχετίσεων (correlation matrix). Στόχος της διαδικασίας αυτής ήταν η επιλογή των πιο σημαντικών χαρακτηριστικών, με σκοπό τη μείωση του συνολικού αριθμού τους σε περίπου 100. Αφού επιλέχθηκαν τα περίπου 100 πιο σημαντικά χαρακτηριστικά, πραγματοποιήθηκε εκ νέου η δημιουργία πίνακα συσχετίσεων, αυτή τη φορά εστιάζοντας αποκλειστικά στο χαρακτηριστικό-στόχο, το οποίο είναι η γνωστική κατάσταση (Cognitive State).

Όπως προκύπτει από τον πίνακα συσχετίσεων, το χαρακτηριστικό με τη μεγαλύτερη συσχέτιση με την κλάση είναι το Experienced Cognitive Decline. Το συγκεκριμένο χαρακτηριστικό αποτυπώνει την υποκειμενική εμπειρία του ασθενούς σε σχέση με τα γνωστικά συμπτώματα που βιώνει, γεγονός που δικαιολογεί την ισχυρή συσχέτισή του με τη γνωστική κατάσταση, καθώς η προσωπική αντίληψη του ασθενούς αποτελεί συχνά αξιόπιστο δείκτη της συνολικής του κατάστασης. Το δεύτερο σημαντικότερο χαρακτηριστικό είναι το Functional Impairment Due to Cognitive, το οποίο αναφέρεται στη λειτουργική δυσλειτουργία του ασθενούς ως αποτέλεσμα της γνωστικής επιδείνωσης που προκαλεί η νόσος.

Το συγκεκριμένο χαρακτηριστικό θεωρείται ιδιαίτερα σημαντικό, καθώς η γενική κατάσταση του ασθενούς και η επιδείνωσή της εκδηλώνονται άμεσα μέσω τέτοιου είδους συμπτωμάτων. Στα υπόλοιπα χαρακτηριστικά δεν παρατηρείται υψηλός βαθμός συσχέτισης σε σύγκριση με τα δύο προαναφερθέντα, γεγονός που υποδηλώνει τη μικρότερη συμβολή τους στην πρόβλεψη της γνωστικής κατάστασης. Με την ολοκλήρωση του σταδίου ανάλυσης και συσχέτισης των χαρακτηριστικών, η διαδικασία προχώρησε στο επόμενο στάδιο, το οποίο περιλάμβανε την εφαρμογή και εκτέλεση αλγορίθμων, με σκοπό την εξαγωγή των τελικών συμπερασμάτων της έρευνας.



Εικόνα 14. Πίνακας συσχέτισης με την κλάση

7.4 Αλγόριθμοι για αξιολόγηση

Στο στάδιο της εξαγωγής των τελικών συμπερασμάτων, επιλέχθηκε η χρήση τριών αλγορίθμων επιβλεπόμενης μηχανικής μάθησης, με στόχο την αξιόπιστη αξιολόγηση και ταξινόμηση των δεδομένων στις κατηγορίες της γνωστικής κατάστασης (*Cognitive State*). Οι αλγόριθμοι που χρησιμοποιήθηκαν είναι οι εξής: ο RUS Boost Classifier, ο Ada Boost Classifier και ο XGB Classifier. Για την εφαρμογή των παραπάνω αλγορίθμων, τα δεδομένα χωρίστηκαν σε σύνολο εκπαίδευσης (training set) σε ποσοστό 70% και σε σύνολο ελέγχου (testing set) σε ποσοστό 30%. Η διαδικασία εκπαίδευσης πραγματοποιήθηκε με ακριβώς τα ίδια σύνολα εκπαίδευσης και για τους τρεις αλγορίθμους, προκειμένου να διασφαλιστεί μια αντικειμενική και συγκρίσιμη αξιολόγηση των αποτελεσμάτων.

Κατά τη διάρκεια των πειραμάτων, διαπιστώθηκε ότι η τρίτη κατηγορία της γνωστικής κατάστασης περιλάμβανε μόλις 15 δείγματα, γεγονός που δεν επέτρεπε την εξαγωγή ασφαλών και αξιόπιστων συμπερασμάτων. Παράλληλα, η δεύτερη κατηγορία περιείχε επίσης σημαντικά λιγότερα δείγματα σε σύγκριση με την πρώτη. Για την αντιμετώπιση αυτών των προβλημάτων, η τρίτη κατηγορία συγχωνεύθηκε με τη δεύτερη και στη συνέχεια εφαρμόστηκε η τεχνική εξισορρόπησης δεδομένων SMOTE, με σκοπό την εξισορρόπηση των δύο τελικών κατηγοριών που χρησιμοποιήθηκαν στην ανάλυση. Τα αποτελέσματα της εκτέλεσης των αλγορίθμων παρουσιάζονται στους πίνακες που ακολουθούν.

Πίνακας 2. Αποτελέσματα ταξινομητών.

Cognitive State	**0**	**1**	**Macro avg**	**Weight avg**	**Accuracy**
Precision	0.92	0.97	0.94	0.95	
Recall	0.96	0.93	0.95	0.95	
F1-Score	0.94	0.95	0.95	0.95	0.95
Support	199	247	446	446	446
Cognitive State	**0**	**1**	**Macro avg**	**Weight avg**	**Accuracy**
Precision	0.92	0.94	0.93	0.93	
Recall	0.93	0.93	0.93	0.93	
F1-Score	0.93	0.93	0.93	0.93	0.93
Support	209	237	446	446	446
Cognitive State	**0**	**1**	**Macro avg**	**Weight avg**	**Accuracy**

Precision	0.96	0.92	0.94	0.94	
Recall	0.91	0.97	0.94	0.94	
F1-Score	0.93	0.94	0.94	0.94	0.93
Support	209	237	446	446	446

Με βάση τις μετρήσεις των αποτελεσμάτων που παρουσιάστηκαν, διαπιστώθηκε ότι οι αλγόριθμοι εμφάνισαν ιδιαίτερα υψηλή απόδοση, καθώς και οι τρεις πέτυχαν υψηλά ποσοστά ακρίβειας (Precision). Ειδικότερα, για την κλάση 1, ο αλγόριθμος RUS Boost παρουσίασε την υψηλότερη απόδοση με ποσοστό 97%, ακολούθησε ο Ada Boost με 94%, ενώ ο XGB Classifier πέτυχε ποσοστό 92%. Η επιλογή του ποσοστού ακρίβειας στην κλάση 1 ως βασικού μέτρου σύγκρισης των αλγορίθμων σχετίζεται άμεσα με τη φύση και τη σημασία της συγκεκριμένης κλάσης. Η κλάση αυτή αντιπροσωπεύει τη δεύτερη κατάσταση του ασθενούς στη μεταβλητή Cognitive State, η οποία ουσιαστικά αποτυπώνει την επιδείνωση της γνωστικής του κατάστασης. Κατά συνέπεια, η σωστή ταξινόμηση της συγκεκριμένης κλάσης είναι ιδιαίτερα κρίσιμη, καθώς μπορεί να συμβάλει ουσιαστικά στην έγκαιρη πρόβλεψη της επιδείνωσης ενός ασθενούς βάσει συγκεκριμένων συμπτωμάτων. Ωστόσο, για την επίτευξη μιας πιο λεπτομερούς και αξιόπιστης πρόβλεψης της πορείας της νόσου, θα ήταν απαραίτητη η ύπαρξη επιπλέον δεδομένων για όλες τις καταστάσεις του ασθενούς, και κυρίως για την τρίτη κατάσταση, για την οποία υπήρχε περιορισμένος αριθμός δειγμάτων. Ο μικρός όγκος δεδομένων για τη συγκεκριμένη κατάσταση δεν επέτρεψε την εφαρμογή τεχνικών εξισορρόπησης, καθώς κάτι τέτοιο θα οδηγούσε σε μη αξιόπιστα αποτελέσματα. Παράλληλα, η μετάβαση από την πρώτη στη δεύτερη κατάσταση της νόσου φαίνεται να είναι ιδιαίτερα κρίσιμη, καθώς στο μεταβατικό αυτό στάδιο υπάρχει επαρκές χρονικό περιθώριο για την εφαρμογή κατάλληλης θεραπευτικής αγωγής, με στόχο την επιβράδυνση της εξέλιξης προς το τρίτο και τελικό στάδιο της ασθένειας.

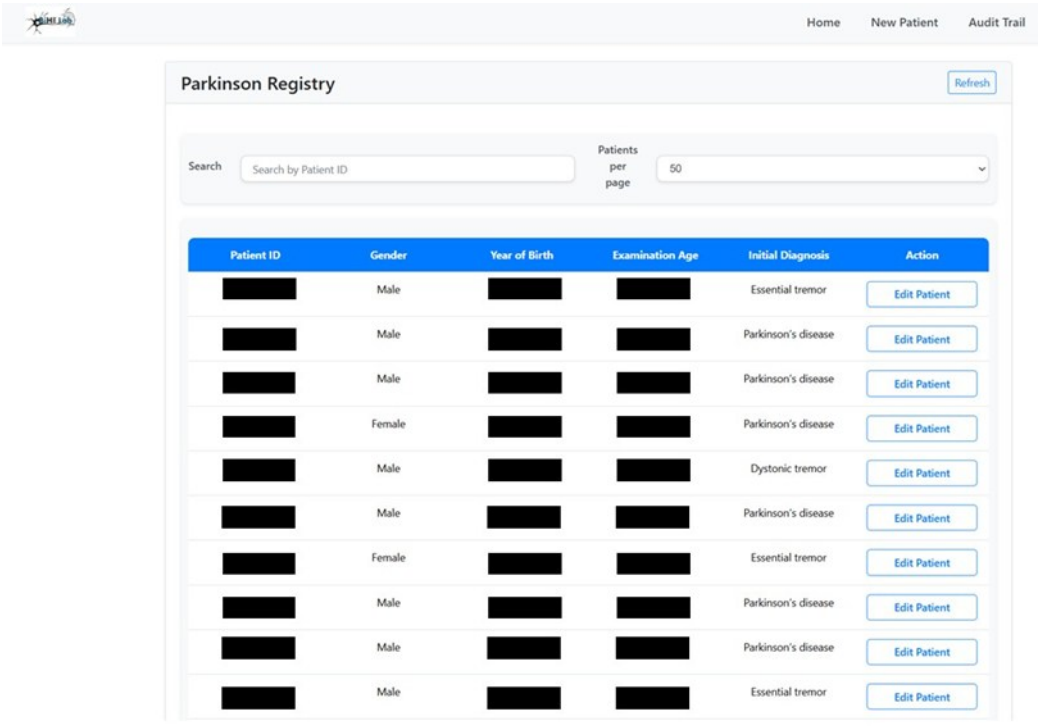
8. Ανάπτυξη πλατφόρμας συλλογής και διαχείρισης δεδομένων για κλινικές μελέτες σε ασθενείς με διαταραχές της νόσου Πάρκινσον «Parkinson Registry»

Αναπτύχθηκε ειδική πλατφόρμα «Parkinson Registry», με στόχο την καταγραφή δεδομένων σχετικά με το νευρολογικό προφίλ ασθενών που πάσχουν από τη νόσο Πάρκινσον ή από συναφή παρκινσονικά σύνδρομα, στο πλαίσιο κλινικής μελέτης. Η πλατφόρμα συλλέγει ένα ευρύ φάσμα δημογραφικών, κλινικών, γνωστικών, νευρολογικών και απεικονιστικών δεδομένων για την παρακολούθηση της κλινικής εικόνας του πληθυσμού της μελέτης σε βάθος χρόνου, καθώς και για τη μεταγενέστερη ανάλυση των δεδομένων με σκοπό την καλύτερη

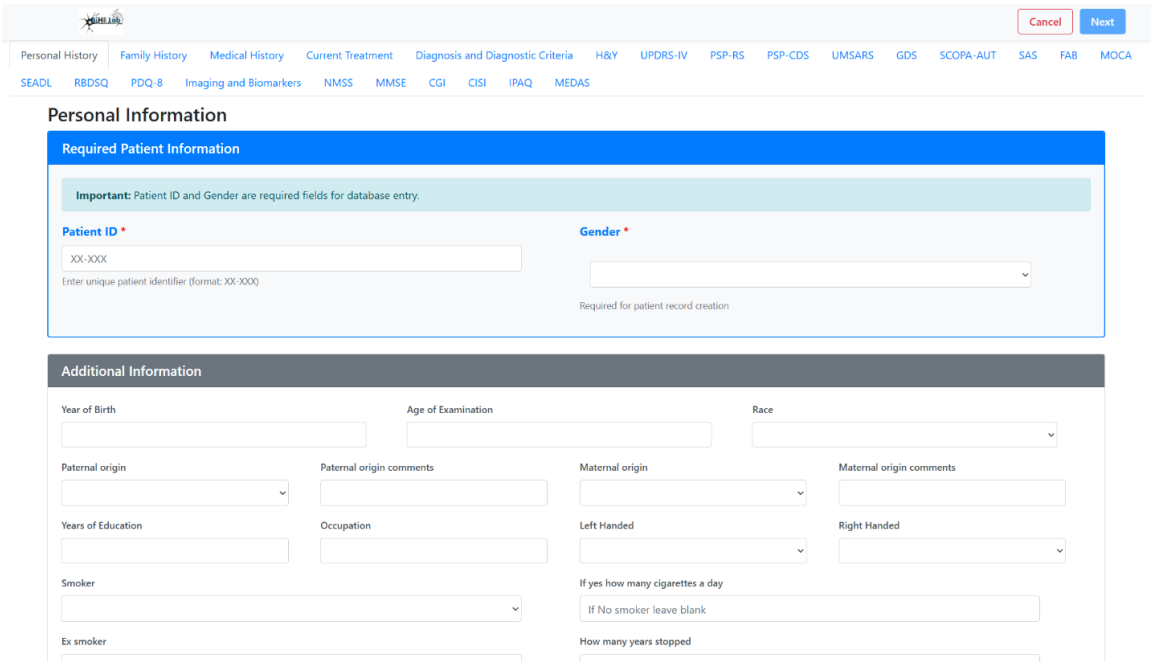
κατανόηση της νόσου και της ανταπόκρισης σε ενδεχόμενες παρεμβάσεις. Μερικές από τις κλινικές κλίμακες που συλλέγονται είναι οι εξής:

1. H&Y: Hoehn and Yahr Scale
2. UPDRS-IV: Unified Parkinson's Disease Rating Scale (Part IV: Motor Complications)
3. PSP-RS: Progressive Supranuclear Palsy Rating Scale
4. PSP-CDS: Progressive Supranuclear Palsy Clinical Deficits Scale
5. UMSARS: Unified Multiple System Atrophy Rating Scale
6. GDS: Geriatric Depression Scale
7. SCOPA-AUT: Scales for Outcomes in Parkinson's Disease - Autonomic
8. SAS: Starkstein Apathy Scale (or Zung Self-Rating Anxiety Scale)
9. FAB: Frontal Assessment Battery
10. MOCA: Montreal Cognitive Assessment
11. SEADL: Schwab and England Activities of Daily Living
12. RBDSQ: REM Sleep Behavior Disorder Screening Questionnaire
13. PDQ-8: Parkinson's Disease Questionnaire-8

Η πλατφόρμα υποστηρίζει καταγραφή δεδομένων σε πολλαπλά χρονικά σημεία, ανάλογα με τη συχνότητα των επισκέψεων των συμμετεχόντων στη νευρολογική κλινική του νοσοκομείου. Επιπλέον, παρέχεται η δυνατότητα προσθήκης ή αφαίρεσης των ερωτηματολογίων και των δεικτών που συλλέγονται ανάλογα με τις ανάγκες και τις οδηγίες του υπευθύνου της μελέτης, καθώς και εξαγωγής των δεδομένων σε οποιοδήποτε στάδιο της μελέτης μετά από σχετικό αίτημα. Τα δεδομένα συλλέγονται και επεξεργάζονται σε ανωνυμοποιημένη μορφή, με κάθε ασθενή να διαθέτει μοναδικό αναγνωριστικό συστήματος (patient ID). Παράλληλα, υποστηρίζεται η δημιουργία πολλαπλών προφίλ χρηστών σε περίπτωση που η μελέτη απαιτεί την εμπλοκή περισσότερων ατόμων για την εισαγωγή δεδομένων, ενώ τηρείται αρχείο καταγραφής και επεξεργασίας (audit trail) για πλήρη διαφάνεια σε κάθε στάδιο της μελέτης. Η πλατφόρμα «Parkinson Registry» λειτουργεί σαν ένα πλήρες μητρώο καταγραφής και βοηθά συνολικά την κλινική μελέτη, κάνοντας πιο εύκολη την ανάλυση των δεδομένων μέσα στο πλαίσιο των εκάστοτε πρωτόκολλων συλλογής, ενώ μπορεί να προσαρμοστεί σχετικά εύκολα στις ανάγκες κάθε κλινικής μελέτης.



Εικόνα 15. Η αρχική οθόνη της πλατφόρμας «Parkinson Registry»



Εικόνα 16. Η οθόνη προσθήκης νέου ασθενούς στην πλατφόρμα «Parkinson Registry»

Βιβλιογραφία

1. Akter, S., Liu, Z., Simoes, E. J., and Rao, P. (2025). Using machine learning and electronic health record (ehr) data for the early prediction of alzheimer's disease and related dementias. *The Journal of Prevention of Alzheimer's Disease*, 100169
2. Al-Sahab, B., Leviton, A., Loddenkemper, T., Paneth, N., and Zhang, B. (2024). Biases in electronic health records data for generating real-world evidence: an overview. *Journal of Healthcare Informatics Research* 8, 121–139
3. Amrollahi, F., Shashikumar, S. P., Holder, A. L., and Nemati, S. (2022). Leveraging clinical data across healthcare institutions for continual learning of predictive risk models. *Scientific reports* 12, 8380
4. Bakouny, Z. and Patt, D. A. (2021). Machine learning and real-world data: More than just buzzwords. *JCO Clinical Cancer Informatics*, 811–813doi:10.1200/CCI.21.00092. PMID: 34383581
5. Bastarache, L., Brown, J. S., Cimino, J. J., Dorr, D. A., Embi, P. J., Payne, P. R., et al. (2022). Developing real-world evidence from real-world data: Transforming raw data into analytical datasets. *Learning Health Systems* 6, e10293
6. Boustani, M., Perkins, A. J., Khandker, R. K., Duong, S., Dexter, P. R., Lipton, R., et al. (2020). Passive digital signature for early identification of alzheimer's disease and related dementia. *Journal of the American Geriatrics Society* 68, 511–518
7. Chauhan, V. K., Thakur, A., O'Donoghue, O., Rohanian, O., Molaei, S., and Clifton, D. A. (2024). Continuous patient state attention model for addressing irregularity in electronic health records. *BMC Medical Informatics and Decision Making* 24, 117
8. Collins, F. and Tabak, L. (2014). Using machine learning to identify health outcomes from electronic health record data. *Nature* 505, 612–613
9. [Dataset] DeMers, D. and Wachs, D. (2024). Physiology, mean arterial pressure
10. Fisher, S., Manuel, D. G., Hsu, A. T., Bennett, C., Tuna, M., Eddeen, A. B., et al. (2021). Development and validation of a predictive algorithm for risk of dementia in the community setting. *J Epidemiol Community Health* 75, 843–853
11. Ford, E., Rooney, P., Oliver, S., Hoile, R., Hurley, P., Banerjee, S., et al. (2019). Identifying undetected dementia in uk primary care patients: a retrospective case-control study comparing machine-learning and standard epidemiological approaches. *BMC medical informatics and decision*
12. *making* 19, 1–9
13. Gao, X. R., Chiariglione, M., Qin, K., Nuytemans, K., Scharre, D. W., Li, Y.-J., et al. (2023). Explainable machine learning aggregates polygenic risk scores and electronic health records for alzheimer's disease prediction. *Scientific reports* 13, 450
14. Goldstein, B. A., Navar, A. M., Pencina, M. J., and Ioannidis, J. P. A. (2017). Preprocessing structured clinical data for predictive modeling and decision support. *Annals of Internal Medicine* 167, 361–368
15. Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology* 143, 29–36
16. Huang, H., Ren, Y., Wang, J., Zhang, Z., Zhou, J., Chang, S., et al. (2024). Renal function and risk of dementia: a mendelian randomization study. *Renal Failure* 46, 2411856
17. Ibrahim, J. G., Chu, H., and Chen, M.-H. (2012). Missing data in clinical studies: issues and methods. *Journal of clinical oncology* 30, 3297–3303

18. Jammeh, E. A., Camille, B. C., Stephen, W. P., Escudero, J., Anastasiou, A., Zhao, P., et al. (2018). Machine-learning based identification of undiagnosed dementia in primary care: a feasibility study. *BJGP open* 2
19. Kennelly, S. P., Lawlor, B. A., and Kenny, R. A. (2009). Blood pressure and dementia—a comprehensive review. *Therapeutic advances in neurological disorders* 2, 241–260
20. Kim, M. K., Rouphael, C., McMichael, J., Welch, N., and Dasarathy, S. (2023). Challenges in and opportunities for electronic health record-based data analysis and interpretation. *Gut and Liver* 18, 201
21. Kim, S. and Min, W.-K. (2025). Toward high-quality real-world laboratory data in the era of healthcare big data. *Annals of Laboratory Medicine* 45, 1–11
22. Laabs, B.-H., Westenberger, A., and K¨onig, I. R. (2024). Identification of representative trees in random forests based on a new tree-based distance measure. *Advances in Data Analysis and Classification* 18, 363–380
23. Lebedev, A., Westman, E., Van Westen, G., Kramberger, M., Lundervold, A., Aarsland, D., et al. (2014). Random forest ensembles for detection and prediction of alzheimer's disease with a good between-cohort robustness. *NeuroImage: Clinical* 6, 115–125
24. Li, J., Yan, X. S., Chaudhary, D., Avula, V., Mudiganti, S., Husby, H., et al. (2021). Imputation of missing values for electronic health record laboratory data. *NPJ digital medicine* 4, 147
25. Li, Q., Yang, X., Xu, J., Guo, Y., He, X., Hu, H., et al. (2023). Early prediction of alzheimer's disease and related dementias using real-world electronic health records. *Alzheimer's & Dementia* 19, 3506–3518
26. Lu, Y., Pike, J. R., Selvin, E., Mosley, T., Palta, P., Sharrett, A. R., et al. (2021). Low liver enzymes and risk of dementia: the atherosclerosis risk in communities (aric) study. *Journal of Alzheimer's Disease* 79, 1775–1784
27. Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30
28. Moore, A. and Bell, M. (2022). Xgboost, a novel explainable ai technique, in the prediction of myocardial infarction: a uk biobank cohort study. *Clinical Medicine Insights: Cardiology* 16, 11795468221133611
29. Nori, V. S., Hane, C. A., Crown, W. H., Au, R., Burke, W. J., Sanghavi, D. M., et al. (2019).
30. Machine learning models to predict onset of dementia: a label learning approach. *Alzheimer's & Dementia: Translational Research & Clinical Interventions* 5, 918–925
31. Office of the National Coordinator for Health Information Technology (2022). National trends in hospital and physician adoption of electronic health records. *HealthIT. gov*
32. Powers, D. and Ailab (2011). Evaluation: From precision, recall and f-measure to roc, informedness, markedness & correlation. *J. Mach. Learn. Technol* 2, 2229–3981. doi:10.9735/2229-3981
33. Qiang, Y.-X., Deng, Y.-T., Zhang, Y.-R., Wang, H.-F., Zhang, W., Dong, Q., et al. (2023). Associations of blood cell indices and anemia with risk of incident dementia: a prospective cohort study of 313,448 participants. *Alzheimer's & Dementia* 19, 3965–3976
34. Qizilbash, N., Gregson, J., Johnson, M. E., Pearce, N., Douglas, I., Wing, K., et al. (2015). Bmi and risk of dementia in two million people over two decades: a retrospective cohort study. *The*
35. *lancet Diabetes & endocrinology* 3, 431–436

36. Rashidisabet, H., Sethi, A., Jindarak, P., Edmonds, J., Chan, R. P., Leiderman, Y. I., et al. (2023). Validating the generalizability of ophthalmic artificial intelligence models on real-world clinical data. *Translational Vision Science & Technology* 12, 8–8
37. Ren, L., Wang, T., Seklouli, A. S., Zhang, H., and Bouras, A. (2023). A review on missing values for main challenges and methods. *Information Systems* 119, 102268
38. Ren, W., Liu, Z., Wu, Y., Zhang, Z., Hong, S., Liu, H., et al. (2024). Moving beyond medical statistics: A systematic review on missing data handling in electronic health records. *Health Data Science* 4, 0176
39. Shickel, B., Tighe, P. J., Bihorac, A., and Rashidi, P. (2018). Deep representation learning of patient data from electronic health records (ehr): A systematic review. *Journal of Biomedical Informatics* 83, 36–46
40. Tang, A. S., Rankin, K. P., Cerono, G., Miramontes, S., Mills, H., Roger, J., et al. (2024). Leveraging electronic health records and knowledge networks for alzheimer's disease prediction and sex-specific biological insights. *Nature Aging* 4, 379–395
41. Wang, L., Wang, F., Liu, J., Zhang, Q., and Lei, P. (2018). Inverse relationship between baseline serum albumin levels and risk of mild cognitive impairment in elderly: a seven-year retrospective cohort study. *The Tohoku journal of experimental medicine* 246, 51–57
42. World Health Organization (1992). *International Statistical Classification of Diseases and Related Health Problems. Tenth Revision (ICD-10)* (Geneva: World Health Organization)
43. Xiao, Y., Devakumar, V., Xu, L., Liu, L., Mo, H., and Hong, X. (2023). Elevated serum creatinine levels and risk of cognitive impairment in older adults with diabetes: a nhanes study from 2011-2014. *Frontiers in Endocrinology* 14, 1149084
44. Tsiouris, K.M., Konitsiotis, S., Koutsouris, D.D. and Fotiadis, D.I., 2020. Prognostic factors of rapid symptoms progression in patients with newly diagnosed Parkinson's disease. *Artificial intelligence in medicine*, *103*, p.101807
45. Smith, N., Williams, O., Ricciardi, L., Morgante, F., Barrick, T.R., Edwards, M. and Lambert, C., 2021. Predicting future cognitive impairment in de novo parkinson's disease using clinical data and structural MRI. *medRxiv*, pp.2021-08.
46. K. Leng, E. Li, R. Eser, A. Piergies, R. Sit, M. Tan, N. Neff, S. H. Li, R. D. Rodriguez, C. K. Suemoto et al., Molecular characterization of selectively vulnerable neurons in alzheimers diseases, Nature neuroscience, vol. 24, no. 2, pp. 276–287, 2021.
47. G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu et al., Massively parallel digital transcriptional profiling of single cells, Nature communications, vol. 8, no. 1, p. 14049, 2017.
48. N. J. Bernstein, N. L. Fong, I. Lam, M. A. Roy, D. G. Hendrickson, and D. R. Kelley, "Solo: doublet identification in single-cell rna-seq via semi-supervised deep learning," Cell systems, vol. 11, no. 1, pp. 95–101, 2020

Επιστημονικές Δημοσιεύσεις

1. Exarchos, T.P., Dimakopoulos, G.A., Lazaros, K., Krokidis, M., Vrahatis, A., Grammenos, G., Avramouli, A., Skolariki, K., Adams, R., Mahairaki, V., Oh, E.S., et al. 2025. 5-year dementia prediction and decision support system based on real-world data. *Frontiers in aging neuroscience*, *17*, p.1670609

2. Lazaros, K., Adam, S., Krokidis, M.G., Exarchos, T., Vlamos, P. and Vrahatis, A.G., 2025. Non-invasive biomarkers in the era of big data and machine learning. *Sensors*, *25*(5), p.1396
3. Krokidis, M.G., Koumadorakis, D.E., Lazaros, K., Ivantsik, O., Exarchos, T.P., Vrahatis, A.G., Kotsiantis, S. and Vlamos, P., 2025. AlphaFold3: an overview of applications and performance insights. *International Journal of Molecular Sciences*, *26*(8), p.3671.
4. Lazaros, K., Gonidi, M., Kontara, N., Krokidis, M.G., Vrahatis, A.G., Exarchos, T. and Vlamos, P., 2024. Exploring the Association between Pro-Inflammation and the Early Diagnosis of Alzheimer's Disease in Buccal Cells Using Immunocytochemistry and Machine Learning Techniques. *Applied Sciences*, *14*(18), p.8372.

Δημοσιεύσεις και αναφορές σε πρακτικά συνεδρίων (Conference papers and proceedings)

1. Lazaros, K., Krokidis, M.G., Grammenos, G., Adam, S., Exarchos, T. Vlamos, P. and Vrahatis, A.G. 2025. Uncovering Molecular Insights into Blood–Brain Barrier Dysfunction in Alzheimer's Disease. 2025 IEEE 25th International Conference on Bioinformatics and Bioengineering (BIBE), 6-8 November 2025, Athens, Greece, S1.5.
2. Papikinos, T., Krokidis, M.G., Vrahatis, A.G. Vlamos, P. and Exarchos, T. 2025. Investigating protein folding as a target for neurodegenerative disorders using a biological activity prediction model based on deep learning, 3rd International Conference on Chemo and Bioinformatics, 25-26 September, 2025. Kragujevac, Serbia (pp. 319-322).
3. Papatheodorou M-C, Exarchos, Karanikolaos, P., Lazaros, K, Vrahatis, A.G. and Exarchos, T.P., Kapogiannis, D., Vlamos, P and Krokidis, M.G. 2025. Advanced Computational Approaches Identify Gene Expression and Extracellular Vesicle Lipid markers in Alzheimer's Disease, 75th Conference of the HSBMB, 5-7 December 2025, Athens, Greece, P212 – 277