ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Διάγνωση εγκεφαλικού επεισοδίου με χρήση μεθόδων Μηχανικής Μάθησης

Διπλωματική Εργασία

Τζουρμανά Ελευθερία

Επιβλέπων : Γεώργιος Ματσόπουλος
Καθηγητής ΕΜΠ

Αθήνα, Φεβρουάριος 2025

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Διάγνωση εγκεφαλικού επεισοδίου με χρήση μεθόδων Μηχανικής Μάθησης

Διπλωματική Εργασία

Τζουρμανά Ελευθερία

Επιβλέπων : Γεώργιος Ματσόπουλος
Καθηγητής ΕΜΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 21η Φεβρουαρίου 2025.

Γ.Ματσόπουλος Καθηγητής ΕΜΠ	Π.Τσανάκας Καθηγητής ΕΜΠ	Αθ.Παναγόπουλος Καθηγητής ΕΜΠ

Αθήνα, Φεβρουάριος 2025

.....
Τζουρμανά Ελευθερία

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

Περίληψη

Τα εγκεφαλικά επεισόδια συνιστούν μια σοβαρή ιατρική κατάσταση που προκαλείται από τη διαταραχή της ροής του αίματος προς συγκεκριμένες περιοχές του εγκεφάλου, με αποτέλεσμα τη βλάβη στα εγκεφαλικά κύτταρα. Η έγκαιρη διάγνωση του εγκεφαλικού επεισοδίου είναι ζωτικής σημασίας, καθώς η καθυστέρηση στην ιατρική παρέμβαση αυξάνει σημαντικά τον κίνδυνο μη αναστρέψιμων εγκεφαλικών βλαβών.

Η παρούσα εργασία επικεντρώνεται στην εξερεύνηση του ρόλου των εξωκυτταρικών κυστιδίων (EVs) ως μη επεμβατικοί βιοδείκτες για την έγκαιρη διάγνωση και παρακολούθηση του εγκεφαλικού λόγω της δυνατότητάς τους να αντανακλούν την κυτταρική δραστηριότητα και να διαπερνούν τον αιματοεγκεφαλικό φραγμό.

Η μελέτη βασίζεται σε δεδομένα που συλλέχθηκαν από το Πανεπιστημιακό Ιατρικό Κέντρο του Άμστερνταμ, περιλαμβάνοντας κλινικές και δημογραφικές πληροφορίες των ασθενών καθώς χαρακτηριστικά των EVs. Η μεθοδολογία της εργασίας επικεντρώνεται σε δύο άξονες: την επιλογή χαρακτηριστικών για την αναγνώριση των κρίσιμων παραμέτρων του εγκεφαλικού και την εφαρμογή αλγορίθμων μηχανικής μάθησης Gradient Boosting, XGBoost, Random Forest, Naïve Bayes, SVM, KNN και Decision Trees) για την ταξινόμηση των ασθενών.

Στο πρώτο μέρος της ανάλυσης, εξετάζεται η διάγνωση του εγκεφαλικού επεισοδίου. Τα αποτελέσματα δείχνουν ότι ο βιοδείκτης CD235a-PE+ είναι ο πιο αποδοτικός, έχοντας επιλέξει τα εξής χαρακτηριστικά: Age, RR syst., Thrombocytes, wbc, glucose, Total, Total Concentration, Mean, Median, kurtosis και σημειώνοντας ακρίβεια 88%, recall 97%, precision 89%, f1 – score 93% και τιμή AUC 80%. Στο δεύτερο στάδιο, η διάκριση των τύπων εγκεφαλικού παρουσίασε ακρίβεια που κυμαίνεται από 40.5% έως 52.4%, με κοινά χαρακτηριστικά όπως Age, RR syst., Thrombocytes, wbc, και glucose.

Στο τρίτο μέρος της έρευνας, εφαρμόζεται η μέθοδος "Two-Stage Hybrid Data Classifiers" για διάγνωση σε δύο στάδια: το πρώτο στάδιο κατηγοριοποιεί τους ασθενείς ανάλογα αν έχουν υποστεί εγκεφαλικό ή όχι, ενώ το δεύτερο εστιάζει στην ανίχνευση του τύπου εγκεφαλικού. Εξετάζονται δύο προσεγγίσεις: η επιλογή χαρακτηριστικών μόνο στο πρώτο στάδιο ή η επιλογή χαρακτηριστικών και στα δύο στάδια για την βελτίωση της απόδοσης του μοντέλου. Στην πρώτη περίπτωση, ο βιοδείκτης CD235a-PE+ παρουσίασε την υψηλότερη ακρίβεια (73%), με τιμές Precision 57% και Recall 50%. Στη δεύτερη περίπτωση, για τον ίδιο βιοδείκτη παρατηρήθηκε η μεγαλύτερη ακρίβεια (59%). Τέλος στο 4^{ο} στάδιο, εξετάζεται μια εναλλακτική ταξινόμηση σε τρεις κατηγορίες (Control, Ischemic Stroke, Bleeding).. Η μετάβαση στη νέα ταξινόμηση έδειξε σημαντική βελτίωση στην ακρίβεια και στη διαγνωστική αξία των μοντέλων, με τον βιοδείκτη CD235a-PE+ να επιτυγχάνει ακρίβεια 73.8% και F1-score 61.1%, ενώ ο CD14-PB+ παρουσίασε το υψηλότερο AUC (73%) για τη διάκριση μεταξύ ισχαιμικού και αιμορραγικού εγκεφαλικού επεισοδίου.

Λέξεις κλειδιά:

Εγκεφαλικό επεισόδιο, εξωκυτταρικά κυστίδια (EVs), διάγνωση, επιλογή χαρακτηριστικών, μηχανική μάθηση

Abstract

Stroke is a serious medical condition caused by disruption of blood flow to specific areas of the brain, resulting in damage to brain cells. Early diagnosis of stroke is crucial, as delay in medical intervention significantly increases the risk of irreversible brain damage.

This study focuses on exploring the role of extracellular vesicles (EVs) as non-invasive biomarkers for the early diagnosis and monitoring of strokes, owing to their ability to reflect cellular activity and traverse the blood-brain barrier.

The study is based on data collected from the Amsterdam University Medical Center, including clinical and demographic information from patients, as well as characteristics of EVs.The methodology of this study is structured around two primary axes: the selection of features to identify critical biological and clinical parameters associated with stroke and the application of machine learning algorithms (Gradient Boosting, XGBoost, Random Forest, Naïve Bayes, SVM, KNN, and Decision Trees) for the classification of patients.

In the first stage of the analysis, stroke diagnosis is examined. The results indicate that the biomarker CD235a-PE+ is the most efficient, with selected features including Age, RR syst., Thrombocytes, wbc, glucose, Total, Total Concentration, Mean, Median, and kurtosis, achieving an accuracy of 88%, recall of 97%, precision of 89%, an F1-score of 93%, and an AUC of 80%. In the second stage, the differentiation of stroke types presented accuracy ranging from 40.5% to 52.4%, with common features such as Age, RR syst., Thrombocytes, wbc, and glucose.

In the third part of the study, the "Two-Stage Hybrid Data Classifiers" method is employed for two-stage diagnosis: the first stage categorizes patients based on whether they have experienced a stroke, while the second stage focuses on detecting the type of stroke. Two approaches are examined: feature selection only in the first stage, and feature selection in both stages for enhanced model performance. In the first approach, the biomarker CD235a-PE+ achieved the highest accuracy (73%), with Precision at 57% and Recall at 50%. In the second approach, the same biomarker showed the highest accuracy of 59%. Finally, in the fourth stage, an alternative classification into three categories (Control, Ischemic Stroke, Bleeding) was explored. The transition to this new classification demonstrated a significant improvement in accuracy and diagnostic value, with the biomarker CD235a-PE+ achieving an accuracy of 73.8% and an F1-score of 61.1%, while CD14-PB+ presented the highest AUC (73%) for distinguishing between ischemic and hemorrhagic strokes.

Keywords:

Stroke, extracellular vesicles (EVs), diagnosis, feature selection, machine learning

Ευχαριστίες

Σε αυτό το σημείο, θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στον καθηγητή μου, κ. Γεώργιο Ματσόπουλο, για την εμπιστοσύνη που μου έδειξε αναθέτοντάς μου αυτό το ιδιαίτερα ενδιαφέρον θέμα. Ευχαριστώ θερμά επίσης τον κ. Βασίλη Κατσιγιάννη για τις πολύτιμες συμβουλές και καθοδήγηση κατά τη διάρκεια της έρευνας και συγγραφής αυτής της διπλωματικής εργασίας.

Θα ήθελα, επιπλέον, να ευχαριστήσω τα μέλη της επιτροπής της διπλωματικής μου, τον Καθηγητή Παναγιώτη Τσανάκα και τον Καθηγητή Αθανάσιο Δ. Παναγόπουλο, για τις σημαντικές τους συνεισφορές.

Οι ειλικρινείς μου ευχαριστίες απευθύνονται στην οικογένειά μου και στους φίλους μου, οι οποίοι με στήριξαν αμέριστα καθ' όλη τη διάρκεια των σπουδών μου.

Πίνακας Περιεχομένων

Πίνακας Εικόνων & Πινάκων

1. Κεφάλαιο: Εισαγωγή

1.1 Αντικείμενο Διπλωματικής Εργασίας

Η παρούσα διπλωματική εργασία διερευνά τη δυνατότητα διάγνωσης του εγκεφαλικού επεισοδίου μέσω αλγορίθμων μηχανικής μάθησης, αξιοποιώντας εξωκυτταρικά κυστίδια (EVs) ως πιθανούς βιοδείκτες. Ο πρωταρχικός στόχος είναι η ανάπτυξη αξιόπιστων διαγνωστικών μοντέλων που θα επιτρέψουν την έγκαιρη και ακριβή ανίχνευση του εγκεφαλικού επεισοδίου, με σκοπό τη βελτίωση της διαγνωστικής διαδικασίας και τη διευκόλυνση έγκαιρων θεραπευτικών παρεμβάσεων. Τα EVs επιλέχθηκαν ως αντικείμενο μελέτης λόγω της ικανότητάς τους να διαπερνούν τον αιματοεγκεφαλικό φραγμό (BBB), γεγονός που τα καθιστά ιδανικούς υποψήφιους για μη επεμβατική διάγνωση μέσω αιματολογικών εξετάσεων.

Η μελέτη βασίζεται σε δεδομένα που συγκεντρώθηκαν στο πλαίσιο έρευνας του Πανεπιστημιακού Ιατρικού Κέντρου του Άμστερνταμ και περιλαμβάνουν κλινικές και δημογραφικές πληροφορίες των ασθενών, καθώς και μετρήσεις κυτταρομετρίας ροής σε δείγματα αίματος. Οι εν λόγω μετρήσεις αφορούν διάφορους βιοδείκτες, από τους οποίους απομονώθηκαν τα εξωκυτταρικά κυστίδια, τα οποία αποτέλεσαν τον κεντρικό άξονα της ανάλυσης. Στη συνέχεια, εξήχθησαν στατιστικά χαρακτηριστικά που σχετίζονται με το μέγεθος και τη συγκέντρωσή τους, τα οποία σε συνδυασμό με τα κλινικά και δημογραφικά στοιχεία των ασθενών χρησιμοποιήθηκαν ως σύνολο δεδομένων για την εκπαίδευση και την αξιολόγηση των μοντέλων μηχανικής μάθησης.

Η μεθοδολογία της παρούσας εργασίας εστιάζει σε δύο βασικούς άξονες. Ο πρώτος αφορά τη διαδικασία επιλογής χαρακτηριστικών, με στόχο την ταυτοποίηση των πλέον κρίσιμων βιολογικών και κλινικών παραμέτρων που σχετίζονται με το εγκεφαλικό επεισόδιο. Η αναγνώριση αυτών των χαρακτηριστικών μπορεί να συμβάλει στη βελτίωση της διαγνωστικής ακρίβειας και στην καλύτερη κατανόηση των υποκείμενων παθοφυσιολογικών μηχανισμών της νόσου. Ο δεύτερος άξονας επικεντρώνεται στην εφαρμογή και αξιολόγηση αλγορίθμων μηχανικής μάθησης, προκειμένου να προσδιοριστεί η αποτελεσματικότητα διαφόρων προσεγγίσεων στην ταξινόμηση των ασθενών. Οι αλγόριθμοι που εξετάζονται περιλαμβάνουν τα Gradient Boosting, XGBoost, Random Forest, Naïve Bayes, SVM, KNN και Decision Trees.

Η ανάλυση οργανώνεται σε τρία διακριτά στάδια: (i) αρχική διάγνωση του εγκεφαλικού επεισοδίου, (ii) ταξινόμηση σε τέσσερις διαγνωστικές κατηγορίες (Control, Vessel Occlusion, Bleeding, LVO) και (iii) εναλλακτική ταξινόμηση σε τρεις κατηγορίες (Control, Ischemic Stroke, Bleeding), η οποία εξετάζεται ως πιθανή βελτιστοποίηση της αρχικής ταξινόμησης, με στόχο την ενίσχυση της απόδοσης των μοντέλων.

Συνοψίζοντας, η παρούσα εργασία αποσκοπεί στη διερεύνηση της αξιοπιστίας των EVs ως βιοδεικτών για την έγκαιρη διάγνωση του εγκεφαλικού επεισοδίου, καθώς και στην αξιολόγηση της εφαρμοσιμότητας των αλγορίθμων μηχανικής μάθησης στη διαγνωστική διαδικασία. Μέσω αυτής της προσέγγισης, η μελέτη επιδιώκει να συμβάλει στην ανάπτυξη προηγμένων διαγνωστικών εργαλείων, ενισχύοντας τη δυνατότητα ταχείας και ακριβούς διάγνωσης, γεγονός που θα μπορούσε να έχει σημαντικό αντίκτυπο στη διαχείριση και αντιμετώπιση των ασθενών με εγκεφαλικό επεισόδιο.

1.2 Βιβλιογραφική Ανασκόπηση

Τα τελευταία χρόνια, η χρήση μηχανικής μάθησης έχει αναδειχθεί ως μια καινοτόμος προσέγγιση στη διαγνωστική διαδικασία, ενώ τα εξωκυτταρικά κυστίδια (EVs) έχουν προσελκύσει επιστημονικό ενδιαφέρον ως πιθανοί βιοδείκτες. Η παρούσα βιβλιογραφική ανασκόπηση εξετάζει προηγούμενες μελέτες που εστιάζουν στη διάγνωση του εγκεφαλικού επεισοδίου μέσω μηχανικής μάθησης, είτε με τη χρήση EVs είτε με άλλους αιματολογικούς ή μοριακούς βιοδείκτες, αναδεικνύοντας τις τεχνικές, τους αλγόριθμους και τα ερευνητικά ευρήματα που συμβάλλουν στη βελτίωση της διαγνωστικής ακρίβειας και στην πιθανή κλινική εφαρμογή τους.

Η μελέτη των Sherif & Ahmed (2024) [1] διερεύνησε τη χρήση μοντέλων μηχανικής μάθησης για τη διαφορική διάγνωση του εγκεφαλικού επεισοδίου με βάση αιματολογικούς βιοδείκτες. Οι ερευνητές χρησιμοποίησαν δεδομένα ασθενών για την ανάπτυξη ενός ταξινομητή που διακρίνει το ισχαιμικό από το αιμορραγικό εγκεφαλικό επεισόδιο, εφαρμόζοντας αλγορίθμους όπως Support Vector Machines (SVM), Adaptive Neuro-Fuzzy Inference System (ANFIS), K-Nearest Neighbor (KNN) και Decision Tree (DT). Τα αποτελέσματα ανέδειξαν τον SVM ως τον πιο αποδοτικό αλγόριθμο, με ακρίβεια 99,1%. Η μελέτη αυτή επιβεβαίωσε ότι συγκεκριμένοι αιματολογικοί δείκτες, όπως τα Absolute Neutrophil Count (ANC), Creatine Phosphokinase (CPK), Neutro/Neutrophils και White Blood Cell (WBC) Count, μπορούν να διαδραματίσουν κρίσιμο ρόλο στη διάγνωση του εγκεφαλικού επεισοδίου, προσφέροντας μια μη επεμβατική εναλλακτική διάγνωσης.

Παρόμοια, η έρευνα "Machine Learning Analysis of MicroRNA Expression Data" [2] εξερεύνησε τη χρήση microRNA (miRNAs) ως διαγνωστικών βιοδεικτών για το ισχαιμικό εγκεφαλικό επεισόδιο. Αναλύοντας δείγματα από 50 ασθενείς με εγκεφαλικό και 50 υγιείς, οι ερευνητές κατέληξαν σε τέσσερα miRNAs (miR-19a, miR-148a, miR-320d, miR-342-3p), των οποίων η έκφραση παρουσίασε διακριτές μεταβολές μεταξύ των δύο ομάδων. Τα καλύτερα αποτελέσματα προέκυψαν από το μοντέλο SVM, το οποίο με συνδυασμό miR-148a, miR-342-3p και miR-19a πέτυχε AUC=0.958.

Μια ακόμα σημαντική μελέτη, με τίτλο "New strategy for clinical etiologic diagnosis of acute ischemic stroke and blood biomarker discovery based on machine learning" [3], εστίασε στην αιτιολογική διάγνωση του οξέος ισχαιμικού εγκεφαλικού επεισοδίου (AIS) μέσω ανάλυσης αιματολογικών παραμέτρων. Εξετάστηκαν 64 κλινικά και αιματολογικά χαρακτηριστικά σε 456 ασθενείς με AIS και 28 υγιείς μάρτυρες, με στόχο την ταξινόμηση υποτύπων του ισχαιμικού εγκεφαλικού, όπως η αθηροσκλήρωση μεγάλων αρτηριών (LAA), το καρδιοεμβολικό (CE) και η απόφραξη μικρών αγγείων (SVO). Οι αλγόριθμοι μηχανικής μάθησης πέτυχαν ακρίβεια άνω του 73%, υποδεικνύοντας ότι η ανάλυση αιματολογικών βιοδεικτών μπορεί να προσφέρει μια αξιόπιστη εναλλακτική στην κλινική διάγνωση.

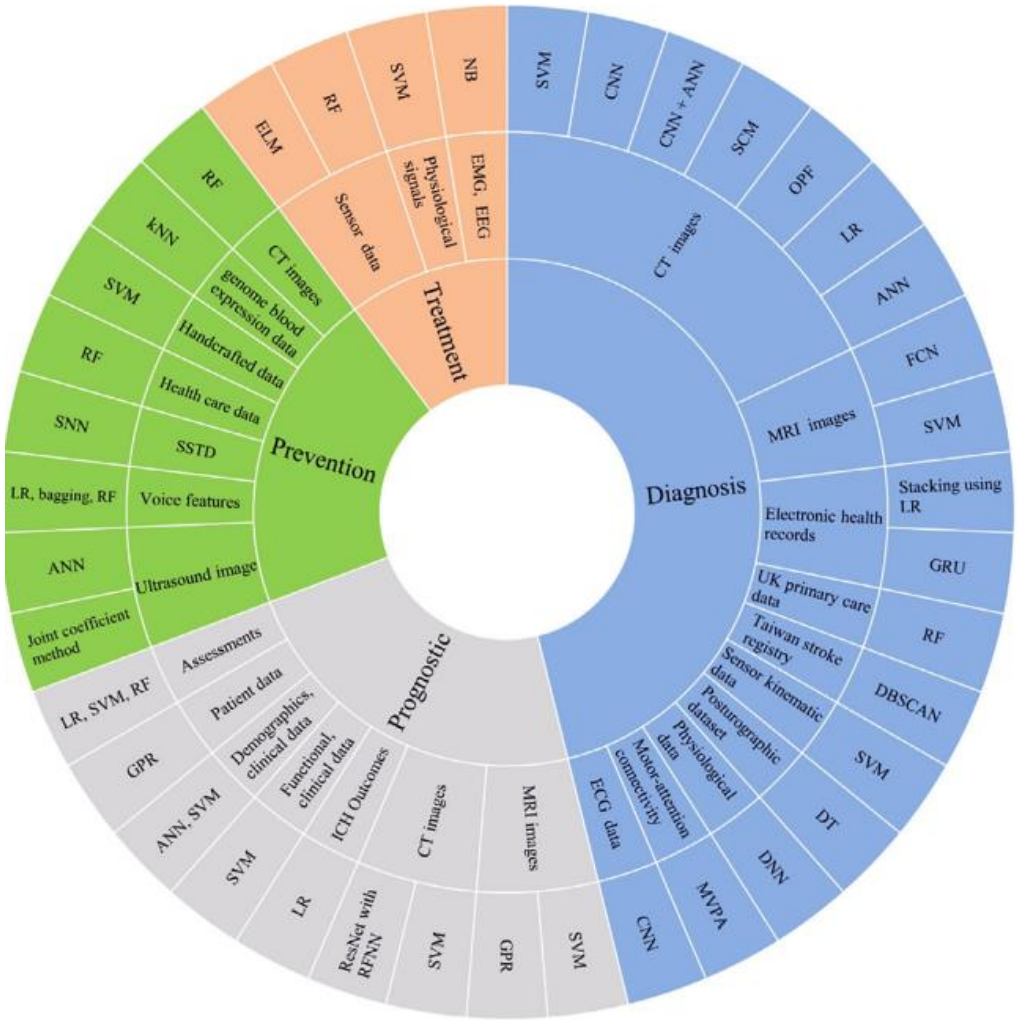
Η Εικόνα 1 παρέχει μια συνοπτική παρουσίαση των εφαρμογών της μηχανικής μάθησης στη διαχείριση του εγκεφαλικού επεισοδίου, καταγράφοντας τη χρήση διαφόρων αλγορίθμων σε τέσσερις κύριους τομείς: διάγνωση, πρόγνωση, πρόληψη και θεραπεία. Η απεικόνιση βασίζεται στην ανάλυση 39 ερευνητικών μελετών που καλύπτουν την περίοδο 2007-2019 [3]. Όπως προκύπτει από τα δεδομένα, το μεγαλύτερο ποσοστό των ερευνών που αφορά την διάγνωση εστιάζει σε απεικονιστικές μεθόδους (CT, MRI), ηλεκτρονικά ιατρικά αρχεία και φυσιολογικά δεδομένα, με αλγορίθμους όπως Convolutional Neural Networks (CNN), Support Vector Machines (SVM) και Random Forest (RF) να επιτυγχάνουν υψηλά επίπεδα διαγνωστικής ακρίβειας.

Παράλληλα, η χρήση των EVs ως βιοδεικτών για τη διάγνωση νευρολογικών παθήσεων, συμπεριλαμβανομένου του εγκεφαλικού επεισοδίου, έχει αρχίσει να αναδεικνύεται ως πολλά υποσχόμενη προσέγγιση. Σύμφωνα με τη μελέτη των Bang et al. (2024) [4], η ανάλυση EV-miRNAs επέτρεψε την ταξινόμηση υποτύπων του ισχαιμικού εγκεφαλικού επεισοδίου μέσω next-generation sequencing (NGS), επιτυγχάνοντας ακρίβεια 92%. Η έρευνα αυτή έδειξε ότι ο συνδυασμός πολλαπλών microRNAs στα EVs παρέχει πιο αξιόπιστες διαγνωστικές πληροφορίες σε σύγκριση με τη χρήση μεμονωμένων βιοδεικτών. Αντίστοιχα, άλλες μελέτες έχουν διερευνήσει τη χρήση μηχανικής μάθησης για την ταξινόμηση δεδομένων

EVs, εφαρμόζοντας αλγορίθμους όπως Support Vector Machines (SVM), Random Forest (RF) και XGBoost, επιτυγχάνοντας υψηλή ακρίβεια στη διαφοροποίηση μεταξύ ασθενών και υγιών ατόμων [5]. Επιπλέον, υπάρχουν μελέτες που επικεντρώνονται στη χρήση EVs για την ταξινόμηση μιας μόνο κατηγορίας εγκεφαλικού, όπως η διάκριση του LVO (Large Vessel Occlusion) [6] ή του ισχαιμικού εγκεφαλικού [7]. Οι έρευνες αυτές καταδεικνύουν ότι η μηχανική μάθηση μπορεί να αξιοποιηθεί για την ανάλυση των EVs και την εξαγωγή προγνωστικών μοντέλων, βελτιώνοντας την ακρίβεια της διάγνωσης.

Η παρούσα εργασία υιοθετεί μια διαφορετική προσέγγιση, εξετάζοντας τη χρήση εξωκυτταρικών κυστιδίων (EVs) ως βιοδεικτών σε συνδυασμό με μεθόδους μηχανικής μάθησης. Σε αντίθεση με προηγούμενες μελέτες που εστιάζουν μόνο σε μια κατηγορία εγκεφαλικού επεισοδίου, η παρούσα έρευνα στοχεύει στη διάγνωση τόσο του εγκεφαλικού επεισοδίου όσο και στη διαστρωμάτωσή του σε διαφορετικές κατηγορίες. Ειδικότερα, η μελέτη οργανώνεται σε τρία στάδια ταξινόμησης: 1) αρχική διάγνωση του εγκεφαλικού επεισοδίου, 2) διάκριση σε τέσσερις διαγνωστικές κατηγορίες (Control, Vessel Occlusion, Bleeding, LVO) και 3) εναλλακτική ταξινόμηση σε τρεις κατηγορίες (Control, Ischemic Stroke, Bleeding) για τη βελτιστοποίηση της διαγνωστικής ακρίβειας.

Συμπερασματικά, η χρήση EVs σε συνδυασμό με μηχανική μάθηση αποτελεί μια καινοτόμο εναλλακτική στις παραδοσιακές διαγνωστικές μεθόδους, προσφέροντας προοπτικές για μη επεμβατική, έγκαιρη και ακριβή διάγνωση του εγκεφαλικού επεισοδίου. Η συμβολή της παρούσας μελέτης έγκειται στη συστηματική αξιολόγηση της αποτελεσματικότητας των EVs ως διαγνωστικών βιοδεικτών και στη διερεύνηση του βέλτιστου συνδυασμού αλγορίθμων μηχανικής μάθησης για τη βελτίωση της ακρίβειας των διαγνωστικών μοντέλων.



Εικόνα 1.1: Εφαρμογές μηχανικής μάθησης στη διαχείριση του εγκεφαλικού επεισοδίου με βάση 39 μελέτες που καλύπτουν την περίοδο 2007-2019 [3]

1.3 Δομή Διπλωματικής Εργασίας

Η παρούσα διπλωματική εργασία αποτελείται από έξι κεφάλαια, τα οποία έχουν διαρθρωθεί με σκοπό την παροχή μιας ολοκληρωμένης ανάλυσης σχετικά με τη διάγνωση του εγκεφαλικού επεισοδίου μέσω εξωκυτταρικών κυστιδίων (EVs) και αλγορίθμων μηχανικής μάθησης.

Το πρώτο κεφάλαιο εισάγει το αντικείμενο της μελέτης, αναλύοντας τη σημασία της έγκαιρης διάγνωσης του εγκεφαλικού επεισοδίου. Επιπλέον, περιλαμβάνει μια βιβλιογραφική ανασκόπηση προηγούμενων ερευνών που έχουν αξιοποιήσει EVs και τεχνικές τεχνητής νοημοσύνης στη διάγνωση του εγκεφαλικού, καθώς και τη συνοπτική παρουσίαση της δομής της εργασίας.

Το δεύτερο κεφάλαιο επικεντρώνεται στη Μηχανική Μάθηση, περιγράφοντας τις θεμελιώδεις έννοιες και τις μεθόδους που χρησιμοποιούνται στην παρούσα μελέτη. Αρχικά, γίνεται εισαγωγή στις διαφορετικές προσεγγίσεις εκπαίδευσης των μοντέλων μηχανικής μάθησης, ενώ στη συνέχεια αναλύονται οι αλγόριθμοι που εφαρμόστηκαν, συμπεριλαμβανομένων των Gradient Boosting, XGBoost, Random Forest, Naïve Bayes, SVM, KNN και Decision Trees. Επιπλέον, αναφέρονται οι διαδικασίες επιλογής χαρακτηριστικών και αναλύεται η μέθοδος που επιλέχθηκε για τη συγκεκριμένη μελέτη, η Recursive Feature Elimination (RFE) καθώς και οι τεχνικές που εφαρμόστηκαν για την αντιμετώπιση των προκλήσεων.

Το τρίτο κεφάλαιο εστιάζει στη διάγνωση του εγκεφαλικού επεισοδίου. Παρουσιάζονται οι βασικοί τύποι εγκεφαλικών επεισοδίων, οι αιτίες και οι παράγοντες κινδύνου, καθώς και οι διαγνωστικές μέθοδοι που χρησιμοποιούνται στην κλινική πρακτική. Ιδιαίτερη έμφαση δίνεται στη σημασία της έγκαιρης διάγνωσης και στη χρήση των εξωκυτταρικών κυστιδίων ως βιοδεικτών, χάρη στην ικανότητά τους να διαπερνούν τον αιματοεγκεφαλικό φραγμό, καθιστώντας τα ιδανικούς υποψηφίους για μη επεμβατική διάγνωση μέσω αιματολογικών εξετάσεων. Το κεφάλαιο ολοκληρώνεται με την παρουσίαση του τρόπου με τον οποίο οι μέθοδοι μηχανικής μάθησης, σε συνδυασμό με τους επιλεγμένους βιοδείκτες, μπορούν να αξιοποιηθούν στη διάγνωση του εγκεφαλικού.

Το τέταρτο κεφάλαιο περιλαμβάνει την παρουσ των αποτελεσμάτων, η οποία διαρθρώνεται σε τρία μέρη. Το πρώτο μέρος αφορά τη διάγνωση του εγκεφαλικού επεισοδίου, παρουσιάζοντας τα αποτελέσματα της επιλογής χαρακτηριστικών και την απόδοση των μοντέλων μηχανικής μάθησης. Το δεύτερο μέρος εστιάζει στη διάκριση του τύπου του εγκεφαλικού επεισοδίου, εξετάζοντας την ικανότητα των μοντέλων να ταξινομήσουν τους ασθενείς σε τέσσερις κατηγορίες (Control, Vessel Occlusion, Bleeding και LVO). Στο τρίτο μέρος, προτείνεται μια εναλλακτική προσέγγιση ταξινόμησης, όπου οι κατηγορίες Vessel Occlusion και LVO συγχωνεύονται σε μία ενιαία κλάση (Ischemic Stroke), επιτρέποντας μια πιο αποτελεσματική διαφοροποίηση μεταξύ Ισχαιμικού και Αιμορραγικού εγκεφαλικού επεισοδίου.

Το πέμπτο κεφάλαιο περιλαμβάνει την παρουσίαση των βασικών συμπερασμάτων, τη συγκριτική ανάλυση των αποτελεσμάτων της εργασίας σε σχέση με προηγούμενες μελέτες καθώς και τη διερεύνηση των περιορισμών που ενδέχεται να επηρέασαν την απόδοση των μοντέλων. Παράλληλα, διατυπώνονται προτάσεις για μελλοντικές ερευνητικές κατευθύνσεις, οι οποίες δύνανται να ενισχύσουν την ακρίβεια και την αξιοπιστία της διάγνωσης του εγκεφαλικού επεισοδίου με τη χρήση μεθόδων μηχανικής μάθησης.

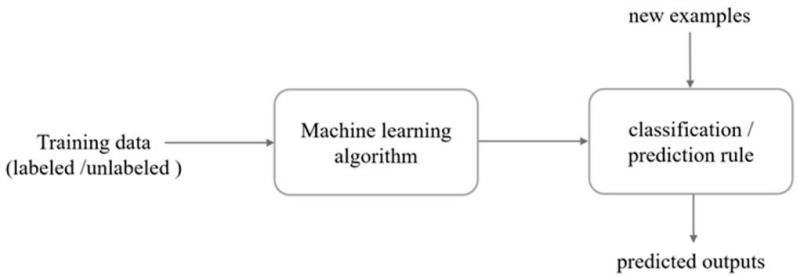
Τέλος, το έκτο κεφάλαιο περιλαμβάνει τη βιβλιογραφία που χρησιμοποιήθηκε στην παρούσα μελέτη, παρέχοντας τις απαραίτητες αναφορές στις επιστημονικές εργασίες που αποτέλεσαν τη βάση της ερευνητικής διαδικασίας.

2. Κεφάλαιο: Μηχανική Μάθηση

2.1 Εισαγωγή στην Μηχανική Μάθηση

Είναι γεγονός ότι η Τεχνητή Νοημοσύνη και η Μηχανική Μάθηση εισχωρούν ολοένα και περισσότερο στην καθημερινή ζωή, βρίσκοντας εφαρμογές σε πολλούς τομείς και αποτελώντας αναπόσπαστο μέρος κάθε επιστημονικού κλάδου. Το 1959, ο Arthur Samuel, ο επιστήμονας που επινόησε τον όρο «Μηχανική Μάθηση» την περιέγραψε ως τον «κλάδο που δίνει στους υπολογιστές την ικανότητα να μαθαίνουν χωρίς να προγραμματίζονται ρητά». Ο ίδιος κατέληξε ότι το να προγραμματίζονται οι υπολογιστές να μαθαίνουν από την εμπειρία θα εξαλείψει τελικά την ανάγκη για τόσο λεπτομερή προγραμματισμό. Ο Tom M. Mitchell πρότεινε έναν πιο επίσημο ορισμό που χρησιμοποιείται ευρέως: «Ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από την εμπειρία E σχετικά με μια κατηγορία εργασιών T και μετρητή απόδοσης P, αν η απόδοσή του στις εργασίες T, όπως μετράται από το P, βελτιώνεται με την εμπειρία E» [8].

Οι μεθοδολογίες μηχανικής μάθησης περιλαμβάνουν συνήθως μια διαδικασία εκμάθησης με στόχο την απόκτηση γνώσης από την «εμπειρία» (εκπαιδευτικά δεδομένα) για την εκτέλεση μιας συγκεκριμένης εργασίας. Στην πράξη, η «εμπειρία» μεταφράζεται στη χρήση δεδομένων [9]. Τα δεδομένα στη μηχανική μάθηση αποτελούνται από ένα σύνολο παραδειγμάτων, όπου κάθε παράδειγμα περιγράφεται από ένα σύνολο χαρακτηριστικών ή μεταβλητών. Τα χαρακτηριστικά μπορεί να είναι ονομαστικά (π.χ., καταχωρίσεις), δυαδικά (0 ή 1), ταξινομικά (π.χ., A+ ή B−) ή αριθμητικά (π.χ., ακέραιοι ή πραγματικοί αριθμοί). Η απόδοση ενός μοντέλου μηχανικής μάθησης σε μια συγκεκριμένη εργασία μετριέται μέσω ενός δείκτη απόδοσης, ο οποίος βελτιώνεται με την εμπειρία κατά τη διάρκεια του χρόνου. Μετά την ολοκλήρωση της διαδικασίας εκμάθησης, το εκπαιδευμένο μοντέλο μπορεί να χρησιμοποιηθεί για ταξινόμηση, πρόβλεψη ή ομαδοποίηση νέων παραδειγμάτων (δοκιμαστικά δεδομένα) αξιοποιώντας την εμπειρία που αποκτήθηκε κατά τη φάση εκπαίδευσης.



Εικόνα 2.1: Διαγραμματική απεικόνιση διαδικασίας Μηχανικής Μάθησης [10]

Οι εργασίες μηχανικής μάθησης (ML) ταξινομούνται συνήθως σε διάφορες ευρείες κατηγορίες, ανάλογα με τον τύπο εκμάθησης (επιβλεπόμενη ή μη επιβλεπόμενη), τα μοντέλα εκμάθησης (όπως η ταξινόμηση, η παλινδρόμηση, η ομαδοποίηση και η μείωση διαστάσεων) ή τα μοντέλα που χρησιμοποιούνται για την υλοποίηση της επιλεγμένης εργασίας [10].

Η Μηχανική Μάθηση βασίζεται σε διάφορους αλγόριθμους για την επίλυση προβλημάτων που σχετίζονται με δεδομένα. Η επιλογή του κατάλληλου αλγορίθμου εξαρτάται από το είδος του προβλήματος που χρειάζεται να επιλυθεί, τον αριθμό των μεταβλητών, τον τύπο του μοντέλου που ταιριάζει καλύτερα, τα επιθυμητά αποτελέσματα καθώς και τους διαθέσιμους υπολογιστικούς πόρους.

2.1.1 Χρήση Μηχανικής Μάθησης σε ιατρικά δεδομένα

Η τεχνητή νοημοσύνη και η Μηχανική Μάθηση χρησιμοποιούνται ευρέως για την ανάλυση ιατρικών δεδομένων με στόχο την ανάπτυξη αποτελεσματικών διαγνωστικών εργαλείων. Με την επεξεργασία μεγάλου όγκου δεδομένων, τα μοντέλα μηχανικής μάθησης μπορούν να αναγνωρίσουν μοτίβα σε μεγάλες βάσεις δεδομένων και να κάνουν ακριβείς προβλέψεις, βοηθώντας τους ιατρούς να εντοπίσουν πρώιμα σημάδια ασθενειών όπως οι καρδιαγγειακές παθήσεις και το εγκεφαλικό. Η ανάλυση ιατρικών δεδομένων μέσω Μηχανικής Μάθησης περιλαμβάνει τεχνικές όπως η εξόρυξη δεδομένων (data mining), οι αλγόριθμοι ταξινόμησης (classification algorithms) και οι μοντελοποιήσεις πρόβλεψης (predictive modeling). Τα μοντέλα αυτά χρησιμοποιούν κλινικά δεδομένα, όπως ηλικία, αρτηριακή πίεση, επίπεδα χοληστερόλης, οικογενειακό ιστορικό, διαβήτη και συνήθειες του ασθενούς (π.χ. κάπνισμα, διατροφή), προκειμένου να προβλέψουν την πιθανότητα εμφάνισης εγκεφαλικού επεισοδίου. Επιπλέον, προηγμένες τεχνικές όπως τα Generative Adversarial Networks (GANs) και οι αλγόριθμοι αναγνώρισης εικόνας (image recognition) χρησιμοποιούνται για την αυτόματη ανάλυση απεικονιστικών εξετάσεων (π.χ. MRI, CT scans) και την ανίχνευση εγκεφαλικών βλαβών με μεγαλύτερη ακρίβεια.

Η αυξανόμενη διαθεσιμότητα δεδομένων στον τομέα της υγείας επιτρέπει την ανάπτυξη πιο ακριβών και αποδοτικών διαγνωστικών εργαλείων. Ωστόσο, η προεπεξεργασία των δεδομένων είναι απαραίτητη για τη βελτίωση της ακρίβειας και τη μείωση του χρόνου εκτέλεσης των αλγορίθμων. Η εξέλιξη των υπολογιστικών δυνατοτήτων και η ανάπτυξη νέων μοντέλων μηχανικής μάθησης δημιουργούν νέες ευκαιρίες στον τομέα της ιατρικής, βελτιώνοντας την έγκαιρη διάγνωση και αυξάνοντας τα ποσοστά επιβίωσης. Μελλοντικές έρευνες επικεντρώνονται στη βελτιστοποίηση των αλγορίθμων και στην εφαρμογή τους στην καθημερινή κλινική πρακτική, με στόχο την πρόληψη και την καλύτερη διαχείριση των εγκεφαλικών επεισοδίων και άλλων σοβαρών παθήσεων [11].

2.2 Εκπαίδευση ML

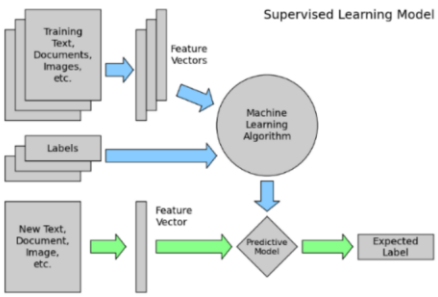
Η διαδικασία εκπαίδευσης αποτελεί θεμελιώδες βήμα για τη δημιουργία μοντέλων Μηχανικής Μάθησης, καθώς καθορίζει την ικανότητά τους να μαθαίνουν από δεδομένα και να αποδίδουν στις αντίστοιχες εργασίες. Ανάλογα με τη φύση των δεδομένων και των στόχων, η εκπαίδευση μπορεί να πραγματοποιηθεί με επιβλεπόμενη ή μη επιβλεπόμενη μέθοδο. Η ειδοποιός διαφορά μεταξύ των δύο βρίσκεται στον τρόπο που εκπαιδεύεται ένας αλγόριθμος μηχανικής μάθησης. Η επιβλεπόμενη μάθηση αξιοποιεί δεδομένα που περιέχουν γνωστές ετικέτες (labels) για να προβλέψει τα επιθυμητά αποτελέσματα, ενώ η μη επιβλεπόμενη μάθηση εστιάζει στην ανακάλυψη υποκείμενων προτύπων ή σχέσεων σε δεδομένα χωρίς ετικέτες. Στις επόμενες ενότητες, περιγράφονται λεπτομερώς τα χαρακτηριστικά, οι μέθοδοι και οι εφαρμογές των δυο αυτών προσεγγίσεων.

2.2.1 Επιβλεπόμενη εκπαίδευση

Η επιβλεπόμενη μάθηση (supervised learning) αποτελεί την πιο κοινή τεχνική σε προβλήματα ταξινόμησης και παλινδρόμησης. Ο στόχος της είναι η δημιουργία ενός συστήματος πρόβλεψης, όπου τα δεδομένα εισόδου συνοδεύονται από γνωστές ετικέτες (labels). Η διαδικασία μάθησης ακολουθεί 2 βήματα: την εκπαίδευση (training) και την δοκιμή (test). Κατά τη διαδικασία εκπαίδευσης, το μοντέλο λαμβάνει ως είσοδο ένα σύνολο δεδομένων που περιλαμβάνει χαρακτηριστικά (features) και τις αντίστοιχες σωστές εξόδους. Το μοντέλο μαθαίνει συγκρίνοντας τις προβλέψεις του με τις πραγματικές ετικέτες και εντοπίζοντας τα σφάλματα. Στη συνέχεια, τροποποιεί τις παραμέτρους του ώστε να βελτιώσει την ακρίβειά του. Μετά την εκπαίδευση, το μοντέλο μπορεί να χρησιμοποιηθεί για την πρόβλεψη ετικετών σε νέα, άγνωστα δεδομένα, παρέχοντας αποτελέσματα που βασίζονται στην εμπειρία που αποκτήθηκε από τα δεδομένα εκπαίδευσης.

Οι διάφοροι αλγόριθμοι της επιβλεπόμενης μάθησης δημιουργούν μια συνάρτηση που αντιστοιχίζει εισόδους σε επιθυμητές εξόδους. Με άλλα λόγια, κάθε αλγόριθμος χρησιμοποιεί τις ήδη υπάρχουσες ετικέτες για να ερμηνεύσει την σχέση που υπάρχει μεταξύ των δεδομένων εισόδου και εξόδου. Η επιβλεπόμενη μάθηση είναι η πιο κοινή τεχνική για προβλήματα ταξινόμησης, καθώς ο στόχος είναι να «μάθει» το σύστημα ταξινόμησης που έχει ήδη δημιουργηθεί από τον χρήστη. Το σύστημα αυτό συνήθως λαμβάνει ένα σύνολο χαρακτηριστικών (features) και ετικετών (labels), με τον κύριο στόχο να κατασκευάσει έναν εκτιμητή που μπορεί να προβλέψει την ετικέτα ενός αντικειμένου με βάση τα χαρακτηριστικά του. Ο αλγόριθμος συγκρίνει την πραγματική έξοδο με την αναμενόμενη, εντοπίζει τα σφάλματα και τροποποιεί το μοντέλο ανάλογα [12].

Όπως αναφέρθηκε προηγουμένως, οι εργασίες επιβλεπόμενης μάθησης χωρίζονται σε δύο κατηγορίες: την ταξινόμηση και την παλινδρόμηση. Στο παρακάτω σχήμα, ο αλγόριθμος επεξεργάζεται δεδομένα εκπαίδευσης (X), τα οποία χρησιμοποιούνται για την εκπαίδευση του μοντέλου πρόβλεψης. Μετά την εκπαίδευση, το μοντέλο εφαρμόζεται σε νέα δεδομένα για να προβλέψει πιθανές ετικέτες (y) στο σύνολο δοκιμής. Αν η ετικέτα (y) περιλαμβάνει διακριτές κατηγορίες, τότε η εργασία ονομάζεται ταξινόμηση. Αντίθετα, αν η ετικέτα (y) περιλαμβάνει αριθμητικές τιμές, τότε πρόκειται για εργασία παλινδρόμησης.



Εικόνα 2.2: Διάγραμμα ροής για επιβλεπόμενη μάθηση [12]

Οι πιο δημοφιλείς μέθοδοι που χρησιμοποιούνται στην επιβλεπόμενη μάθηση κατά την βιβλιογραφία είναι: η γραμμική παλινδρόμηση (Linear regression), τα δένδρα απόφασης (Decision trees), οι Μηχανές Διανυσμάτων Υποστήριξης (Support vector machine-SVM), οι Κ-Κοντινότερων Γειτόνων (k-Nearest Neighbors), οι Ταξινομητές Naïve Bayes (Naive Bayes classifiers), Τυχαία Δάση (Random forest), τα Νευρωνικά δίκτυα (Neural Networks) καθώς και το Πολυεπίπεδο Περσέπτρον (Multilayer Perceptron-MLP).

2.2.2 Μη επιβλεπόμενη μάθηση

Η μη επιβλεπόμενη μάθηση είναι ένας από τους κύριους τομείς της Μηχανικής Μάθησης, που επικεντρώνεται στη μοντελοποίηση δεδομένων όπου δεν υπάρχουν διαθέσιμες ετικέτες ή κατηγοριοποιήσεις για τις εισόδους. Αυτό σημαίνει ότι ο αλγόριθμος λαμβάνει μόνο τα χαρακτηριστικά των δεδομένων και προσπαθεί να εντοπίσει πρότυπα, δομές ή σχέσεις μέσα στο σύνολο των δεδομένων χωρίς καθοδήγηση. Ο αλγόριθμος που θα επεξεργαστεί τα δεδομένα εξετάζει τα δεδομένα εισόδου προκειμένου να βρει τις ομοιότητες μεταξύ τους και ανάλογα με τις μεταξύ τους ομοιότητες να τα κατατάξει σε ομάδες. Με αυτήν την μέθοδο μπορούμε να κατηγοριοποιήσουμε δεδομένα που δεν έχουν ετικέτες (unlabeled). Η πιο διαδεδομένη τεχνική μη επιβλεπόμενης μάθησης είναι η ομαδοποίηση (clustering). Άλλες γνωστές τεχνικές είναι η ανίχνευση ανωμαλιών (anomaly detection), τα νευρωνικά δίκτυα (neural networks), οι κανόνες συσχετίσεων (association rules), η μείωση διάστασης (dimensionality reduction) και οι χάρτες αυτό-οργάνωσης (self-organizing maps) [13, 14].

2.2.3 Ενισχυτική μάθηση

Η ενισχυτική μάθηση (Reinforcement Learning - RL) είναι ένας τύπος μηχανικής μάθησης κατά τον οποίο ένας παράγοντας (agent) μαθαίνει να αλληλεπιδρά με ένα δυναμικό περιβάλλον μέσω μιας διαδικασίας δοκιμής και σφάλματος (trial-and-error), λαμβάνοντας ανατροφοδότηση με τη μορφή ανταμοιβών ή ποινών. Ο κύριος στόχος της ενισχυτικής μάθησης είναι να μεγιστοποιήσει τη συνολική μακροπρόθεσμη ανταμοιβή, διαμορφώνοντας μια πολιτική (policy) που καθορίζει ποιες ενέργειες πρέπει να επιλέγονται σε κάθε κατάσταση.

Η RL διαφέρει από την επιβλεπόμενη μάθηση καθώς δεν παρέχεται στον παράγοντα ένα σύνολο δεδομένων εισόδου-εξόδου για εκπαίδευση. Αντίθετα, ο παράγοντας πρέπει να ανακαλύψει αυτόνομα τις βέλτιστες ενέργειες, αλληλεπιδρώντας με το περιβάλλον του και προσαρμόζοντας τη στρατηγική του βάσει των ανταμοιβών που λαμβάνει. Η διαδικασία αυτή συνοδεύεται από δύο βασικές προκλήσεις. Πρώτον, το δίλημμα της εξερεύνησης έναντι της εκμετάλλευσης (exploration vs. exploitation), όπου ο παράγοντας πρέπει να εξισορροπήσει την αναζήτηση νέων στρατηγικών με την αξιοποίηση των ήδη γνωστών ενεργειών που οδηγούν σε υψηλές ανταμοιβές. Δεύτερον, η μάθηση από καθυστερημένες ανταμοιβές (delayed reward learning), καθώς οι συνέπειες μιας δράσης μπορεί να γίνουν εμφανείς σε μεταγενέστερο χρόνο, καθιστώντας δύσκολη τη συσχέτιση μεταξύ της ενέργειας και του τελικού αποτελέσματος.

Η ενισχυτική μάθηση βασίζεται σε μαθηματικά μοντέλα όπως οι Μαρκοβιανές Διαδικασίες Απόφασης (Markov Decision Processes), ενώ συχνά χρησιμοποιεί αλγορίθμους όπως Q-learning, Deep Q-Networks (DQN) και Policy Gradient Methods. Έχει εφαρμογές σε τομείς όπως η ρομποτική, τα παιχνίδια, η βελτιστοποίηση διαδικασιών και η αυτόνομη οδήγηση [15].

2.3 Αλγόριθμοι Μηχανικής Μάθησης

Στην παρούσα ενότητα παρουσιάζονται μερικοί από τους πιο δημοφιλείς αλγόριθμους που χρησιμοποιούνται στη Μηχανική Μάθηση. Η επιλογή του κατάλληλου αλγορίθμου εξαρτάται από τη φύση του προβλήματος, τη διαθέσιμη ποσότητα και ποιότητα των δεδομένων, καθώς και τον στόχο της εκάστοτε μελέτης. Οι αλγόριθμοι που παρουσιάζονται παρακάτω αποτέλεσαν αντικείμενο διερεύνησης και εφαρμογής στην παρούσα εργασία.

2.3.1 Gradient Boosting

Το μοντέλο Gradient Boosting είναι ένα ισχυρό μοντέλο μηχανικής μάθησης που μπορεί να εφαρμοστεί σε αρκετά προβλήματα. Ανήκει στην οικογένεια των «μεθόδων boosting [9]» και είναι βασισμένο στην κλίση (gradient descent). Συνδυάζει δύο βασικές τεχνικές: την ενίσχυση (boosting), όπου διαδοχικά αδύναμα μοντέλα συνδυάζονται για να δημιουργήσουν ένα ισχυρότερο μοντέλο, και τη χρήση της κλίσης (gradient) για τη βελτιστοποίηση της συνάρτησης απώλειας, μειώνοντας σταδιακά το σφάλμα. Αυτός ο συνδυασμός αποτελεί τη βάση της λειτουργίας και της ονομασίας του μοντέλου.

Συγκεκριμένα, η βασική ιδέα πίσω από αυτόν τον αλγόριθμο είναι να προσθέτει νέα μοντέλα (ή «βάσεις») με την πάροδο του χρόνου. Κάθε νέο μοντέλο εκπαιδεύεται για να διορθώσει τα σφάλματα που έγιναν από τα προηγούμενα μοντέλα του συνόλου. Ο Gradient Boosting χρησιμοποιεί «αδύναμους μαθητές» (weak learners), δηλαδή μοντέλα που είναι σχετικά απλά και έχουν μικρή ακρίβεια πρόβλεψης από μόνα τους προκειμένου να δημιουργήσουν έναν «ισχυρό μαθητή». Αυτά τα μοντέλα είναι αρκετά για να κάνουν μικρές διορθώσεις κάθε φορά.

Κάθε νέο μοντέλο εκπαιδεύεται προσπαθώντας να προσαρμοστεί στην «αρνητική κλίση» της συνάρτησης απώλειας. Στην περίπτωση ενός Gradient Boosting Classifier, η συνάρτηση απώλειας συνήθως είναι μια logarithmic loss function (ή log loss), η οποία μετρά την απόκλιση μεταξύ των προβλέψεων του μοντέλου και των πραγματικών κλάσεων. Σε κάθε επανάληψη, το νέο μοντέλο εκπαιδεύεται για να ελαχιστοποιήσει το σφάλμα του συνολικού συνόλου. Η προσθήκη των νέων μοντέλων είναι διαδοχική και, συνήθως, τα μοντέλα προστίθενται σταδιακά (π.χ., σε 100 ή 1000 επαναλήψεις). Κάθε νέο μοντέλο δεν έχει ίση βαρύτητα, αλλά το βάρος του μοντέλου εξαρτάται από το πόσο καλά διορθώνει τα σφάλματα των προηγούμενων μοντέλων. Η συνολική πρόβλεψη του αλγορίθμου προκύπτει από το σύνολο των προβλέψεων όλων των μοντέλων [16]. Παρατίθενται παρακάτω ένας ψευδοκώδικας που περιγράφει την χρήση του αλγορίθμου Gradient Boosting [17].

Πίνακας 2.1: Ψευδοκώδικας για το μοντέλο του Gradient Boosting [17]

Βήμα 1: Αρχικοποίηση Το αρχικό μοντέλο προβλέψεων F_o (x) υπολογίζεται ελαχιστοποιώντας τη συνολική συνάρτηση απώλειας: $$F_O(x) = argmin_\rho \sum_{i=1}^{N} L(y_i, \rho)$$
Βήμα 2: Επαναληπτική Διαδικασία Για κάθε επανάληψη m=1 έως M (όπου M είναι ο συνολικός αριθμός επαναλήψεων):
2.1 Υπολογίζεται η αρνητική κλίση της συνάρτησης απώλειας για κάθε παρατήρηση i, η οποία λειτουργεί ως τα «υπολείμματα» (residuals): $$\tilde{y}_i = -\left[\frac{\theta L\big(y_i, F(x_i)\big)}{\theta F(x_i)}\right]_{F(x)=F_{m-1}(x)}, i = 1, \dots, N$$
2.2 Προσαρμόζεται ένας «base learner» h(x;a) για να προσεγγίσει τα υπολείμματα $\tilde{y}_i$. Αυτό γίνεται μέσω ελαχιστοποίησης του τετραγωνικού σφάλματος. Στην πράξη ο «weak learner» είναι ένα μικρό δέντρο απόφασης.

$$a_m = argmin \sum_{i=1}^{N} [\widetilde{y_i} - \beta h(x_i; a)]^2$$
2.3 Υπολογίζεται ένας συντελεστής βήματος ρ_m που ελαχιστοποιεί τη συνολική συνάρτηση απώλειας, με το νέο base learner να προστίθεται στο προηγούμενο μοντέλο: $$\rho_m = argmin \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + \rho h(x_i; a_m))$$
2.4 Το νέο μοντέλο δημιουργείται προσθέτοντας το αποτέλεσμα του base learner με το βέλτιστο βάρος στο τρέχον μοντέλο: $$F_m(x) = F_{m-1}(x) + \rho_m h(x_i; a_m)$$
Βήμα 3: Τέλος επανάληψης Η διαδικασία επαναλαμβάνεται για M βήματα. Το τελικό μοντέλο $F_m(x)$ είναι ένας γραμμικός συνδυασμός όλων των base learners. $$F_m(x) = F_o(x) + \sum_{m=1}^{M} \rho_m h(x_i; a_m)$$

Για την ανάλυση στην παρούσα εργασία, χρησιμοποιήθηκε ο αλγόριθμος Gradient Boosting Classifier από τη βιβλιοθήκη scikit-learn για την ταξινόμηση δεδομένων. Η βελτιστοποίηση των υπερπαραμέτρων πραγματοποιήθηκε μέσω της μεθόδου Grid Search με πενταπλή διασταυρούμενη επικύρωση (5-fold cross-validation). Ο χώρος αναζήτησης των υπερπαραμέτρων περιλάμβανε τον αριθμό των εκτιμητών (n_estimators), το μέγιστο βάθος των δέντρων (max_depth) και τον ρυθμό εκμάθησης (learning_rate). Ως μετρική αξιολόγησης χρησιμοποιήθηκε η ακρίβεια (accuracy). Με την ολοκλήρωση της διαδικασίας Grid Search, εκτυπώθηκε το καλύτερο εκτιμηθέν μοντέλο, οι βέλτιστες τιμές των υπερπαραμέτρων καθώς η καλύτερη βαθμολογία διασταυρούμενης επικύρωσης.

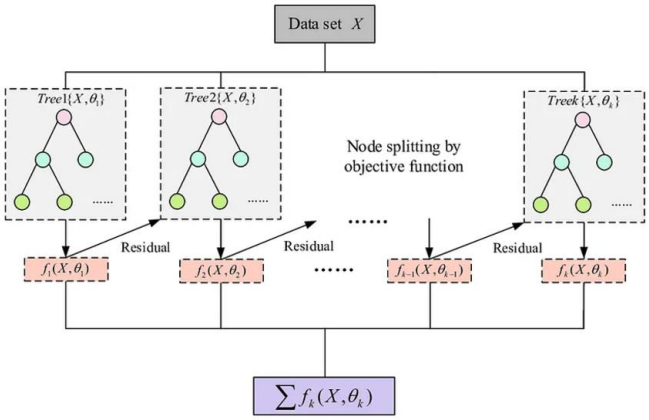
Το βέλτιστο μοντέλο που προέκυψε από τη διαδικασία χρησιμοποιήθηκε στη συνέχεια για την πρόβλεψη των δεδομένων του συνόλου δοκιμής. Η αξιολόγηση της απόδοσής του πραγματοποιήθηκε με τη χρήση των μετρικών accuracy, precision, recall και F1 score μέσω της συνάρτησης evaluate_model, η οποία αναπτύχθηκε στο πλαίσιο της εργασίας και υπολογίζει και εκτυπώνει τις βασικές μετρικές απόδοσης κάθε μοντέλου ταξινόμησης καθώς επίσης εμφανίζει και τον πίνακα σύγχυσης. Τα αποτελέσματα εκτυπώθηκαν για να διευκολύνουν την κατανόηση της αποδοτικότητας του μοντέλου. Επιπλέον, κατασκευάστηκε η καμπύλη ROC (Receiver Operating Characteristic) χρησιμοποιώντας τη συνάρτηση plot_roc_curve, επιτρέποντας την ανάλυση της διαχωριστικής ικανότητας του ταξινομητή.

2.3.2 XG Boost

Μία παραλλαγή και επέκταση του μοντέλου Gradient Boosting αποτελεί το XG Boost (eXtreme Gradient Boosting Decision Tree). Προτάθηκε από τους Chen και Guestrin το 2016 στο Πανεπιστήμιο της Ουάσιγκτον [18] και εισάγει σημαντικές βελτιώσεις που το διαφοροποιούν από τον απλό αλγόριθμο Gradient Boosting. Συγκεκριμένα, ενώ στον Gradient Boosting οι «αδύναμοι μαθητές» προστίθενται σειριακά, το XGBoost αξιοποιεί πολυνηματική προσέγγιση, κάνοντας χρήση των πολλαπλών πυρήνων της CPU, γεγονός που οδηγεί σε σημαντικά αυξημένη ταχύτητα και απόδοση. Επιπλέον, το XGBoost περιλαμβάνει υλοποίηση sparse aware, η οποία διαχειρίζεται αυτόματα τις ελλείπουσες τιμές στα δεδομένα καθώς δομή block για να υποστηρίξει την παραλληλοποίηση στη διαδικασία κατασκευής των δέντρων απόφασης (decision trees). Ένα άλλο χαρακτηριστικό είναι η δυνατότητα συνεχούς εκπαίδευσης, που επιτρέπει την περαιτέρω ενίσχυση ενός ήδη εκπαιδευμένου μοντέλου με νέα δεδομένα. Το XGBoost έχει αποδειχθεί εξαιρετικά αποτελεσματικό, κυριαρχώντας σε προβλήματα ταξινόμησης, παλινδρόμησης και γενικότερα σε προβλήματα προγνωστικής μοντελοποίησης [19].

Αναλυτικότερα, το XG Boost συνδυάζει πολλά απλά μοντέλα για να κατασκευάσει ένα ισχυρότερο συνολικό μοντέλο μέσω της τεχνικής της ενίσχυσης βαθμιδωτών δέντρων. Η διαδικασία περιλαμβάνει τη σταδιακή προσθήκη απλών παραμετρικών συναρτήσεων (base learners), οι οποίες προσαρμόζονται στα «ψευδό-υπολείμματα» (residuals). Τα ψευδό-υπολείμματα είναι οι διαφορές μεταξύ των πραγματικών τιμών και των τρεχουσών προβλέψεων του μοντέλου. Ουσιαστικά, είναι οι κλίσεις της συνάρτησης απώλειας, που υπολογίζονται για κάθε σημείο δεδομένων. Σε κάθε επανάληψη, υπολογίζεται η κλίση της συνάρτησης απώλειας και παράλληλα μια απλή συνάρτηση προσαρμόζεται σε αυτά τα δεδομένα, ελαχιστοποιώντας την τετραγωνική απόκλιση.

Το ιδιαίτερο χαρακτηριστικό του XGBoost είναι ότι δεν απορρίπτει το προηγούμενο δέντρο. Αντίθετα, χρησιμοποιεί τις πληροφορίες από όλα τα υπάρχοντα δέντρα στο ensemble και προσθέτει ένα νέο δέντρο που εστιάζει στη διόρθωση των υπολειπόμενων σφαλμάτων. Με αυτόν τον τρόπο, το μοντέλο βελτιώνεται σταδιακά. Κατά τη διάρκεια όλης της διαδικασίας, το XGBoost βελτιστοποιεί μια συνάρτηση απώλειας. Αυτή η συνάρτηση αξιολογεί πόσο καλά οι προβλέψεις του μοντέλου ταιριάζουν με τις πραγματικές τιμές. Με την ελαχιστοποίηση της συνάρτησης αυτής, διασφαλίζεται ότι το ensemble γίνεται πιο ακριβές και αποδοτικό [20, 21].



Εικόνα 2.3: Διάγραμμα ροής του μοντέλου XGBoost [20]

Στον παρακάτω πίνακα, απεικονίζεται η διαδικασία εκπαίδευσης του μοντέλου XG Boost, η οποία βασίζεται στη βελτιστοποίηση της συνάρτησης απώλειας μέσω αδύναμων μοντέλων, όπως δέντρα απόφασης. Ξεκινά με αρχικοποίηση του μοντέλου και συνεχίζεται με επαναληπτικό υπολογισμό παραγώγων (gradient, hessian), εκπαίδευση νέου μοντέλου και ενημέρωση του συνολικού μοντέλου με συντελεστή εκμάθησης. Το τελικό μοντέλο προκύπτει ως το άθροισμα όλων των βημάτων, εξασφαλίζοντας υψηλή απόδοση.

Πίνακας 2.2: Αλγόριθμος XG Boost [21, 22, 23]

Είσοδος: Σύνολο εκπαίδευσης $\{(x_i, y_i)\}_{i=1}^{N}$, συνάρτηση απώλειας $L\big(y_i, F(x)\big)$, αριθμός αδύναμων μαθητών M, ρυθμός εκμάθησης α.
<u>Βήμα 1:</u> Αρχικοποίηση Το αρχικό μοντέλο προβλέψεων ξεκινάει με μια σταθερή τιμή θ, η οποία ελαχιστοποιεί το άθροισμα της συνάρτησης απώλειας $L(y_i, \theta)$. $$F_O(x) = argmin_\theta \sum_{i=1}^{N} L(y_i, \theta)$$
<u>Βήμα 2:</u> Επαναληπτική Διαδικασία Για κάθε επανάληψη m=1 έως M (όπου M είναι ο συνολικός αριθμός επαναλήψεων):
2.1 Υπολογίζεται το gradient $\widehat{g_m}(x_i)$, δηλαδή η παράγωγος της συνάρτησης απώλειας και το hessian $\widehat{h_m}(x_i)$, η δεύτερη παράγωγος. $$\widehat{g_m}(x_i) = -\left[\frac{\theta L\big(y_i, F(x_i)\big)}{\theta F(x_i)}\right]_{F(x)=F_{m-1}(x)}$$

$$\widehat{h_m}(x_i) = -\left[\frac{\theta^2 L(y_i, F(x_i))}{\theta F(x_i)^2}\right]_{F(x)=F_{m-1}(x)}$$
2.2 Το αδύναμο μοντέλο $\widehat{\varphi_m}(x)$ εκπαιδεύεται για να προβλέπει το πηλίκο gradient/hessian για κάθε δείγμα, ελαχιστοποιώντας το εξής πρόβλημα βελτιστοποίησης: $$\widehat{\varphi_m} = argmin_{\varphi\in\Phi}\sum_{i=1}^{N}[\varphi(x_i) - \frac{\widehat{g_m}(x_i)}{\widehat{h_m}(x_i)}]^2$$
2.3 Το νέο μοντέλο προστίθενται στο προηγούμενο με ένα συντελεστή εκμάθησης α: $$\widehat{f_m}(x) = \alpha\widehat{\varphi_m}(x)$$ Το συνολικό μοντέλο ενημερώνεται: $$\widehat{f_m}(x) = \widehat{f_{m-1}}(x) - \widehat{f_m}(x)$$
Βήμα 3: Τέλος επανάληψης Το τελικό μοντέλο μετά από M επαναλήψεις είναι το άθροισμα όλων των αδύναμων μοντέλων: $$\hat{f}(x) = \widehat{f_M}(x) = \sum_{m=0}^{M}\widehat{f_m}(x)$$

Λαμβάνοντας υπόψη όλα τα παραπάνω, μπορούν να συνοψιστούν οι βασικές διαφορές μεταξύ του Gradient Boosting και του XG Boost όπως φαίνονται στον παρακάτω πίνακα.

Πίνακας 2.3: Σύνοψη βασικών διαφορών GB και XG Boost [22, 24, 25]

	Gradient Boosting	**XG Boost**
Περιγραφή	Ensemble τεχνική που μετατρέπει πολλαπλούς weak learners σε ισχυρό μοντέλο μέσω της βελτιστοποίησης σφαλμάτων.	Προηγμένη υλοποίηση του gradient boosting.
Απόδοση και Ταχύτητα	Αρκετά αποδοτικό αλλά πιο αργό από το XG Boost λόγω της διαδοχικής εκπαίδευσης των weak learners.	Υψηλά βελτιστοποιημένο, γρήγορο με παράλληλη επεξεργασία.
Βελτιστοποίηση	Ελαχιστοποίηση του σφάλματος μέσω του gradient της συνάρτησης απώλειας.	Κανονικοποιημένη αντικειμενική συνάρτηση με ενσωματωμένη L1/L2 κανονικοποίηση.
Κανονικοποίηση	Υποστηρίζει βασικές παραμέτρους για τακτική κανονικοποίηση, αλλά χρειάζεται προσεκτική ρύθμιση.	Ενσωματωμένη κανονικοποίηση L1 (Lasso) και L2 (Ridge).

Απουσιάζουσες τιμές	Χειρίζεται κενά δεδομένα μέσω surrogate splits χωρίς προεπεξεργασία.	Χειρίζεται κενά δεδομένα μέσω surrogate splits χωρίς προεπεξεργασία.
Επεξεργασία πολλών δεδομένων	Λειτουργεί καλά αλλά μπορεί να απαιτεί μεγαλύτερη υπολογιστική ισχύ σε μεγάλα σύνολα δεδομένων.	Πολύ αποτελεσματικό με παράλληλη επεξεργασία και αποδοτική χρήση μνήμης.
Πρόληψη Overfitting	Ευαίσθητο στο overfitting, απαιτεί προσεκτική ρύθμιση hyperparameters (regularization, early stopping).	Κανονικοποίηση L1/L2 και δυνατότητες early stopping.
Υπολογιστική Πολυπλοκότητα	Μικρότερη υπολογιστική πολυπλοκότητα σε μικρότερα σύνολα δεδομένων, καλύτερο για γρήγορη εκπαίδευση.	Χρειάζεται ισχυρό υπολογιστικό εξοπλισμό και παραλληλία για μεγάλες εφαρμογές.

Στο πλαίσιο της παρούσας μελέτης εξετάστηκε το μοντέλο XGBoost, αξιοποιώντας τη βιβλιοθήκη XGBoost της Python. Τα αποτελέσματα που προέκυψαν αξιολογήθηκαν με βάση ένα σύνολο μετρικών, όπως η ακρίβεια, η ακρίβεια πρόβλεψης (precision), η ανάκληση (recall), το F1-score και ο πίνακας σύγχυσης.

2.3.3 Random Forest

Ο ταξινομητής Random Forest είναι ένας ευρέως χρησιμοποιούμενος αλγόριθμος Μηχανικής Μάθησης ο οποίος συνδυάζει τα αποτελέσματα πολλαπλών δέντρων απόφασης για να καταλήξει σε ένα ενιαίο, ισχυρότερο αποτέλεσμα με την χρήση των τεχνικών bootstrapping και aggregation. Κάθε δέντρο απόφασης εκπαιδεύεται χρησιμοποιώντας ένα bootstrap sample, το οποίο είναι ένα τυχαίο δείγμα του ίδιου μεγέθους με το σύνολο εκπαίδευσης, αλλά δημιουργείται μέσω τυχαίας δειγματοληψίας από το ίδιο το σύνολο εκπαίδευσης. Για να αυξηθεί η τυχαιότητα και να μειωθεί η συσχέτιση μεταξύ των δέντρων, εφαρμόζεται feature bagging, όπου κάθε δέντρο χρησιμοποιεί τυχαίο υποσύνολο χαρακτηριστικών κατά την κατασκευή του.

Ένα χαρακτηριστικό του Random Forest είναι η χρήση των out-of-bag (OOB) δειγμάτων, τα οποία είναι παρατηρήσεις που δεν περιλαμβάνονται στο bootstrap sample κάθε δέντρου και χρησιμοποιούνται ως σύνολο δοκιμής. Η εκτίμηση του σφάλματος γίνεται με βάση τα OOB δείγματα, και αποδεικνύεται ότι αυτή η εκτίμηση είναι ακριβής, όπως και η χρήση ενός ξεχωριστού συνόλου δοκιμής. Αυτή η μέθοδος μειώνει την ανάγκη για ξεχωριστό σύνολο δοκιμής και παρέχει γρήγορη και αξιόπιστη εκτίμηση του σφάλματος γενίκευσης του μοντέλου.

Η τελική πρόβλεψη του Random Forest γίνεται μέσω της «ψηφοφορίας» των επιμέρους δέντρων του μοντέλου. Κάθε δέντρο «ψηφίζει» για μια κατηγορία, και η κατηγορία με τις περισσότερες ψήφους επιλέγεται ως η τελική πρόβλεψη του μοντέλου. Αυτός ο μηχανισμός αυξάνει την ακρίβεια του μοντέλου, καθώς ενσωματώνει την ποικιλία και τη δύναμη πολλών δέντρων απόφασης. Η διαδικασία αυτή επαναλαμβάνεται για αρκετές

επαναλήψεις (συνήθως από 100 έως 1000 φορές), δημιουργώντας το τελικό Random Forest που είναι ιδιαίτερα ανθεκτικό στην υπερπροσαρμογή (overfitting) [26, 27].

Ο αλγόριθμος Random Forest προσφέρει αρκετά πλεονεκτήματα κυρίως λόγω της συνδυαστικής φύσης του. Πρώτον, μπορεί να χειρίζεται αποτελεσματικά την μη ισορροπημένη ταξινόμηση, προσφέροντας καλή ανοχή στο θόρυβο και υψηλή ακρίβεια στην ταξινόμηση. Επίσης, ο αλγόριθμος μπορεί να διαχειριστεί μεγάλους όγκους δεδομένων και να προσφέρει καλές επιδόσεις σε κατανεμημένα και παράλληλα περιβάλλοντα [28]. Παράλληλα, απαιτεί λιγότερη προεπεξεργασία δεδομένων σε σχέση με άλλους αλγορίθμους καθώς δεν απαιτεί κανονικοποίηση ή κωδικοποίηση των χαρακτηριστικών, και μπορεί να χειριστεί κατηγορηματικά δεδομένα χωρίς προηγούμενη μετατροπή. Ωστόσο, η διαδικασία εκπαίδευσης μπορεί να είναι χρονοβόρα, καθώς απαιτεί την εκπαίδευση πολλών δέντρων απόφασης, γεγονός που συνεπάγεται και απαίτηση για περισσότερους υπολογιστικούς πόρους.

Στην παρούσα εργασία, ο αλγόριθμος Random Forest χρησιμοποιήθηκε με 2 τρόπους: για την επιλογή χαρακτηριστικών μέσω της τεχνικής RFE (Recursive Feature Elimination) καθώς και για την ταξινόμηση των δεδομένων.

Η διαδικασία επιλογής των κατάλληλων χαρακτηριστικών για το Random Forest ακολούθησε μια προσέγγιση παρόμοια με αυτή του Gradient Boosting. Αρχικά, εφαρμόστηκε η τεχνική υπερδειγματοληψίας (oversampling) με τη βοήθεια του RandomOverSampler για να αντιμετωπιστεί το πρόβλημα της ανισορροπίας κλάσεων στο σύνολο εκπαίδευσης. Με αυτή τη μέθοδο, αυξήθηκε η ποσότητα των δειγμάτων της μειονοτικής κλάσης, ενισχύοντας το σύνολο εκπαίδευσης και βελτιώνοντας τη μάθηση του μοντέλου. Στη συνέχεια, το GridSearchCV χρησιμοποιήθηκε για την αναζήτηση των καλύτερων υπερπαραμέτρων του μοντέλου, όπως ο αριθμός των δέντρων (n_estimators), το μέγιστο βάθος των δέντρων (max_depth), οι ελάχιστες τιμές για το διαχωρισμό (min_samples_split), οι ελάχιστες τιμές για το φύλλο (min_samples_leaf) και τα χαρακτηριστικά που χρησιμοποιούνται σε κάθε διαχωρισμό (max_features). Η διαδικασία εκτελέστηκε με πενταπλή διασταυρούμενη επικύρωση (5-fold cross-validation) και ορίζοντας ως μετρική αξιολόγησης την ακρίβεια. Το βέλτιστο μοντέλο που προέκυψε από αυτή την διαδικασία χρησιμοποιήθηκε στη συνέχεια για την πρόβλεψη των ετικετών.

Στο Scikit-Learn, η σημασία των χαρακτηριστικών εκτιμάται αυτόματα μετά την εκπαίδευση του μοντέλου για τον αλγόριθμο Random Forest. Ο υπολογισμός βασίζεται στη μέση μείωση της "ακαθαρσίας" (impurity) που προκαλείται από τους κόμβους που χρησιμοποιούν το κάθε χαρακτηριστικό για τη διάσπαση. Η μείωση αυτή υπολογίζεται ως σταθμισμένος μέσος όρος, όπου το βάρος κάθε κόμβου καθορίζεται από τον αριθμό των δειγμάτων εκπαίδευσης που περνούν από αυτόν. Στο τέλος, οι τιμές κανονικοποιούνται, ώστε το συνολικό άθροισμα της σημασίας όλων των χαρακτηριστικών να ισούται με 1, διευκολύνοντας τη σύγκριση της σχετικής συμβολής κάθε χαρακτηριστικού στο μοντέλο.

2.3.4 Naive Bayes

Ο ταξινομητής Naive Bayes (Αφελής Bayes) είναι ένας πιθανοθεωρητικός ταξινομητής που βασίζεται στην εφαρμογή του θεωρήματος του Bayes με ισχυρές παραδοχές ανεξαρτησίας μεταξύ των χαρακτηριστικών. Ένα μοντέλο Naive Bayesian είναι εύκολο να κατασκευαστεί, χωρίς πολύπλοκες επαναληπτικές εκτιμήσεις παραμέτρων, κάτι που το καθιστά ιδιαίτερα χρήσιμο και στον τομέα της ιατρικής για τη διάγνωση ασθενειών. Παρά την απλότητά του, ο ταξινομητής Naive Bayes συχνά αποδίδει εντυπωσιακά καλά και χρησιμοποιείται ευρέως, καθώς πολλές φορές ξεπερνά σε απόδοση πιο σύνθετες μεθόδους ταξινόμησης [29].

Το Θεώρημα του Bayes παρέχει έναν τρόπο υπολογισμού της μεταγενέστερης πιθανότητας, P(c|x), με βάση τις P(c), P(x), και P(x|c). Ο ταξινομητής Naive Bayes υποθέτει ότι η επίδραση της τιμής ενός προβλεπτικού χαρακτηριστικού (x) σε μια δεδομένη κατηγορία (c) είναι ανεξάρτητη από τις τιμές των άλλων προβλεπτικών χαρακτηριστικών. Αυτή η

παραδοχή ονομάζεται υπό συνθήκη ανεξαρτησία κατηγορίας (class conditional independence). Η μαθηματική σχέση που περιγράφει το θεώρημα Bayes είναι η εξής:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$, όπου

P(c|x): μεταγενέστερη πιθανότητα της κατηγορίας (στόχος) δεδομένου ενός χαρακτηριστικού,
P(c): προηγούμενη πιθανότητα της κατηγορία,
P(x|c): πιθανότητα (likelihood), δηλαδή η πιθανότητα του χαρακτηριστικού δεδομένης της κατηγορίας,
P(x): εκ των προτέρων πιθανότητα του χαρακτηριστικού [29, 30].

Αξίζει να σημειωθεί ότι ονομάζεται αφελής Bayes επειδή ο υπολογισμός των πιθανοτήτων για κάθε υπόθεση είναι απλοποιημένος και εύκολα εφαρμόσιμος. Δηλαδή, θεωρούμε ότι οι τιμές κάθε τιμής χαρακτηριστικού είναι ανεξάρτητες υπό όρους, δεδομένης της τιμής στόχου. Αυτή είναι μια πολύ ισχυρή υπόθεση που είναι πιο απίθανη σε πραγματικά δεδομένα, δηλαδή ότι τα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους. Παρόλα αυτά, η προσέγγιση αποδίδει εκπληκτικά καλά σε δεδομένα όπου αυτή η υπόθεση ισχύει. Η εκπαίδευση είναι γρήγορη γιατί μόνο η πιθανότητα κάθε τάξης και η πιθανότητα κάθε κατηγορίας με διαφορετικές τιμές εισόδου πρέπει να υπολογιστούν. Παράλληλα, δεν χρειάζεται να προσαρμόζονται συντελεστές με διαδικασίες βελτιστοποίησης. Οι πιθανότητες της κλάσης είναι απλώς η συχνότητα των περιπτώσεων που ανήκουν σε κάθε κατηγορία διαιρεμένη με τον συνολικό αριθμό των περιπτώσεων. Οι πιθανότητες υπό όρους είναι η συχνότητα κάθε τιμής χαρακτηριστικού για μια δεδομένη κλάση διαιρεμένη με τη συχνότητα των περιπτώσεων με αυτήν την κλάση [lab ML]. Παρακάτω παρατίθενται ο ψευδό – κώδικας τον οποίο ακολουθεί ο ταξινομητής Naïve Bayes.

Πίνακας 2.4: Ψευδό - αλγόριθμος για το μοντέλο Naive Bayes [30]

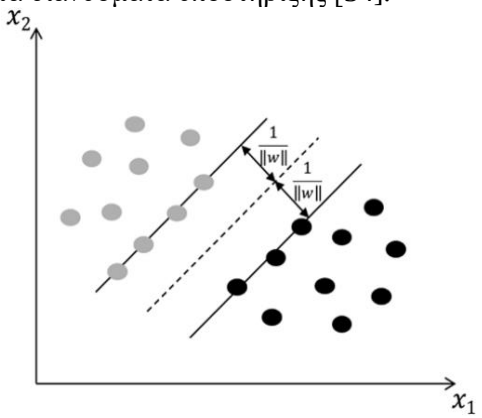
Είσοδος: Σύνολο εκπαίδευσης T, F = (f1, f2, f3, … , fn) οι τιμές των μεταβλητών πρόβλεψης στο σύνολο δοκιμής
Έξοδος: Η προβλεπόμενη κλάση του συνόλου δοκιμής
Βήμα 1: Διαβάζεται το σύνολο εκπαίδευσης T.
Βήμα 2: Για κάθε κλάση στο σύνολο εκπαίδευσης: 2.1 Υπολογίζεται ο μέσος όρος και την τυπική απόκλιση για κάθε μεταβλητή πρόβλεψης (fi) στη συγκεκριμένη κλάση.
Βήμα 3: Για κάθε δείγμα στο σύνολο δοκιμής: 3.1 Για κάθε κλάση: 3.1.1 Για κάθε μεταβλητή πρόβλεψης fi υπολογίζεται η πιθανότητα της fi χρησιμοποιώντας την εξίσωση της κανονικής Gaussian κατανομής. 3.1.2 Πολλαπλασιάζονται οι πιθανότητες όλων των μεταβλητών προκειμένου να υπολογιστεί η πιθανοφάνεια της κλάσης. 3.2 Συγκρίνονται οι πιθανοφάνειες για όλες τις κλάσεις.
Βήμα 4: Η κλάση με τη μεγαλύτερη πιθανοφάνεια επιλέγεται ως η τελική πρόβλεψη.

Για την ανάλυση που απαιτείται στην εργασία, εξετάστηκε ο αλγόριθμος Naive Bayes με τον παρακάτω τρόπο. Αρχικά, προκειμένου να διασφαλιστεί ότι τα δεδομένα μπορούν να τροφοδοτηθούν σωστά στον αλγόριθμο, εφαρμόστηκε κανονικοποίηση των δεδομένων μέσω της τεχνικής Standard Scaler. Αυτό σημαίνει ότι τα χαρακτηριστικά των δεδομένων εκπαίδευσης και δοκιμής κλιμακώθηκαν έτσι ώστε να έχουν μηδενική μέση τιμή και μια τυπική απόκλιση 1, εξασφαλίζοντας τη σωστή κλιμάκωση για τον αλγόριθμο Naive Bayes, ο οποίος βασίζεται σε ανεξάρτητες κατανομές χαρακτηριστικών. Στη συνέχεια, το μοντέλο Gaussian Naive Bayes εκπαιδεύτηκε με τα κανονικοποιημένα δεδομένα εκπαίδευσης και τα αποτελέσματα του χρησιμοποιήθηκαν για την πρόβλεψη στις τιμές δοκιμής.

2.3.5 SVM (Support vector machine)

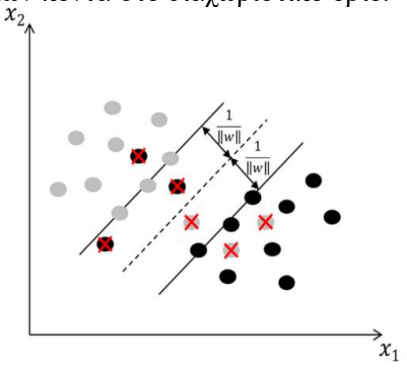
Ο SVM (Support vector machine) ή αλλιώς Μηχανή Διανυσμάτων Υποστήριξης είναι ένας αλγόριθμος για επιβλεπόμενη μάθηση, που χρησιμοποιείται για ταξινόμηση και παλινδρόμηση [31, 32]. Ο στόχος του SVM είναι να βρει το καταλληλότερο υπερεπίπεδο (hyperplane) με το υψηλότερο περιθώριο που μπορεί να διαιρέσει τις κλάσεις γραμμικά με την βοήθεια διανυσμάτων υποστήριξης. Το SVM διακρίνεται σε 2 κατηγορίες: hard margin και soft margin.

Η πρώτη προσέγγιση στοχεύει στον εντοπισμό ενός υπερεπιπέδου το οποίο διαχωρίζει πλήρως τα δεδομένα διαφορετικών κλάσεων χωρίς σφάλματα. Το υπερεπίπεδο αυτό εξασφαλίζει τη μέγιστη δυνατή απόσταση (margin) από τα κοντινότερα σημεία κάθε κλάσης, γνωστά ως support vectors. Η εξίσωση του υπερεπιπέδου ορίζει το όριο διαχωρισμού των κλάσεων, με στόχο το μέγιστο περιθώριο από τα δεδομένα, ενώ διασφαλίζει ότι όλα τα σημεία ταξινομούνται σωστά. Τα διανύσματα υποστήριξης είναι εκείνα που βρίσκονται πιο κοντά στο υπερεπίπεδο και επηρεάζουν άμεσα την απόφαση ταξινόμησης. Δεν επιτρέπονται οι παραβιάσεις του περιθωρίου, γεγονός που την καθιστά την συγκεκριμένη προσέγγιση κατάλληλη μόνο όταν τα δεδομένα είναι γραμμικά διαχωρίσιμα. Όπως φαίνεται και στην παρακάτω εικόνα, το διακεκομμένο ευθύγραμμο τμήμα στο κέντρο της εικόνας είναι το διαχωριστικό υπερεπίπεδο, τα γκρι και τα μαύρα σημεία που ακουμπούν τις γραμμές με απόσταση $\frac{1}{\|w\|}$ από το υπερεπίπεδο είναι τα support vectors ενώ το περιθώριο είναι η απόσταση από το υπερεπίπεδο μέχρι τα διανύσματα υποστήριξης [34].



Εικόνα 2.4: Hard- maximum margin separating hyperplane [34]

Όταν τα δεδομένα δεν είναι πλήρως διαχωρίσιμα, όπως τα σημεία που σημειώνονται με "X" στην επόμενη φωτογραφία, εισάγονται μεταβλητές x_i στη συνάρτηση στόχου του SVM για να επιτραπεί σφάλμα στην λανθασμένη ταξινόμηση. Στην περίπτωση αυτή, το SVM δεν αναζητά το «hard margin» που θα ταξινομούσε όλα τα δεδομένα αλάνθαστα. Αντίθετα, το SVM λειτουργεί ως ταξινομητής μαλακού περιθωρίου (soft-margin classifier), δηλαδή ταξινομεί τα περισσότερα δεδομένα σωστά, επιτρέποντας ωστόσο στο μοντέλο να κάνει λάθη στην ταξινόμηση λίγων σημείων κοντά στο διαχωριστικό όριο.



Εικόνα 2.5: Μερικές λανθασμένες ταξινομήσεις, μέρος του soft - margin SVM

Ωστόσο, όταν ένα πρόβλημα δεν είναι γραμμικά διαχωρίσιμο στον χώρο εισόδου, το soft – margin SVM δεν μπορεί να βρει ένα σταθερό υπερεπίπεδο διαχωρισμού που να ελαχιστοποιεί τον αριθμό των λανθασμένων ταξινομήσεων και να γενικεύει καλά. Για να ξεπεραστεί αυτό το πρόβλημα, μπορεί να χρησιμοποιηθεί ένας πυρήνας (kernel) για να μετασχηματίσει τα δεδομένα σε έναν υψηλότερης διάστασης χώρο, τον οποίο ονομάζουμε χώρο πυρήνα (kernel space), όπου τα δεδομένα θα είναι γραμμικά διαχωρίσιμα. Στον χώρο πυρήνα, μπορεί να αποκτηθεί ένα γραμμικό υπερεπίπεδο για να διαχωρίσει τις διάφορες κλάσεις που εμπλέκονται στην ταξινόμηση, αντί να λυθεί ένα υπερεπίπεδο υψηλής τάξης στον χώρο εισόδου.

Στο πλαίσιο της μελέτης, χρησιμοποιήθηκε ο αλγόριθμος Support Vector Machine (SVM). Πραγματοποιήθηκε συγκριτική αξιολόγηση της απόδοσης του μοντέλου με και χωρίς την εφαρμογή της τεχνικής SMOTE (Synthetic Minority Oversampling Technique), με στόχο να εκτιμηθεί η αποτελεσματικότητά της στη βελτίωση της απόδοσης. Η μέθοδος SMOTE εφαρμόστηκε για την παραγωγή συνθετικών δειγμάτων στη μειοψηφική κλάση, εξισορροπώντας την κατανομή στο σύνολο εκπαίδευσης. Τα αποτελέσματα της ανάλυσης έδειξαν ότι η χρήση του SMOTE προσφέρει σημαντική βελτίωση, ενισχύοντας τόσο τη συνολική ακρίβεια όσο και τις επιδόσεις σε κρίσιμες μετρικές απόδοσης. Έπειτα, εφαρμόστηκε Min-Max Scaling και βελτιστοποίηση των υπερπαραμέτρων μέσω Grid Search με δεκαπλή διασταυρούμενη επικύρωση. Ο χώρος αναζήτησης περιλάμβανε τις παραμέτρους C (παράγοντας κανονικοποίησης), kernel (τύπος πυρήνα: linear, rbf, poly) και gamma (συντελεστής για μη γραμμικούς πυρήνες). Ως μετρική αξιολόγησης χρησιμοποιήθηκε η ακρίβεια. Το βέλτιστο μοντέλο που προέκυψε με την εφαρμογή του SMOTE χρησιμοποιήθηκε για την πρόβλεψη των δεδομένων δοκιμής.

2.3.6 KNN (k-nearest neighbors) classifier

Ένας ακόμα αρκετά διαδεδομένος αλγόριθμος Μηχανικής Μάθησης που χρησιμοποιήθηκε στην παρούσα εργασία είναι ο αλγόριθμος Κ πλησιέστερων γειτόνων - K-Nearest Neighbors (KNN). Ο συγκεκριμένος αλγόριθμος αποτελεί μία γενίκευση των κανόνων του πιο κοντινού γείτονα. Συγκριτικά με τον κανόνα του πλησιέστερου γείτονα, η βασική διαφορά είναι ότι ο KNN επεκτείνει την ανάλυση στους k πλησιέστερους γείτονες κατά τη φάση λήψης αποφάσεων. Δηλαδή, δεν περιορίζεται μόνο στο να εξετάσει το πιο κοντινό σημείο για να καθορίσει την κατηγορία (ετικέτα κλάσης) ενός άγνωστου σημείου, αλλά λαμβάνει υπόψη τα k πλησιέστερα σημεία. Αυτή η επέκταση επιτρέπει στον KNN αλγόριθμο να αντλεί και να αξιοποιεί περισσότερες πληροφορίες.

Σε αντίθεση με άλλους αλγόριθμους ταξινόμησης, όπως οι Random Forest ή SVM, οι οποίοι περιλαμβάνουν φάση εκπαίδευσης για την δημιουργία ενός μοντέλου από τα δεδομένα εκπαίδευσης, ο αλγόριθμος KNN παραλείπει τη διαδικασία εκμάθησης. Αντί να δημιουργεί ένα μοντέλο, αποθηκεύει απλώς τα δεδομένα εκπαίδευσης και όταν καλείται να ταξινομήσει ένα νέο δείγμα, υπολογίζει άμεσα την απόσταση του από όλα τα σημεία του συνόλου εκπαίδευσης. Χρησιμοποιεί τις πλησιέστερες γειτονιές για να προσδιορίσει την ετικέτα κατηγορίας του νέου δείγματος. Αυτό τον καθιστά έναν απλό αλγόριθμο που δεν απαιτεί εκτενή προεπεξεργασία, κάνοντάς τον κατάλληλο για προβλήματα όπου τα δεδομένα είναι περιορισμένα και εύκολα διαχειρίσιμα. Ωστόσο, η μέθοδος αυτή μπορεί να είναι λιγότερο αποδοτική για πολύ μεγάλα σύνολα δεδομένων, καθώς απαιτεί σύγκριση κάθε νέου δείγματος με όλα τα αποθηκευμένα δείγματα, κάτι που αυξάνει τον υπολογιστικό φόρτο [35].

Για να προσδιοριστούν οι πλησιέστερες ομάδες ή τα πλησιέστερα σημεία, χρειαζόμαστε κάποια μέθοδο μέτρησης απόστασης. Η πιο ευρέως χρησιμοποιούμενη είναι η Ευκλείδεια απόσταση η οποία είναι η καρτεσιανή απόσταση μεταξύ δύο σημείων που βρίσκονται σε επίπεδο ή υπερεπίπεδο και υπολογίζεται ως εξής:

$$distance(x, Xi) = \sqrt{\sum_{j=1}^{d} \left(x_j - X_{i,j}\right)^2}$$

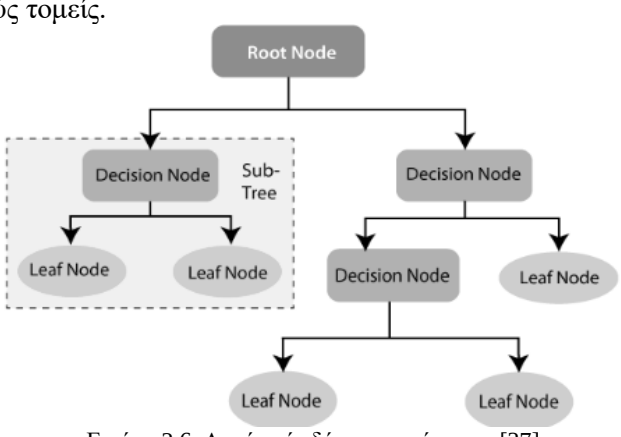
. Επιπλέον ως μετρική για τον υπολογισμό της απόστασης χρησιμοποιείται και η Taxicab metric. Η απόσταση Taxicab (Cityblock) d1 μεταξύ δυο διανυσμάτων p και q σε έναν n – διάστατο πραγματικό διανυσματικό χώρο με σταθερό καρτεσιανό σύστημα συντεταγμένων είναι το άθροισμα των μηκών των προβολών του τμήματος ευθείας που συνδέει τα σημεία στους άξονες των συντεταγμένων. Συγκεκριμένα, η απόσταση δίνεται από τον τύπο [46]:

$$d1(p,q) = \|p-q\|_1 = \sum |p_i - q_i|$$

Στο πλαίσιο της εργασίας, χρησιμοποιήθηκε ο αλγόριθμος k-Nearest Neighbors (kNN) από τη βιβλιοθήκη scikit-learn για την ταξινόμηση δεδομένων. Για την προεπεξεργασία των δεδομένων, εφαρμόστηκε κλιμάκωση μέσω της τεχνικής Min-Max Scaling ώστε τα χαρακτηριστικά να περιοριστούν στο διάστημα [0, 1], διασφαλίζοντας κατά αυτόν τον τρόπο τη σωστή λειτουργία του αλγορίθμου. Επιπλέον, πραγματοποιήθηκε βελτιστοποίηση των υπερπαραμέτρων του μοντέλου μέσω Grid Search με πενταπλή διασταυρούμενη επικύρωση. Ο χώρος αναζήτησης περιλάμβανε τον αριθμό των γειτόνων (n_neighbors) και τη στρατηγική βαρών (weights), όπου εξετάστηκαν οι επιλογές 'uniform' (όλοι οι γείτονες συνεισφέρουν ισοδύναμα) και 'distance' (οι γείτονες συνεισφέρουν με βάρη ανάλογα με την απόστασή τους). Ως μετρική αξιολόγησης χρησιμοποιήθηκε η ακρίβεια. Το βέλτιστο μοντέλο που προέκυψε χρησιμοποιήθηκε στη συνέχεια για την πρόβλεψη των δεδομένων του συνόλου δοκιμής.

2.3.7 Decision Tree (Δέντρα Αποφάσεων)

Τα Δέντρα Αποφάσεων είναι μια από τις μεθόδους που χρησιμοποιούνται ευρέως στον τομέα της Μηχανικής Μάθησης. Πρόκειται για μια δομή που μοιάζει με διάγραμμα ροής και αποτελείται από κόμβους και κλάδους. Κάθε κόμβος αντιπροσωπεύει χαρακτηριστικά μιας κατηγορίας που πρέπει να ταξινομηθεί, ενώ κάθε υποσύνολο αντιπροσωπεύει μια πιθανή τιμή που μπορεί να λάβει ο κόμβος. Χάρη στην απλότητα της ανάλυσης που προσφέρουν και την ακρίβειά τους σε πολλαπλές μορφές δεδομένων, τα δέντρα αποφάσεων έχουν βρει ευρεία εφαρμογή σε πολλούς τομείς.



Εικόνα 2.6: Δομή ενός δέντρου απόφασης [37]

Δυο σημαντικές έννοιες για τον αλγόριθμο των Δέντρων Αποφάσεων αποτελούν η Entropy (Εντροπία) και το Information Gain (Κέρδος Πληροφορίας). Η εντροπία χρησιμοποιείται για να μετρήσει την τυχαιότητα σε ένα σύνολο δεδομένων. Η τιμή της εντροπίας κυμαίνεται πάντα μεταξύ 0 και 1. Όσο μικρότερη είναι η τιμή της (ιδανικά 0), τόσο πιο καλά διαχωρισμένο θεωρείται το σύνολο δεδομένων. Αντίθετα, υψηλότερες τιμές υποδηλώνουν μεγαλύτερη «ασάφεια» στη διάκριση των δεδομένων. Η εντροπία ενός συνόλου δεδομένων S με στόχο G, υπολογίζεται με βάση την ακόλουθη εξίσωση:

$$Entropy(S) = -\sum_{i=1}^{c} P_i * log_2(P_i)\,, [45]$$

όπου Pi είναι η αναλογία του αριθμού δειγμάτων του υποσυνόλου προς την συνολική τιμή του i-ου χαρακτηριστικού.

Το Κέρδος Πληροφορίας είναι το μέτρο που χρησιμοποιείται για τη διαίρεση των δεδομένων σε σύνολα και συχνά αναφέρεται και ως αμοιβαία πληροφορία (mutual information). Όσο υψηλότερη είναι η τιμή του, τόσο καλύτερη είναι η πληροφορία που παρέχει ένα χαρακτηριστικό για το διαχωρισμό των δεδομένων. Ουσιαστικά, η τιμή του Information Gain προσδιορίζει πόσο η γνώση μιας τυχαίας μεταβλητής συμβάλλει στη μείωση της αβεβαιότητας σε ένα σύνολο δεδομένων. Ορίζεται όπως παρακάτω όπου A: το χαρακτηριστικό που εξετάζεται, V(A): το εύρος τιμών του χαρακτηριστικού A, S: το συνολικό σύνολο δεδομένων, Sv: το υποσύνολο των δεδομένων όπου το χαρακτηριστικό A έχει την τιμή v.

$$Gain(S, A) = \sum_{v \in V(A)} \frac{|S_V|}{|S|} \, Entropy\,(S_V)\,, [37]$$

Υπάρχουν διάφοροι αλγόριθμοι δέντρων αποφάσεων. Ένας από αυτούς είναι ο ID3 (Iterative Dichotomiser 3) ο οποίος αναπτύχθηκε το 1986 από τον Ross Quinlan. Βασίζεται σε μια απληστική μέθοδο από πάνω προς τα κάτω (top-down greedy search) για να επιλέγει το καλύτερο χαρακτηριστικό σε κάθε κόμβο του δέντρου, χρησιμοποιώντας το Information Gain ως μέτρο. Ωστόσο, δέχεται μόνο κατηγορηματικά χαρακτηριστικά για την κατασκευή του δέντρου και δεν είναι ανθεκτικός στον θόρυβο, κάτι που σημαίνει ότι απαιτεί εντατική προεπεξεργασία δεδομένων [38].

Το C4.5 είναι ο διάδοχος του ID3 και αφαίρεσε τον περιορισμό ότι τα χαρακτηριστικά πρέπει να είναι κατηγορικά ορίζοντας δυναμικά ένα διακριτό χαρακτηριστικό (με βάση αριθμητικές μεταβλητές) που διαιρεί την τιμή συνεχούς χαρακτηριστικού σε ένα διακριτό σύνολο διαστημάτων. Το C4.5 μετατρέπει τα εκπαιδευμένα δέντρα (δηλαδή την έξοδο του αλγορίθμου ID3) σε σύνολα κανόνων if-then. Στη συνέχεια, αξιολογείται η ακρίβεια κάθε κανόνα για να καθοριστεί η σειρά με την οποία θα πρέπει να εφαρμοστούν. Το κλάδεμα γίνεται αφαιρώντας την προϋπόθεση ενός κανόνα, εάν η ακρίβεια του κανόνα βελτιωθεί χωρίς αυτόν.

Το CART (Δένδρα ταξινόμησης και παλινδρόμησης) είναι πολύ παρόμοιο με το C4.5, αλλά διαφέρει στο ότι υποστηρίζει αριθμητικές μεταβλητές στόχου (παλίνδρομο) και δεν υπολογίζει σύνολα κανόνων. Το CART κατασκευάζει δυαδικά δέντρα χρησιμοποιώντας το χαρακτηριστικό και το όριο που αποδίδουν το μεγαλύτερο κέρδος πληροφοριών σε κάθε κόμβο. Το scikit-learn χρησιμοποιεί μια βελτιστοποιημένη έκδοση του αλγορίθμου CART. Ωστόσο, η εφαρμογή scikit-learn δεν υποστηρίζει κατηγορικές μεταβλητές προς το παρόν.

Η μέθοδος DecisionTreeClassifier του sklearn δέχεται ως όρισμα το κριτήριο επιλογή χαρακτηριστικών "gini" ή "entropy" καθώς και το μέγιστο βάθος που επιθυμούμε να έχει το δένδρο (max_depth). Το Scikit-Learn χρησιμοποιεί τον αλγόριθμο CART, ο οποίος παράγει μόνο δυαδικά δέντρα: οι μη φυλλικοί κόμβοι έχουν πάντα δύο παιδιά (δηλαδή, οι ερωτήσεις έχουν μόνο ναι/όχι απαντήσεις). Ωστόσο, άλλοι αλγόριθμοι όπως ο ID3 μπορούν να παράγουν Δέντρα απόφασης με κόμβους που έχουν περισσότερα από δύο παιδιά. Η κλάση DecisionTreeClassifier έχει μερικές άλλες παραμέτρους που περιορίζουν ομοίως το σχήμα του Δέντρου Αποφάσεων: criterion: κριτήριο επιλογή χαρακτηριστικών "gini" ή "entropy", max_depth (η προεπιλεγμένη τιμή είναι None, που σημαίνει απεριόριστο), min_samples_split (ο ελάχιστος αριθμός δειγμάτων που πρέπει να έχει ένας κόμβος για να μπορέσει να διαχωριστεί), min_samples_leaf (ο ελάχιστος αριθμός δειγμάτων που πρέπει να έχει ένας κόμβος φύλλου), min_weight_fraction_leaf (ίδιο με το min_samples_leaf αλλά εκφράζεται ως κλάσμα του συνολικού αριθμού σταθμισμένων δειγμάτων, max_leaf_nodes (μέγιστος αριθμός κόμβων φύλλου), max_features (μέγιστος αριθμός χαρακτηριστικών που αξιολογούνται για διαχωρισμό σε κάθε κόμβο).

Για την συγκεκριμένη εργασία, εφαρμόστηκε ο αλγόριθμος Decision Tree, αφού προηγουμένως πραγματοποιήθηκε διαδικασία Grid Search για τον εντοπισμό των βέλτιστων υπερπαραμέτρων του. Στο πλαίσιο αυτό, δοκιμάστηκαν διαφορετικές τιμές για τις παραμέτρους max_depth, min_samples_split και min_samples_leaf. Το βέλτιστο μοντέλο που προέκυψε από την αναζήτηση χρησιμοποιήθηκε για την πρόβλεψη στα δεδομένα του συνόλου δοκιμής.

2.4 Επιλογή χαρακτηριστικών

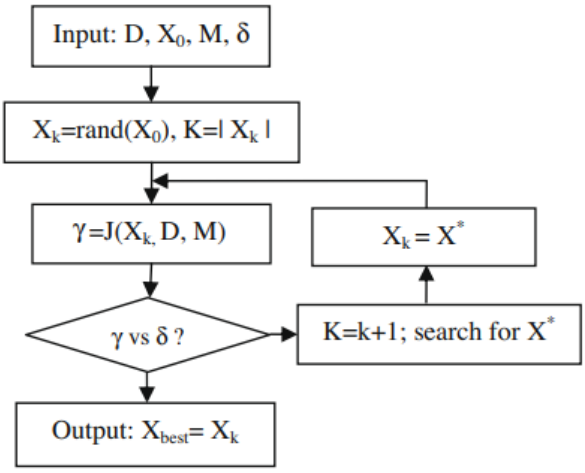
Ένα από τα βασικά ζητήματα στην ανάλυση και εξόρυξη βιοϊατρικών δεδομένων είναι η λεγόμενη «κατάρα της διαστατικότητας» (curse of dimensionality) [40-42], ιδιαίτερα τα βιοϊατρικά δεδομένα έχουν συχνά πολλά χαρακτηριστικά αλλά σχετικά λίγα δείγματα. Είναι αντιληπτό ότι η πολλή πληροφορία δεν συνεπάγεται ποιοτική πληροφορία. Επομένως, η παρουσία άσχετων χαρακτηριστικών σε έναν μεγάλο, πολυδιάστατο χώρο δεδομένων μειώνει την ακρίβεια των ταξινομητών και δυσχεραίνει την εξαγωγή χρήσιμης πληροφορίας από τα δεδομένα αυτά. Για αυτό τον λόγο, είναι απαραίτητο να αποκλείονται τα μη σχετικά χαρακτηριστικά. Η διαδικασία αυτή ονομάζεται επιλογή χαρακτηριστικών (feature selection) και έχει εξελιχθεί σε σημαντικό κομμάτι της εξόρυξης δεδομένων στον τομέα Βιοϊατρικής. Μέσω αυτής, επιλέγονται οι στήλες που θα τροφοδοτήσουν τα μοντέλα μηχανικής μάθησης, διευκολύνοντας κατά αυτόν τον τρόπο την οπτικοποίηση των δεδομένων και την καλύτερη κατανόηση των μοντέλων αυτών. Παράλληλα, μειώνονται οι απαιτήσεις μέτρησης και αποθήκευσης των δεδομένων και ελαχιστοποιείται το κόστος διαχείρισης των βάσεων δεδομένων.

Ο βασικός στόχος της επιλογής χαρακτηριστικών είναι η εύρεση ενός βέλτιστου υποσυνόλου χαρακτηριστικών από το αρχικά σύνολο που να οδηγεί στην δημιουργία ενός αποδοτικού μοντέλου ταξινόμησης. Στην ουσία επιλέγεται ένας αριθμός από τα ήδη υπάρχοντα χαρακτηριστικά, ελαχιστοποιώντας κάποια συνάρτηση κόστους. Η επιλογή χαρακτηριστικών με υψηλή διακριτική ικανότητα είναι κρίσιμη για τη σωστή λειτουργία ενός ταξινομητή, καθώς επηρεάζει άμεσα την απόδοσή του. Τα χαρακτηριστικά πρέπει να διαχωρίζουν αποτελεσματικά τις διαφορετικές τάξεις, λαμβάνοντας μακρινές τιμές μεταξύ διαφορετικών κατηγοριών και κοντινές τιμές μέσα στην ίδια τάξη, ώστε να εξασφαλίζεται σαφής διάκριση.

Τις τελευταίες δεκαετίες, έχει διεξαχθεί εκτενής σχετική έρευνα από ερευνητές από πολυεπιστημονικούς τομείς, συμπεριλαμβανομένων των στατιστικών, της αναγνώρισης προτύπων, της μηχανικής μάθησης και της εξόρυξης δεδομένων [43]. Δεκάδες μέθοδοι επιλογής χαρακτηριστικών έχουν αναπτυχθεί τα τελευταία χρόνια και αυτές μπορούν να χωριστούν σε τρεις βασικές κατηγορίες: (i) filter methods – μέθοδοι φιλτραρίσματος, (ii) wrapper methods – μέθοδοι περιτυλίγματος και (iii) hybrid methods – υβριδικές μέθοδοι. [44-46 & 20]. Παρακάτω, παρουσιάζεται η διαδικασία για κάθε μια εκ των 3 προσεγγίσεων επιλογής χαρακτηριστικών, όπου το D είναι το σύνολο δεδομένων εκπαίδευσης με το αρχικό σύνολο χαρακτηριστικών, το X best είναι το βέλτιστο υποσύνολο χαρακτηριστικών που πρέπει να επιλεγεί και το J(Xk) υποδηλώνει μια συνάρτηση αξιολόγησης για τη μέτρηση της απόδοσης ενός υποσυνόλου χαρακτηριστικών Xk, με βάση την ανεξάρτητη δοκιμή (M) ή τον αλγόριθμο μηχανικής μάθησης (A), αντίστοιχα, σε μεθόδους φίλτρου ή περιτυλίγματος.

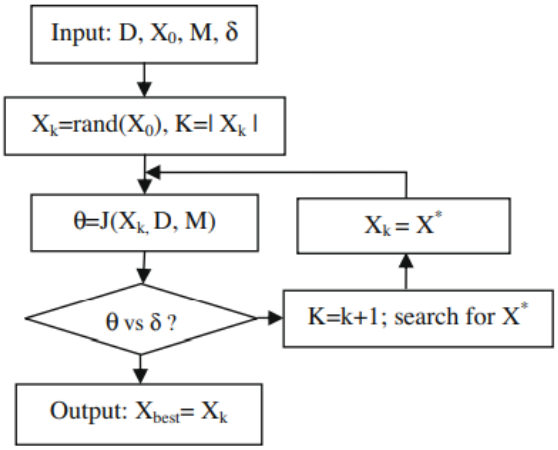
Η μέθοδος φιλτραρίσματος αναζητά σημαντικά χαρακτηριστικά εξετάζοντας τα χαρακτηριστικά κάθε μεμονωμένου feature χρησιμοποιώντας μια ανεξάρτητη δοκιμή. Όπως φαίνεται στην εικόνα 8, ο αλγόριθμος φίλτρου ξεκινά την αναζήτηση από ένα υποσύνολο χαρακτηριστικών X0 δηλαδή ένα κενό σύνολο ή οποιοδήποτε τυχαία επιλεγμένο υποσύνολο και ακολουθεί τα επόμενα βήματα: (i) αξιολόγηση του τρέχοντος υποσυνόλου χαρακτηριστικών X* χρησιμοποιώντας μια ανεξάρτητη μέθοδο δοκιμής (M) και (ii) σύγκριση με το καλύτερο υποσύνολο χαρακτηριστικών που ελήφθη στο προηγούμενο βήμα, το Xk1. Εάν διαπιστωθεί ότι είναι καλύτερο, θεωρείται ως το τρέχον καλύτερο υποσύνολο. Εδώ k = |Xk| είναι ο αριθμός των χαρακτηριστικών στο Xk που αντιπροσωπεύει την ιδιότητα του υποσυνόλου χαρακτηριστικών.

Η διαδικασία αυτή επαναλαμβάνεται μέχρι να ικανοποιηθεί ένα προκαθορισμένο κριτήριο τερματισμού δ1, όπως: (1) η προσθήκη ή διαγραφή χαρακτηριστικών να μην βελτιώνει περαιτέρω την ποιότητα του υποσυνόλου χαρακτηριστικών, (2) να επιτευχθεί το επιθυμητό επίπεδο απόδοσης του μοντέλου, ή (3) να φτάσει η διαδικασία σε ένα όριο, όπως ο μέγιστος αριθμός επαναλήψεων ή ο ελάχιστος αριθμός χαρακτηριστικών. Όταν ικανοποιηθεί κάποιο από αυτά τα κριτήρια, ο αλγόριθμος σταματά και εξάγει το καλύτερο υποσύνολο χαρακτηριστικών που εντόπισε κατά τη διάρκεια της διαδικασίας (Xbest), το οποίο θεωρείται το βέλτιστο βάσει των προκαθορισμένων απαιτήσεων [47].



Εικόνα 2.7: Μέθοδος φιλτραρίσματος για επιλογή χαρακτηριστικών [47]

Η μέθοδος περιτυλίγματος, όπως φαίνεται στο Σχήμα 1(β), εφαρμόζει έναν συγκεκριμένο αλγόριθμο μηχανικής μάθησης όπως το δέντρο αποφάσεων ή support vector machine (που συμβολίζεται ως A). Για κάθε επανάληψη της αναζήτησης χαρακτηριστικών, η απόδοση του υποσυνόλου χαρακτηριστικών X* αξιολογείται από την ποιότητα του μοντέλου ταξινόμησης που αντιστοιχεί στα χαρακτηριστικά του X*. Η απόδοση ταξινόμησης του X* συγκρίνεται με το καλύτερο υποσύνολο χαρακτηριστικών που ελήφθη κατά τα προηγούμενα βήματα. Εάν η απόδοση είναι καλύτερη από την προηγούμενη, τότε Xk = X*. Αυτή η διαδικασία διεξάγεται βήμα προς βήμα μέχρι να εκπληρωθεί ένα προκαθορισμένο κριτήριο δ2, όπως συζητήθηκε παραπάνω.

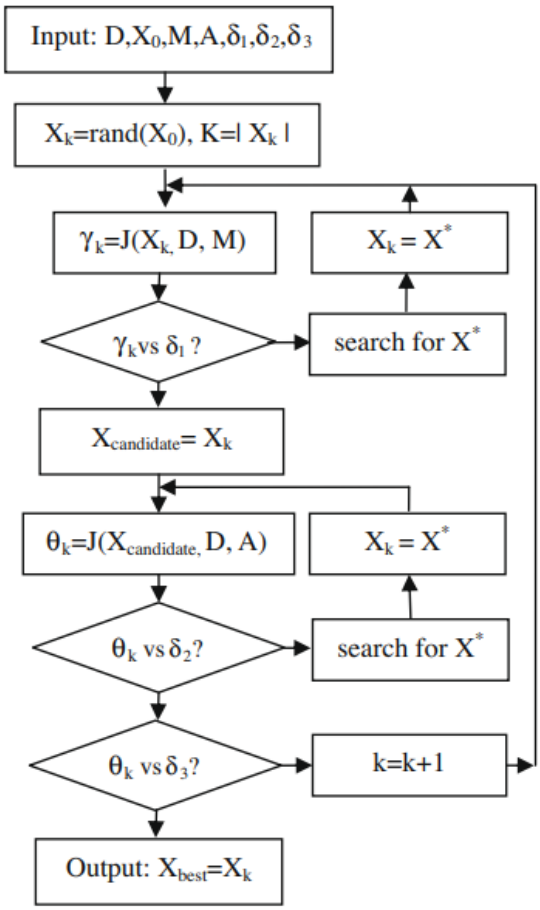


Εικόνα 2.8: Μέθοδος περιτυλίγματος για επιλογή χαρακτηριστικών [47]

Η τεχνική φιλτραρίσματος πλεονεκτεί στο γεγονός ότι μπορεί να επεξεργαστεί αποτελεσματικά δεδομένα που περιέχουν πολλά χαρακτηριστικά, ότι είναι γρήγορη υπολογιστικά και ανεξάρτητη από τον αλγόριθμο μάθησης. Ωστόσο, αγνοεί την αλληλεπίδραση με τον ταξινομητή και την εξάρτηση μεταξύ των χαρακτηριστικών, γεγονός που μπορεί να οδηγήσει σε μεταβαλλόμενη απόδοση ταξινόμησης σε διαφορετικούς αλγόριθμους. Από την άλλη πλευρά, η μέθοδος wrapper παράγει συχνά καλύτερους ταξινομητές, καθώς λαμβάνει υπόψη τις εξαρτήσεις μεταξύ χαρακτηριστικών και τη συλλογική συνεισφορά τους στο μοντέλο. Ωστόσο, η τελευταία έχει αυξημένο κίνδυνο υπερπροσαρμογής (overfitting) ενώ παράλληλα είναι υπολογιστικά πιο απαιτητική και ειδικά σε μεγάλα σύνολα χαρακτηριστικών.

Για να αξιοποιηθούν τα πλεονεκτήματα και των δύο προσεγγίσεων, η υβριδική τεχνική συνδυάζει τις παραπάνω 2 μεθόδους. Η διαδικασία ξεκινά από ένα αρχικό σύνολο χαρακτηριστικών (X0) και χρησιμοποιεί τη μέθοδο filter για να επιλέξει υποψήφια χαρακτηριστικά βασιζόμενη σε ανεξάρτητα κριτήρια. Στη συνέχεια, η μέθοδος wrapper

αξιολογεί αυτά τα υποψήφια χαρακτηριστικά χρησιμοποιώντας έναν συγκεκριμένο αλγόριθμο μάθησης (A) και το αντίστοιχο κριτήριο δ2. Όταν βρεθεί το καλύτερο υποσύνολο χαρακτηριστικών για συγκεκριμένο αριθμό χαρακτηριστικών (k), η συνολική απόδοση της ταξινόμησης αξιολογείται βάσει ενός τρίτου κριτηρίου (δ3). Εάν η απόδοση πληροί το συγκεκριμένο κριτήριο δ3, η διαδικασία επιλογής χαρακτηριστικών ολοκληρώνεται και επιστρέφει το τρέχον καλύτερο υποσύνολο ως το βέλτιστο. Διαφορετικά, η διαδικασία συνεχίζεται, αυξάνοντας το πλήθος των χαρακτηριστικών (k+1) με την προσθήκη ενός νέου χαρακτηριστικού από τα εναπομείναντα, και επαναλαμβάνεται η παραπάνω διαδικασία.



Εικόνα 2.9: Υβριδική μέθοδος για επιλογή χαρακτηριστικών [47]

2.4.1 Recursive Feature Elimination (RFE)

Ένας αλγόριθμος επιλογής χαρακτηριστικών είναι ο Recursive Feature Elimination (RFE). Ο RFE ακολουθεί την wrapper μέθοδο και βασίζεται σε επαναλαμβανόμενη αφαίρεση χαρακτηριστικών (backward selection). Στόχος του είναι να εντοπίσει τον βέλτιστο συνδυασμό χαρακτηριστικών για την επίτευξη υψηλής ακρίβειας σε ένα μοντέλο πρόβλεψης. Η διαδικασία ξεκινά με το πλήρες σύνολο χαρακτηριστικών και χρησιμοποιεί επαναληπτική αφαίρεση χαρακτηριστικών που δεν συμβάλλουν θετικά ή ακόμα και επιβαρύνουν την ακρίβεια του μοντέλου.

Στη συγκεκριμένη εργασία, ο RFE υλοποιήθηκε μαζί με έναν αλγόριθμο Random Forest, RF-RFE. Το Random Forest αξιολογεί τη σημασία κάθε χαρακτηριστικού, και στη συνέχεια το RFE αφαιρεί διαδοχικά τα λιγότερο σημαντικά χαρακτηριστικά, έως ότου βρεθεί το βέλτιστο υποσύνολο. Η μέθοδος αυτή είναι ιδανική για την αναγνώριση αλληλεπιδράσεων μεταξύ χαρακτηριστικών, αλλά είναι υπολογιστικά δαπανηρή, καθώς απαιτεί εκπαίδευση του μοντέλου σε κάθε επανάληψη. Ωστόσο, η χρήση του RFE με Random Forest παρέχει έναν

αξιόπιστο και προσαρμόσιμο τρόπο για τη βελτιστοποίηση της ακρίβειας σε σύνολα δεδομένων με μεγάλο αριθμό χαρακτηριστικών [48].

Η επιλογή του RFE σε συνδυασμό με τον αλγόριθμο Random Forest κρίθηκε ως η καταλληλότερη μέθοδος για την παρούσα εργασία, δεδομένης της φύσης των ιατρικών δεδομένων. Ο RFE, ως wrapped method, προσαρμόζεται στον ίδιο τον αλγόριθμο ταξινόμησης, εξασφαλίζοντας ότι η επιλογή χαρακτηριστικών βασίζεται στην απόδοση του μοντέλου. Σε αντίθεση με τις filter methods, που δεν λαμβάνουν υπόψη τις σχέσεις μεταξύ των χαρακτηριστικών, ο RFE εντοπίζει και διατηρεί κρίσιμα χαρακτηριστικά που σχετίζονται με τη δομή των δεδομένων, μειώνοντας ταυτόχρονα τον θόρυβο και βελτιώνοντας τη γενίκευση. Επιπλέον, είναι ιδιαίτερα αποτελεσματικός για την αναγνώριση μη γραμμικών σχέσεων και αλληλεπιδράσεων μεταξύ χαρακτηριστικών, γεγονός που είναι σημαντικό στις ιατρικές εφαρμογές. Συγκριτικά, οι hybrid methods μπορεί να είναι υπερβολικά πολύπλοκες, ενώ ο RFE προσφέρει μια ισορροπία μεταξύ ακρίβειας, αποδοτικότητας και ερμηνευσιμότητας. Ως εκ τούτου, ο RFE με Random Forest επιλέχθηκε για να παρέχει αξιόπιστα αποτελέσματα, διατηρώντας την ακρίβεια και την ερμηνευσιμότητα που απαιτείται σε εφαρμογές ιατρικής ανάλυσης δεδομένων.

2.5 Τρόποι Βελτίωσης Απόδοσης

Η αποτελεσματική εκπαίδευση και αξιολόγηση μοντέλων Μηχανικής Μάθησης απαιτεί την υιοθέτηση κατάλληλων τεχνικών που βελτιστοποιούν την απόδοση τους και εξασφαλίζουν τη γενίκευση σε νέα δεδομένα. Στο παρόν κεφάλαιο εξετάζονται διάφορες τεχνικές που συμβάλλουν στην αντιμετώπιση προκλήσεων όπως η υπερπροσαρμογή, η υποεκπαίδευση και η ανισοκατανομή κλάσεων. Οι τεχνικές αυτές περιλαμβάνουν, μεταξύ άλλων, τη χρήση διασταυρούμενης επικύρωσης, τη ρύθμιση υπερπαραμέτρων, την κανονικοποίηση δεδομένων και την εφαρμογή μεθόδων επαναδειγματοληψίας. Επιπλέον, εξετάζονται πιο εξειδικευμένες τεχνικές όπως η πρώιμη διακοπή, η μείωση διαστάσεων, η κανονικοποίηση παραμέτρων καθώς και η αύξηση δεδομένων.

Ο συνδυασμός αυτών των μεθόδων επιτρέπει την καλύτερη προσαρμογή του μοντέλου στα δεδομένα, τη διατήρηση της ισορροπίας ανάμεσα στη μεροληψία και τη διακύμανση και τη βελτίωση της συνολικής απόδοσης, ακόμα και σε περιβάλλοντα με περιορισμένους πόρους ή με ανισόρροπες κατανομές δεδομένων.

2.5.1 Υπερπροσαρμογή και Υποπροσαρμογή

Η αποδοτικότητα ενός μοντέλου μηχανικής μάθησης εξαρτάται από την αντιμετώπιση κρίσιμων ζητημάτων κατά τη διάρκεια της εκπαίδευσής του, όπως η επιλογή του κατάλληλου μοντέλου, η ρύθμιση των παραμέτρων, η σωστή διαίρεση του συνόλου δεδομένων, η εφαρμογή διασταυρούμενης επικύρωσης, ισορροπία μεταξύ μεροληψίας και διακύμανσης (bias-variance trade-off) καθώς και η αποφυγή προβλημάτων όπως η υπερπροσαρμογή (overfitting) και η υποεκπαίδευση (underfitting).

Τα σφάλματα σε ένα μοντέλο επιβλεπόμενης μάθησης διαχωρίζονται σε δύο βασικές κατηγορίες: σφάλματα λόγω μεροληψίας (bias) και λόγω διακύμανσης (variance). Ως «bias error» ορίζεται ως η διαφορά μεταξύ της αναμενόμενης (ή μέσης) πρόβλεψης του μοντέλου μας και της σωστής τιμής την οποία προσπαθεί να προβλέψει. Από την άλλη, ως «variance error» αναφέρεται η μεταβλητότητα μιας πρόβλεψης ενός μοντέλου για ένα συγκεκριμένο σύνολο δεδομένων [48].

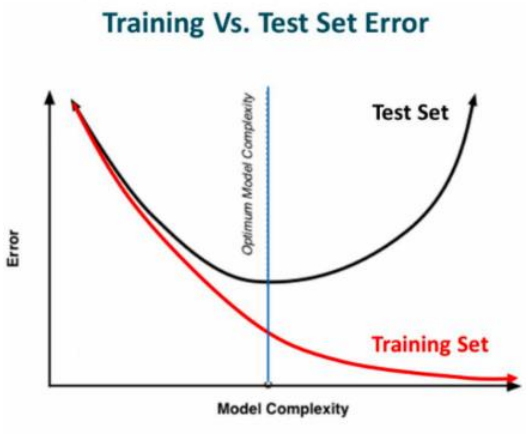
Η εμφάνιση των προαναφερθέντων σφαλμάτων εξαρτάται άμεσα από την πολυπλοκότητα του μοντέλου. Ο τρόπος με τον οποίο εκτιμάται η πολυπλοκότητα ενός μοντέλου είναι η χωρητικότητά (capacity) του. Η χωρητικότητα είναι μια έννοια που καθορίζει την ικανότητα του μοντέλου να γενικεύει και σχετίζεται άμεσα με το χώρο υποθέσεων (hypothesis space) του μοντέλου [49].

Όταν ένα μοντέλο είναι πιο πολύπλοκο από όσο χρειάζεται για να επιλύσει το πρόβλημα, τείνει να "απομνημονεύει" τις λεπτομέρειες και τις ιδιαιτερότητες των δεδομένων εκπαίδευσης. Αυτό έχει ως αποτέλεσμα να μην μπορεί να γενικεύσει σωστά σε νέα δεδομένα, κάτι που φαίνεται από τη μεγάλη διαφορά μεταξύ του σφάλματος εκπαίδευσης και του σφάλματος ελέγχου, όπως παρατηρείται στο δεξί μέρος της παρακάτω εικόνας. Σε αυτή την περίπτωση, λέμε ότι το μοντέλο παρουσιάζει φαινόμενα υπερπροσαρμογής (overfitting). Σε τέτοιες περιπτώσεις, το μοντέλο μπορεί να κάνει διαφορετικές προβλέψεις για παραπλήσια δεδομένα, εμφανίζοντας έτσι υψηλή διακύμανση (high variance).

Από την άλλη, όταν η πολυπλοκότητα του μοντέλου είναι μικρή για το πρόβλημα που καλείται να μάθει, τότε το σφάλμα εκπαίδευσης είναι μεγάλο και το μοντέλο δεν μπορεί να ανακαλύψει τις στατιστικές ιδιότητες των δεδομένων (αριστερό τμήμα της παρακάτω εικόνας). Σε αυτή την περίπτωση λέμε ότι το μοντέλο εμφανίζει χαρακτηριστικά υποπροσαγμογής (underfitting) καθώς έχει υπεραπλουστεύσει τα δεδομένα, εμφανίζοντας έτσι υψηλή μεροληψία (high bias).

Στην ιδανική περίπτωση, το μοντέλο έχει την πολυπλοκότητα που απαιτεί το πρόβλημα, οπότε και μπορεί να μάθει τα δεδομένα εκπαίδευσης (εμφανίζοντας χαμηλό σφάλμα εκπαίδευσης) αλλά μπορεί ταυτόχρονα και να γενικεύσει (εμφανίζοντας παραπλήσιο σφάλμα

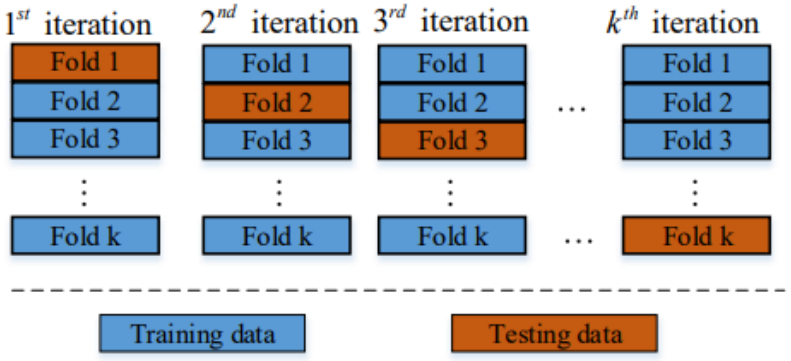
ελέγχου). Αυτή η κατάσταση απεικονίζεται στο κέντρο του σχήματος (κάθετη μπλε γραμμή), οπότε και θεωρούμε ότι το μοντέλο έχει προσαρμοστεί (fit) στο πρόβλημα.



Εικόνα 2.10: Bias – variance tradeoff [50]

2.5.2 Διασταυρούμενη επικύρωση (kCV)

Η τεχνική k-Fold Cross Validation (KCV) είναι μία από τις πιο δημοφιλείς μεθόδους επαναδειγματοληψίας (resampling techniques) στη μηχανική μάθηση. Αξιοποιείται για την εκτίμηση της απόδοσης ενός μοντέλου και θεωρείται απλή, αποτελεσματική και αξιόπιστη [51]. Στο k-fold CV, το σύνολο δεδομένων διαιρείται σε k υποσύνολα (folds). Σε κάθε επανάληψη, ένα από τα υποσύνολα χρησιμοποιείται ως σύνολο δοκιμής (testing data), ενώ τα υπόλοιπα k-1 υποσύνολα χρησιμοποιούνται ως σύνολα εκπαίδευσης (training data) [52]. Η διαδικασία αυτή επαναλαμβάνεται έως ότου αξιολογηθεί ολόκληρο το σύνολο δεδομένων. Τα αποτελέσματα του k-fold CV υπολογίζονται συνήθως ως ο μέσος όρος των επιδόσεων του μοντέλου σε όλες τις επαναλήψεις. Η παρακάτω εικόνα παρουσιάζει μια επισκόπηση της διαδικασίας αυτής.



Εικόνα 2.11: Τεχνική k-Fold Cross Validation [53]

Μια συνηθισμένη επιλογή για την τιμή του k είναι το 10, καθώς θεωρείται αποτελεσματική [53]. Σε μικρά σύνολα δεδομένων, όμως, μπορεί να είναι προτιμότερο να αυξήσουμε την τιμή του k, ώστε περισσότερα δεδομένα να χρησιμοποιούνται για την εκπαίδευση σε κάθε επανάληψη, γεγονός που μειώνει τη μεροληψία (bias) κατά την εκτίμηση της απόδοσης. Από την άλλη, υψηλές τιμές k αυξάνουν τον χρόνο υπολογισμού και μπορεί να προκαλέσουν μεγαλύτερη διασπορά (variance) στις εκτιμήσεις, επειδή τα folds θα διαφέρουν περισσότερο μεταξύ τους [52].

Η χρήση του k-fold cross-validation είναι απαραίτητη προκειμένου να διασφαλιστεί ότι το μοντέλο αξιολογείται αξιόπιστα και μπορεί να γενικεύσει καλά σε νέα δεδομένα. Με τη μέθοδο αυτή, όλα τα δεδομένα συμμετέχουν τόσο στην εκπαίδευση όσο και στην επικύρωση, μειώνοντας τον κίνδυνο υπερπροσαρμογής και παράλληλα εξασφαλίζοντας ότι τα αποτελέσματα δεν επηρεάζονται από τυχαίες αποκλίσεις στα δεδομένα. Επιπλέον, η τεχνική αυτή είναι ιδιαίτερα χρήσιμη όταν έχουμε περιορισμένο μέγεθος δεδομένων καθώς τα αξιοποιεί πλήρως, παρέχοντας μια πιο ρεαλιστική εκτίμηση της απόδοσης του μοντέλου.

2.5.3 Ρύθμιση Υπερπαραμέτρων

Η ρύθμιση των υπερπαραμέτρων (Hyperparameter Tuning) αποτελεί κρίσιμο βήμα για την επίτευξη υψηλής απόδοσης στα μοντέλα μηχανικής μάθησης, ιδιαίτερα στα μη παραμετρικά μοντέλα [54]. Οι προκαθορισμένες τιμές των υπερπαραμέτρων (default settings) σπάνια εγγυώνται τη βέλτιστη απόδοση, καθώς τα μοντέλα αυτά συχνά απαιτούν εξατομικευμένες ρυθμίσεις για να προσαρμοστούν στις ιδιαιτερότητες των δεδομένων. Εργαλεία όπως το grid search σε συνδυασμό με διασταυρούμενη επικύρωση επιτρέπουν τη συστηματική εξερεύνηση των τιμών των υπερπαραμέτρων, παρέχοντας τη δυνατότητα βελτιστοποίησης της ακρίβειας και της γενίκευσης του μοντέλου.

Στην παρούσα εργασία, χρησιμοποιήθηκε το GridSearchCV της βιβλιοθήκης scikit-learn το οποίο εκτελεί αναζήτηση για τις καλύτερες υπερπαραμέτρους ενός αλγορίθμου εκμάθησης μέσω διασταυρούμενης επικύρωσης (cross-validation). Χρησιμοποιεί τον αλγόριθμο μηχανικής μάθησης (estimator) για τον οποίο αναζητούνται οι βέλτιστες παράμετροι και το param_grid, το οποίο είναι ένα λεξικό ή λίστα λεξικών που περιέχει τα ονόματα των παραμέτρων και τις λίστες των τιμών που πρέπει να εξεταστούν. Εν συνεχεία, για κάθε συνδυασμό παραμέτρων, υπολογίζει την απόδοση του μοντέλου χρησιμοποιώντας την παράμετρο scoring (στρατηγική αξιολόγησης της απόδοσης) και την cv (στρατηγική διασταυρούμενης επικύρωσης). Ανάλογα με την τιμή του refit, το GridSearchCV μπορεί να επανεκπαιδεύσει το μοντέλο με τις καλύτερες παραμέτρους χρησιμοποιώντας το σύνολο των δεδομένων.

2.5.4 Κανονικοποίηση τιμών

Το scaling ή κανονικοποίηση είναι μία διαδικασία προεπεξεργασίας δεδομένων που χρησιμοποιείται για να τυποποιήσει τα χαρακτηριστικά σε ένα σύνολο δεδομένων. Βασικός στόχος της κανονικοποίησης είναι να εξασφαλίσει ότι όλα τα χαρακτηριστικά βρίσκονται στο ίδιο εύρος τιμών, επιτρέποντας στους αλγορίθμους μηχανικής μάθησης να λειτουργούν πιο αποτελεσματικά. Ορισμένοι αλγόριθμοι, όπως οι K-Nearest Neighbors (KNN) και Support Vector Machines (SVM), είναι ιδιαίτερα ευαίσθητοι στις κλίμακες των χαρακτηριστικών, και χαρακτηριστικά με μεγάλες διαφορές στα μεγέθη τους μπορεί να επηρεάσουν αρνητικά την απόδοσή τους [55].

Στην παρούσα εργασία, έγινε χρήση του StandardScaler() για τους αλγορίθμους Naive Bayes, K-Nearest Neighbors (KNN) και Support Vector Machines (SVM). Το StandardScaler στο Scikit-Learn είναι μια κλάση που χρησιμοποιείται για την κανονικοποίηση δεδομένων (scaling), με στόχο να τυποποιήσει τα χαρακτηριστικά σε ένα σύνολο δεδομένων αφαιρώντας τον μέσο όρο και κλίνοντας τις τιμές ώστε να έχουν μονάδα διακύμανση (unit variance). Αυτή η τυποποίηση (standardization) είναι απαραίτητη για αλγορίθμους μηχανικής μάθησης που επηρεάζονται από την κλίμακα των χαρακτηριστικών. Ο StandarScaler υπολογίζει το z-score για κάθε χαρακτηριστικό του συνόλου δεδομένων, εφαρμόζοντας τον τύπο $z = \frac{x-\mu}{\sigma}$, όπου είναι x: η τιμή ενός χαρακτηριστικού, μ: ο μέσος όρος του χαρακτηριστικού (υπολογίζεται από το training set), σ: η τυπική απόκλιση του χαρακτηριστικού (υπολογίζεται από το training set). Μετά την χρήση του, κάθε χαρακτηριστικό θα έχει μέση τιμή ίση με 0 και τυπική απόκλιση

ίση με 1. Επιπλέον, τα χαρακτηριστικά κανονικοποιούνται ξεχωριστά, ώστε να μην επηρεάζονται το ένα από το άλλο.

Σύμφωνα με την βιβλιογραφία, ο Naive Bayes βασίζεται στην κατανομή των δεδομένων και όχι στις αποστάσεις ή στις σχετικές κλίμακες των χαρακτηριστικών. Επομένως, η κανονικοποίηση συνήθως δεν επηρεάζει την απόδοσή του [46]. Ωστόσο, για το συγκεκριμένο σύνολο δεδομένων, εξετάστηκε και κρίθηκε απαραίτητο να εφαρμοστεί κανονικοποίηση. Παράλληλα, ο KNN βασίζεται σε μετρικές αποστάσεων για να βρει τα κοντινότερα σημεία. Αν τα χαρακτηριστικά έχουν διαφορετικές κλίμακες, τα χαρακτηριστικά με μεγαλύτερο εύρος τιμών θα κυριαρχούν στον υπολογισμό των αποστάσεων, επηρεάζοντας αρνητικά την ακρίβεια. Τέλος, το SVM χρησιμοποιεί πυρήνες (kernels) οι οποίοι βασίζονται σε υπολογισμούς αποστάσεων. Παρόμοια με τον KNN, οι διαφορές στις κλίμακες των χαρακτηριστικών μπορεί να επηρεάσουν τη βελτιστοποίηση της υπερεπίπεδης διαχωρισμού, μειώνοντας την ακρίβεια του μοντέλου. Επομένως, κανονικοποίηση διασφαλίζει ότι τα χαρακτηριστικά βρίσκονται στην ίδια κλίμακα και βελτιώνει την απόδοση.

2.5.5 Αντιμετώπιση Ανισοκατανομής Κλάσεων

Η ανισοκατανομή κλάσεων αποτελεί μια σημαντική πρόκληση στη Μηχανική Μάθηση, ειδικά σε προβλήματα ταξινόμησης. Εμφανίζεται όταν ορισμένες κλάσεις (πλειοψηφικές) κυριαρχούν στο σύνολο δεδομένων, ενώ άλλες (μειοψηφικές) είναι υποεκπροσωπούμενες. Αυτή η ανισορροπία μπορεί να προκαλέσει μεροληψία στα μοντέλα, τα οποία συχνά ευνοούν την πλειοψηφική κλάση εις βάρος της μειοψηφικής. Επιπλέον, οι συνήθεις μετρικές, όπως η ακρίβεια, μπορεί να οδηγήσουν σε λανθασμένα συμπεράσματα, καθώς αποκρύπτουν την κακή απόδοση στις μειοψηφικές κλάσεις. Άλλες προκλήσεις περιλαμβάνουν την επικάλυψη των κλάσεων, την ύπαρξη θορύβου στα δεδομένα και την υψηλή διαστατικότητα σε συνδυασμό με τον μικρό αριθμό δειγμάτων στις μειοψηφικές κλάσεις. Για την αντιμετώπιση αυτών των προκλήσεων, έχουν αναπτυχθεί τεχνικές, όπως η επαναδειγματοληψία (oversampling/undersampling), η εκμάθηση με κόστος (cost-sensitive learning) και οι μέθοδοι συνόλων (ensemble methods), οι οποίες επιτρέπουν την αποτελεσματική διαχείριση ανισοκατανεμημένων δεδομένων [56].

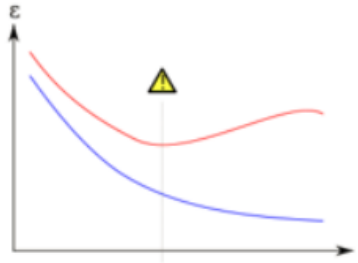
Στο oversampling, δημιουργούνται νέα δείγματα της μειοψηφικής κλάσης είτε μέσω τυχαίας αντιγραφής είτε με τη χρήση μεθόδων όπως το SMOTE, το οποίο δημιουργεί συνθετικά δείγματα βασισμένα σε γειτονικά δεδομένα της μειοψηφικής κλάσης. Από την άλλη, το undersampling περιλαμβάνει την αφαίρεση δειγμάτων από την πλειοψηφική κλάση για τη μείωση της ανισορροπίας. Παρότι το undersampling μειώνει το μέγεθος του συνόλου δεδομένων, είναι υπολογιστικά πιο αποδοτικό σε σύνολα δεδομένων με μεγάλο πλήθος παρατηρήσεων.

Μία από τις συχνά χρησιμοποιούμενες μεθόδους για την αντιμετώπιση της ανισοκατανομής κλάσεων είναι η Synthetic Minority Over-sampling Technique (SMOTE). Η συγκεκριμένη μέθοδος δημιουργεί νέα συνθετικά δείγματα για τη μειοψηφική κλάση, μέσω παρεμβολής ανάμεσα σε υπάρχοντα δείγματα της μειοψηφικής κλάσης. Με άλλα λόγια, το SMOTE προσθέτει νέα δεδομένα για την υποεκπροσωπούμενη κλάση, προκειμένου να βελτιώσει την ισορροπία στο σύνολο δεδομένων και να ενισχύσει την απόδοση των αλγορίθμων μηχανικής μάθησης [57]. Επίσης, το RandomOverSampler είναι μια τεχνική υπερδειγματοληψίας η οποία χρησιμοποιείται για την αντιμετώπιση προβλημάτων ανισοκατανομής κλάσεων στα δεδομένα. Πρόκειται για μέθοδο που ανήκει στη βιβλιοθήκη imbalanced-learn στο Python και εφαρμόζεται για να αυξηθεί το μέγεθος της μειοψηφικής κλάσης, δημιουργώντας αντίγραφα των υπαρχόντων δειγμάτων της μειοψηφικής κλάσης με τυχαίο τρόπο [57]. Αυτές οι δυο μέθοδοι εφαρμόστηκαν στα δεδομένα της παρούσης μελέτης ανάλογα με τις ανάγκες του συνόλου δεδομένων σε συνδυασμό με το εκάστοτε μοντέλο Μηχανικής Μάθησης.

2.5.6 Πρόσθετες τεχνικές βελτίωσης

2.5.6.1 Πρώιμος Τερματισμός

Ο Πρώιμος Τερματισμός (Early Stopping) είναι μια τεχνική που χρησιμοποιείται στη μηχανική μάθηση για να αποτρέψει την υπερπροσαρμογή, σταματώντας την εκπαίδευση ενός μοντέλου όταν η απόδοσή του στα δεδομένα επικύρωσης αρχίσει να επιδεινώνεται. Κατά τη διάρκεια της εκπαίδευσης, το μοντέλο βελτιώνει την ακρίβειά του στο σύνολο εκπαίδευσης, αλλά μετά από ένα σημείο μπορεί να αρχίσει να μαθαίνει και τον θόρυβο των δεδομένων, μειώνοντας την ικανότητά του να γενικεύει σε νέα δεδομένα. Το Early Stopping παρακολουθεί την απόδοση στο validation set και σταματά την εκπαίδευση όταν δεν υπάρχει περαιτέρω βελτίωση για έναν προκαθορισμένο αριθμό επαναλήψεων (patience rounds). [58].



Εικόνα 2.12: Validation error vs test error [58]

Χρησιμοποιείται ευρέως τόσο σε νευρωνικά δίκτυα όσο και σε Boosting αλγορίθμους (όπως Gradient Boosting και AdaBoost), όπου η συνεχής προσθήκη νέων βαρών ή δέντρων μπορεί να οδηγήσει σε υπερπροσαρμογή. Αντί να εκπαιδεύεται επ' αόριστον, το boosting συνεχίζει να προσθέτει δέντρα ή βάσεις μοντέλων μέχρι να βρει το ιδανικό σημείο διακοπής, το οποίο συνήθως καθορίζεται μέσω Cross-Validation ή προκαθορισμένων patience rounds. Έρευνες έχουν δείξει ότι το Early Stopping βελτιώνει τη γενικευσιμότητα, εξασφαλίζει σύγκλιση με συνέπεια και μειώνει το υπολογιστικό κόστος, καθιστώντας το απαραίτητο εργαλείο στα Boosting μοντέλα [59]. Στην πράξη, υποστηρίζεται από δημοφιλείς βιβλιοθήκες όπως XGBoost, LightGBM, CatBoost και Gradient Boosting (sklearn's Gradient Boosting Classifier/Regressor), μέσω της παραμέτρου early_stopping_rounds στις πρώτες και των validation_fraction, n_iter_no_change και tol στη sklearn, επιτρέποντας την αυτοματοποιημένη επιλογή του βέλτιστου αριθμού iterations και διασφαλίζοντας ότι η εκπαίδευση σταματά όταν δεν υπάρχει περαιτέρω βελτίωση στην απόδοση του μοντέλου.

2.5.6.2 Επαύξηση Δεδομένων

Η Επαύξηση Δεδομένων (Data Augmentation) είναι μια τεχνική που χρησιμοποιείται στη μηχανική μάθηση για τη δημιουργία επιπλέον δεδομένων από ένα υπάρχον σύνολο, με σκοπό τη βελτίωση της απόδοσης των μοντέλων και την αποφυγή της υπερπροσαρμογής [60]. Στα εικονογραφικά ή κειμενικά δεδομένα, αυτό επιτυγχάνεται με μετασχηματισμούς, όπως περιστροφή εικόνων, αλλαγές φωτεινότητας ή μετάθεση λέξεων. Στα δομημένα δεδομένα, όπως αυτά της ιατρικής, χρησιμοποιούνται στατιστικές μέθοδοι, προσθήκη θορύβου ή συνθετικά δεδομένα για να εμπλουτίσουν το σύνολο δεδομένων.

Ωστόσο, στα ιατρικά δεδομένα, η εφαρμογή αυτής της τεχνικής παρουσιάζει σημαντικούς περιορισμούς. Αρχικά, οι τυχαίες μεταβολές στα δεδομένα, όπως η αλλαγή ηλικίας ή αρτηριακής πίεσης, μπορεί να οδηγήσουν σε μη ρεαλιστικά κλινικά σενάρια, επηρεάζοντας την ακρίβεια του μοντέλου. Επιπλέον, οι ιατρικές βάσεις δεδομένων είναι συχνά περιορισμένες, ειδικά σε περιπτώσεις σπάνιων παθήσεων, καθιστώντας δύσκολη τη

δημιουργία συνθετικών δεδομένων που αντικατοπτρίζουν με ακρίβεια την πραγματικότητα. Παράλληλα, η διαχείριση αυτών των δεδομένων πρέπει να συμμορφώνεται με αυστηρούς κανονισμούς προστασίας προσωπικών δεδομένων, όπως οι GDPR και HIPAA, περιορίζοντας τη χρήση μεθόδων όπως τα Generative Adversarial Networks (GANs). Επιπλέον, η δημιουργία νέων δειγμάτων μέσω τεχνικών όπως το SMOTE ή η προσθήκη θορύβου απαιτεί ιδιαίτερη προσοχή, ώστε να μην αλλοιώνεται η κλινική σημασία των δεδομένων. Σε αντίθεση με εικόνες ή κείμενα, όπου οι μικρές τροποποιήσεις δεν επηρεάζουν το συνολικό νόημα, στα ιατρικά δεδομένα ακόμη και μια μικρή αλλαγή μπορεί να επηρεάσει δραστικά τη διάγνωση και τη θεραπεία. Επομένως, παρόλο που το Data Augmentation μπορεί να συμβάλει στη βελτίωση των μοντέλων, απαιτείται προσεκτική εφαρμογή ώστε να διατηρείται η ιατρική εγκυρότητα και αξιοπιστία των δεδομένων.

2.5.6.3 Μείωση Διαστάσεων

Η Μείωση Διαστάσεων (Dimensionality Reduction) είναι μια τεχνική στη μηχανική μάθηση που στοχεύει στη μείωση του αριθμού των χαρακτηριστικών ενός συνόλου δεδομένων, διατηρώντας παράλληλα τις πιο σημαντικές πληροφορίες. Όταν τα δεδομένα έχουν υψηλή διάσταση, η πολυπλοκότητα των μοντέλων αυξάνεται, ενώ παράλληλα επηρεάζεται η απόδοσή τους λόγω του φαινομένου της «κατάρας της διάστασης» (curse of dimensionality). Η μείωση διάστασης συμβάλλει στην αντιμετώπιση αυτών των προβλημάτων, βελτιώνοντας την αποδοτικότητα και την ακρίβεια των αλγορίθμων. Υπάρχουν δύο βασικές κατηγορίες τεχνικών μείωσης διάστασης. Η πρώτη κατηγορία είναι η μέθοδος επιλογής χαρακτηριστικών (Feature Selection), όπου επιλέγονται τα πιο σημαντικά χαρακτηριστικά με βάση συγκεκριμένα κριτήρια. Η συγκεκριμένη μέθοδος αναλύθηκε εκτενώς στο Κεφάλαιο 2.4, καθώς αποτελεί κεντρικό στοιχείο της παρούσας έρευνας. Η επιλογή χαρακτηριστικών συμβάλλει στη μείωση του όγκου των δεδομένων, βελτιώνοντας την αποδοτικότητα των αλγορίθμων μηχανικής μάθησης, ενώ παράλληλα διατηρεί τις κρίσιμες πληροφορίες που είναι απαραίτητες για την ακριβή διάγνωση. Η δεύτερη κατηγορία είναι η μέθοδος εξαγωγής χαρακτηριστικών (Feature Extraction), όπου νέα χαρακτηριστικά δημιουργούνται από τα υπάρχοντα δεδομένα, συχνά μέσω μετασχηματισμών ή συνδυασμού των αρχικών χαρακτηριστικών.

Μεταξύ των πιο διαδεδομένων τεχνικών μείωσης διάστασης είναι η Ανάλυση Κύριων Συνιστωσών (PCA), η οποία μετασχηματίζει τα δεδομένα σε ένα νέο σύνολο ανεξάρτητων μεταβλητών, διατηρώντας όσο το δυνατόν περισσότερη από την αρχική πληροφορία. Επίσης, η Ανάλυση Διακριτών Συνιστωσών (LDA) χρησιμοποιείται κυρίως για προβλήματα ταξινόμησης, βελτιώνοντας τη διαχωριστική ικανότητα των δεδομένων. Η μείωση διάστασης προσφέρει πολλαπλά πλεονεκτήματα. Πρώτον, επιταχύνει την εκπαίδευση και την εκτέλεση των αλγορίθμων, καθώς η επεξεργασία μικρότερου όγκου δεδομένων είναι πιο αποδοτική. Δεύτερον, βελτιώνει την οπτικοποίηση δεδομένων, επιτρέποντας την απεικόνιση πολύπλοκων δεδομένων σε δύο ή τρεις διαστάσεις, αποκαλύπτοντας κρυμμένα μοτίβα. Τέλος, συμβάλλει στην αποτροπή της υπερπροσαρμογής. αφού η μείωση των χαρακτηριστικών μειώνει την πιθανότητα το μοντέλο να μάθει θόρυβο αντί για χρήσιμα πρότυπα [61].

2.5.6.4 Κανονικοποίηση δεδομένων

Η κανονικοποίηση (Regularization) είναι μια θεμελιώδης τεχνική στη μηχανική μάθηση που στοχεύει στη βελτίωση της γενίκευσης ενός μοντέλου, μειώνοντας τον κίνδυνο υπερπροσαρμογής. Το overfitting προκύπτει όταν ένα μοντέλο απομνημονεύει τα δεδομένα εκπαίδευσης αντί να μάθει τα υποκείμενα μοτίβα, με αποτέλεσμα να έχει χαμηλή απόδοση σε νέα, μη ορατά δεδομένα. Η κανονικοποίηση αντιμετωπίζει αυτό το πρόβλημα προσθέτοντας όρους ποινής στη συνάρτηση κόστους, οι οποίοι περιορίζουν την πολυπλοκότητα του μοντέλου. Οι πιο διαδεδομένες τεχνικές κανονικοποίησης περιλαμβάνουν την L1 κανονικοποίηση (Lasso), την L2 κανονικοποίηση (Ridge) και τη συνδυαστική τους μορφή, γνωστή ως Elastic Net.

Στην πράξη, η L2 κανονικοποίηση προσθέτει έναν όρο ποινής που τιμωρεί μεγάλα βάρη, ενθαρρύνοντας πιο ομαλές και σταθερές προβλέψεις. Αντίθετα, η L1 κανονικοποίηση ενισχύει την αραιότητα του μοντέλου, μηδενίζοντας ορισμένα βάρη και επιτρέποντας την αυτόματη επιλογή των πιο σημαντικών χαρακτηριστικών. Αυτές οι τεχνικές είναι ιδιαίτερα χρήσιμες σε προβλήματα υψηλής διάστασης, όπου ο μεγάλος αριθμός χαρακτηριστικών μπορεί να οδηγήσει σε υπερπροσαρμογή. Η εφαρμογή κανονικοποίησης είναι κρίσιμη σε πολύπλοκα προβλήματα ταξινόμησης και παλινδρόμησης, βελτιώνοντας τη σταθερότητα, την ερμηνευσιμότητα και την αξιοπιστία των μοντέλων μηχανικής μάθησης. Επιπλέον, η κανονικοποίηση χρησιμοποιείται σε συνδυασμό με άλλες τεχνικές, όπως η διασταυρούμενη επικύρωση (cross-validation), ώστε να εξασφαλίζεται η βέλτιστη ισορροπία μεταξύ απόδοσης και γενίκευσης του μοντέλου [62].

3. Κεφάλαιο: Διάγνωση Εγκεφαλικού με Μηχανική Μάθηση

3.1 Σημασία διάγνωσης εγκεφαλικού

3.1.1 Ορισμός και Κατηγορίες Εγκεφαλικού Επεισοδίου

Το εγκεφαλικό επεισόδιο είναι η βλάβη που προκαλείται όταν η παροχή αίματος σε μία περιοχή του εγκεφάλου σταματήσει με αποτέλεσμα τα κύτταρα που δεν λαμβάνουν οξυγόνο να πεθαίνουν [63, 64]. Κατά την διάρκεια των αιώνων, οι επιστήμονες έχουν προσπαθήσει να δώσουν έναν επαρκή ορισμό για την έννοια του «εγκεφαλικού επεισοδίου». Ο ορισμός που δόθηκε από τον Παγκόσμιο Οργανισμό Υγείας-Π.Ο.Υ. (World Health Organization-WHO) το 1970 για την επεξήγηση του όρου είναι ως εξής: «τα ταχέως αναπτυγμένα κλινικά σημεία εστιακής (ή γενικευμένης) διαταραχής της εγκεφαλικής λειτουργίας, που διαρκούν περισσότερο από 24 ώρες ή οδηγούν στον θάνατο, χωρίς προφανή, αιτία πέραν της αγγειακής προέλευσης». Για να είναι δυνατόν να περιοριστούν οι αδυναμίες αυτού του ορισμού που είναι βασισμένος σε συμπτωματολογία, ο «American Heart Association and American Stroke Association» συμπεριέλαβε στο νέο ορισμό του την «σιωπηρή παθολογία» (silent pathology) τα «σιωπηρά» εμφράγματα του εγκεφάλου, του αμφιβληστριωδούς και του νωτιαίου μυελού καθώς και σιωπηρές εγκεφαλικές αιμορραγίες (‘silent’ brain, retinal and spinal infarcts and silent cerebral haemorrhages) [65].

Τα εγκεφαλικά επεισόδια αποτελούν μία από τις κυριότερες αιτίες αναπηρίας και θανάτου παγκοσμίως. Συγκεκριμένα, το εγκεφαλικό επεισόδιο αποτελεί τη δεύτερη μεγαλύτερη αιτία θνησιμότητας παγκοσμίως, μετά από τη στεφανιαία νόσο. Σύμφωνα με τον Παγκόσμιο Οργανισμό Εγκεφαλικού (World Stroke Organization), καταγράφονται πάνω από 13,7 εκατομμύρια εγκεφαλικά επεισόδια κάθε χρόνο [66]. Οι πιο συχνές μορφές είναι το ισχαιμικό εγκεφαλικό επεισόδιο (Ischemic Stroke) και το αιμορραγικό εγκεφαλικό επεισόδιο (Hemorrhagic Stroke), καθώς και ειδικές υποκατηγορίες όπως το εγκεφαλικό λόγω απόφραξης μεγάλου αγγείου (Large Vessel Occlusion - LVO), η υπαραχνοειδής αιμορραγία (Subarachnoid Hemorrhage - SAH) καθώς και το παροδικό ισχαιμικό επεισόδιο (Transient Ischemic Attack - TIA).

Το ισχαιμικό εγκεφαλικό επεισόδιο συμβαίνει όταν η ροή του αίματος προς τον εγκέφαλο διακόπτεται λόγω απόφραξης ή στένωσης μιας αρτηρίας, προκαλώντας ισχαιμία και ενδεχομένως μόνιμη νευρωνική βλάβη. Ο πιο συχνός μηχανισμός είναι η εμβολή, που προέρχεται κυρίως από καρδιογενείς θρόμβους λόγω αρρυθμιών, όπως η κολπική μαρμαρυγή, ή από αρτηριο-αρτηριακά έμβολα που αποσπώνται από αθηρωματικές πλάκες. Επιπλέον, μια σοβαρή μορφή ισχαιμικού εγκεφαλικού είναι η απόφραξη μεγάλου αγγείου (Large Vessel Occlusion - LVO Stroke), όπου η μέση εγκεφαλική αρτηρία (MCA) ή η εσωτερική καρωτίδα (ICA) φράσσεται, προκαλώντας εκτεταμένη εγκεφαλική βλάβη [67].

Το αιμορραγικό εγκεφαλικό επεισόδιο συμβαίνει όταν ένα αιμοφόρο αγγείο στον εγκέφαλο ρήγνυται και προκαλεί αιμορραγία στον εγκεφαλικό ιστό ή στον περιβάλλοντα χώρο. Αυτή η αιμορραγία μπορεί να οδηγήσει σε αυξημένη ενδοκρανιακή πίεση, οίδημα, και θάνατο νευρικών κυττάρων λόγω της τοξικής επίδρασης του αίματος εκτός των αγγείων [68]. Στην ίδια κατηγορία ανήκει και η υπαραχνοειδής αιμορραγία η οποία ορίζεται η εξαγγείωση αίματος από ένα μεγάλο αγγείο του εγκεφάλου (συνήθως από μια αρτηρία) προς τον υπαραχνοειδή χώρο, δηλαδή τον χώρο που παρεμβάλλεται μεταξύ των περιβλημάτων του εγκεφάλου και συγκεκριμένα της χοριοειδούς και της αραχνοειδούς μήνιγγας.

Το παροδικό ισχαιμικό επεισόδιο είναι ένα προειδοποιητικό εγκεφαλικό, που συμβαίνει όταν ένας θρόμβος μπλοκάρει προσωρινά μια αρτηρία του εγκεφάλου, διακόπτοντας τη ροή του αίματος για μικρό χρονικό διάστημα. Τα συμπτώματά του εμφανίζονται ξαφνικά και μοιάζουν με εκείνα ενός εγκεφαλικού επεισοδίου, αλλά διαρκούν μόνο για λίγα λεπτά ή ώρες, χωρίς να αφήνουν μόνιμη εγκεφαλική βλάβη. Η βασική διαφορά του από ένα ισχαιμικό εγκεφαλικό είναι ότι η απόφραξη είναι παροδική, και η ροή του αίματος αποκαθίσταται από

μόνη της. Παρόλα αυτά, το ΤΙΑ θεωρείται σημαντικό προειδοποιητικό σημάδι για μελλοντικό σοβαρό εγκεφαλικό, καθώς δεν υπάρχει τρόπος να προβλεφθεί αν τα συμπτώματα θα επαναληφθούν με μόνιμη εγκεφαλική βλάβη [69].

3.1.2 Αιτίες και Παράγοντες Κινδύνου

Το εγκεφαλικό επεισόδιο είναι μια σοβαρή ιατρική κατάσταση που μπορεί να προκληθεί από διάφορους παράγοντες κινδύνου, οι οποίοι επηρεάζουν τη ροή του αίματος προς τον εγκέφαλο. Αυτοί οι παράγοντες χωρίζονται σε τροποποιήσιμους και μη τροποποιήσιμους, με τους πρώτους να σχετίζονται κυρίως με τον τρόπο ζωής και τις ιατρικές παθήσεις, ενώ οι δεύτεροι αφορούν γενετικούς και βιολογικούς παράγοντες [70, 71]. Στον παρακάτω πίνακα, συνοψίζονται οι παράγοντες στους οποίους οφείλεται η εμφάνιση εγκεφαλικού επεισοδίου.

Πίνακας 3.1: Παράγοντες κινδίνου για την εμφάνιση εγκεφαλικού [70-72]

Μη τροποποιήσιμοι παράγοντες	
Παράγοντας	**Επεξήγηση**
Ηλικία	Ένας από τους πιο σημαντικούς παράγοντες πρόκλησής του (για Ηλικία>55 ετών διπλασιάζεται η πιθανότητα πρόκλησής του).
Φύλο	Λόγω των κινδύνων που συνδέονται με την εγκυμοσύνη και τη χρήση αντισυλληπτικών χαπιών, οι προ εμμηνοπαυσιακές γυναίκες έχουν κίνδυνο εγκεφαλικού επεισοδίου εξίσου υψηλό ή και υψηλότερο από εκείνον των ανδρών. Σε μεγαλύτερες ηλικίες, τα ποσοστά εγκεφαλικού είναι ελαφρώς υψηλότερα στους άνδρες.
Γενετικοί παράγοντες	Εντοπισμός ορισμένων γονιδιακών διαταραχών που συσχετίζονται με την εμφάνιση εγκεφαλικού (π.χ. νόσος Farby, δρεπανοκυτταρική αναιμία).
Οικογενειακό ιστορικό εγκεφαλικού	Η ύπαρξη συγγενών πρώτου βαθμού που έχουν υποστεί εγκεφαλικό αυξάνει σημαντικά την πιθανότητα εμφάνισής του, πιθανώς λόγω γενετικών και περιβαλλοντικών παραγόντων.
Προηγούμενο εγκεφαλικό επεισόδιο ή παροδικό ισχαιμικό επεισόδιο (ΤΙΑ)	Τα άτομα που έχουν υποστεί ένα εγκεφαλικό ή ένα παροδικό ισχαιμικό επεισόδιο έχουν αυξημένο κίνδυνο επανεμφάνισης.
Τροποποιήσιμοι παράγοντες	
Παράγοντας	**Επεξήγηση**
Υπέρταση	Ο σημαντικότερος τροποποιήσιμος παράγοντας κινδύνου για εγκεφαλικό. Περίπου 50% των ασθενών με εγκεφαλικό και ακόμη περισσότεροι με ICH έχουν ιστορικό υπέρτασης. Ακόμη και σε μη υπερτασικούς, όσο υψηλότερη η πίεση, τόσο μεγαλύτερος ο κίνδυνος. Η σημασία της υπέρτασης μειώνεται μετά τα 60 έτη και γίνεται μη σημαντική στα 80 έτη.
Σακχαρώδης διαβήτης	Ύπαρξη του διπλασιάζει την πιθανότητα εμφάνισής του εγκεφαλικού. Μάλιστα, το 41,3% των ασθενών που

	εμφάνισαν εγκεφαλικό είχαν σακχαρώδη διαβήτη όπως αποτυπώνεται σε σχετική νοσοκομειακή μελέτη.
Καρδιακοί παράγοντες	Το καρδιοεμβολικό έμφραγμα από κατά βάση κολπική μαρμαρυγή αποτελεί το βασικό αίτιο του ισχαιμικού εγκεφαλικού. Από νοσοκομειακή μελέτη η γενικότερη αυτή κατηγορία εμφανίζεται στο 29,1%.
Κάπνισμα	Ύπαρξη του διπλασιάζει την πιθανότητα εμφάνισής του εγκεφαλικού (43% των ασθενών σε σχετική νοσοκομειακή μελέτη).
Κατανάλωση αλκοόλ και άλλων ουσιών	Κατανάλωση ναρκωτικών ουσιών και αλκοόλ έχει παρατηρηθεί πως προκαλούν κίνδυνο εμφάνισης εγκεφαλικού
Παχυσαρκία και καθιστική ζωή	Τιμές στο αίμα που συσχετίζονται με αυτόν τρόπο ζωής αυξάνουν τον κίνδυνο εγκεφαλικού, όπως υψηλή χοληστερόλη (25,5% των ασθενών σε σχετική νοσοκομειακή μελέτη).
Φλεγμονή	Εμφάνιση μέτριας σχέσης για αυξημένους φλεγμονώδεις βιοδείκτες για εμφάνισή του.
Παχυσαρκία	Το αυξημένο σωματικό βάρος αυξάνει τον κίνδυνο εμφάνισης εγκεφαλικού επεισοδίου.
Διατροφή	Μια ανθυγιεινή διατροφή (υψηλή σε αλάτι, λιπαρά, και επεξεργασμένα τρόφιμα) σχετίζεται με υψηλότερο κίνδυνο εγκεφαλικού.
Φυσική αδράνεια	Η έλλειψη άσκησης αυξάνει τον κίνδυνο, καθώς συνδέεται με υπέρταση, διαβήτη και παχυσαρκία.
Δυσλιπιδαιμία	Υψηλή LDL και χαμηλή HDL σχετίζονται με αυξημένο κίνδυνο ισχαιμικού εγκεφαλικού.
Ψυχολογικό στρες και κατάθλιψη	Αυξάνουν τον κίνδυνο μέσω της επίδρασης στο αυτόνομο νευρικό σύστημα και την αρτηριακή πίεση.
Οξυγόνωση	Χαμηλό O_2 (<90%) σχετίζεται με υπνική άπνοια, αυξάνοντας τον κίνδυνο εγκεφαλικού.
Θρομβοκύτταρα	Αυξημένα PLT σχετίζονται με θρομβώσεις, χαμηλά PLT με αιμορραγίες.
Λευκά αιμοσφαίρια (WBC)	Αυξημένα WBC υποδηλώνουν χρόνια φλεγμονή, αυξάνοντας τον κίνδυνο αθηροσκλήρωσης.
Γλυκόζη	Υψηλά επίπεδα σακχάρου αυξάνουν τον κίνδυνο εγκεφαλικού και επιδεινώνουν την πρόγνωση.
C-αντιδρώσα πρωτεΐνη (CRP)	Δείκτης φλεγμονής. Υψηλά επίπεδα σχετίζονται με αυξημένο κίνδυνο αθηροθρόμβωσης.
Νάτριο (Na^+)	Υπονατριαιμία σχετίζεται με χειρότερη πρόγνωση μετά από εγκεφαλικό.

3.1.3 Συμπτώματα και Κλινικά Χαρακτηριστικά

Το εγκεφαλικό επεισόδιο μπορεί να εκδηλωθεί με τυπικά και άτυπα συμπτώματα, τα οποία επηρεάζουν τη διάγνωση και την ταχύτητα της ιατρικής παρέμβασης. Τα τυπικά συμπτώματα περιλαμβάνουν αιφνίδια αδυναμία ή παράλυση στη μία πλευρά του σώματος, η οποία μπορεί να επηρεάζει το πρόσωπο, το χέρι ή το πόδι, διαταραχές ομιλίας, όπως δυσαρθρία ή αφασία, καθώς και αιφνίδια απώλεια ή θάμβος όρασης σε ένα ή και στα δύο μάτια. Επιπλέον, μπορεί να εμφανιστεί απώλεια ισορροπίας ή συντονισμού, συνοδευόμενη από ζάλη, καθώς και ξαφνικός, έντονος πονοκέφαλος χωρίς εμφανή αιτία, ο οποίος είναι πιο συχνός σε περιπτώσεις αιμορραγικού εγκεφαλικού. Η γρήγορη αναγνώριση των τυπικών συμπτωμάτων βασίζεται στη μέθοδο FAST, η οποία περιλαμβάνει την ασυμμετρία στο πρόσωπο (Face), την αδυναμία στα

χέρια (Arms), τη διαταραγμένη ομιλία (Speech) και την ανάγκη για άμεση ιατρική παρέμβαση (Time).

Παρόλο που τα τυπικά συμπτώματα είναι τα πιο αναγνωρίσιμα, υπάρχουν και άτυπα συμπτώματα, τα οποία μπορεί να προκαλέσουν καθυστέρηση στη διάγνωση. Αυτά περιλαμβάνουν ξαφνική σύγχυση ή αλλαγή συμπεριφοράς, η οποία μπορεί να μοιάζει με ψυχιατρική διαταραχή, καθώς και λιποθυμία ή διαταραχές συνείδησης. Επίσης, σε ορισμένες περιπτώσεις, οι ασθενείς εμφανίζουν ναυτία και έμετο καθώς και αιφνίδιο λόξυγκα. Επιπλέον, μπορεί να εμφανιστούν δυσκολία στην αναπνοή, αίσθημα παλμών, αιφνίδια κόπωση ή ανεξήγητη αδυναμία, που μπορεί να οδηγήσουν σε εσφαλμένη διάγνωση ή καθυστέρηση στην παροχή θεραπείας.

Πέρα από τα τυπικά και άτυπα συμπτώματα, σύμφωνα με την βιβλιογραφία [75] εμφανίζονται και τα λειτουργικά συμπτώματα εγκεφαλικού (Functional Stroke Symptoms - FSMs), τα οποία μιμούνται εγκεφαλικό επεισόδιο αλλά δεν προέρχονται από πραγματική αγγειακή βλάβη. Τα FSMs συχνά συνοδεύονται από ασυνέπεια στα κινητικά ή αισθητικά συμπτώματα, όπως αδυναμία που αλλάζει ανάλογα με το περιβάλλον ή την προσοχή του ασθενούς, καθώς και ταχεία αποκατάσταση των συμπτωμάτων, σε αντίθεση με τα πραγματικά εγκεφαλικά, όπου η αποκατάσταση διαρκεί περισσότερο. Άλλα χαρακτηριστικά των FSMs περιλαμβάνουν συμπτώματα που δεν συμβαδίζουν με την εγκεφαλική ανατομία, όπως αδυναμία σε όλο το πόδι αλλά όχι στο χέρι, κάτι που δεν αντιστοιχεί σε συγκεκριμένη νευρολογική βλάβη. Επίσης, παρατηρείται ιστορικό ψυχολογικού στρες ή αγχωδών διαταραχών, το οποίο μπορεί να συνδέεται με την εμφάνιση αυτών των συμπτωμάτων.

Σε κάθε περίπτωση, η κατανόηση των συμπτωμάτων είναι απαραίτητη για τη σωστή και έγκαιρη διάγνωση του εγκεφαλικού επεισοδίου. Η διαφορική διάγνωση μεταξύ πραγματικού εγκεφαλικού και FSMs είναι κρίσιμη, καθώς μπορεί να επηρεάσει σημαντικά τη θεραπευτική προσέγγιση. Συνεπώς, η αναγνώριση των διαφορετικών εκδηλώσεων του εγκεφαλικού επεισοδίου συμβάλλει στη βελτίωση της ιατρικής διαχείρισης, στη μείωση της θνησιμότητας και στην πρόληψη μακροχρόνιων επιπλοκών [73].

3.1.4 Διαγνωστικές Μέθοδοι

Η τυπική διαδικασία αξιολόγησης του εγκεφαλικού επεισοδίου βασίζεται στην απεικόνιση του εγκεφάλου και των εγκεφαλικών αγγείων, σε συνδυασμό με την κλινική εκτίμηση της βαρύτητας του εγκεφαλικού μέσω της κλίμακας NIHSS (National Institute of Health Stroke Scale). Η κλίμακα NIHSS αποτελεί ένα σημαντικό εργαλείο για την αξιολόγηση της βαρύτητας του εγκεφαλικού επεισοδίου και την καθοδήγηση της θεραπείας. Βάσει της βαθμολογίας, τα εγκεφαλικά επεισόδια κατηγοριοποιούνται ως εξής: 0–5 βαθμοί αντιστοιχούν είτε σε παροδικό ισχαιμικό επεισόδιο (TIA) (όταν NIHSS = 0 και δεν υπάρχουν ευρήματα στην εξέταση) είτε σε ήπιο εγκεφαλικό επεισόδιο (NIHSS 1–5), που συνήθως έχει καλή πρόγνωση και ελάχιστη αναπηρία. Βαθμολογία 6–10 υποδηλώνει μέτριας βαρύτητας εγκεφαλικό, το οποίο μπορεί να επηρεάσει τη λειτουργικότητα και να απαιτήσει αποκατάσταση. Βαθμολογία 11–20 αντιστοιχεί σε μέτρια έως σοβαρή αναπηρία, με αυξημένο κίνδυνο επιπλοκών και ανάγκη για εντατική θεραπεία. Τέλος, βαθμολογία ≥20 σημαίνει βαρύ, απειλητικό για τη ζωή εγκεφαλικό επεισόδιο, με μεγάλη πιθανότητα σοβαρής αναπηρίας ή θνησιμότητας, απαιτώντας άμεση ιατρική παρέμβαση [74]. Ο παρακάτω πίνακας παρουσιάζει συνοπτικά τις διαγνωστικές μεθόδους που χρησιμοποιούνται για την ανίχνευση του εγκεφαλικού επεισοδίου.

Πίνακας 3.2:Μέθοδοι απεικόνισης για διάγνωση εγκεφαλικού [66]

Μέθοδος απεικόνισης	**Χρήση**
Αξονική τομογραφία χωρίς σκιαγραφικό (Non contrast computed tomography – NCCT)	• Διάγνωση σοβαρού εγκεφαλικού. • Διάκριση παθήσεων που μιμούνται το εγκεφαλικό, όπως όγκοι εγκεφάλου.

Αγγειογραφία με αξονική τομογραφία (CTA - Computed Tomography Angiography)	• Ανίχνευση απόφραξης μεγάλου αγγείου (LVO).
Αγγειογραφία μονής φάσης με αξονική τομογραφία (sCTA – Single-phase CT Angiography)	• Επιτρέπει την εκτίμηση της συνολικής παράπλευρης κυκλοφορίας.
Αγγειογραφία πολλαπλών φάσεων με αξονική τομογραφία (mCTA – Multiphase CTA)	• Παρέχει χρονοεξαρτώμενες εικόνες της εγκεφαλικής αγγείωσης με τρεις διαδοχικές λήψεις.
Αξονική τομογραφία αιμάτωσης (CTP – CT Perfusion) ή Αξονική τομογραφία διάχυσης (PCT – Perfusion CT)	• Παρέχει χρονοεξαρτώμενες εικόνες της αιμάτωσης του εγκεφάλου. • Χρησιμοποιείται για τον εντοπισμό και την ποσοτικοποίηση του όγκου του ισχαιμικού πυρήνα και της πνευμονίας.
Ψηφιακή αφαιρετική αγγειογραφία (DSA – Digital Subtraction Angiography)	• Παραδοσιακή απεικόνιση της κατάστασης της κυκλοφορίας του αίματος.
Μαγνητική τομογραφία (MRI – Magnetic Resonance Imaging)	• Η πιο ευαίσθητη τεχνική για την ανίχνευση του οξέος ισχαιμικού εγκεφαλικού. • Χρησιμοποιείται για τη διάγνωση μικρών εγκεφαλικών και συνήθως κατά την παρακολούθηση των ασθενών.
Αγγειογραφία μαγνητικού συντονισμού (MRA – MR Angiography)	• Χρησιμοποιείται σε ασθενείς που έχουν αντένδειξη στη χρήση σκιαγραφικού ή όταν η CTA δεν δίνει διαγνωστικά αποτελέσματα.

Η Αξονική Τομογραφία (CT) και η Μαγνητική Τομογραφία (MRI) είναι οι δύο κύριες διαγνωστικές μέθοδοι που χρησιμοποιούνται για την ανίχνευση του εγκεφαλικού επεισοδίου, με την CT να αποτελεί την πιο συχνά χρησιμοποιούμενη επιλογή στην οξεία φάση λόγω της ταχύτητας και της ευρείας διαθεσιμότητάς της. Η NCCT (Αξονική χωρίς σκιαγραφικό) είναι η πρώτη εξέταση που πραγματοποιείται στα επείγοντα περιστατικά, καθώς μπορεί να ανιχνεύσει αιμορραγικά εγκεφαλικά και να αποκλείσει άλλες καταστάσεις που μιμούνται το εγκεφαλικό, όπως όγκοι. Επιπλέον, η CTA χρησιμοποιείται για την ανίχνευση αποφράξεων μεγάλων αγγείων, ενώ η CTP βοηθά στον καθορισμό της περιοχής του εγκεφαλικού ιστού που μπορεί να σωθεί, ειδικά σε ασθενείς που φτάνουν αργότερα από το θεραπευτικό "παράθυρο". Αντίθετα, η MRI, και ειδικότερα η DWI-MRI (Μαγνητική Τομογραφία Διάχυσης), είναι η πιο ευαίσθητη εξέταση για την ανίχνευση πρώιμων ισχαιμικών αλλοιώσεων, ανιχνεύοντας εγκεφαλική ισχαιμία ακόμη και μέσα στα πρώτα λεπτά από την έναρξη των συμπτωμάτων. Επιπλέον, η MRA (Μαγνητική Αγγειογραφία) επιτρέπει την απεικόνιση των εγκεφαλικών αγγείων χωρίς σκιαγραφικό, ενώ η SWI (Susceptibility Weighted Imaging) χρησιμοποιείται για την ανίχνευση μικροαιμορραγιών. Ωστόσο, η MRI απαιτεί περισσότερο χρόνο, δεν είναι πάντα διαθέσιμη στα επείγοντα και είναι ευαίσθητη στις κινήσεις του ασθενούς, γεγονός που την καθιστά λιγότερο πρακτική για την άμεση διάγνωση του εγκεφαλικού. Συνολικά, η CT αποτελεί την πρώτη γραμμή διάγνωσης λόγω της ταχύτητάς της, ενώ η MRI χρησιμοποιείται σε περιπτώσεις μικρών ισχαιμικών εγκεφαλικών ή όταν η CT δεν είναι διαγνωστική.

Παρόλο που η Αξονική Τομογραφία και η Μαγνητική Τομογραφία είναι οι πιο ευρέως χρησιμοποιούμενες μέθοδοι για τη διάγνωση ενός εγκεφαλικού επεισοδίου ενδέχεται να μην είναι διαθέσιμες σε όλες τις υγειονομικές δομές. Γι’ αυτόν τον λόγο, διενεργείται εντατική έρευνα και ανάπτυξη νέων, πρακτικών, φορητών, οικονομικών και συμπληρωματικών διαγνωστικών εργαλείων που μπορούν να χρησιμοποιηθούν σε μονάδες επειγόντων περιστατικών. Ενδεικτικά αναφέρονται μερικά παραδείγματα από την ανάπτυξη φορητών και ταχύτερων διαγνωστικών τεχνολογιών. Αρχικά, το Strokefinder MD100, που αναπτύχθηκε από την Medfield Diagnostics AB (Σουηδία), είναι μια μικροκυματική συσκευή ικανή να ανιχνεύσει ενδοκρανιακή αιμορραγία με ευαισθησία 100% και ειδικότητα 75%. Η τεχνολογία αυτή χρησιμοποιεί μη ιονίζουσα μικροκυματική ακτινοβολία, και μέσα σε 45 δευτερόλεπτα

μπορεί να διαφοροποιήσει το 65% των ασθενών με ισχαιμικό εγκεφαλικό (AIS) από εκείνους με αιμορραγία.Επιπλέον, υπάρχει το Lucid Robotic System (Neural Analytics, ΗΠΑ) αξιοποιεί διακρανιακό υπέρηχο Doppler (TCD) για τη μέτρηση της εγκεφαλικής αιματικής ροής, με 91% διαγνωστική ακρίβεια στην ανίχνευση αποφράξεων μεγάλων αγγείων (LVO). Παράλληλα, το SONAS (BURL Concepts, ΗΠΑ) χρησιμοποιεί υπέρηχο και σκιαγραφικές μικροφυσαλίδες για την εκτίμηση της εγκεφαλικής αιμάτωσης, και βρίσκεται σε κλινικές δοκιμές. Παράλληλα, το Visor (Cerebrotech, ΗΠΑ) χρησιμοποιεί την τεχνολογία VIPS (Volumetric Impedance Phase Shift Spectroscopy) για τη διαφοροποίηση σοβαρού από ήπιο εγκεφαλικό επεισόδιο με ευαισθησία 93% και ειδικότητα 92%. Πρόκειται για μη επεμβατική συσκευή που τοποθετείται στο κεφάλι του ασθενούς και ανιχνεύει μεταβολές στη βιοεμπέδηση, οι οποίες συνδέονται με την αιματική ροή στον εγκέφαλο. Τέλος, η Ηλεκτρική Τομογραφία Εμπέδησης (EIT) αποτελεί φορητή, ασφαλή και οικονομική τεχνική, η οποία χρησιμοποιεί ραδιοσυχνότητες για τη μέτρηση της ηλεκτρικής εμπέδησης του εγκεφαλικού ιστού. Μελετάται για την ταχεία διάγνωση του εγκεφαλικού, συμπεριλαμβανομένης της ανίχνευσης τύπου εγκεφαλικού (ισχαιμικό ή αιμορραγικό), καθώς και της παρακολούθησης εγκεφαλικού οιδήματος. Πρόσφατες κλινικές δοκιμές δείχνουν υποσχόμενα αποτελέσματα για την ενσωμάτωση της EIT στη διάγνωση του οξέος ισχαιμικού εγκεφαλικού [66].

3.1.5 Θεραπεία και Αποκατάσταση

Μέχρι σήμερα, οι μόνες εγκεκριμένες θεραπευτικές προσεγγίσεις για τη θεραπεία του οξέος ισχαιμικού εγκεφαλικού είναι η χορήγηση θρομβολυτικής αγωγής και η μηχανική απομάκρυνση του αποφρακτικού θρόμβου μέσω ενδοαγγειακής μηχανικής θρομβεκτομής (MT).

Συγκεκριμένα, η χορήγηση θρομβολυτικής αγωγής είναι μια φαρμακευτική θεραπεία που χρησιμοποιείται για τη διάλυση των θρόμβων που προκαλούν απόφραξη των εγκεφαλικών αγγείων σε ασθενείς με οξύ ισχαιμικό εγκεφαλικό. Ο ιστικός ενεργοποιητής του πλασμινογόνου (tPA) και η ανασυνδυασμένη μορφή του (rtPA - recombinant tPA, alteplase) είναι οι κύριες θρομβολυτικές ουσίες που χρησιμοποιούνται. Το rtPA ενεργοποιεί το πλασμινογόνο, το οποίο μετατρέπεται σε πλασμίνη, ένα ένζυμο που διασπά την ινική του θρόμβου, αποκαθιστώντας έτσι τη ροή αίματος στον εγκέφαλο. Η θρομβόλυση είναι πιο αποτελεσματική όταν χορηγείται εντός 4.5 ωρών από την έναρξη των συμπτωμάτων, καθώς μετά από αυτό το χρονικό διάστημα αυξάνεται ο κίνδυνος αιμορραγίας. Η έγκαιρη χορήγηση μειώνει τον κίνδυνο αναπηρίας κατά 30% στους 3 μήνες, αλλά η επιτυχής διάλυση του θρόμβου επιτυγχάνεται σε λιγότερο από το 50% των ασθενών, καθώς η αποτελεσματικότητά της εξαρτάται από τη σύσταση του θρόμβου. Παρότι αποτελεί την πρώτη θεραπευτική επιλογή για το ισχαιμικό εγκεφαλικό, η αποτελεσματικότητα και η ασφάλειά της εξαρτώνται από την ταχεία διάγνωση και τη σωστή επιλογή ασθενών.

Η μηχανική θρομβεκτομή είναι μια ελάχιστα επεμβατική ενδοαγγειακή θεραπεία που χρησιμοποιείται για την αφαίρεση θρόμβων αίματος από αποφραγμένα εγκεφαλικά αγγεία σε ασθενείς με οξύ ισχαιμικό εγκεφαλικό, ιδιαίτερα σε περιπτώσεις αποφράξεων μεγάλων αγγείων, όπου η θρομβολυτική θεραπεία συχνά δεν επαρκεί. Η διαδικασία περιλαμβάνει την εισαγωγή ενός καθετήρα μέσω της μηριαίας αρτηρίας, ο οποίος καθοδηγείται μέχρι το σημείο της απόφραξης στον εγκέφαλο, ώστε ο θρόμβος να αφαιρεθεί είτε με συσκευή ανάκτησης τύπου stentriever είτε μέσω αναρρόφησης (aspiration thrombectomy). Στην πρώτη περίπτωση, ένας μεταλλικός νάρθηκας (stent) εγκλωβίζει τον θρόμβο και τον απομακρύνει, ενώ στη δεύτερη, ένας καθετήρας εφαρμόζει αρνητική πίεση, αναρροφώντας τον θρόμβο. Η μηχανική θρομβεκτομή θεωρείται πλέον "gold standard" θεραπεία για ασθενείς με σοβαρό ισχαιμικό εγκεφαλικό, καθώς οι κλινικές μελέτες του 2015 επιβεβαίωσαν ότι βελτιώνει σημαντικά τα ποσοστά επαναγγείωσης και τη νευρολογική αποκατάσταση, καθιερώνοντάς την ως την κύρια θεραπεία για τη διάσωση του εγκεφαλικού ιστού όταν η θρομβόλυση δεν είναι επαρκής ή δυνατή [66].

Η θεραπεία του αιμορραγικού εγκεφαλικού διαφέρει από αυτή του ισχαιμικού, καθώς το ζήτημα σε αυτή την περίπτωση δεν είναι η απόφραξη των αγγείων αλλά η αιμορραγία εντός του εγκεφάλου. Η θεραπεία του αιμορραγικού εγκεφαλικού στοχεύει στη διαχείριση της αιμορραγίας, τη σταθεροποίηση του ασθενούς και την υποστήριξη της νευρολογικής αποκατάστασης. Ανάλογα με τη σοβαρότητα της κατάστασης, εφαρμόζονται φαρμακευτικές και χειρουργικές παρεμβάσεις. Στην φαρμακευτική αντιμετώπιση, βασική προτεραιότητα είναι ο έλεγχος της αρτηριακής πίεσης, καθώς η υπέρταση μπορεί να επιδεινώσει την αιμορραγία, με στόχο τη μείωσή της κάτω από 140 mmHg. Για ασθενείς που λαμβάνουν αντιπηκτικά φάρμακα, η αναστροφή της αντιπηκτικής αγωγής γίνεται με προθρομβινικά συμπλέγματα (PCC), βιταμίνη K ή idarucizumab, ώστε να αποτραπεί περαιτέρω αιμορραγία. Επιπλέον, η διαχείριση της ενδοκρανιακής πίεσης μπορεί να απαιτεί χορήγηση ωσμωτικών διαλυμάτων για τη μείωση του εγκεφαλικού οιδήματος, ενώ σε επιλεγμένους ασθενείς χορηγούνται αντιεπιληπτικά φάρμακα για την πρόληψη κρίσεων. Στις χειρουργικές επιλογές, η αφαίρεση του αιματώματος είναι απαραίτητη σε περιπτώσεις μεγάλου όγκου (>30 mL), αιμορραγίας στην παρεγκεφαλίδα (>3 cm) ή σοβαρής νευρολογικής επιδείνωσης λόγω αυξημένης ενδοκρανιακής πίεσης. Νεότερες ελάχιστα επεμβατικές τεχνικές (MISTIE, ICES) συνδυάζουν χειρουργική αφαίρεση του αιματώματος με τη χορήγηση θρομβολυτικών ουσιών, ώστε να επιταχυνθεί η απορρόφηση του υπολειπόμενου αιματώματος. Σε περιπτώσεις υπαραχνοειδούς αιμορραγίας ή αποφρακτικής υδροκεφαλίας, μπορεί να τοποθετηθεί παροχέτευση εγκεφαλονωτιαίου υγρού (EVD) για τη μείωση της πίεσης στον εγκέφαλο [75].

Η αποκατάσταση ενός ασθενούς μετά από εγκεφαλικό είναι μια πολύπλευρη και εξατομικευμένη διαδικασία, που στοχεύει στην ανάκτηση της λειτουργικότητας και τη βελτίωση της ποιότητας ζωής. Ξεκινά άμεσα, μόλις σταθεροποιηθεί η κατάσταση του ασθενούς, και περιλαμβάνει φυσικοθεραπεία για την ενδυνάμωση των μυών, εργοθεραπεία για την εκμάθηση καθημερινών δραστηριοτήτων, και λογοθεραπεία για την αποκατάσταση της ομιλίας και της κατάποσης. Παράλληλα, η ψυχολογική υποστήριξη είναι καθοριστική για την αντιμετώπιση συναισθηματικών δυσκολιών, ενώ η θεραπευτική άσκηση και η υδροθεραπεία ενισχύουν την κινητικότητα και τη συνολική ευεξία. Η διατροφή, προσαρμοσμένη στις ανάγκες του ασθενούς, παίζει σημαντικό ρόλο στην ανάρρωση, ενώ τεχνικές όπως η εικαστική θεραπεία μπορούν να βοηθήσουν στην έκφραση συναισθημάτων και την ενίσχυση της γνωστικής λειτουργίας. Η συνολική επανένταξη του ασθενούς στην κοινωνική και επαγγελματική ζωή επιτυγχάνεται μέσα από ένα ολοκληρωμένο πρόγραμμα αποκατάστασης, προσαρμοσμένο στις ατομικές του ανάγκες, με στόχο τη μέγιστη δυνατή αυτονομία και λειτουργικότητα.

3.1.6 Σημασία έγκαιρης διάγνωσης – Ρόλος των Βιοδεικτών

Η έγκαιρη διάγνωση του εγκεφαλικού επεισοδίου είναι ζωτικής σημασίας για τη μείωση της θνησιμότητας, της αναπηρίας και τη βελτίωση των θεραπευτικών αποτελεσμάτων. Δεδομένου ότι το εγκεφαλικό είναι μια επείγουσα ιατρική κατάσταση, κάθε λεπτό καθυστέρησης αυξάνει τον κίνδυνο μη αναστρέψιμης εγκεφαλικής βλάβης, καθώς περίπου 1,9 εκατομμύρια νευρώνες χάνονται κάθε λεπτό χωρίς αποκατάσταση της ροής του αίματος. Στην περίπτωση ενός ισχαιμικού εγκεφαλικού, η θρομβολυτική θεραπεία είναι αποτελεσματική μόνο αν χορηγηθεί εντός 4.5 ωρών από την έναρξη των συμπτωμάτων, ενώ η μηχανική θρομβεκτομή είναι αποτελεσματική εντός 6 ωρών, με δυνατότητα εφαρμογής έως 24 ώρες σε επιλεγμένους ασθενείς. Ωστόσο, η επιτυχία αυτών των θεραπειών εξαρτάται από την ταχεία διαφοροποίηση του ισχαιμικού από το αιμορραγικό εγκεφαλικό, καθώς η χορήγηση θρομβολυτικών σε περίπτωση αιμορραγίας μπορεί να αποβεί μοιραία. Η άμεση αποκατάσταση της ροής του αίματος περιορίζει την έκταση της νευρωνικής βλάβης, συμβάλλοντας στη διατήρηση περισσότερων νευρολογικών λειτουργιών. Έρευνες δείχνουν ότι οι ασθενείς που λαμβάνουν θεραπεία έγκαιρα έχουν πολύ καλύτερη νευρολογική αποκατάσταση συγκριτικά με όσους καθυστερούν να λάβουν ιατρική φροντίδα.

Τα σύγχρονα διαγνωστικά εργαλεία, όπως το CT και το MRI συμβάλλουν στην ταχεία και ακριβή διαφοροποίηση αυτών των καταστάσεων. Ωστόσο, η διαθεσιμότητα αυτών των απεικονιστικών μεθόδων είναι περιορισμένη, ειδικά σε οικονομικά ασθενέστερες περιοχές, μικρά νοσοκομεία και κλινικές. Οι καθυστερήσεις στην αξιολόγηση και τη θεραπεία του εγκεφαλικού επιδεινώνουν την πρόγνωση του ασθενούς [66]. Επομένως, εκτός από τις απεικονιστικές μεθόδους, υπάρχει ενδιαφέρον για τη χρήση βιοδεικτών αίματος ως συμπληρωματικό διαγνωστικό εργαλείο. Οι ιδανικοί βιοδείκτες εγκεφαλικού θα πρέπει να έχουν υψηλή ευαισθησία και ειδικότητα, ώστε να διακρίνουν το εγκεφαλικό από παθήσεις που το «μιμούνται» (stroke mimics) και να διαφοροποιούν το αιμορραγικό από το ισχαιμικό εγκεφαλικό.

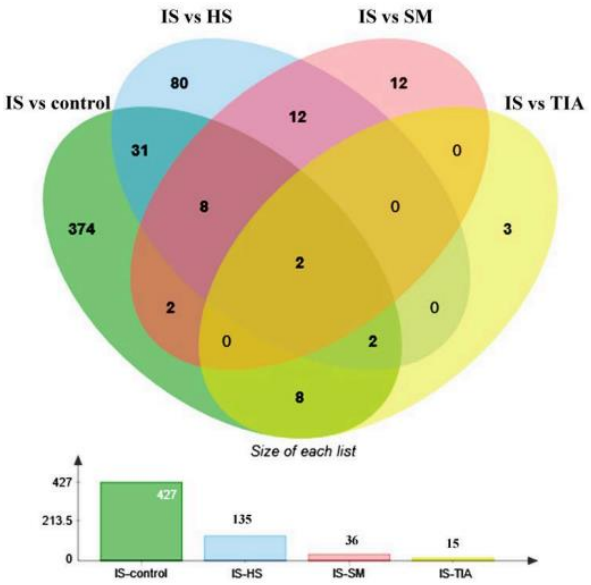
Από υπάρχουσα έρευνα [76], έχουν αναλυθεί 182 μελέτες, οι οποίες περιλάμβαναν συνολικά 30.446 συμμετέχοντες από 38 χώρες. Συγκεκριμένα, οι συμμετέχοντες κατανέμονταν ως εξής: 15.675 ασθενείς με ισχαιμικό εγκεφαλικό (IS), 2.317 με αιμορραγικό εγκεφαλικό (HS), 1.798 με μιμητικά εγκεφαλικού επεισοδίου (SM), 846 με παροδικό ισχαιμικό επεισόδιο (TIA) και 9.810 υγιείς μάρτυρες (controls). Η πλειονότητα των μελετών αξιολόγησε πρωτεϊνικούς βιοδείκτες (n = 127), ακολουθούμενες από μελέτες σε RNA (n = 41), μεταβολίτες (n = 11) και γονιδιακή έκφραση (n = 9). Επιπλέον, 9 μελέτες ανέφεραν βιοδείκτες από διαφορετικά μοριακά επίπεδα ταυτόχρονα. Μόνο το 10% των μελετών (19 μελέτες) περιλάμβανε ομάδες επικύρωσης ή αναπαραγωγής δεδομένων για την επιβεβαίωση της διαγνωστικής τους αξίας. Συνολικά, εντοπίστηκαν 518 βιοδείκτες, συμπεριλαμβανομένων 203 πρωτεϊνών, 114 γονιδίων, 108 μεταβολιτών και 88 μεταγραφικών βιοδεικτών (transcripts). Οι μεταγραφικοί βιοδείκτες περιλάμβαναν 70 microRNAs, 7 circular RNAs, 6 long-non-coding RNAs, 3 mRNAs και 3 tRNAs. Από τους συνολικά εντοπισμένους βιοδείκτες, 427 προέρχονταν από συγκρίσεις IS-control, 135 από IS-HS, 36 από IS-SM και 15 από IS-TIA.

Πίνακας 3.3: Βιοδείκτες με τη μεγαλύτερη ευαισθησία και ειδικότητα για τη διάγνωση του ισχαιμικού εγκεφαλικού σε πρώιμο στάδιο [76]

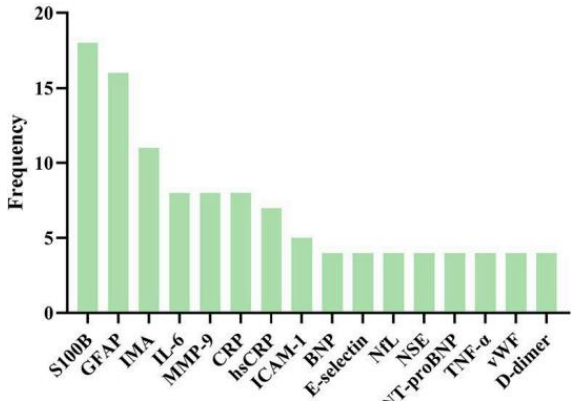
Biomarkers	**Levels**	**Sampling Time**	**Sensitivity**	**Specificity**	**AUC**	**Biological Process**
Antibody to NR2A/2B	Protein	3 h	97%	98%	0.99	Immune response
BDNF	Protein	4 h	100%	92%	0.983	Neuronal survival and growth
IMA index	Protein	6 h	95.8%	96.4%	0.99	Ischemia response
NR2 peptide	Protein	12 h	92.1%	96.5%	0.92	Brain cell damage
GPBB	Protein	12 h	93%	93%	0.96	Ischaemia response
ADAMTS13	Proteomics	24 h	90%	98%	0.96	Blood homeostasis and endothelial function regulation
S100A7	Proteomics	24 h	97%	91%	0.912	Blood homeostasis and endothelial function regulation
VILIP-1	Protein	3 h	100%	100%	1.0	Brain cell damage
miR-107	Transcript	24 h	93.8%	92.2%	0.97	Cerebral ischaemic injury

miR-124	Transcript	24 h	91.6%	93.52%	0.9527	Anti-inflammation
APOA1-UP	Protein	24 h	90.3%	97.1%	0.975	Cholesterol metabolism
lncRNAs LINK-A	Transcript	24 h	92%	94%	0.914	Angiogenesis
ANTXR2 + STX3+ PDK4 + CD163 + MAL4 + GRAP+	Gene	5 h, 24 h	95.7%	95.7%	0.997	Immune response
S100B + BNGF+ vWF +	Protein	3 h	98.1%	91.7%	N/A	Brain cell damage, neuronal survival, coagulation, inflammation
MMP-9+ MCP-1	Protein	6 h	98.1%	93.1%	N/A	Brain cell damage, neuronal survival, coagulation, inflammation

Δύο βιοδείκτες, το ασβέστιο-δεσμευτικό πρωτεϊνικό σύμπλεγμα B (S100B) και η μεταλλοπρωτεϊνάση-9 (MMP-9), εμφανίστηκαν σε όλες τις συγκριτικές ομάδες, υποδεικνύοντας ότι μπορούν να διακρίνουν το ισχαιμικό εγκεφαλικό από όλες τις άλλες καταστάσεις. Ωστόσο, το 89% των βιοδεικτών (444 από τους 518) αξιολογήθηκε μόνο μία φορά, ενώ μόνο ένας μικρός αριθμός βιοδεικτών μελετήθηκε τέσσερις ή περισσότερες φορές. Από τις 182 μελέτες, οι 72 (39,6%) ανέφεραν τόσο την ευαισθησία όσο και την ειδικότητα των βιοδεικτών και εξήχθησαν οι βιοδείκτες με την υψηλότερη διαγνωστική ακρίβεια (ευαισθησία και ειδικότητα >90%), οι οποίοι παρουσιάζονται στον παρακάτω πίνακα. Αξίζει να σημειωθεί ότι οι περισσότεροι από αυτούς δεν έχουν επικυρωθεί ή έχουν ελεγχθεί μόνο σε λίγες ανεξάρτητες ομάδες.



Εικόνα 3.1 Διάγραμμα Venn για κοινούς βιοδείκτες που επικαλύπτονται μεταξύ IS και των ομάδων ελέγχου, HS, SM και TIA [76]



Εικόνα 3.2: Συχνότητα εμφάνισης βιοδεικτών [76]

Όπως παρατηρείται από την εικόνα 15, συνολικά, 16 βιοδείκτες μελετήθηκαν τουλάχιστον 4 φορές μεταξύ των μελετών που συμπεριλήφθηκαν (Β). Αυτοί περιλαμβάνουν S100B (n = 18), GFAP (Glial fibrillary acidic protein) (n = 16), IMA (Ischaemia-modified albumin) (n = 11), IL-6 (Interleukin 6) (n = 8), MMP-9 (Matrix metalloproteinase-9) (n = 8), CRP (C-reactive protein) (n = 8), hsCRP (High-sensitive C-reactive protein) (n = 7), ICAM-1 (Intercellular adhesion molecule 1) (n = 5), BNP (Natriuretic peptides B) (n = 4), NfL (Neurofilament light chain) (n = 4), NSE (Neuron-specific enolase) (n = 4), NT-proBNP (N-terminal pro-natriuretic peptide B) (n = 4), TNF-α (Tumor necrosis factor α) (n = 4) και vWF (Von Willebrand factor) (n = 4) [76].

Οι δύο συχνότερα εμφανιζόμενοι βιοδείκτες στο εγκεφαλικό επεισόδιο είναι το GFAP και το S100β. Κατά τη διάρκεια ενός εγκεφαλικού επεισοδίου, η δυσλειτουργία του αιματοεγκεφαλικού φραγμού επιτρέπει τη διαρροή πρωτεϊνών του εγκεφάλου στο αίμα, οδηγώντας στην απελευθέρωση γλοιακών δομικών πρωτεϊνών, όπως το GFAP και το S100β, σε σημαντικά υψηλότερα επίπεδα από το φυσιολογικό. Αυτές οι πρωτεΐνες έχουν αναδειχθεί ως υποσχόμενοι βιοδείκτες για τη διάγνωση του εγκεφαλικού επεισοδίου, καθώς σχετίζονται άμεσα με την οξεία νευρολογική δυσλειτουργία. Η αυξημένη συγκέντρωσή τους στο αίμα μπορεί να βοηθήσει στη διαφοροποίηση μεταξύ ισχαιμικού και αιμορραγικού εγκεφαλικού, βελτιώνοντας έτσι την έγκαιρη και ακριβή διάγνωση. Συγκεκριμένα, το GFAP θεωρείται ο πιο αξιόπιστος βιοδείκτης για τη διάκριση ισχαιμικού από αιμορραγικό εγκεφαλικό, καθώς τα επίπεδά του στο αίμα αυξάνονται καθυστερημένα και φτάνουν στο μέγιστο 2–4 ημέρες μετά το ισχαιμικό εγκεφαλικό επεισόδιο, ενώ η αύξησή του στην εγκεφαλική αιμορραγία είναι άμεση. Αντίθετα, το S100β, αν και έχει δείξει ελπιδοφόρα αποτελέσματα στη διαφοροποίηση του ισχαιμικού από το αιμορραγικό εγκεφαλικό, δεν είναι τόσο εξειδικευμένο, καθώς αυξάνεται και σε άλλες νευρολογικές παθήσεις [77].

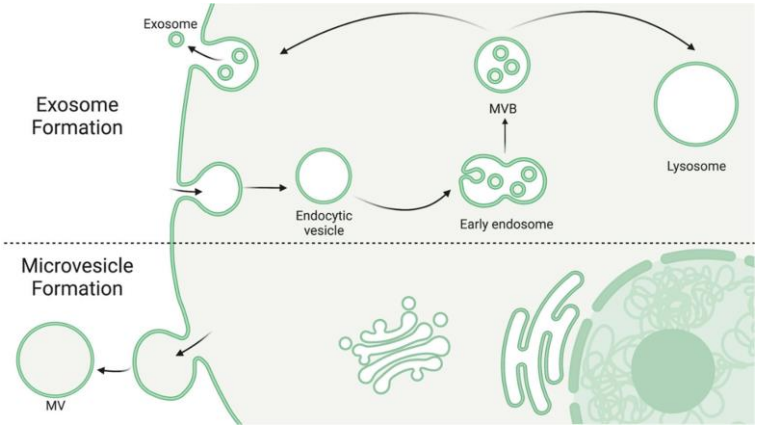
Παράλληλα, η IMA (Ischaemia-Modified Albumin) εμφανίζεται συχνά ως βιοδείκτης εγκεφαλικού επεισοδίου λόγω της αλλαγμένης δομής της αλβουμίνης υπό συνθήκες ισχαιμίας. Κατά τη διάρκεια ενός ισχαιμικού εγκεφαλικού, η ανεπάρκεια οξυγόνου και η οξειδωτική βλάβη μεταβάλλουν τη μεταλλοδεσμευτική ικανότητα της αλβουμίνης, οδηγώντας στη δημιουργία της IMA. Η IL-6 (Interleukin-6) είναι ένας φλεγμονώδης δείκτης, ο οποίος αυξάνεται μετά το εγκεφαλικό και συνδέεται με κακή πρόγνωση. Τέλος, η MMP-9 (Matrix Metalloproteinase-9) σχετίζεται με τη διάσπαση του αιματοεγκεφαλικού φραγμού, και τα αυξημένα επίπεδά της υποδηλώνουν αυξημένο κίνδυνο αιμορραγικής μετατροπής σε ασθενείς με ισχαιμικό εγκεφαλικό [77].

Παρόλο που οι βιοδείκτες εγκεφαλικού αποτελούν ένα ελπιδοφόρο εργαλείο, δεν είναι ακόμα επαρκώς αξιόπιστοι για να αντικαταστήσουν τις απεικονιστικές τεχνικές. Η μελλοντική έρευνα τείνει να επικεντρώνεται στην ανακάλυψη νέων βιοδεικτών και στη βελτίωση των εργαστηριακών τεχνικών, ώστε να δημιουργηθεί ένας αξιόπιστος συνδυασμός διαγνωστικών δεικτών που θα μπορεί να εφαρμοστεί στην κλινική πράξη για την έγκαιρη και ακριβή διάγνωση του εγκεφαλικού.

3.1.7 Εξωκυτταρικά Κυστίδια (EVs) ως Διαγνωστική Λύση

3.1.7.1 Ορισμός και Ιδιότητες ως πιθανοί Βιοδείκτες

Η παρούσα εργασία εστιάζει στη διερεύνηση του ρόλου των εξωκυττάριων κυστιδίων (EVs) ως βιοδεικτών για την έγκαιρη διάγνωση του εγκεφαλικού επεισοδίου. Τα EVs είναι μικροσκοπικά κυστίδια που εκκρίνονται από όλα τα κύτταρα του Κεντρικού Νευρικού Συστήματος (ΚΝΣ) και φέρουν βιολογικές πληροφορίες, όπως πρωτεΐνες, λιπίδια και νουκλεϊκά οξέα. Διακρίνονται σε εξωσώματα (30–150 nm), τα οποία σχηματίζονται μέσω της ενδοσωμικής οδού, και σε μικροκυστίδια (50–1000 nm), τα οποία προέρχονται από την εκβλάστηση της κυτταρικής μεμβράνης. Η δημιουργία εξωσωμάτων και μικροκυστιδίων (MVs) ακολουθεί διαφορετικούς μηχανισμούς. Τα εξωσώματα σχηματίζονται μέσω της εσωτερικής εκβλάστησης των πρώιμων ενδοσωμάτων καθώς αυτά ωριμάζουν σε πολυκυστικά σωμάτια (MVBs). Τα MVBs είτε κατευθύνονται προς τα λυσοσώματα για αποικοδόμηση, όπου μαζί με αυτά διασπώνται και τα εξωσώματα, είτε συγχωνεύονται με την κυτταρική μεμβράνη απελευθερώνοντας το περιεχόμενό τους. Αντίθετα, τα μικροκυστίδια είναι μεγαλύτερης διαμέτρου και σχηματίζονται από την εκβλάστηση της εξωτερικής κυτταρικής μεμβράνης. Τα EVs συμμετέχουν σε βασικές φυσιολογικές διεργασίες, όπως η απομάκρυνση κυτταρικών αποβλήτων και η κυτταρική επικοινωνία, ενώ εμφανίζουν ειδικούς επιφανειακούς δείκτες που αντανακλούν την προέλευσή τους και τη φυσιολογική ή παθολογική κατάσταση των κυττάρων από τα οποία προέρχονται.



Εικόνα 3.3: Σχηματισμός εξωσωμάτων και μικροκυστιδίων (MVs) [78]

Ένα από τα πιο σημαντικά χαρακτηριστικά τους είναι η ικανότητά τους να διαπερνούν τον αιματοεγκεφαλικό φραγμό (BBB), καθιστώντας τα ιδανικούς υποψηφίους για μη επεμβατική διάγνωση μέσω απλών αιματολογικών εξετάσεων. Στα πρώτα στάδια ενός ισχαιμικού εγκεφαλικού επεισοδίου, η μειωμένη παροχή οξυγόνου και θρεπτικών ουσιών στον εγκεφαλικό ιστό οδηγεί σε ενεργειακή ανεπάρκεια (ATP depletion) και διαταραχή της ομοιόστασης. Ως αποτέλεσμα, τα ενδοθηλιακά κύτταρα διογκώνονται, προκαλώντας αυξημένη διαπερατότητα του αιματοεγκεφαλικού φραγμού και διαρροή βιολογικών μορίων στην κυκλοφορία του αίματος. Καθώς η διάσπαση του BBB προηγείται από συγκεκριμένες φυσιολογικές διεργασίες, τα EVs που απελευθερώνονται νωρίς από τα προσβεβλημένα κύτταρα μπορούν να προσφέρουν ένα πρώιμο προφίλ έκκρισης. Αυτό σημαίνει ότι μπορούν να παρέχουν πολύτιμες πληροφορίες για την κατάσταση του εγκεφάλου πριν η βλάβη γίνει μη αναστρέψιμη, επιτρέποντας την έγκαιρη διάγνωση και ταχύτερη έναρξη της θεραπείας.

Επομένως, λόγω της δυνατότητάς τους να αντανακλούν την κυτταρική δραστηριότητα και να διαπερνούν τον αιματοεγκεφαλικό φραγμό (BBB), τα EVs έχουν προταθεί ως μη επεμβατικοί βιοδείκτες για την έγκαιρη διάγνωση και παρακολούθηση του εγκεφαλικού επεισοδίου και άλλων νευρολογικών διαταραχών. Η χρήση των EVs ως βιοδεικτών σε αιματολογικές εξετάσεις μπορεί να συμβάλει στην ελαχιστοποίηση των καθυστερήσεων στη

θεραπεία ασθενών με εγκεφαλικό, κάτι που είναι ζωτικής σημασίας, δεδομένου ότι η αποτελεσματικότητα των θεραπειών εξαρτάται άμεσα από τον χρόνο παρέμβασης. Συνεπώς, η αξιοποίηση των EVs για τη διάγνωση του εγκεφαλικού επεισοδίου μπορεί να οδηγήσει σε καινοτόμες, μη επεμβατικές και ταχύτερες διαγνωστικές στρατηγικές, βελτιώνοντας τα θεραπευτικά αποτελέσματα και την πρόγνωση των ασθενών.

Αυξημένα επίπεδα ενδοθηλιακών EVs μετά από οξύ ισχαιμικό εγκεφαλικό (AIS) καταγράφηκαν για πρώτη φορά το 2006. Τα EVs με θετικά CD105, CD54 και CD45 σχετίστηκαν με τον όγκο του εμφράκτου και την κλινική έκβαση των ασθενών. Μελέτες έδειξαν ότι τα EVs μπορούν να διακρίνουν σοβαρά από ελαφρά εγκεφαλικά επεισόδια, αν και δεν διαχωρίζουν πάντα τα ήπια περιστατικά από υγιείς ελέγχους. Εντός 48 ωρών από ένα εγκεφαλικό επεισόδιο ή παροδικό ισχαιμικό επεισόδιο (TIA), παρατηρείται αύξηση των CD146+, CD62E+ και annexin V+ EVs, ενώ τις επόμενες ημέρες οι τιμές τους συνεχίζουν να αυξάνονται ειδικά σε περιπτώσεις AIS. Επιπλέον, συγκεκριμένα EVs που φέρουν το microRNA-155 (miRNA-155) έχουν θετική συσχέτιση με τη βαρύτητα του εγκεφαλικού επεισοδίου και τον όγκο του εμφράκτου.

Όσον αφορά το αιμορραγικό εγκεφαλικό επεισόδιο (ICH), μελέτες έχουν δείξει ότι τα EVs από ενδοθηλιακά κύτταρα (CD105+, CD106+, CD54+, CD62E+) εμφανίζονται αυξημένα τόσο στο AIS όσο και στο ICH. Παρότι τα EVs δεν διαφοροποιούν ακόμη αξιόπιστα τους τύπους εγκεφαλικού επεισοδίου, οι μελλοντικές μελέτες ενδέχεται να ανακαλύψουν μοναδικά προφίλ EVs που αντανακλούν τις διαφορετικές παθοφυσιολογικές οδούς του AIS και του ICH.

Εκτός από τη διαγνωστική τους αξία, τα ενδοθηλιακά EVs μπορούν να προσφέρουν πληροφορίες για την πρόγνωση και την αποκατάσταση μετά από εγκεφαλικό επεισόδιο. Αυξημένα επίπεδα CD62E+ EVs σχετίζονται με σοβαρότερα εγκεφαλικά επεισόδια και υψηλότερες βαθμολογίες στη NIHSS, ενώ οι μεταβολές στα επίπεδα EVs μπορεί να αντικατοπτρίζουν την αποτελεσματικότητα της θεραπείας. Συγκεκριμένα, μετά από AIS, τα επίπεδα των E-selectin και P-selectin EVs μειώνονται μέσα σε 3-6 μήνες, ενώ τα EVs που προέρχονται από ενεργοποιημένα αιμοπετάλια παραμένουν υψηλά, αντικατοπτρίζοντας συνεχιζόμενη ενεργοποίηση της πήξης κατά τη φάση ανάρρωσης. Κατά συνέπεια, τα ενδοθηλιακά EVs μπορούν να αποτελέσουν ισχυρούς βιοδείκτες τόσο για την εκτίμηση της σοβαρότητας του εγκεφαλικού επεισοδίου όσο και για την παρακολούθηση της θεραπευτικής ανταπόκρισης [78].

3.1.7.2 Ανίχνευση EVs

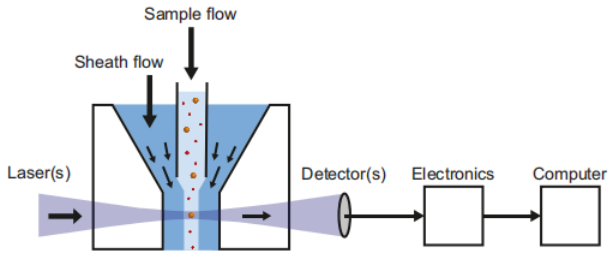
Για να έχουν κλινική εφαρμογή, οι EV βιοδείκτες πρέπει να ανιχνεύονται στην υπεροξεία φάση, να διαφοροποιούν αξιόπιστα τους τύπους εγκεφαλικού επεισοδίου και να μπορούν να μετρηθούν εύκολα με απλές διαγνωστικές συσκευές. Ωστόσο, η τρέχουσα τεχνολογία δεν επιτρέπει ακόμα την ανάλυση EVs σε πραγματικό χρόνο σε επείγοντα περιστατικά. Αντιθέτως, οι EV βιοδείκτες που αφορούν την πρόγνωση του εγκεφαλικού δεν έχουν περιορισμούς ως προς τον χρόνο απόκρισης, επιτρέποντας έτσι μεγαλύτερη πολυπλοκότητα στις μελέτες τους. Δύο πιθανές κλινικές εφαρμογές των EVs μετά από εγκεφαλικό επεισόδιο είναι: (1) ως προγνωστικοί δείκτες της μακροπρόθεσμης κλινικής έκβασης και (2) ως δείκτες παρακολούθησης της θεραπευτικής ανταπόκρισης κατά την αποκατάσταση. Ενώ οι αλλαγές στα EVs 48 ώρες μετά το εγκεφαλικό δείχνουν ελπιδοφόρα αποτελέσματα ως προγνωστικοί βιοδείκτες, είναι απαραίτητες περαιτέρω μελέτες που να επιβεβαιώνουν την ειδικότητά τους σε σύγκριση με άλλους τύπους τραυματισμών του ΚΝΣ [78].

Παράλληλα, η αξιοποίησή τους είναι δύσκολη λόγω δύο βασικών προκλήσεων: (1) Το πολύ μικρό μέγεθος των EVs καθιστά δύσκολη την ανίχνευση και τον χαρακτηρισμό τους και (2) τα EVs από διαφορετικούς κυτταρικούς τύπους συνυπάρχουν με άλλα μη-EV σωματίδια. Η ανίχνευση τους μπορεί να γίνει μέσω της κυτταρομετρίας ροής (Flow Cytometry - FCM) η οποία είναι μια πολυπαραμετρική τεχνολογία που μπορεί να ανιχνεύσει μεμονωμένα EVs και να καθορίσει την κυτταρική τους προέλευση. Συγκεκριμένα, η τεχνολογία αυτή μετρά και στη συνέχεια αναλύει πολλαπλά φυσικά χαρακτηριστικά μεμονωμένων σωματιδίων καθώς αυτά

ρέουν σε ένα υγρό ρεύμα μέσα από μια δέσμη φωτός. Οι ιδιότητες οι οποίες μετρούνται περιλαμβάνουν το σχετικό μέγεθος ενός σωματιδίου, τη σχετική κοκκίωση ή εσωτερική πολυπλοκότητα, και τη σχετική ένταση φθορισμού. Αυτά τα χαρακτηριστικά προσδιορίζονται με τη χρήση ενός συστήματος ζεύξης οπτικής-ηλεκτρονικής, το οποίο καταγράφει πως το κύτταρο ή το σωματίδιο σκεδάζει το προσπίπτον φως laser.

Οι φυσικές ιδιότητες όπως είναι το μέγεθος το οποίο αντιπροσωπεύεται από την εμπρόσθια σκέδαση του φωτός [Forward Light Scatter, FSC] και η εσωτερική πολυπλοκότητα (για παράδειγμα, σχήμα του πυρήνα, αριθμός κυτταροπλασματικών σωματιδίων ή αδρότητα κυτταρικής μεμβράνης) που αντιπροσωπεύεται από την πλάγια σκέδαση [Side Light Scatter, SSC], μπορούν να καθορίσουν συγκεκριμένους κυτταρικούς πληθυσμούς.

Όταν ενωμένα με αντισώματα κύτταρα περνούν από μία πηγή φωτός, τα φθορίζοντα μόρια διεγείρονται και μεταβαίνουν σε μία υψηλότερη ενεργειακά κατάσταση. Με την επιστροφή τους στις καταστάσεις ηρεμίας τους, τα φθοροχρώματα εκπέμπουν φωτεινή ενέργεια σε υψηλότερα μήκη κύματος. Η χρήση πολλαπλών φθοροχρωμάτων, καθένα με παρόμοια μήκη κύματος διέγερσης και διαφορετικά μήκη κύματος εκπομπής (ή «χρώματα»), επιτρέπει να μετρηθούν ταυτόχρονα αρκετές κυτταρικές ιδιότητες.



Εικόνα 3.4: Σχηματική αναπαράσταση μιας κυτταρομετρίας ροής [79]

Μέσα σε έναν κυτταρομετρητή ροής, τα κύτταρα σε εναιώρημα οδηγούνται μέσα σε ένα ρεύμα, δημιουργούμενο από έναν περιβάλλοντα μανδύα ισοτονικού υγρού (sheath flow), το οποίο δημιουργεί γραμμική ροή επιτρέποντας στα κύτταρα να περνούν μεμονωμένα μέσω ενός σημείου (interrogation point). Σε αυτό το σημείο, μία δέσμη μονοχρωματικού φωτός, συνήθως από ένα laser, διατέμνει τα κύτταρα. Το εκπεμπόμενο φως εκπέμπεται προς όλες τις κατευθύνσεις και συλλέγεται μέσω των οπτικών συστημάτων τα οποία κατευθύνουν το φως σε μια σειρά φίλτρων και διχρωικών καθρεπτών, τα οποία απομονώνουν συγκεκριμένου μήκους κύματος μπάντες. Τα φωτεινά σήματα ανιχνεύονται από λυχνίες φωτοπολλαπλασιασμού και ψηφιοποιούνται για ανάλυση στον ηλεκτρονικό υπολογιστή. Η προκύπτουσα πληροφορία συνήθως 6 παρουσιάζεται σε ιστόγραμμα ή σε μορφή δισδιάστατου διαγράμματος σημείων (dot-plot).

Επειδή όμως δεν μπορούμε να ξεχωρίσουμε τα κύτταρα απευθείας από το εξωτερικό τους, τότε τα «συνδέουμε» με τα αντίστοιχα αντισώματα. Συγκεκριμένα, αντισώματα συζευγμένα με φθορίζουσες χρωστικές, μπορούν να δένουν σε συγκεκριμένες πρωτεΐνες στις κυτταρικές μεμβράνες ή στο εσωτερικό των κυττάρων. Η χρώση με αντισώματα γίνεται μέσα σε ισοτονικό διάλυμα. Η χρώση μπορεί να γίνεται είτε απευθείας, είτε μέσω ενός δεύτερου αντισώματος, το οποίο είναι συζευγμένο με φθοροχρώματα [80].

Η διάμετρος των εξωκυττάριων κυστιδίων (EVs) αποτελεί μια σημαντική παράμετρο που μπορεί να παρέχει πολύτιμες πληροφορίες για τη βιολογική τους προέλευση και λειτουργία. Στην κυτταρομετρία ροής (FCM), η εκτίμηση της διαμέτρου βοηθά στην επαλήθευση ότι τα ανιχνευόμενα σωματίδια είναι πράγματι EVs και όχι άλλες μη-κυτταρικές δομές, όπως λιποπρωτεΐνες ή κυτταρικά θραύσματα. Στην περίπτωση του εγκεφαλικού επεισοδίου, η ανάλυση της διαμέτρου των EVs ενδέχεται να προσφέρει πληροφορίες για τη σοβαρότητα και τον τύπο του εγκεφαλικού, καθώς οι αλλαγές στη δομή και τη σύσταση των EVs μπορούν να αντανακλούν τη νευροφλεγμονή, την ενδοθηλιακή δυσλειτουργία και την ενεργοποίηση κυττάρων του Κεντρικού Νευρικού Συστήματος. Για παράδειγμα, αυξημένες συγκεντρώσεις EVs συγκεκριμένου μεγέθους μπορεί να υποδηλώνουν ενεργοποίηση ενδοθηλιακών κυττάρων μετά από ισχαιμική βλάβη, ενώ διαφορετικά μεγέθη EVs μπορεί να προέρχονται από νευρώνες ή μικρογλοιακά κύτταρα που ανταποκρίνονται στη βλάβη [81].

3.2 Διάγνωση Εγκεφαλικού με Μηχανική Μάθηση

Η παρούσα μελέτη επικεντρώνεται στη διάγνωση του εγκεφαλικού επεισοδίου, μιας σοβαρής ιατρικής κατάστασης που απαιτεί άμεση αναγνώριση και ταχεία αντιμετώπιση. Ειδικά για το εγκεφαλικό, η χρήση αλγορίθμων επιβλεπόμενης μάθησης, βαθιάς μάθησης (deep learning) και νευρωνικών δικτύων μπορεί να αποδειχθεί αποτελεσματική στη διαφοροποίηση μεταξύ ισχαιμικού και αιμορραγικού εγκεφαλικού επεισοδίου, στην πρόβλεψη του κινδύνου εγκεφαλικού και στην εξατομικευμένη θεραπεία των ασθενών.

Τα μοντέλα μηχανικής μάθησης μπορούν να λειτουργήσουν ως συστήματα υποβοήθησης λήψης αποφάσεων για τους γιατρούς, παρέχοντας πρόσθετες πληροφορίες και ανάλυση δεδομένων σε πραγματικό χρόνο. Οι αλγόριθμοι αυτοί μπορούν να συνδυάζουν δεδομένα από ιατρικές εξετάσεις, ιστορικά ασθενών και δημογραφικά χαρακτηριστικά για να προσφέρουν πιθανές διαγνώσεις ή να προτείνουν εξατομικευμένες θεραπευτικές επιλογές. Αυτό δεν σημαίνει ότι αντικαθιστούν την κρίση του ιατρού, αλλά ότι ενισχύουν τη διαδικασία λήψης αποφάσεων, μειώνοντας το περιθώριο ανθρώπινου λάθους και επιταχύνοντας την έναρξη θεραπείας. Στους ασθενείς που έχουν υποστεί εγκεφαλικό, ο χρόνος είναι κρίσιμος και ως εκ τούτο η έγκαιρη αναγνώριση και ταξινόμηση του τύπου εγκεφαλικού (σε πρώτη φάση σε ισχαιμικό ή αιμορραγικό) μέσω μοντέλων μηχανικής μάθησης μπορεί να συμβάλει στη βέλτιστη επιλογή θεραπείας, βελτιώνοντας σημαντικά τα ποσοστά επιβίωσης και αποκατάστασης των ασθενών.

Στην παρούσα εργασία, η διάγνωση εγκεφαλικού μέσω μηχανικής μάθησης γίνεται με την επεξεργασία ιατρικών δεδομένων και την εκπαίδευση μερικών μοντέλων πρόβλεψης. Αρχικά, φορτώνεται το dataset το οποίο περιέχει ορισμένες εγγραφές με χαρακτηριστικά που αφορούν ασθενείς, περιλαμβάνοντας δημογραφικά, κλινικά και αιματολογικά δεδομένα. Αφού γίνει μια αρχική επιλογή χαρακτηριστικών, τα επιλεγμένα δεδομένα χρησιμοποιούνται για την εκπαίδευση μοντέλων μηχανικής μάθησης με σκοπό την πρόβλεψη και τη διάγνωση εγκεφαλικού επεισοδίου. Έπειτα, επιλέγονται ορισμένοι ταξινομητές και εκπαιδεύονται ώστε να αναγνωρίζουν μοτίβα που σχετίζονται με την εμφάνιση εγκεφαλικού. Η απόδοσή τους αξιολογείται μέσω μετρικών όπως αναλύεται παρακάτω, ώστε να εξασφαλιστεί η αξιοπιστία του. Τέλος, το πλέον κατάλληλο εκπαιδευμένο μοντέλο μπορεί να χρησιμοποιηθεί για την πρόβλεψη του κινδύνου εγκεφαλικού σε νέους ασθενείς, συμβάλλοντας στην έγκαιρη διάγνωση και την πρόληψη.

3.3 Μετρητικές αξιολόγησης μοντέλων Μηχανικής Μάθησης

Η απόδοση ενός μοντέλου ταξινόμησης καθώς και η ικανότητα γενίκευσης της απόδοσης του μπορούν να αξιολογηθούν με βάση κάποιες μετρικές αξιολόγησης. Σε ένα τυπικό πρόβλημα ταξινόμησης, οι μετρικές αυτές έχουν χρησιμοποιηθεί σε δύο στάδια: αυτό της εκπαίδευσης (learning process) και αυτό της δοκιμής. Στο πρώτο στάδιο, χρησιμοποιούνται για την βελτιστοποίηση του αλγόριθμου προκειμένου να παράγει μια πιο ακριβή πρόβλεψη. Στο 2^{ο} στάδιο, οι μετρικές χρησιμοποιούνται για την αξιολόγηση της αποτελεσματικότητας του ταξινομητή [82]. Για προβλήματα δυαδικής ταξινόμησης, η αξιολόγηση διάκρισης της βέλτιστης λύσης κατά την εκπαίδευση ταξινόμησης μπορεί να οριστεί με βάση τον πίνακα σύγχυσης (confusion matrix) όπως φαίνεται στον παρακάτω πίνακα.

		Προβλεπόμενη κλάση	
		Κλάση 0	**Κλάση 1**
Πραγματική Κλάση	**Κλάση 0**	TN – True Negative	FP – False Positive
	Κλάση 1	FN – False Negative	TP – True Positive

Πίνακας 3.4: Πίνακας σύγχυσης για ένα πρόβλημα δυαδικής ταξινόμησης

Η σειρά του πίνακα αντιπροσωπεύει την πραγματική κλάση, ενώ η στήλη την προβλεπόμενη τάξη. Από τον πίνακα σύγχυσης, οι εγγραφές TP – True Positive και TN – True

Negative υποδηλώνουν τον αριθμό των θετικών και αρνητικών περιπτώσεων που ταξινομούνται σωστά. Όσον αφορά τις υπόλοιπες εγγραφές, τα FP – False Positive και FN – False Negative δηλώνουν τον αριθμό των λανθασμένων αρνητικών και θετικών περιπτώσεων αντίστοιχα.

Με βάση τον συγκεκριμένο πίνακα, μπορούν να δημιουργηθούν αρκετές μετρικές αξιολόγησης για την μέτρηση της απόδοσης του εκάστοτε ταξινομητή, όπως φαίνεται και στον Πίνακα 2. Για την αντιμετώπιση των προβλημάτων πολλαπλών κλάσεων χρησιμοποιούνται κατά κύριο λόγο οι 4 τελευταίες μετρικές [82].

Πίνακας 3.5: Μετρικές αξιολόγησης για μοντέλα ταξινόμησης [82]

Metrics	**Formula**	**Evaluation Focus**
Accuracy (acc)	$\frac{tp + tn}{tp + fp + tn + fn}$	Μετράει την αναλογία των θετικών προβλέψεων επί του συνόλου των περιπτώσεων που αξιολογούνται
Error rate (err)	$\frac{fp + fn}{tp + fp + tn + fn}$	Μετράει την αναλογία των αρνητικών προβλέψεων επί του συνόλου των περιπτώσεων που αξιολογούνται
Sensitivity (sn)	$\frac{tp}{tp + fn}$	Αντιπροσωπεύει το ποσοστό των θετικών περιπτώσεων που ταξινομήθηκαν σωστά
Specificity (sp)	$\frac{tn}{tn + fp}$	Αντιπροσωπεύει το ποσοστό των αρνητικών περιπτώσεων που ταξινομήθηκαν σωστά
Precision (p)	$\frac{tp}{tp + fp}$	Αντιπροσωπεύει την ακρίβεια των προβλέψεων του μοντέλου στη θετική κλάση
Recall (r)	$\frac{tp}{tp + tn}$	Αντιπροσωπεύει πόσες από τις πραγματικές θετικές περιπτώσεις ανιχνεύει το μοντέλο
F-Measure (FM)	$\frac{2 * p * r}{p + r}$	Αντιπροσωπεύει τον αρμονικό μέσο όρο μεταξύ της ανάκλησης (recall) και της ακρίβειας (precision)
Geometric-mean (GM)	$\sqrt{tp * tn}$	Χρησιμοποιείται για τη μεγιστοποίηση του ρυθμού αληθών θετικών (True Positive Rate - TPR) και του ρυθμού αληθών αρνητικών (True Negative Rate - TNR), διατηρώντας τα σχετικά ισορροπημένα
Averaged Accuracy	$\frac{\sum_{i=1}^{l} \frac{tp_i + tn_i}{tp_i + fp_i + tn_i + fn_i}}{l}$	Υπολογίζει την μέση απόδοση του μοντέλου, όπου l ο αριθμός των κλάσεων
Averaged Error Rate	$\frac{\sum_{i=1}^{l} \frac{fp_i + fn_i}{tp_i + fp_i + tn_i + fn_i}}{l}$	Υπολογίζει την μέση τιμή του ποσοστού σφάλματος όλων των κλάσεων
Averaged Precision	$\frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fp_i}}{l}$	Υπολογίζει την μέση τιμή της ακρίβειας

Averaged Recall	$\frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fn_i}}{l}$	Υπολογίζει την μέση τιμή της ανάκλησης
Averaged F-Measure	$\frac{2 * p_M * r_M}{p_M + r_M}$	Υπολογίζει την μέση τιμή του F – Measure, M για macro-averaging

Όπως έχει αποδειχθεί σε προηγούμενες έρευνες [83-84], η ακρίβεια (accuracy) είναι η πιο ευρέως χρησιμοποιούμενη μετρική αξιολόγησης στην πράξης τόσο για δυαδικά όσο και για πολλαπλών κλάσεων προβλήματα ταξινόμησης. Η ακρίβεια μετράει την αναλογία των θετικών προβλέψεων επί του συνόλου των περιπτώσεων που αξιολογούνται. Η συμπληρωματική μετρική που χρησιμοποιείται για την απόδοση ενός μοντέλου είναι το ποσοστό σφάλματος το οποίο κατά αναλογία αντιπροσωπεύει την αναλογία των αρνητικών προβλέψεων επί του συνόλου των περιπτώσεων.

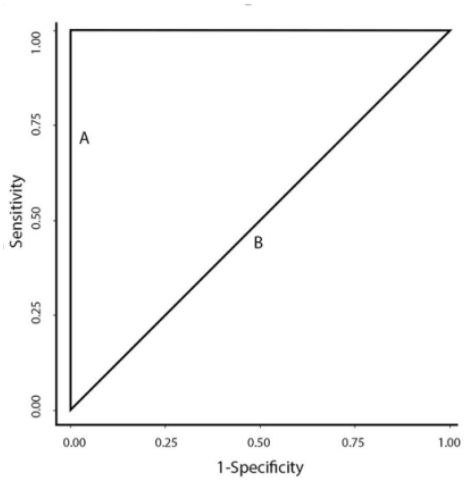
Τα πλεονεκτήματα αυτών των μετρικών είναι ότι αυτή η απόδοση και κατ' επέκταση του μοντέλου είναι εύκολο να υπολογιστεί με λιγότερη πολυπλοκότητα καθώς και να εφαρμοστεί σε προβλήματα ανεξάρτητα από το πλήθος των κλάσεων. Ωστόσο, όπως επισημαίνεται από πολλές μελέτες, η ακρίβεια έχει περιορισμούς στην κατηγοριοποίηση, καθώς παράγει λιγότερο διακριτές και λιγότερο διαχωριστικές τιμές («less distinctive and less discriminable values» [85], [86]). Κατά αυτόν τον τρόπο, καθίσταται δύσκολο η αξιολόγηση της πραγματικής απόδοσης των μοντέλων, ειδικά όταν τα δεδομένα είναι ασαφή ή ανισόρροπα. Αυτό σημαίνει ότι, ακόμα και αν δύο ταξινομητές έχουν παρόμοιες τιμές ακρίβειας, μπορεί να είναι δύσκολο να προσδιοριστεί ποιο είναι καλύτερο, καθώς η ακρίβεια δεν καταγράφει πλήρως τις επιδόσεις σε όλες τις κατηγορίες. Επιπλέον, η ακρίβεια είναι "αδύναμη" ως προς την πληροφοριακή αξία, καθώς δεν παρέχει αρκετές πληροφορίες για το πώς αποδίδει το μοντέλο σε σχέση με τις λιγότερο συχνές ή πιο σημαντικές κατηγορίες, κάνοντάς την λιγότερο κατάλληλη για την επιλογή του βέλτιστου ταξινομητή [82].

Οι υπόλοιπες μετρικές που παρουσιάζονται στον Πίνακα 2 θεωρούνται ανεπαρκείς για την εύρεση της βέλτιστης λύσης καθώς επικεντρώνονται στην αξιολόγηση μόνο μιας κατηγορίας κάθε φορά (ή της θετικής ή της αρνητικής). Αυτές οι μετρικές δεν παρέχουν αρκετές πληροφορίες για να διακρίνουν αποτελεσματικά τα μοντέλα όταν πρέπει να εξεταστούν και οι δύο κατηγορίες ταυτόχρονα [82]. Από την άλλη, οι μετρικές F-Measure (FM) και Geometric-mean (GM) έχουν αναφερθεί ότι είναι πιο χρήσιμες από την ακρίβεια για την αξιολόγηση ταξινομητών σε προβλήματα με δύο κλάσεις [87] ωστόσο δεν έχουν αξιοποιηθεί πλήρως για προβλήματα με περισσότερες κατηγορίες [82].

Επιπλέον, για την αξιολόγηση της απόδοσης ενός μοντέλου χρησιμοποιείται η περιοχή κάτω από την καμπύλη ROC (AUC). Η καμπύλη πιθανότητας ROC (Receiver Operating Characteristic) είναι η γραφική παράσταση της sensitivity στον άξονα y σε σχέση με το specificity στον x άξονα. Δημιουργείται σχεδιάζοντας τον πραγματικό θετικό ρυθμό (TPR – True Positive Rate) συγκριτικά με το ψευδές θετικό ποσοστό (FPR – False Positive Rate). Η περιοχή κάτω από την καμπύλη ονομάζεται AUC score (Area Under Curve) και χρησιμοποιείται σαν μετρικής αξιολόγησης. Όπως είναι γνωστό, όταν αυξάνεται η ευαισθησία, αυξάνεται επίσης και το FPR. Επομένως, η καμπύλη ROC επιτρέπει την παρακολούθηση της μεταβολής των δυο αυτών μετρικών καθώς αλλάζουν οι παράμετροι του μοντέλου. Αυτή η μετρική βρίσκει εφαρμογή και στην ιατρική για την αξιολόγησης της διαγνωστικής ικανότητας διαφόρων μοντέλων στον διαχωρισμό των ασθενών και των υγιών ατόμων [88]. Οι τιμές της περιοχής AUC μπορούν να ερμηνευτούν ως εξής: 0.9 – 1.0 = εξαιρετικό, 0.8 – 0.9 = ικανοποιητικό, 0.70 – 0.80 = μέτριο, 0.6 – 0.7 = κακό ενώ για τιμές 0.5 – 0.6 = αποτυχία καθώς 50% είναι και το αποτέλεσμα επιτυχίας και για μια τυχαία πρόβλεψη [89].

Στην εικόνα 1, η γραμμή Α αντιπροσωπεύει την καμπύλη ROC για μια τέλεια διαγνωστική εξέταση όπου το True Positive Rate (ευαισθησία) και το True Negative Rate (ειδικότητα) είναι 100%. Αυτό σημαίνει ότι το μοντέλο εντοπίζει σωστά όλες τις θετικές και αρνητικές περιπτώσεις χωρίς λάθη. Η AUC για αυτή την καμπύλη είναι 1.0, δηλαδή παίρνει την μέγιστη δυνατή τιμή, γεγονός που είναι πολύ δύσκολο να επιτευχθεί και στην

πραγματικότητα. Από την άλλη, η γραμμή B αντιπροσωπεύει ένα τυχαίο διαγνωστικό τεστ το οποίο και δεν προσφέρει καμία αξιόπιστη πληροφορία για τον διαχωρισμό των περιπτώσεων.



Εικόνα 3.5: Καμπύλη ROC για ένα τεστ με sensitivity and specificity στο 100% (γραμμή A) και για ένα τυχαίο τεστ (γραμμή B) [89].

Για μια ένα πρόβλημα ταξινόμησης 2 κλάσεων, η τιμή της AUC μπορεί να υπολογιστεί ως εξής:

$$AUC = \frac{S_p - \frac{n_p(n_n + 1)}{2}}{n_p * n_n}$$

όπου το Sp είναι το άθροισμα όλων των βαθμολογιών των θετικών παραδειγμάτων (όπως οι εκτιμήσεις ή οι προβλέψεις του μοντέλου), ενώ το το n_p και n_n αναφέρονται στο πλήθος των θετικών και αρνητικών παραδειγμάτων αντίστοιχα στο σύνολο των δεδομένων.

Η AUC έχει αποδειχθεί θεωρητικά και πρακτικά καλύτερη μετρική απόδοση από την ακρίβεια για την αξιολόγηση των ταξινομητών και την διάκριση της βέλτιστης λύσης κατά την εκπαίδευση του ταξινομητή [44]. Ωστόσο, το υπολογιστικό κόστος της είναι υψηλό, ειδικά όταν απαιτείται να διακρίνουμε πολλές λύσεις σε πολυκατηγορητικά προβλήματα. Ενδεικτικά, για το μοντέλο AUC του Provost και Domingos, η χρονική πολυπλοκότητα είναι O(|C|n log n), όπου |C| είναι ο αριθμός των κατηγοριών και n είναι ο αριθμός των δειγμάτων. Ενώ, για το μοντέλο AUC του Hand και Till, η χρονική πολυπλοκότητα είναι O($|C|^2$ n log n), δηλαδή, αυξάνει περισσότερο με τον αριθμό των κατηγοριών (|C|), κάνοντάς το πιο υπολογιστικά ακριβό [89].

3.4 Ανάλυση και Επεξεργασία Δεδομένων

Ο σκοπός της παρούσας εργασίας είναι να εκτιμηθεί, με τη χρήση των προαναφερθέντων μοντέλων Μηχανικής Μάθησης, η πιθανότητα εμφάνισης εγκεφαλικού επεισοδίου σε έναν ασθενή και κατόπιν η διάγνωση του είδους του εγκεφαλικού. Τα δεδομένα που χρησιμοποιήθηκαν, συγκεντρώθηκαν στο πλαίσιο έρευνας του Πανεπιστημιακού Ιατρικού Κέντρου του Άμστερνταμ. Περιλαμβάνουν 140 IDs ασθενών μαζί με αντίστοιχες πληροφορίες για επτά διαφορετικούς βιοδείκτες (CD45-APC+, CD235a-PE+, CD14-PB+, CD31-APC, CD146-PE, CD326-APC+ και Lact-FITC+) καθώς και ορισμένα κλινικά χαρακτηριστικά και αιματολογικές μετρήσεις.

3.4.1 Δεδομένα από EVs

Οι βιοδείκτες αυτοί προέρχονται από την ανίχνευση των εξωκυτταρικών κυστιδίων μέσω ροής κυτταρομετρίας. Η παρουσία των συγκεκριμένων βιοδεικτών στα EVs υποδηλώνει

την κυτταρική τους προέλευση και τον πιθανό ρόλο τους σε φυσιολογικές ή παθολογικές διεργασίες, όπως η φλεγμονή, η αγγειακή δυσλειτουργία και η θρόμβωση. Τα EVs που φέρουν αυτούς τους βιοδείκτες μπορούν να χρησιμοποιηθούν ως μη επεμβατικοί βιοδείκτες για τη διάγνωση και την πρόγνωση του εγκεφαλικού. Η ανάλυσή τους μπορεί να αποκαλύψει ενδοθηλιακή δυσλειτουργία, φλεγμονώδεις αποκρίσεις και κυτταρική βλάβη που σχετίζεται με ισχαιμικό ή αιμορραγικό εγκεφαλικό. Η ροή κυτταρομετρίας επιτρέπει την ταχεία και ακριβή ανίχνευση αυτών των μικροσωματιδίων, καθιστώντας τα EVs μια πολλά υποσχόμενη προσέγγιση για την ανάπτυξη βιοδεικτών εγκεφαλικού και καρδιαγγειακών παθήσεων.

Αναλυτικότερα, παρουσιάζονται στον παρακάτω πίνακα η λειτουργία καθενός από τους συγκεκριμένους βιοδείκτες. Αξίζει να αναλυθεί η σημειογραφία που φαίνεται στον πίνακα. Ενδεικτικά, αναφέρεται ότι το CD45 είναι ο βιοδείκτης, δηλαδή μια πρωτεΐνη επιφανείας των λευκοκυττάρων ενώ το APC+ δηλώνει ότι το αντίσωμα που στοχεύει το CD45 έχει συζευχθεί με τη φθορίζουσα χρωστική APC (Allophycocyanin), επιτρέποντας την ανίχνευσή του με κυτταρομετρία ροής. Μάλιστα, το "+" μετά το APC σημαίνει ότι το εξωκυτταρικό κυστίδιο εκφράζει τον δείκτη CD45 και ανιχνεύεται θετικό στη συγκεκριμένη φθορίζουσα χρωστική.

Πίνακας 3.6: Βιοδείκτες που χρησιμοποιήθηκαν και η κυτταρική τους προέλευση

Βιοδείκτης	**Λειτουργία και κυτταρική προέλευση**	**Πηγές**
CD45-APC+	Δείκτης των λευκοκυττάρων, που υποδηλώνει ότι τα EVs προέρχονται από ανοσοκύτταρα	[78]
CD235a-PE+	Δείκτης των ερυθρών αιμοσφαιρίων, ενδεικτικός της παρουσίας EVs από ερυθροκύτταρα	[91]
CD14-PB+	Δείκτης μονοκυττάρων και μακροφάγων, που συνδέεται με φλεγμονώδεις αποκρίσεις	[92]
CD31-APC	Δείκτης ενδοθηλιακών κυττάρων και αιμοπεταλίων, σημαντικός για τη μελέτη αγγειακής λειτουργίας, αγγειακής φλεγμονής και πήξης	[93]
CD146-PE	Δείκτης ενδοθηλιακών κυττάρων, σχετικός με την αγγειογένεση, τη φλεγμονή και την ενδοθηλιακή δυσλειτουργία	[78]
CD326-APC+	Δείκτης επιθηλιακών κυττάρων, που μπορεί να σχετίζεται με επιθηλιακής προέλευσης EVs	[94]
Lact-FITC+	Δείκτης φλεγμονής και ανοσολογικής απόκρισης	[95]

Η σύνδεση αυτών των βιοδεικτών με το εγκεφαλικό επεισόδιο βασίζεται στη συμμετοχή τους σε κρίσιμες παθοφυσιολογικές διεργασίες, όπως η φλεγμονή, η ενδοθηλιακή δυσλειτουργία, η πήξη του αίματος και η κυτταρική βλάβη. Τα εξωκυτταρικά κυστίδια που φέρουν αυτούς τους βιοδείκτες μπορούν να χρησιμοποιηθούν ως δείκτες για τη διάγνωση, την πρόγνωση και την παρακολούθηση της σοβαρότητας του εγκεφαλικού.

Συγκεκριμένα, η φλεγμονή παίζει κεντρικό ρόλο στο εγκεφαλικό επεισόδιο, τόσο στην οξεία φάση όσο και στη μακροχρόνια αποκατάσταση. Τα CD45+ EVs προέρχονται από λευκοκύτταρα και υποδηλώνουν την ενεργοποίηση του ανοσοποιητικού συστήματος, ενώ τα CD14+ EVs που προέρχονται από μονοκύτταρα και μακροφάγα σχετίζονται με τη φλεγμονώδη απόκριση. Παράλληλα, τα Lact+ EVs υποδηλώνουν την παρουσία ουδετερόφιλων, τα οποία εμπλέκονται στην οξεία φλεγμονή και στην επέκταση της εγκεφαλικής βλάβης. Παράλληλα, η ενδοθηλιακή δυσλειτουργία και η αγγειακή βλάβη είναι βασικοί μηχανισμοί που συνδέονται με το εγκεφαλικό επεισόδιο. Τα CD31+ EVs προέρχονται από ενδοθηλιακά κύτταρα και αιμοπετάλια, υποδηλώνοντας αγγειακή φλεγμονή και βλάβη. Παρόμοια, τα CD146+ EVs σχετίζονται με την αγγειογένεση, ενώ τα ICAM-1+ EVs (CD105+, CD54+, CD45−) έχουν βρεθεί να σχετίζονται με το μέγεθος του εμφράκτου, υποδεικνύοντας σοβαρότερη βλάβη. Οι

διαταραχές στην πήξη του αίματος είναι επίσης σημαντικές στην παθογένεση του εγκεφαλικού. Τα CD41 EVs υποδηλώνουν μη-αιμοπεταλιακής προέλευσης εξωκυτταρικά κυστίδια που σχετίζονται με ενεργοποίηση του πηκτικού μηχανισμού, ενώ τα CD235a+ EVs υποδηλώνουν διάσπαση ερυθροκυττάρων, κάτι που μπορεί να παρατηρηθεί σε αιμορραγικό εγκεφαλικό.

Συνολικά, η ανίχνευση των εξωκυτταρικών κυστιδίων και των βιοδεικτών τους μπορεί να προσφέρει πολύτιμες πληροφορίες για τη διάγνωση, την πρόγνωση και τη θεραπευτική στόχευση του εγκεφαλικού. Η αύξηση συγκεκριμένων EVs σε διαφορετικές χρονικές στιγμές μπορεί να βοηθήσει στον προσδιορισμό της σοβαρότητας του εγκεφαλικού, στον διαχωρισμό του από άλλα αγγειακά συμβάντα, καθώς και στην πρόβλεψη της μακροχρόνιας αποκατάστασης των ασθενών. Επομένως, οι συγκεκριμένοι βιοδείκτες επιλέχθηκαν προσεκτικά, με στόχο την περαιτέρω διερεύνησή τους, προκειμένου να αποκαλυφθεί ο ρόλος τους στις παθοφυσιολογικές διεργασίες και η πιθανή διαγνωστική ή προγνωστική τους αξία.

Για κάθε έναν από τους βιοδείκτες, έχουν καταγραφεί μέσω ανάλυσης flow cytometry για εξωκυτταρικά κυστίδια (EVs) οι εξής μετρικές: Total, Intensity, Total Concentration, Mean, Median, Std, kurtosis, skewness, entropy, καθώς και οι συγκεντρώσεις EVs σε συγκεκριμένα εύρη μεγέθους (100-200 nm, 200-300 nm, 300-400 nm, 400-500 nm, 500-600 nm, 600-700 nm, 700-800 nm, 800-900 nm και 900-1000 nm). Οι συγκεκριμένες μετρικές παρέχουν πολύτιμες πληροφορίες για τις φυσικοχημικές ιδιότητες των EVs, τη διασπορά των βιοδεικτών στην επιφάνειά τους, καθώς και τη συμπεριφορά των αντισωμάτων κατά την ανίχνευση.

Η μέτρηση του συνολικού αριθμού των σωματιδίων (Total) και της συγκέντρωσής τους (Total Concentration) προσφέρει μια γενική εκτίμηση της παρουσίας των EVs στο δείγμα, αντικατοπτρίζοντας πιθανά παθολογικά μοτίβα που σχετίζονται με το εγκεφαλικό επεισόδιο. Η ένταση του σήματος (Intensity) αντιστοιχεί στην έκφραση των επιφανειακών βιοδεικτών, αποκαλύπτοντας πληροφορίες για τη σύνθεση και τη βιολογική δραστικότητα των EVs. Τα στατιστικά μέτρα όπως ο μέσος όρος (Mean), η διάμεσος (Median) και η τυπική απόκλιση (Std) περιγράφουν την κατανομή των τιμών, παρέχοντας ενδείξεις για την ομοιογένεια ή την ετερογένεια των EVs στο δείγμα. Επιπλέον, η κυρτότητα (Kurtosis) και η ασυμμετρία (Skewness) αποτυπώνουν το σχήμα της κατανομής των δεδομένων, εντοπίζοντας πιθανά ακραία μοτίβα που μπορεί να σχετίζονται με παθολογικές μεταβολές. Η εντροπία (Entropy) εκφράζει την πολυπλοκότητα και την ποικιλομορφία του πληθυσμού των EVs, παρέχοντας ενδείξεις για διαφορετικούς κυτταρικούς τύπους ή μοριακές τροποποιήσεις που μπορεί να συσχετίζονται με το εγκεφαλικό. Τέλος, η κατανομή των EVs σε διαφορετικά εύρη μεγέθους (100-1000 nm) προσφέρει κρίσιμες πληροφορίες για την ταξινόμηση του πληθυσμού τους, βοηθώντας στη διαφοροποίηση των τύπων EVs ανάλογα με την κυτταρική προέλευση και τη φυσιολογική ή παθολογική τους λειτουργία.

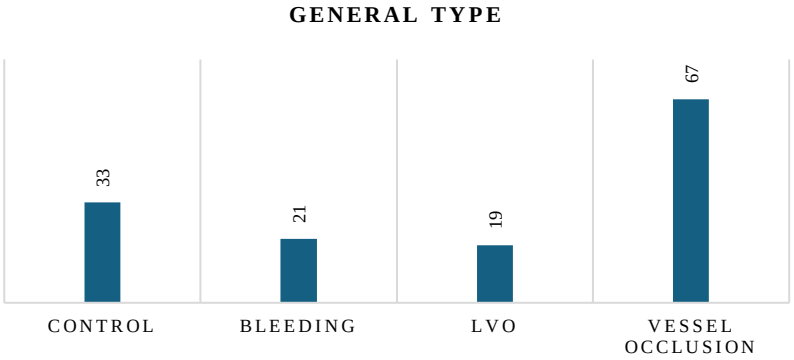
3.4.2 Κλινικά και δημογραφικά χαρακτηριστικά

Παράλληλα, το σύνολο δεδομένων περιείχε τόσο κλινικά χαρακτηριστικά (π.χ., πίεση αίματος, επίπεδα οξυγόνου, ηλικία) όσο και αιματολογικές μετρήσεις (π.χ., αριθμός αιμοπεταλίων, λευκά αιμοσφαίρια) τα οποία και αναλύονται παρακάτω και προσφέρουν μια ολοκληρωμένη βάση για τη μελέτη. Παρά το ενδιαφέρον περιεχόμενο, το dataset δεν ήταν αρκετά μεγάλο, με μόλις 140 παρατηρήσεις, ώστε να αξιοποιηθεί για την εκπαίδευση πιο πολύπλοκων μοντέλων μηχανικής μάθησης, όπως τα πολυεπίπεδα νευρωνικά δίκτυα. Ωστόσο, η μορφή των δεδομένων ως CSV αρχείων διευκόλυνε την επεξεργασία και τη φόρτωσή τους στα μοντέλα. Τα χαρακτηριστικά που αξιοποιήθηκαν και περιγράφουν τη ιατρική κατάσταση των ασθενών παρουσιάζονται παρακάτω.

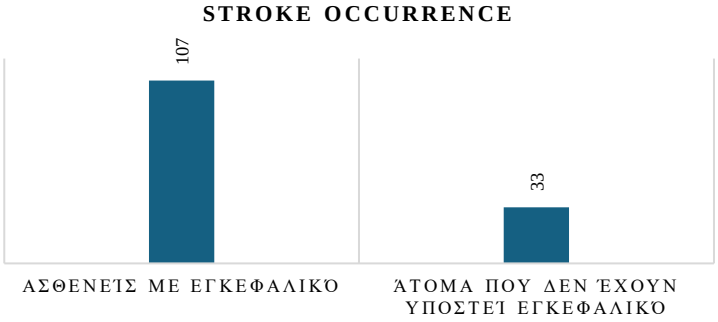
Στο dataset που χρησιμοποιήθηκε, η στήλη General Type αναφέρεται στον γενικό τύπο/είδος του εγκεφαλικού του ασθενούς και περιλαμβάνει τέσσερις βασικές κατηγορίες: Vessel occlusion (Αγγειακή απόφραξη), Control (Υγιής), Bleeding (Αιμορραγία) και LVO (Μεγάλη Αγγειακή Απόφραξη). Στην συγκεκριμένη ανάλυση, το μεγαλύτερο ποσοστό (47%) των περιστατικών αφορά την κατηγορία Vessel occlusion, κάτι που υποδηλώνει ότι αγγειακή απόφραξη είναι ο συνηθέστερος τύπος εγκεφαλικού επεισοδίου στο σύνολο δεδομένων. Η αγγειακή απόφραξη (vessel occlusion) αναφέρεται σε απόφραξη αιμοφόρου αγγείου, που

εμποδίζει την παροχή αίματος και οξυγόνου. Στον εγκέφαλο, είναι κύρια αιτία ισχαιμικού εγκεφαλικού επεισοδίου, είτε μέσω θρόμβου που σχηματίζεται τοπικά είτε εμβολής από άλλο σημείο του σώματος. Η κατηγορία Control αντιπροσωπεύει το 24%, δηλαδή τις ομάδες εκείνες που δεν έχουν υποστεί εγκεφαλικό. Η συγκεκριμένη κατηγορία είναι σημαντική για τη σύγκριση υγιών ατόμων με ασθενείς. Τα περιστατικά Bleeding ανέρχονται στο 15%, δείχνοντας ότι τα εγκεφαλικά επεισόδια λόγω αιμορραγίας είναι λιγότερο συχνά. Ο όρος bleeding stroke αναφέρεται σε αιμορραγικό εγκεφαλικό επεισόδιο (hemorrhagic stroke), το οποίο προκαλείται από αιμορραγία μέσα ή γύρω από τον εγκέφαλο. Αυτό συμβαίνει όταν ένα αιμοφόρο αγγείο στον εγκέφαλο σπάσει, προκαλώντας διαρροή αίματος στον εγκεφαλικό ιστό ή στις γύρω περιοχές. Υπάρχουν δύο κύριοι τύποι αιμορραγικού εγκεφαλικού. Ο πρώτος τύπος είναι η Ενδοεγκεφαλική αιμορραγία (Intracerebral Hemorrhage), δηλαδή η αιμορραγία μέσα στον εγκεφαλικό ιστό, συνήθως λόγω υπέρτασης ή τραυματισμού, ενώ ο δεύτερος είναι η Υπαραχνοειδής αιμορραγία (Subarachnoid Hemorrhage) που συμβαίνει μεταξύ του εγκεφάλου και της λεπτής μεμβράνης που τον καλύπτει, συχνά λόγω ρήξης ανευρύσματος. Τέλος, η κατηγορία LVO (Large Vessel Occlusion) αποτελεί το 14%, γεγονός που υποδηλώνει ότι η απόφραξη μεγάλων αγγείων είναι ο λιγότερο συχνός τύπος εγκεφαλικού σε αυτό το δείγμα. Ο όρος LVO (Large Vessel Occlusion) αναφέρεται σε ισχαιμικά εγκεφαλικά επεισόδια που προκαλούνται από αποφράξεις σε μεγάλες εγκεφαλικές αρτηρίες [96].

Επιπλέον, στο σύνολο των δεδομένων που χρησιμοποιήθηκαν περιλαμβάνεται και η στήλη «Stroke Occurrence». Η συγκεκριμένη στήλη αναφέρεται στο αν έχει συμβεί ή όχι εγκεφαλικό επεισόδιο και καταγράφει την παρουσία ή απουσία του εγκεφαλικού επεισοδίου για κάθε ασθενή. Αυτή η στήλη αποτελεί και τη στήλη – στόχο για τη διάγνωση, δηλαδή είναι η μεταβλητή που θέλουμε να προβλέψουμε, δηλαδή αν κάποιος έχει υποστεί εγκεφαλικό επεισόδιο ή όχι. Στη συγκεκριμένη περίπτωση, έχουμε 107 θετικά δείγματα (άτομα που έχουν υποστεί εγκεφαλικό επεισόδιο) και 33 αρνητικά δείγματα (άτομα που δεν έχουν υποστεί εγκεφαλικό επεισόδιο).



Εικόνα 3.6: Κατανομή ασθενών ανά κατηγορία γενικού τύπου (General Type)



Εικόνα 3.7: Κατανομή ασθενών που έχουν υποστεί εγκεφαλικό

Παράλληλα, περιλαμβάνονται οι στήλες «Gender», «Age», «RR syst», «RR diast», «Oxygen Saturation», «Thrombocytes», «WBC», «CRP», «glucose» και «Nat». Αναλυτικότερα, η στήλη «Gender» καταγράφει το φύλο του ασθενούς, με τους άνδρες να

αντιστοιχούν στην τιμή 0 και τις γυναίκες στην τιμή 1. Στη συγκεκριμένη περίπτωση, υπάρχουν 70 άνδρες και 70 γυναίκες στο σύνολο των δεδομένων. Η στήλη «Age» αναφέρεται στην ηλικία του ασθενούς, η οποία αποτελεί συχνά σημαντικό παράγοντα κινδύνου για την εμφάνιση εγκεφαλικού επεισοδίου όπως δείχνουν και μελέτες. Στο σύνολο των δεδομένων, οι ηλικίες κυμαίνονται από 31 έως 98 ετών, με τον μέσο όρο ηλικίας να είναι τα 67 έτη.

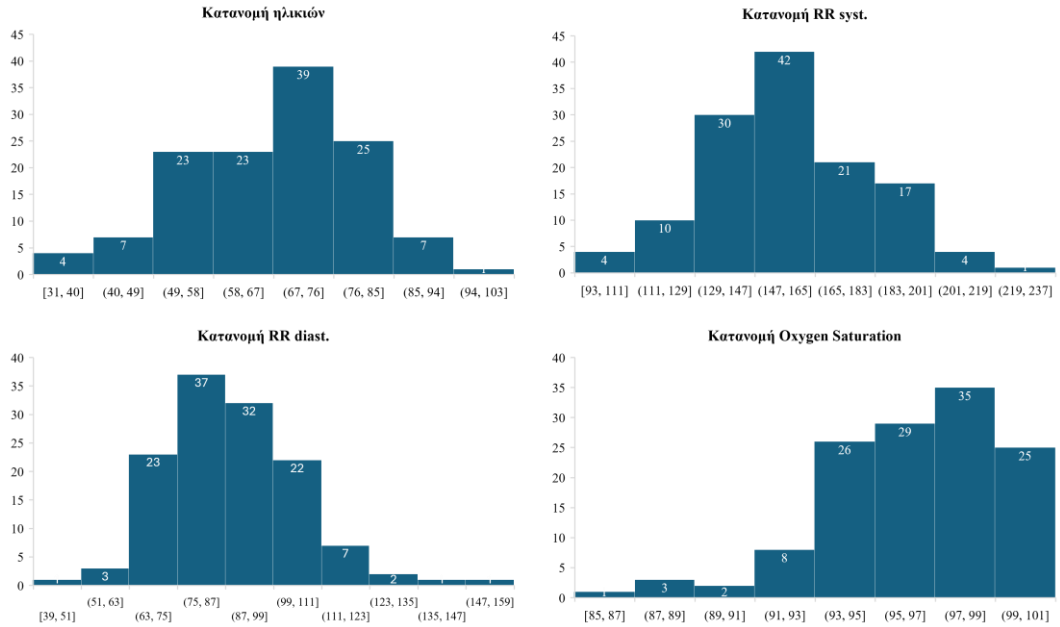
Οι στήλες RR syst. και RR diast. αναφέρονται στην αρτηριακή πίεση του ασθενούς. Η συστολική πίεση (RR syst.) είναι η υψηλότερη τιμή κατά τη διάρκεια της συστολής της καρδιάς, και η διαστολική πίεση (RR diast.) είναι η χαμηλότερη τιμή όταν η καρδιά χαλαρώνει. Στα δεδομένα, η συστολική πίεση κυμαίνεται από 93 έως 224 mmHg με μέσο όρο τα 157,84 mmHg, ενώ η διαστολική πίεση κυμαίνεται από 39 έως 152 mmHg με μέσο όρο τα 89,54 mmHg. Ο κορεσμός οξυγόνου (Oxygen Saturation) δείχνει πόσο καλά οξυγονώνεται ο οργανισμός, με τιμές από 85% έως 100%, και μέσο όρο 96,68%.

Η στήλη Thrombocytes καταγράφει τον αριθμό των αιμοπεταλίων στο αίμα, που είναι κρίσιμα για την πήξη του αίματος, με τιμές από 10 έως 618 x10^3/μL και μέσο όρο τα 257,59 x10^3/μL. Η στήλη WBC (λευκά αιμοσφαίρια) δείχνει την ανοσολογική απόκριση του οργανισμού, με τιμές από 0 έως 18,4 x10^3/μL και μέσο όρο 8,64 x10^3/μL. Η C-αντιδρώσα πρωτεΐνη (CRP) είναι δείκτης φλεγμονής με τιμές από 0,3 έως 51,1 mg/L, και μέσο όρο 5,27 mg/L. Η γλυκόζη (Glucose) αναφέρεται στο επίπεδο γλυκόζης στο αίμα, με τιμές από 4,5 έως 27,2 mmol/L και μέσο όρο 8,06 mmol/L.

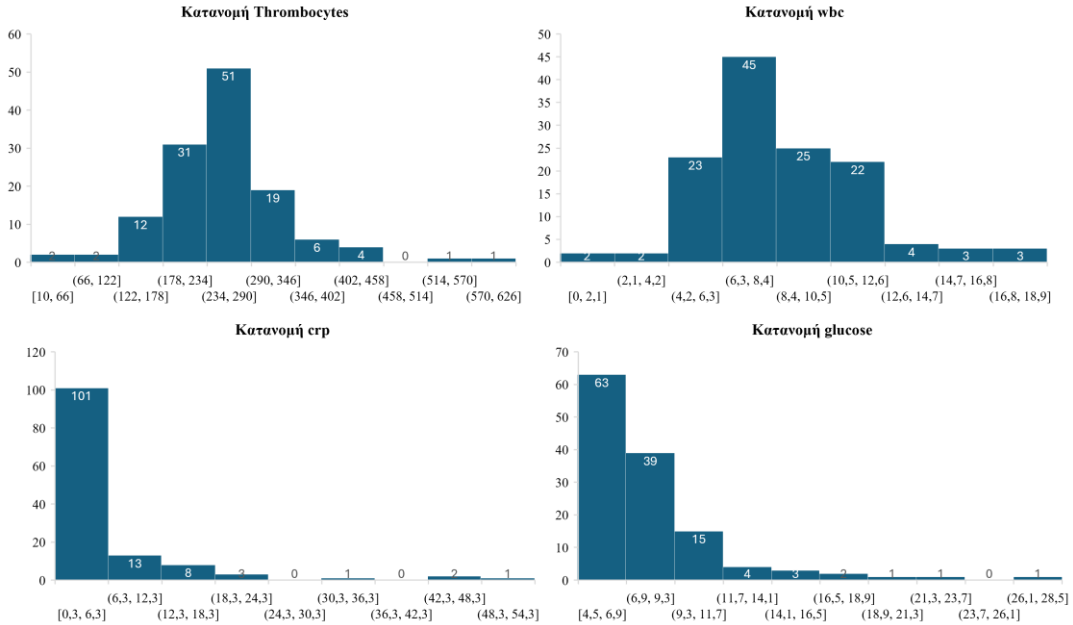
Τέλος, η συγκέντρωση νατρίου (Nat) στο αίμα κυμαίνεται από 127 έως 143 mmol/L, με μέσο όρο 137,13 mmol/L, και είναι σημαντική για τη διατήρηση της ισορροπίας υγρών και ηλεκτρολυτών στον οργανισμό.

Πίνακας 3.7: Στατιστική ανάλυση των τιμών για τα κλινικά χαρακτηριστικά

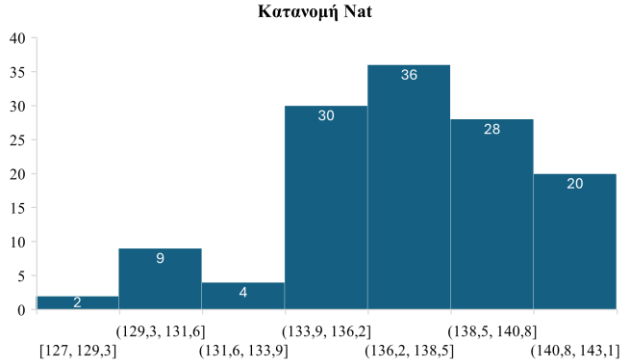
Μεταβλητή	**Μέση Τιμή**	**Διάμεσος**	**Τυπική Απόκλιση**	**Min**	**Max**
Ηλικία (Age)	67,5	69	13	31	98
Συστολική Πίεση (RR syst.)	157.8	156	25	93	224
Διαστολική Πίεση (RR diast.)	89.5	88	16.9	39	153
Οξυγόνωση (Oxygen Sat.)	96.6	97	2.9	85	100
Θρομβοκύτταρα (Thrombocytes)	257.6	255	80.6	10	618
WBC (Λευκά Αιμοσφαίρια)	8.6	8	3.05	0	18.4
CRP (Φλεγμονώδης Δείκτης)	5.27	2	8.7	0.3	51.1
Γλυκόζη (Glucose)	8.1	7.1	3.5	4.5	27.2
Νάτριο (Nat)	137.13	138	3.27	127	143



Εικόνα 3.8: Κατανομή δεδομένων στο σύνολο δεδομένων μας (1)



Εικόνα 3.9: Κατανομή δεδομένων στο σύνολο δεδομένων μας (2)



Εικόνα 3.10: Κατανομή δεδομένων στο σύνολο δεδομένων μας (3)

3.4.3 Επεξεργασία δεδομένων

Για να μπορούν να χρησιμοποιηθούν τα δεδομένα στα μοντέλα μηχανικής μάθησης, απαιτείται μια διαδικασία προεπεξεργασίας που διασφαλίζει την ποιότητά τους και την καταλληλότητά τους για ανάλυση. Η προεπεξεργασία περιλαμβάνει διαχείριση ελλιπών τιμών, μετατροπή κατηγορικών δεδομένων, επιλογή χαρακτηριστικών, κανονικοποίηση αριθμητικών μεταβλητών και εξισορρόπηση κλάσεων.

Αρχικά, πραγματοποιήθηκε διαχείριση των ελλιπών τιμών με τη χρήση της dropna(), αφαιρώντας τυχόν ελλιπείς εγγραφές ώστε να διασφαλιστεί ότι το σύνολο δεδομένων είναι πλήρες. Στη συνέχεια, εφαρμόστηκε η επεξεργασία κατηγορικών δεδομένων, καθώς οι περισσότεροι αλγόριθμοι μηχανικής μάθησης δεν μπορούν να διαχειριστούν μη αριθμητικές τιμές. Για τη μεταβλητή που αφορά την κατάσταση των ασθενών η κατηγορία Vessel occlusion αντιστοιχήθηκε στο 0, η Control στο 1, η Bleeding στο 2 και η LVO στο 3. Παράλληλα, η στήλη "Stroke Occurrence", η οποία δηλώνει αν ένας ασθενής υπέστη εγκεφαλικό επεισόδιο (1) ή όχι (0).

Στη συνέχεια, πραγματοποιήθηκε διαχωρισμός των δεδομένων σε σύνολα εκπαίδευσης (train) και δοκιμής (test) με τη χρήση της train_test_split, κατανέμοντας το 70% των δεδομένων στην εκπαίδευση και το 30% στη δοκιμή. Με αυτόν τον τρόπο, διασφαλίστηκε ότι το μοντέλο δεν υπερεκπαιδεύεται και μπορεί να αξιολογηθεί σωστά σε νέα δεδομένα. Ακολούθησε επιλογή των σημαντικότερων χαρακτηριστικών μέσω της μεθόδου Recursive Feature Elimination (RFE), χρησιμοποιώντας έναν Random Forest Classifier ως εκτιμητή. Το RFE απομάκρυνε σταδιακά τα λιγότερο σημαντικά χαρακτηριστικά, όπως αναλύεται παρακάτω.

Για την καλύτερη προσαρμογή των δεδομένων σε συγκεκριμένους αλγόριθμους μηχανικής μάθησης, πραγματοποιήθηκε κανονικοποίηση (scaling) των αριθμητικών χαρακτηριστικών μέσω του StandardScaler(). Αυτή η τεχνική εφαρμόστηκε σε αλγορίθμους όπως Naive Bayes, K-Nearest Neighbors (KNN) και Support Vector Machines (SVM), καθώς επηρεάζονται από την κλίμακα των χαρακτηριστικών. Το StandardScaler αφαίρεσε τον μέσο όρο από κάθε χαρακτηριστικό και κλιμάκωσε τις τιμές ώστε να έχουν μονάδα διακύμανση (unit variance), διασφαλίζοντας ότι όλα τα χαρακτηριστικά συνεισφέρουν ισότιμα και αποτρέποντας την υπερίσχυση αυτών με μεγαλύτερα αριθμητικά μεγέθη.

Τέλος, όπου απαιτούνταν, εφαρμόστηκε εξισορρόπηση των κλάσεων (class balancing) για να αντιμετωπιστεί τυχόν ανισορροπία μεταξύ των παρατηρήσεων. Σε προβλήματα ταξινόμησης, η ύπαρξη μιας μη ισορροπημένης κατανομής κλάσεων μπορεί να προκαλέσει προκατάληψη του μοντέλου προς την πλειοψηφούσα κλάση, μειώνοντας την ικανότητά του να προβλέψει σωστά τη μειοψηφούσα. Για να διορθωθεί αυτό το ζήτημα, εφαρμόστηκαν τεχνικές όπως oversampling της μειοψηφούσας κλάσης και undersampling της πλειοψηφούσας, καθώς και η μέθοδος SMOTE (Synthetic Minority Over-sampling Technique), η οποία δημιουργεί συνθετικά δείγματα για την ενίσχυση της μειοψηφούσας κλάσης. Αυτή η διαδικασία συνέβαλε στη βελτίωση της ικανότητας του μοντέλου να γενικεύει σωστά και να κάνει αξιόπιστες προβλέψεις.

Η ολοκληρωμένη αυτή προσέγγιση της προεπεξεργασίας διασφάλισε ότι το σύνολο δεδομένων ήταν καθαρό, σωστά κλιμακωμένο και ισορροπημένο, διευκολύνοντας την εκπαίδευση των αλγορίθμων και βελτιώνοντας την απόδοσή τους.

4. Κεφάλαιο: Αποτελέσματα

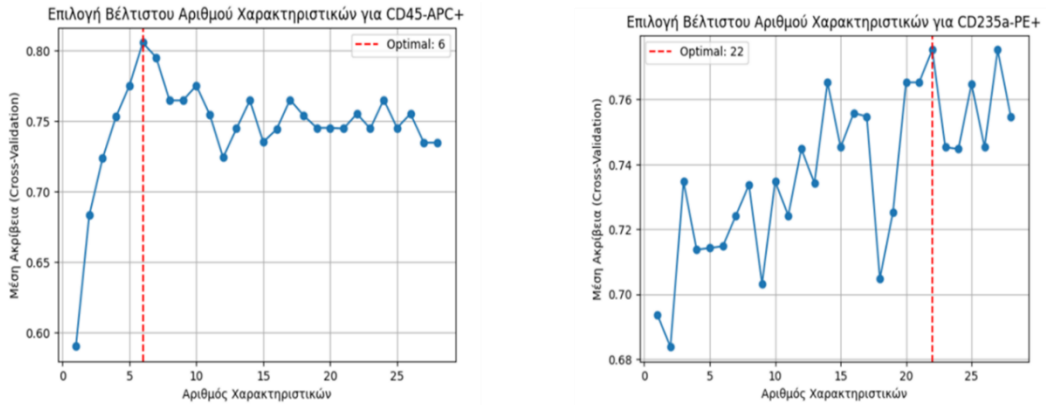
4.1 Μέρος 1ο: Διάγνωση Εγκεφαλικού

4.1.1 Αποτελέσματα από την επιλογή χαρακτηριστικών

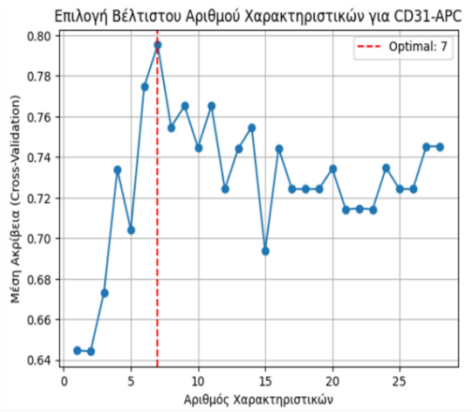
Στο πρώτο μέρος της παρούσας εργασίας, ο στόχος είναι να προβλεφθεί η πιθανότητα εμφάνισης εγκεφαλικού επεισοδίου σε ασθενείς. Το σύνολο δεδομένων που χρησιμοποιήθηκε περιλάμβανε 34 στήλες με χαρακτηριστικά που αντιστοιχούσαν σε κάθε ID ασθενούς. Δεδομένου του μεγάλου αριθμού χαρακτηριστικών, κρίθηκε αναγκαίο να πραγματοποιηθεί διαδικασία επιλογής χαρακτηριστικών (feature selection), ώστε να εντοπιστούν τα πιο σημαντικά χαρακτηριστικά που συμβάλλουν στην πρόβλεψη, να μειωθεί η πολυπλοκότητα του μοντέλου και παράλληλα να βελτιωθεί η ακρίβεια των αποτελεσμάτων.

Η διαδικασία επιλογής χαρακτηριστικών (feature selection) που ακολουθήθηκε είχε στόχο την πρόβλεψη της στήλης «Stroke Occurrence», η οποία υποδεικνύει αν ένας ασθενής εμφάνισε εγκεφαλικό επεισόδιο ή όχι. Αρχικά, από το αρχικό σύνολο δεδομένων, που περιείχε πλήθος χαρακτηριστικών, αφαιρέθηκαν στήλες που δεν θεωρήθηκαν σχετικές ή συμβάλλουσες στη διαδικασία πρόβλεψης, όπως οι στήλες ID, General Type, Class, file, και Diagnosis. Τα δεδομένα στη συνέχεια χωρίστηκαν σε σύνολα εκπαίδευσης και ελέγχου (70%-30%), διασφαλίζοντας την αξιοπιστία της αξιολόγησης του μοντέλου. Ο διαχωρισμός αυτός έγινε πριν από τη διαδικασία επιλογής χαρακτηριστικών, ώστε να αποτραπεί η διαρροή πληροφορίας (data leakage) από το σύνολο ελέγχου στο σύνολο εκπαίδευσης. Με αυτόν τον τρόπο, διασφαλίστηκε ότι η επιλογή χαρακτηριστικών βασίστηκε αποκλειστικά στα δεδομένα του συνόλου εκπαίδευσης, εξασφαλίζοντας ότι το σύνολο ελέγχου παραμένει ανεξάρτητο και ότι η αξιολόγηση του μοντέλου αντιπροσωπεύει ρεαλιστικά τη γενίκευσή του σε νέα δεδομένα.

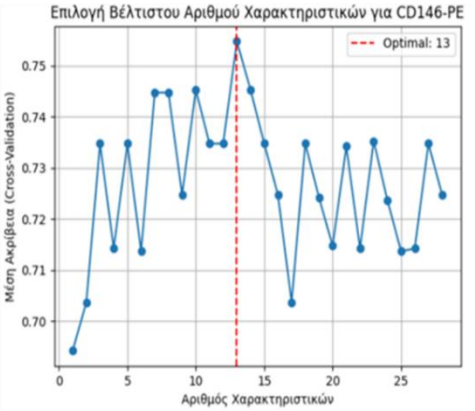
Για την επιλογή των πιο σημαντικών χαρακτηριστικών, χρησιμοποιήθηκε ο αλγόριθμος Recursive Feature Elimination (RFE) σε συνδυασμό με τον Random Forest Classifier να λειτουργεί ως βάση. Ο RFE επανέλαβε τη διαδικασία εκπαίδευσης, αφαιρώντας σταδιακά λιγότερο σημαντικά χαρακτηριστικά. Για κάθε πιθανό αριθμό χαρακτηριστικών, υπολογίστηκε η μέση ακρίβεια μέσω διασταυρωμένης επικύρωσης (5-fold cross-validation) προκειμένου να εκτιμηθεί η απόδοση του μοντέλου. Η ανάλυση ανέδειξε ότι ο βέλτιστος αριθμός χαρακτηριστικών διαφέρει για κάθε βιοδείκτη όπως απεικονίζεται στα διαγράμματα παρακάτω. Ωστόσο, παρατηρήθηκε ότι η απόδοση του μοντέλου, όταν επιλέχθηκαν 10 χαρακτηριστικά, ήταν υψηλή και σταθερή, όπως υποδείχθηκε από τον αλγόριθμο RFE. Ως εκ τούτου, για τη συνέχεια της ανάλυσης και την απλοποίηση της διαδικασίας, αποφασίστηκε να επιλεγούν τα 10 καλύτερα χαρακτηριστικά. Αυτή η επιλογή επιτυγχάνει έναν συμβιβασμό μεταξύ της ακρίβειας του μοντέλου και της απλοποίησης της διαδικασίας, διασφαλίζοντας ότι το μοντέλο παραμένει ερμηνεύσιμο και κατάλληλο για ευρύτερη εφαρμογή.



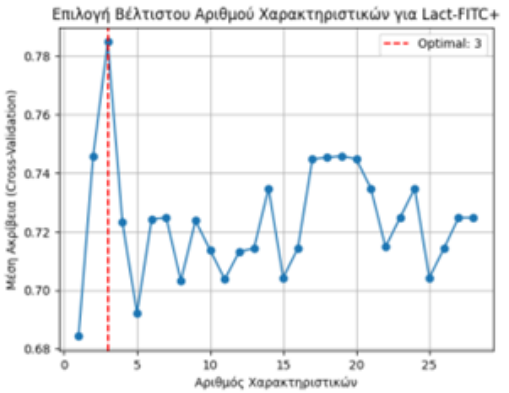
Εικόνα 4.1: Βέλτιστος αριθμός χαρακτηριστικών για διάγνωση εγκεφαλικού (CD45-APC+ - CD235a-PE+)



Εικόνα 4.2: Βέλτιστος αριθμός χαρακτηριστικών για διάγνωση εγκεφαλικού (CD14-PB+ - CD31-APC)



Εικόνα 4.3: Βέλτιστος αριθμός χαρακτηριστικών για διάγνωση εγκεφαλικού (CD326-APC+ - CD146-PE)



Εικόνα 4.4: Βέλτιστος αριθμός χαρακτηριστικών για διάγνωση εγκεφαλικού Lact-FITC+

Όσον αφορά την επιλογή χαρακτηριστικών, στον παρακάτω πίνακα παρουσιάζονται τα επιλεγμένα χαρακτηριστικά όπως αυτά προέκυψαν από την διαδικασία που ακολουθήθηκε.

Πίνακας 4.1: Συγκεντρωτικός πίνακας για την επιλογή χαρακτηριστικών για το 1ο μέρος

Βιοδείκτες	**Επιλεγμένα χαρακτηριστικά για το 1ο μέρος**
CD45-APC+	Age, RR syst., Oxygen saturation, Thrombocytes, wbc, glucose, Intensity, Median, 100-200 nm, 400-500 nm
CD235a-PE+	Age, RR syst., Thrombocytes, wbc, glucose, Total, Total Concentration, Mean, Median, kurtosis

CD14-PB+	Age, RR syst., Thrombocytes, Total, Mean, kurtosis, skewness, 200-300 nm, 400-500 nm, 500-600 nm
CD31-APC	Age, RR syst., Thrombocytes, wbc, glucose, Total, entropy, 100-200 nm, 300-400 nm, 800-900 nm
CD146-PE	Age, RR syst., Thrombocytes, glucose, Median, Std, kurtosis, skewness, 100-200 nm, 900-1000 nm
CD326-APC+	Age, RR syst., Thrombocytes, glucose, Total Concentration, Median, skewness, entropy, 100-200 nm, 200-300 nm
Lact-FITC+	Age, RR syst., Thrombocytes, wbc, glucose, Intensity, Total Concentration, Mean, 400-500 nm, 900-1000 nm

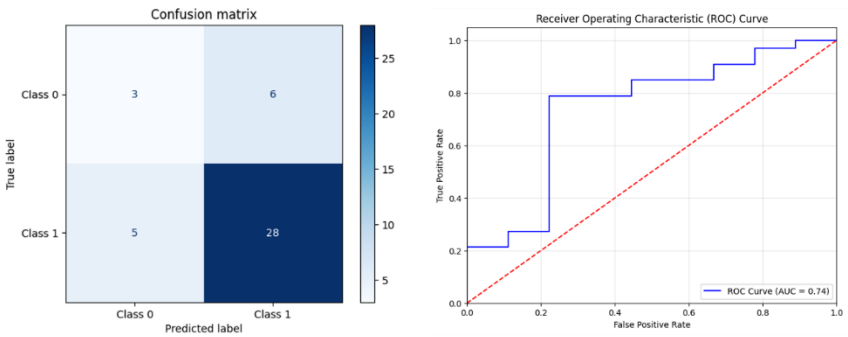
4.1.2 Αποτελέσματα Απόδοσης Μοντέλων Μηχανικής Μάθησης

Παρακάτω παρουσιάζονται τα αποτελέσματα για κάθε δείκτη, ανάλογα με το μοντέλο μηχανικής μάθησης που χρησιμοποιήθηκε. Τα μοντέλα αξιολογήθηκαν με βάση μετρικές όπως ακρίβεια, precision, recall, F1-score και AUC score. Επιπλέον, πραγματοποιήθηκε ανάλυση μέσω πίνακα σύγχυσης και καμπύλης ROC για την καλύτερη κατανόηση της απόδοσης κάθε μοντέλου. Κατά την εκπαίδευση των μοντέλων χρησιμοποιήθηκε η μέθοδος cross-validation, η οποία επιλέγει τυχαία τμήματα του συνόλου δεδομένων για εκπαίδευση και αξιολόγηση. Ως αποτέλεσμα, το άθροισμα των στοιχείων στις γραμμές και τις στήλες του πίνακα σύγχυσης διαφέρει για κάθε μέθοδο επιλογής χαρακτηριστικών. Αυτό συμβαίνει επειδή το μέγεθος και η σύνθεση των δειγμάτων που εξετάζονται και αξιολογούνται δεν είναι σταθερά, αλλά μεταβάλλονται σε κάθε επανάληψη του cross-validation.

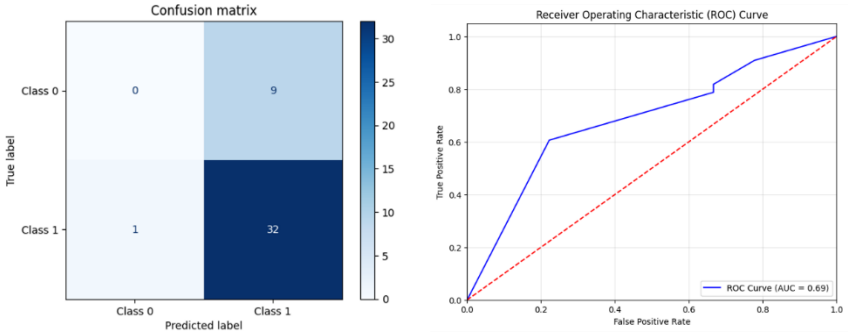
CD45 – APC+

Πίνακας 4.2: Απόδοση των μοντέλων για τον βιοδείκτη CD45 – APC+

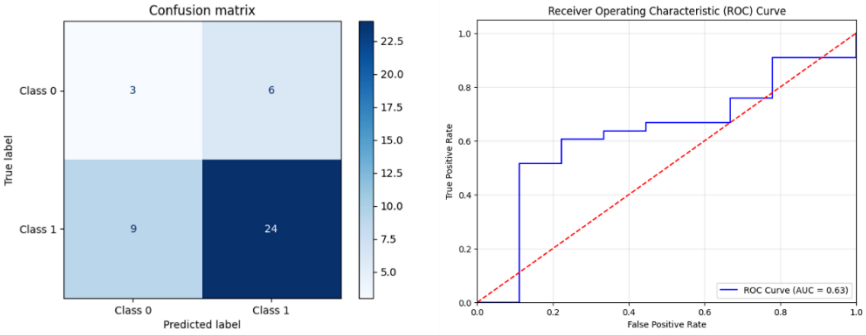
	Accuracy	Precision	Recall	F1-score	AUC score
Gradient Boosting	78.6%	85.3%	87.9%	86.6%	74%
Random Forest	78.7%	86.3%	87.9%	86.6%	81%
KNN	78.6%	78.6%	100%	88%	58%
SVM	59.5%	80.8%	63.6%	71.2%	46%
Naïve Bayes	64.3%	80%	72.7%	76.2%	63%
Decision Tree	76.2%	78%	97%	86.5%	69%
XG Boost	73.8%	82.4%	84.8%	83.6%	74%



Εικόνα 4.5: Confusion matrix & ROC curve (CD45-APC+) – Gradient Boosting

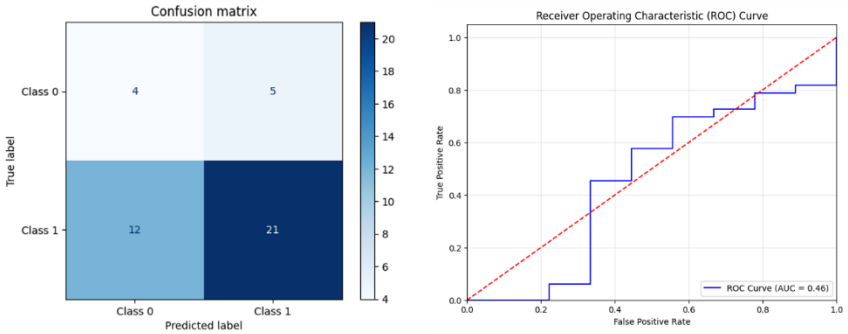


Εικόνα 4.6: Confusion matrix & ROC curve (CD45-APC+) – Random Forest

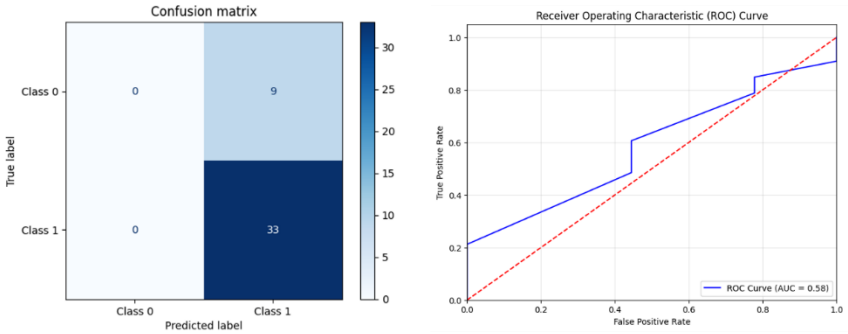


\

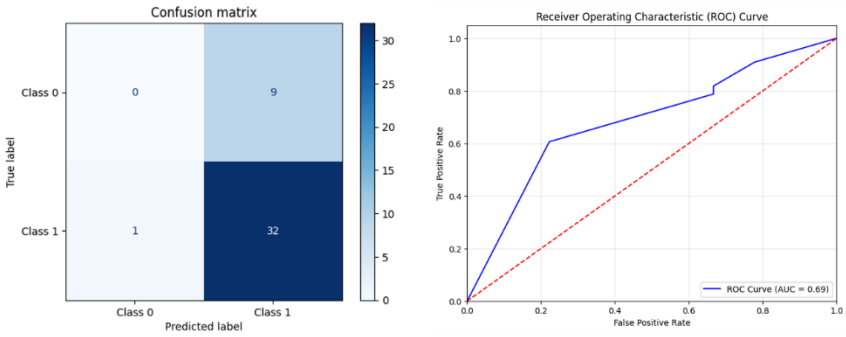
Εικόνα 4.7 Confusion matrix & ROC curve (CD45-APC+) – KNN



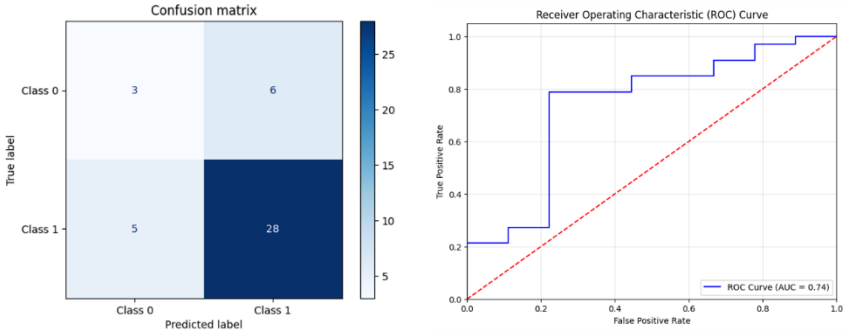
Εικόνα 4.8: Confusion matrix & ROC curve (CD45-APC+) – SVM



Εικόνα 4.9: Confusion matrix & ROC curve (CD45-APC+) – Naive Bayes

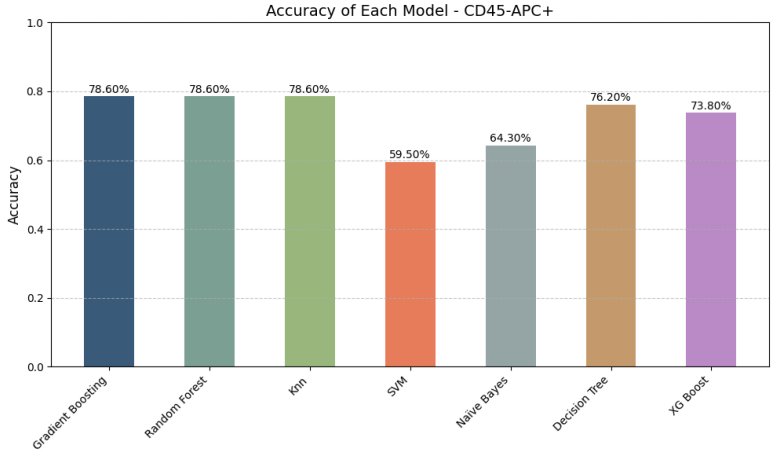


Εικόνα 4.10: Confusion matrix & ROC curve (CD45-APC+) – Decision Tree

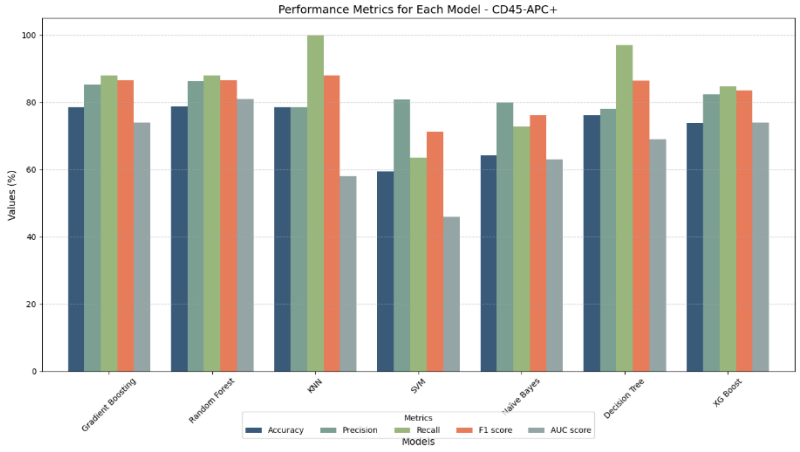


Εικόνα 4.11: Confusion matrix & ROC curve (CD45-APC+) – XG Boost

Παρατηρείται ότι τα μοντέλα Gradient Boosting, Random Forest και KNN έχουν την ίδια υψηλή ακρίβεια (79%), ωστόσο, τα πρώτα δύο μοντέλα υπερτερούν σε Precision με τιμή 86%, υποδεικνύοντας καλύτερη ικανότητα να αποφεύγουν τα ψευδώς θετικά αποτελέσματα. Αντίθετα, το KNN έχει τέλειο Recall (100%), εντοπίζοντας όλα τα θετικά παραδείγματα, γεγονός που το καθιστά εξαιρετικό για εφαρμογές όπου η πλήρης ανίχνευση των θετικών περιπτώσεων είναι κρίσιμη. Ωστόσο, το KNN έχει χαμηλότερη Precision (78.6%), υποδεικνύοντας ότι κάνει περισσότερες ψευδώς θετικές προβλέψεις, κάτι που μπορεί να οφείλεται σε overfitting. Όσον αφορά το AUC score, το Random Forest παρουσιάζει το υψηλότερο, ακολουθούμενο από το Gradient Boosting, δείχνοντας καλύτερη ικανότητα διάκρισης των θετικών από τα αρνητικά παραδείγματα σε σχέση με το KNN, το οποίο έχει το χαμηλότερο AUC score (0.58). Άλλα μοντέλα, όπως το SVM και το Naïve Bayes, εμφανίζουν χαμηλότερη ακρίβεια και διακριτική ικανότητα, ενώ το Naïve Bayes παρουσιάζει ικανοποιητικό Precision και Recall.



Εικόνα 4.12: Ακρίβεια για κάθε μοντέλο - CD45-APC+

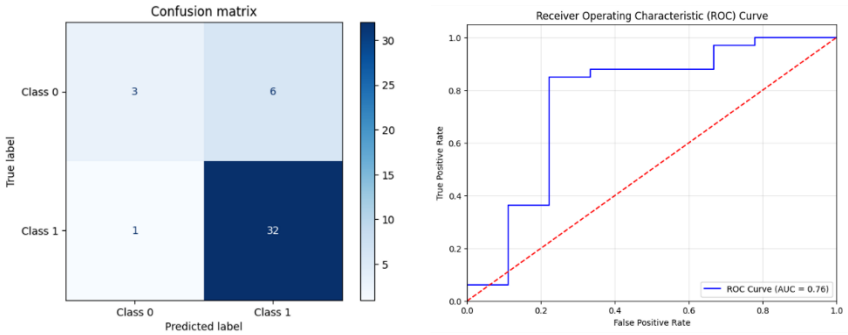


Εικόνα 4.13: Απόδοση όλων των μετρικών για τα μοντέλα - CD45-APC+

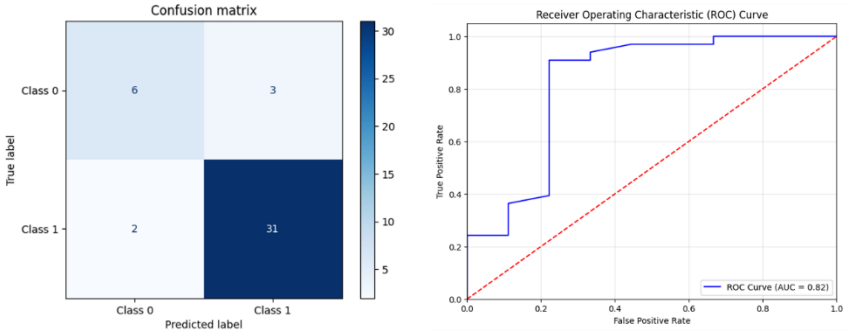
CD235a-PE+

Πίνακας 4.3: Απόδοση των μοντέλων για τον βιοδείκτη CD235a-PE+

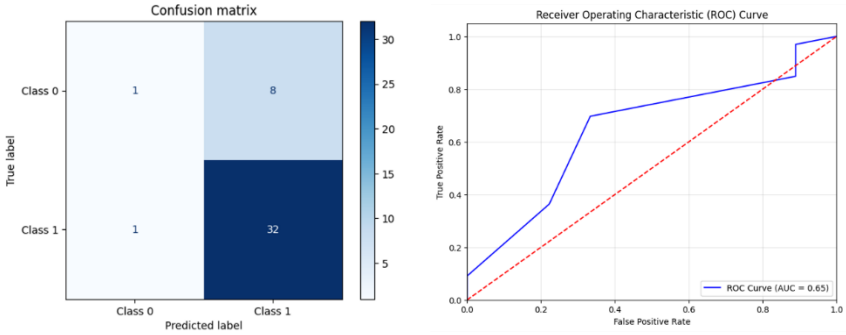
	Accuracy	Precision	Recall	F1-score	AUC score
Gradient Boosting	83.3%	84.2%	97%	90.1%	76%
Random Forest	88.1%	91.2%	93.9%	92.5%	82%
KNN	78.6%	80%	97%	87.7%	65%
SVM	61.9%	81.5%	66.7%	73.3%	54%
Naïve Bayes	73.8%	89.3%	75.8%	82%	63%
Decision Tree	66.7%	78.8%	78.8%	78.8%	65%
XG Boost	88.1%	88.9%	97%	92.8%	80%



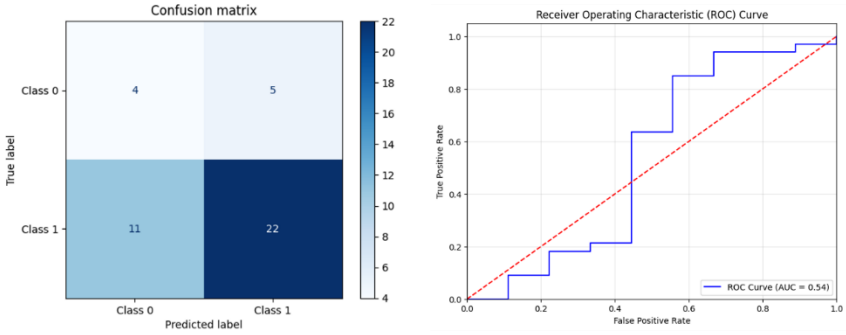
Εικόνα 4.14: Confusion matrix & ROC curve (CD235a-PE+) - Gradient Boosting



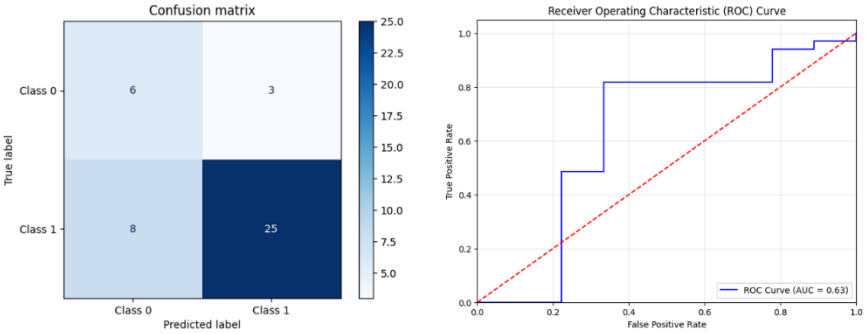
Εικόνα 4.15: Confusion matrix & ROC curve (CD235a-PE+) - Random Forest



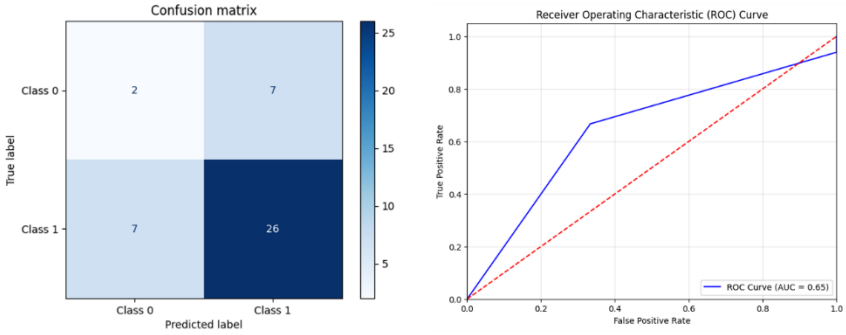
Εικόνα 4.16: Confusion matrix & ROC curve (CD235a-PE+) – KNN



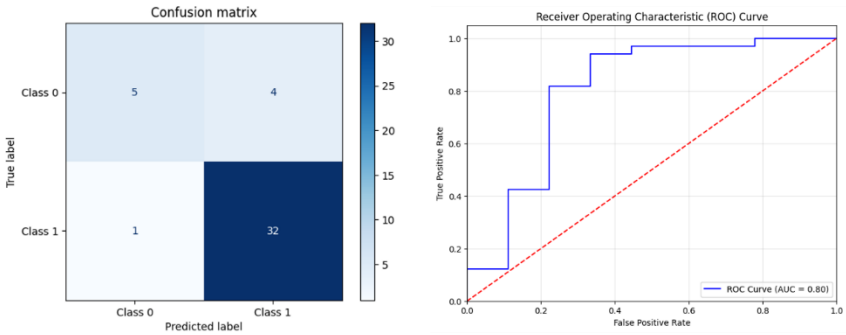
Εικόνα 4.17: Confusion matrix & ROC curve (CD235a-PE+) – SVM



Εικόνα 4.18: : Confusion matrix & ROC curve (CD235a-PE+) – Naive Bayes

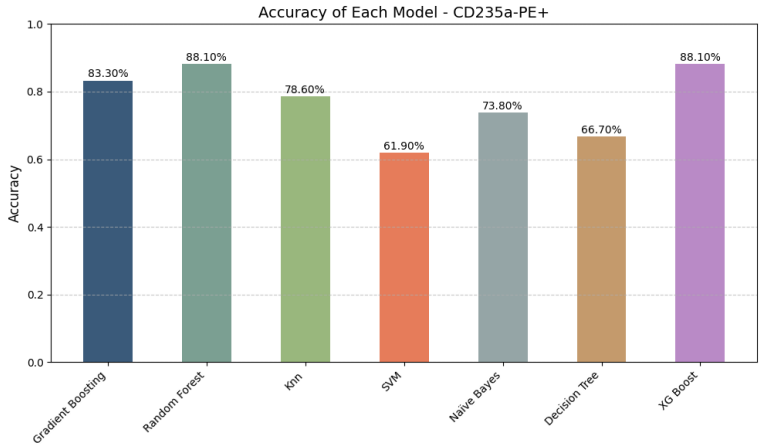


Εικόνα 4.19: : Confusion matrix & ROC curve (CD235a-PE+) – Decision Tree

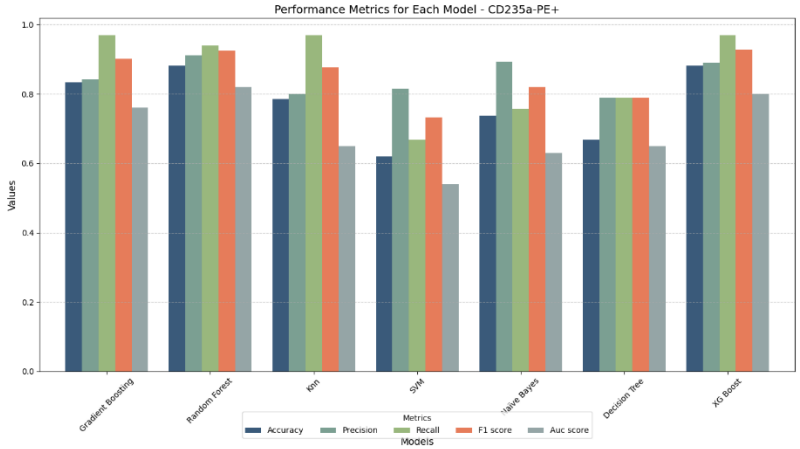


Εικόνα 4.20: : Confusion matrix & ROC curve (CD235a-PE+) – XG Boost

Αναλύοντας τα αποτελέσματα, παρατηρούμε ότι τα μοντέλα Random Forest και XG Boost έχουν την καλύτερη συνολική απόδοση με ακρίβεια 88.1%. Το Random Forest διακρίνεται περαιτέρω με Precision 91.2% και το υψηλότερο AUC score (0.82), δείχνοντας εξαιρετική ικανότητα διάκρισης μεταξύ θετικών και αρνητικών παραδειγμάτων. Το XG Boost ακολουθεί, παρουσιάζοντας υψηλό F1-score (92.8%) και AUC (0.80). Τα μοντέλα Gradient Boosting, KNN και XG Boost έχουν το υψηλότερο Recall (97%), εντοπίζοντας σχεδόν όλα τα θετικά παραδείγματα, γεγονός που τα καθιστά κατάλληλα για εφαρμογές όπου η πλήρης ανίχνευση των θετικών περιπτώσεων είναι κρίσιμη. Ωστόσο, το KNN παρουσιάζει χαμηλότερη Precision (80%) και AUC (0.65), υποδεικνύοντας περισσότερες ψευδώς θετικές προβλέψεις. Από την άλλη, το Naïve Bayes επιδεικνύει ισορροπημένη απόδοση με Precision 89.3%, αλλά χαμηλότερη ακρίβεια (73.8%) και AUC (0.63), ενώ το Decision Tree και το SVM υστερούν συνολικά, με χαμηλότερη ακρίβεια και AUC score. Συνολικά, τα Random Forest και XG Boost ξεχωρίζουν ως τα πιο αξιόπιστα μοντέλα για τη συγκεκριμένη περίπτωση.



Εικόνα 4.21: Ακρίβεια για κάθε μοντέλο - CD235a-PE+

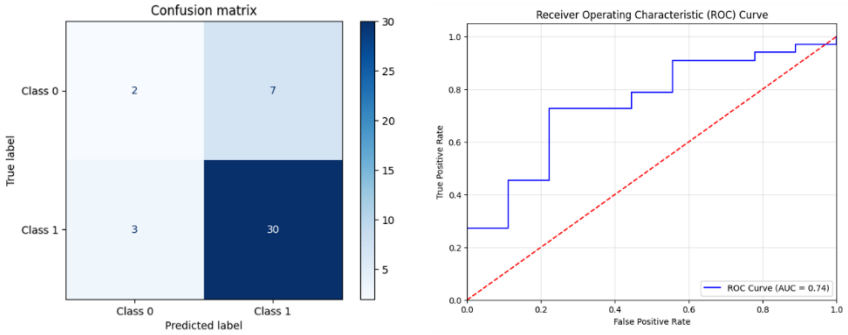


Εικόνα 4.22: Απόδοση όλων των μετρικών για τα μοντέλα – CD235a-PE+

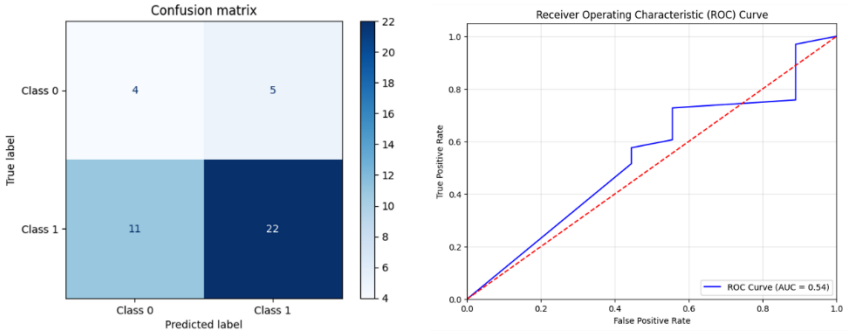
CD14-PB+

Πίνακας 4.4: Απόδοση των μοντέλων για τον βιοδείκτη CD14-PB+

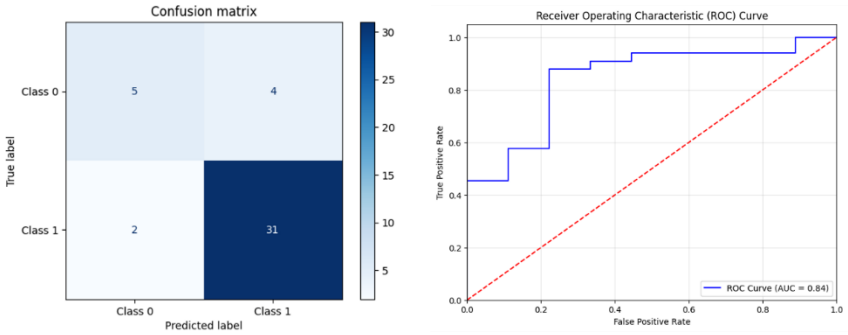
	Accuracy	Precision	Recall	F1-score	AUC score
Gradient Boosting	76.2%	81.1%	90.9%	85.7%	77%
Random Forest	85.7%	86.5%	97%	91.4%	74%
KNN	76.2%	78%	97%	86.5%	73%
SVM	81%	85.7%	90.9%	88.2%	73%
Naïve Bayes	85.7%	88.6%	93.9%	91.2%	84%
Decision Tree	61.9%	81.5%	66.7%	73.3%	54%
XG Boost	76.2%	81.1%	91%	85.7%	74%



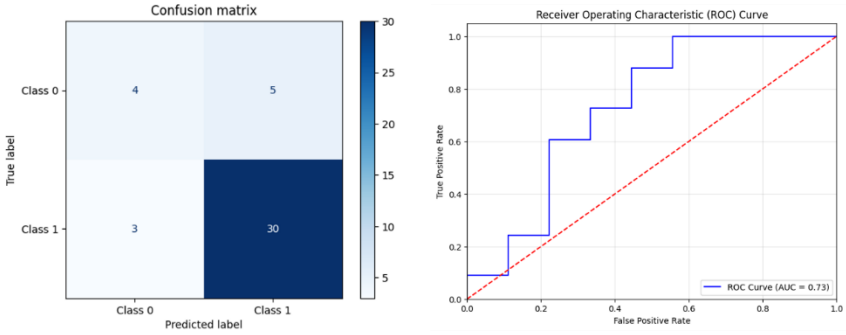
Εικόνα 4.23:Confusion matrix & ROC curve (CD14-PB+) - Gradient Boosting



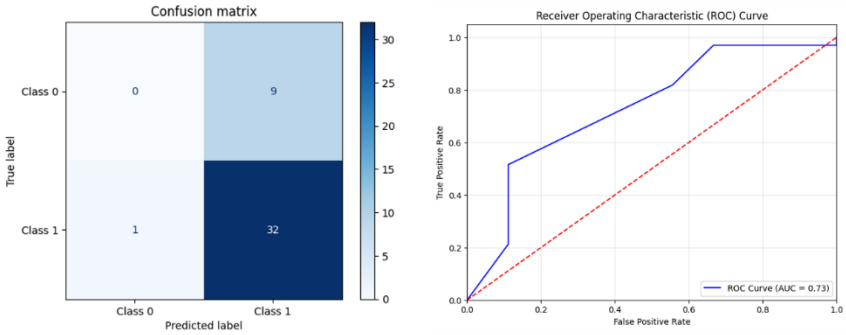
Εικόνα 4.24: Confusion matrix & ROC curve (CD14-PB+) - Random Forest



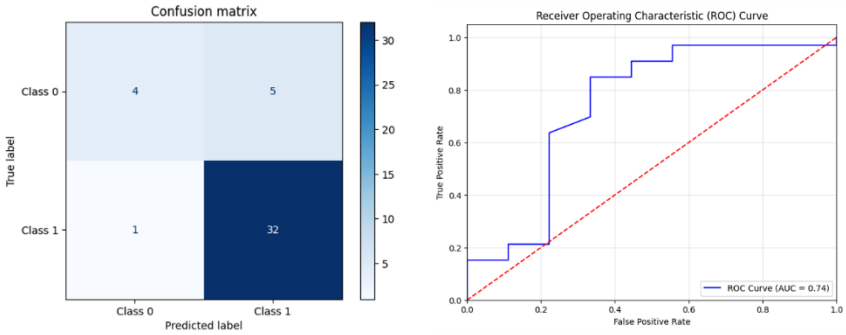
Εικόνα 4.25: Confusion matrix & ROC curve (CD14-PB+) – KNN



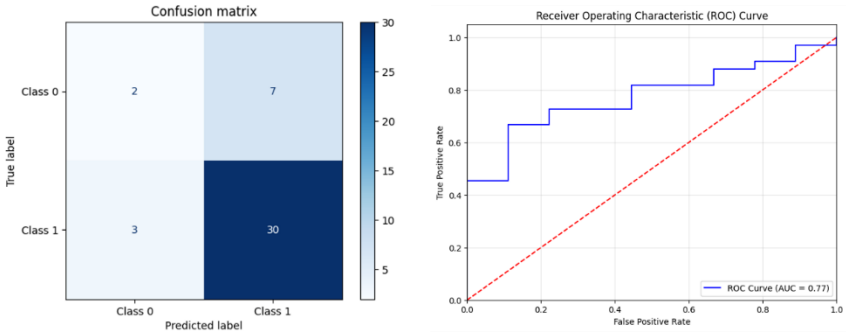
Εικόνα 4.26: Confusion matrix & ROC curve (CD14-PB+) - SVM



Εικόνα 4.27: Confusion matrix & ROC curve (CD14-PB+) - Naive Bayes



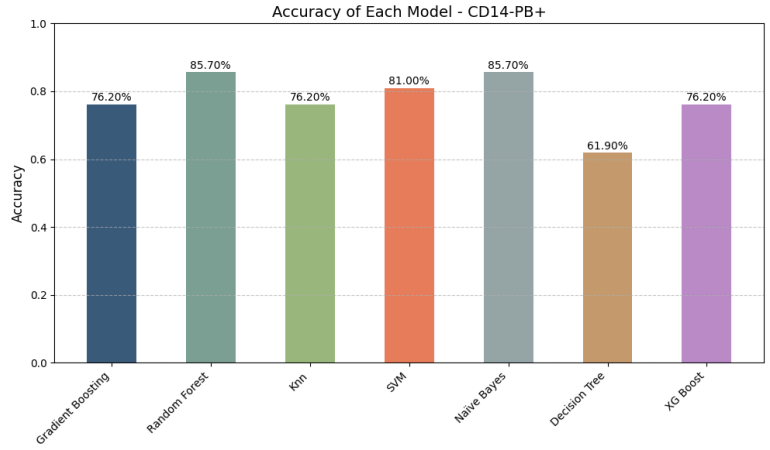
Εικόνα 4.28: Confusion matrix & ROC curve (CD14-PB+) - Decision Tree



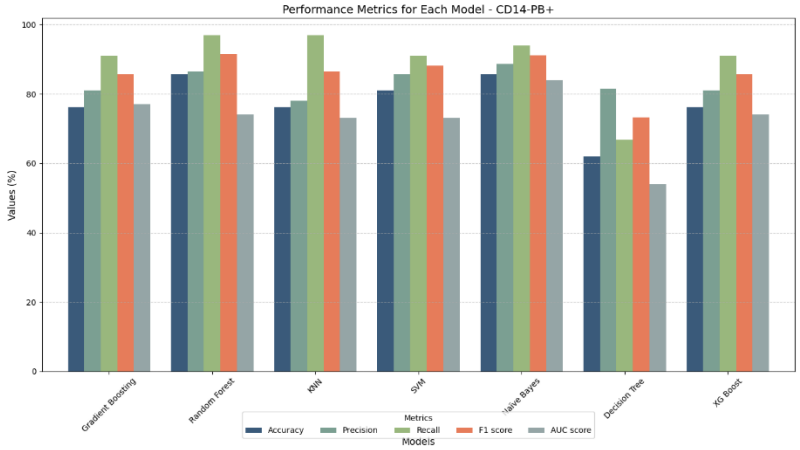
Εικόνα 4.29: Confusion matrix & ROC curve (CD14-PB+) - XG Boost

Αναλύοντας τα αποτελέσματα για τον βιοδείκτη CD14-PB+, παρατηρούμε ότι τα μοντέλα Random Forest και Naïve Bayes έχουν την υψηλότερη ακρίβεια (85.7%) και εξαιρετική συνολική απόδοση. Το Naïve Bayes ξεχωρίζει με υψηλή Precision (88.6%) και το μεγαλύτερο AUC score (0.84), υποδεικνύοντας εξαιρετική ικανότητα διάκρισης μεταξύ

θετικών και αρνητικών παραδειγμάτων. Το Random Forest παρουσιάζει εξαιρετικό Recall (97%), καθιστώντας το ιδανικό για εφαρμογές που απαιτούν πλήρη ανίχνευση των θετικών περιπτώσεων, και υψηλό F1-score (91.4%), που δείχνει καλή ισορροπία μεταξύ Precision και Recall. Τα μοντέλα Gradient Boosting και XG Boost έχουν παρόμοια απόδοση, με ακρίβεια 76.2%, υψηλό Recall (90.9% και 91%, αντίστοιχα) και F1-score (85.7%), αλλά υπολείπονται ελαφρώς στο AUC score (0.77 και 0.74). Το KNN εμφανίζει επίσης υψηλό Recall (97%) αλλά χαμηλότερη Precision (78%) και AUC (0.73), υποδεικνύοντας περισσότερες ψευδώς θετικές προβλέψεις. Το SVM επιδεικνύει καλή ισορροπία με Precision 85.7%, Recall 90.9% και F1-score 88.2%, αν και το AUC score (0.73) είναι μέτριο. Τέλος, το Decision Tree εμφανίζει την χαμηλότερη συνολική απόδοση, με ακρίβεια 61.9% και χαμηλό AUC score (0.54).



Εικόνα 4.30: Ακρίβεια για κάθε μοντέλο - CD14-PB+



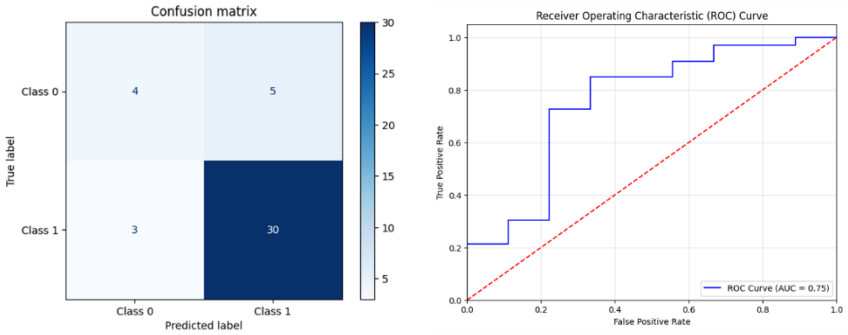
Εικόνα 4.31: Απόδοση όλων των μετρικών για τα μοντέλα - CD14-PB+

CD31-APC

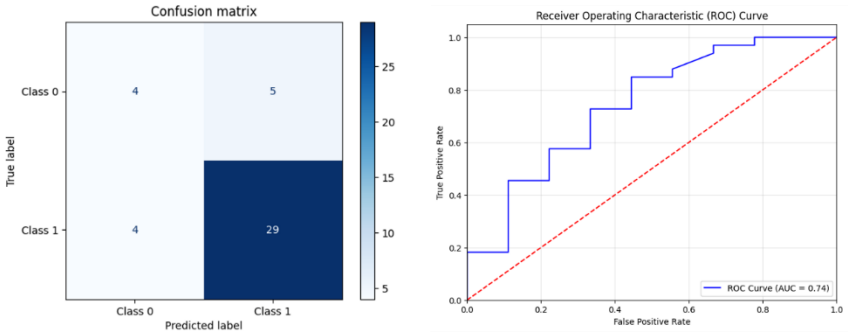
Πίνακας 4.5: Απόδοση των μοντέλων για τον βιοδείκτη CD31-APC

	Accuracy	Precision	Recall	F1-score	AUC score
Gradient Boosting	81%	85.7%	90.9%	88.2%	75%
Random Forest	78.6%	85.3%	87.9%	86.6%	74%
KNN	71.4%	78.4%	87.9%	82.9%	64%
SVM	66. 7%	82.8%	72.7%	77.4%	54%
Naïve Bayes	73.8%	82.4%	84.8%	83.6%	73%
Decision Tree	76.2%	81.1%	90.9%	85.7%	76%

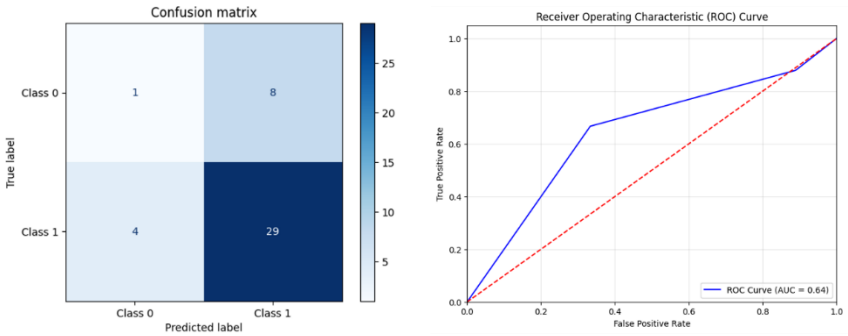
XG Boost	81%	83.8%	93.9%	88.6%	72%



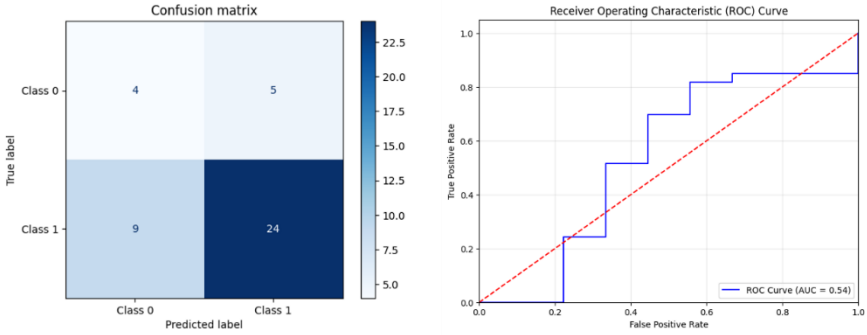
Εικόνα 4.32: Confusion matrix & ROC curve (CD31-APC) - Gradient Boosting



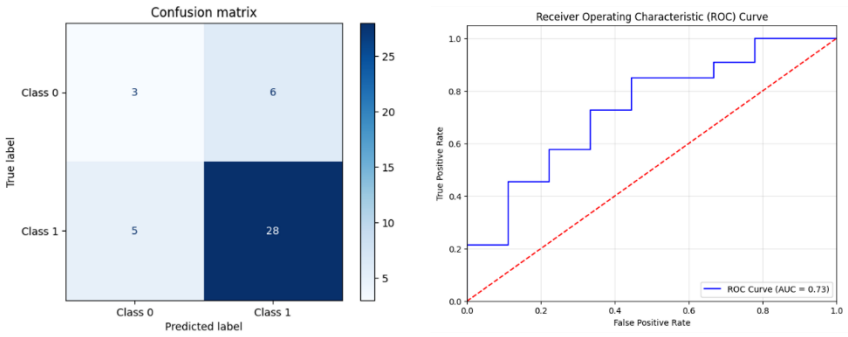
Εικόνα 4.33: Confusion matrix & ROC curve (CD31-APC) - Random Forest



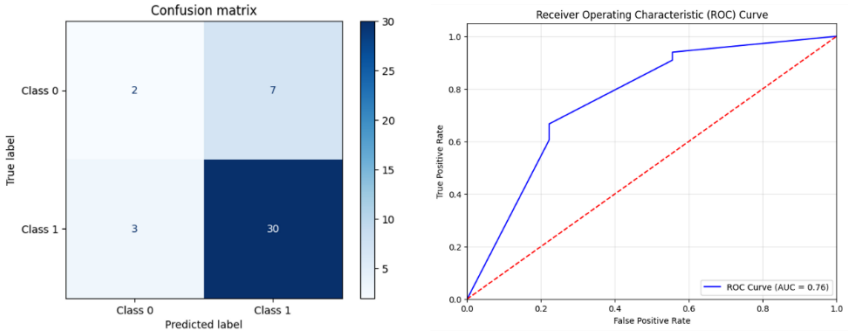
Εικόνα 4.34: Confusion matrix & ROC curve (CD31-APC) – KNN



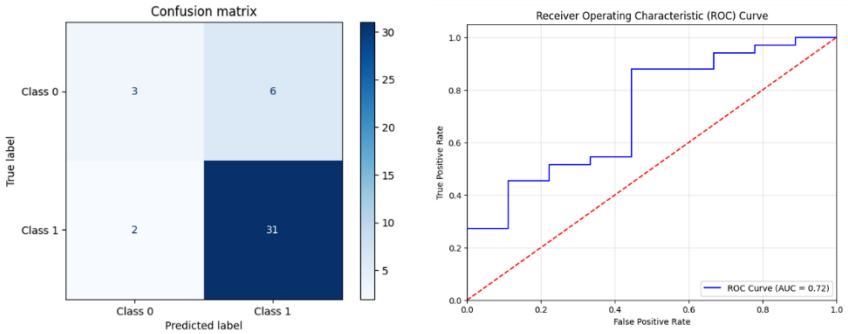
Εικόνα 4.35: Confusion matrix & ROC curve (CD31-APC) – SVM



Εικόνα 4.36: Confusion matrix & ROC curve (CD31-APC) - Naive Bayes



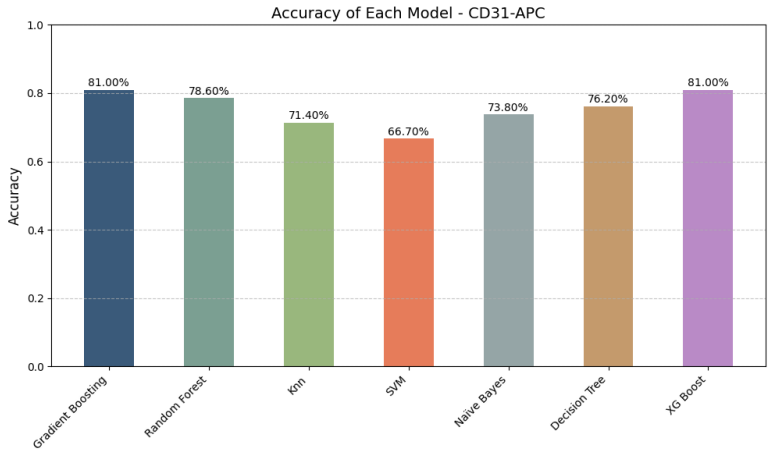
Εικόνα 4.37: Confusion matrix & ROC curve (CD31-APC) - Decision Tree



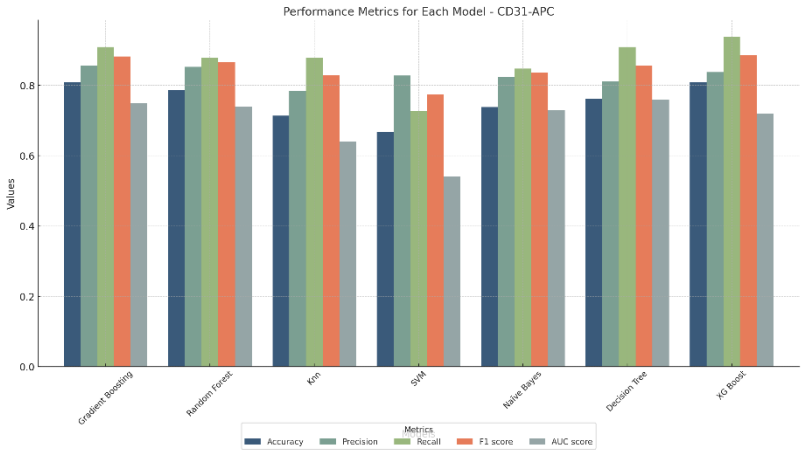
Εικόνα 4.38: Confusion matrix & ROC curve (CD31-APC) - XG Boost

Για το CD31-APC, διακρίνονται ως τα δύο καλύτερα μοντέλα το Gradient Boosting και το XG Boost, καθώς παρουσιάζουν την υψηλότερη ακρίβεια (81%), υποδεικνύοντας ισχυρή συνολική απόδοση. Το Gradient Boosting υπερέχει σε Precision (85.7%) και σημειώνει ισχυρό F1-score (88.2%), καταδεικνύοντας εξαιρετική ισορροπία μεταξύ ακρίβειας (precision) και ευαισθησίας (recall). Επιπλέον, έχει το υψηλότερο AUC score (0.75), γεγονός που επιβεβαιώνει την αξιοπιστία του σε ένα ευρύ φάσμα κατωφλίων. Το XG Boost, από την άλλη, σημειώνει την καλύτερη επίδοση σε Recall (93.9%), καθιστώντας το ιδιαίτερα αποτελεσματικό σε σενάρια όπου η ανίχνευση θετικών περιπτώσεων είναι κρίσιμη. Παράλληλα, το F1-score (88.6%) παραμένει ελαφρώς υψηλότερο από αυτό του Gradient Boosting, καθιστώντας το μια εναλλακτική που δίνει έμφαση στην ευαισθησία. Συγκριτικά, το Random Forest (78.6% ακρίβεια) και το Decision Tree (76.2% ακρίβεια) εμφανίζουν καλή απόδοση, αλλά υπολείπονται έναντι των δύο κορυφαίων μοντέλων στις περισσότερες μετρικές. Το Decision Tree έχει σχετικά υψηλό Recall (90.9%), αλλά υστερεί σε Precision (81.1%), υποδεικνύοντας πιθανώς μεγαλύτερο αριθμό ψευδώς θετικών προβλέψεων. Το Naïve Bayes (73.8% ακρίβεια) και το KNN (71.4% ακρίβεια) εμφανίζουν μέτρια απόδοση, με το KNN να έχει καλό Recall

(87.9%) αλλά χαμηλότερο AUC (0.64), κάτι που περιορίζει την αξιοπιστία του σε διαφορετικά κατώφλια. Παρομοίως, το Naïve Bayes έχει ισορροπημένες τιμές σε Precision (82.4%) και Recall (84.8%), αλλά δεν φτάνει τα επίπεδα των κορυφαίων μοντέλων. Τέλος, το SVM εμφανίζει τη χαμηλότερη ακρίβεια (66.7%) και AUC (0.54), υποδεικνύοντας αδυναμία στο διαχωρισμό των κατηγοριών. Παρόλο που το Precision (82.8%) είναι συγκριτικά υψηλό, το χαμηλό Recall (72.7%) μειώνει τη συνολική αποτελεσματικότητα του μοντέλου.



Εικόνα 4.39: Ακρίβεια για κάθε μοντέλο – CD31-APC

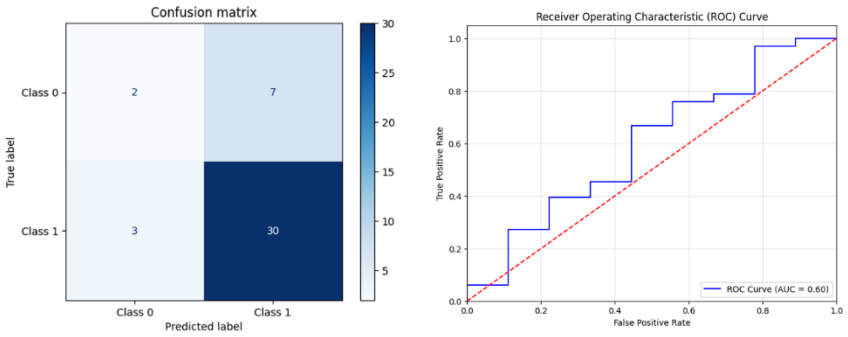


Εικόνα 4.40: Απόδοση όλων των μετρικών για τα μοντέλα – CD31-APC

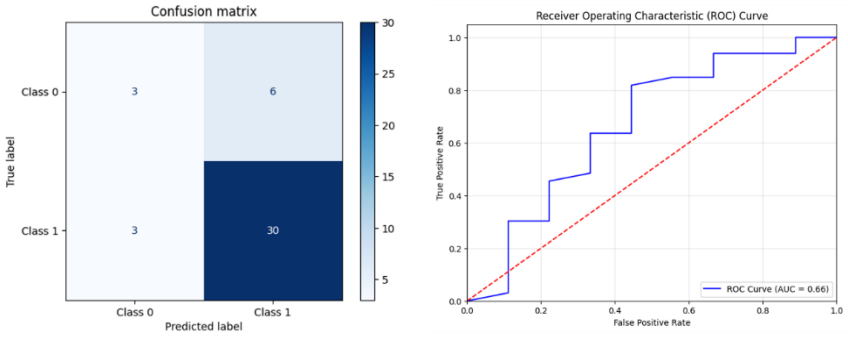
CD146-PE

Πίνακας 4.6: Απόδοση των μοντέλων για τον βιοδείκτη CD146-PE

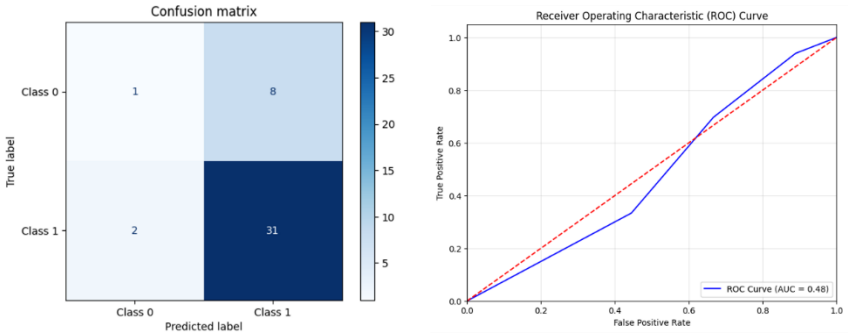
	Accuracy	Precision	Recall	F1-score	AUC score
Gradient Boosting	76.2%	81.1%	90.9%	85.7%	60%
Random Forest	78.6%	83.3%	90.9%	87%	66%
KNN	76.2%	79.5%	93.9%	86.1%	48%
SVM	61.9%	79.3%	69.7%	74.2%	51%
Naïve Bayes	69%	81.2%	78.8%	80%	51%
Decision Tree	71.4%	81.8%	81.8%	81.8%	49%
XG Boost	76.2%	82.9%	87.9%	85.3%	63%



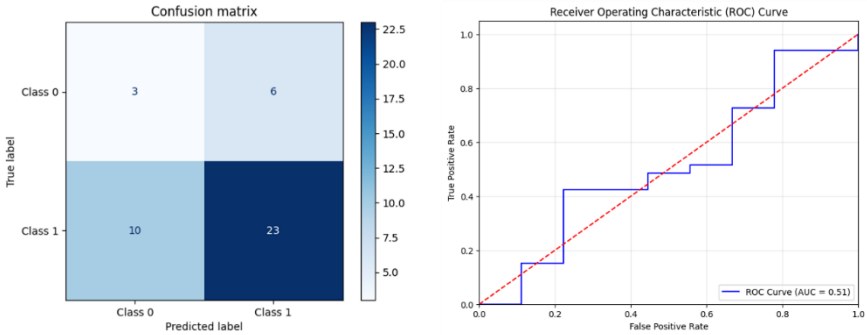
Εικόνα 4.41: Confusion matrix & ROC curve (CD146-PE) – Gradient Boosting



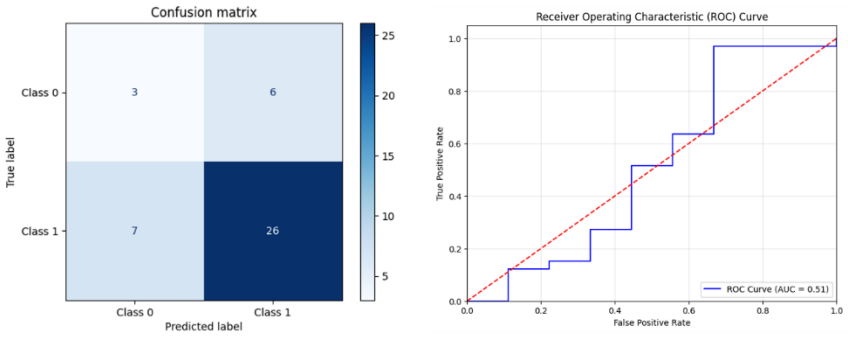
Εικόνα 4.42: Confusion matrix & ROC curve (CD146-PE) – Random Forest



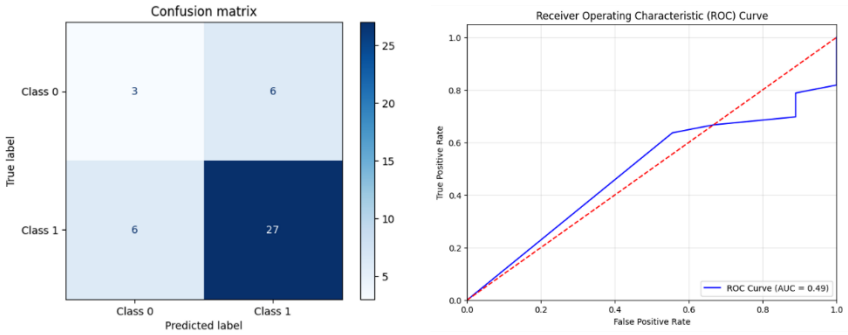
Εικόνα 4.43: Confusion matrix & ROC curve (CD146-PE) – KNN



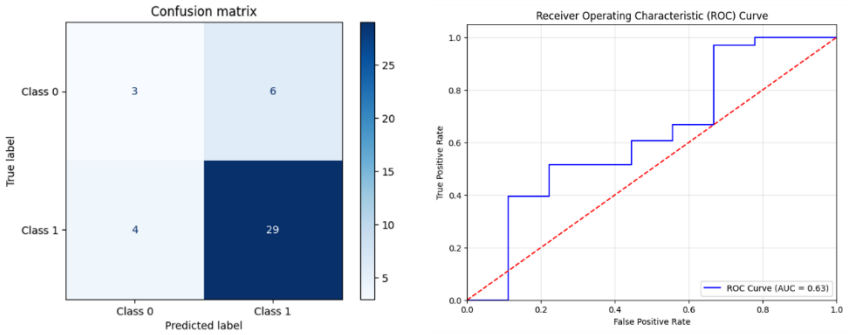
Εικόνα 4.44: Confusion matrix & ROC curve (CD146-PE) – SVM



Εικόνα 4.45: Confusion matrix & ROC curve (CD146-PE) – Naïve Bayes

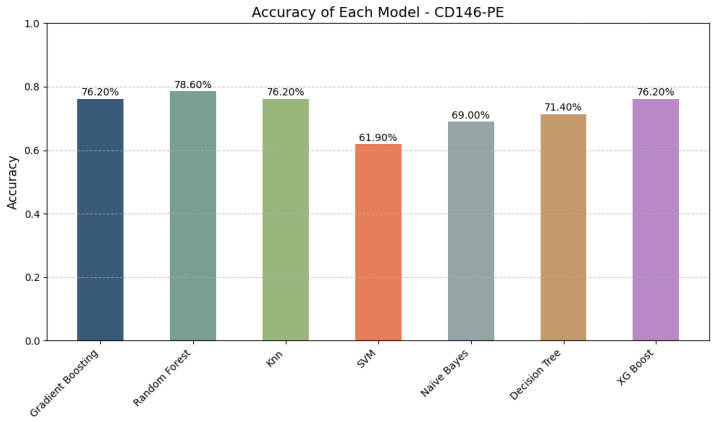


Εικόνα 4.46: Confusion matrix & ROC curve (CD146-PE) - Decision Tree

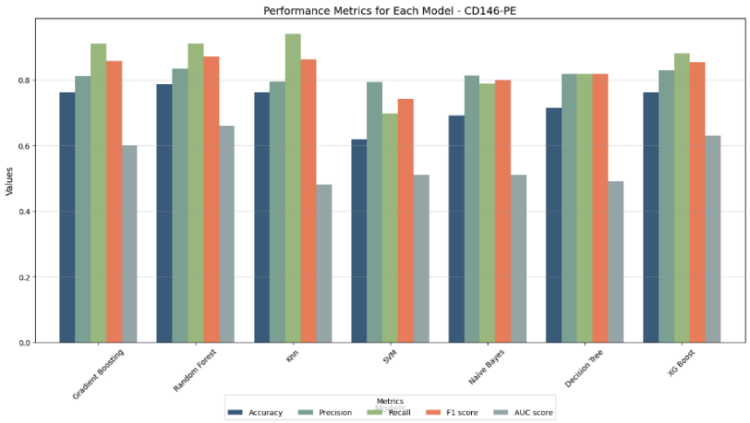


Εικόνα 4.47: Confusion matrix & ROC curve (CD146-PE) - XG Boost

Αναλύοντας τα αποτελέσματα, τα δύο καλύτερα μοντέλα είναι το Random Forest και το Gradient Boosting, με το Random Forest να ξεχωρίζει για την υψηλότερη ακρίβεια (78.6%), Precision (83.3%), Recall (90.9%) και F1-score (87%), ενώ το Gradient Boosting διατηρεί εξαιρετική ισορροπία με ακρίβεια (76.2%), υψηλό Recall (90.9%) και F1-score (85.7%). Το XG Boost παρουσιάζει παρόμοια συνολική απόδοση, με Precision (82.9%), Recall (87.9%) και F1-score (85.3%), καθιστώντας το μια αξιόπιστη εναλλακτική επιλογή. Το KNN υπερέχει στο Recall (93.9%), καθιστώντας το κατάλληλο για εφαρμογές όπου η ανίχνευση θετικών περιπτώσεων είναι κρίσιμη, αλλά υστερεί σε AUC score (0.48). Τα υπόλοιπα μοντέλα, όπως το Naïve Bayes, το Decision Tree και το SVM, εμφανίζουν χαμηλότερη συνολική απόδοση, με μέτρια ή χαμηλά AUC scores, καθιστώντας τα λιγότερο αποδοτικά. Συνολικά, το Random Forest και το Gradient Boosting αποτελούν τις πιο αξιόπιστες επιλογές, με το πρώτο να υπερέχει ελαφρώς σε συνολική απόδοση και το δεύτερο να παρέχει μια σταθερή και αξιόπιστη εναλλακτική.



Εικόνα 4.48: Ακρίβεια για κάθε μοντέλο – CD146-PE

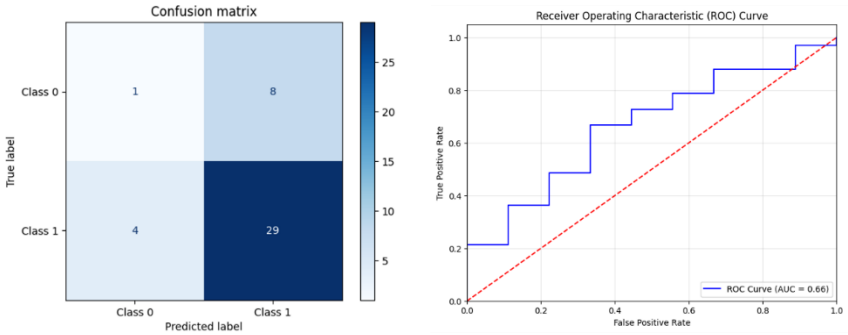


Εικόνα 4.49: Απόδοση όλων των μετρικών – CD146-PE

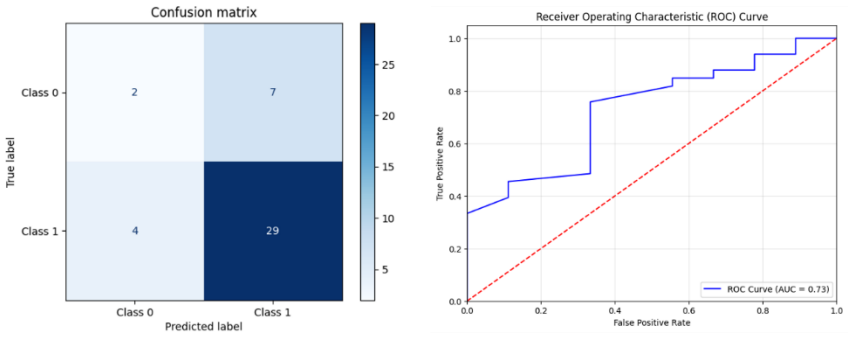
CD326-APC+

Πίνακας 4.7: Απόδοση των μοντέλων για τον βιοδείκτη CD326-APC+

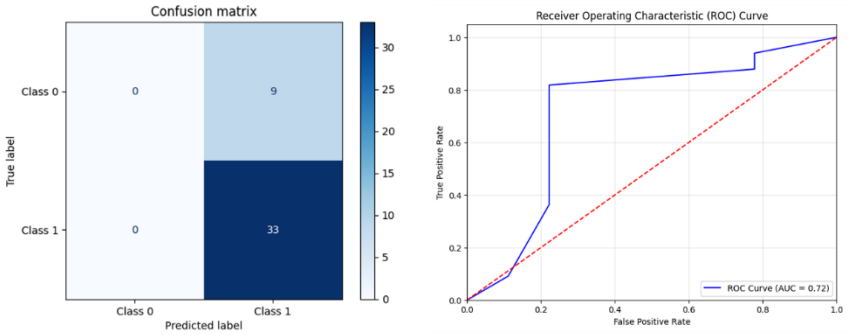
	Accuracy	Precision	Recall	F1-score	AUC score
Gradient Boosting	71.4%	78.4%	87.9%	82.9%	66%
Random Forest	73.8%	80.6%	87.9%	84.1%	73%
KNN	78.6%	78.6%	100%	88%	72%
SVM	66.7%	91.3%	63.6%	75%	72%
Naïve Bayes	54.8%	81.8%	54.5%	65.5%	51%
Decision Tree	66.7%	78.8%	78.8%	78.8%	56%
XG Boost	73.8%	82.4%	84.8%	83.6%	64%



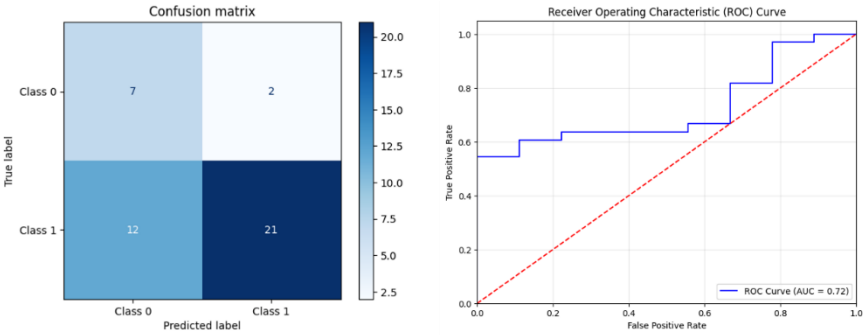
Εικόνα 4.50: Confusion matrix & ROC curve (CD326-APC+) – Gradient Boosting



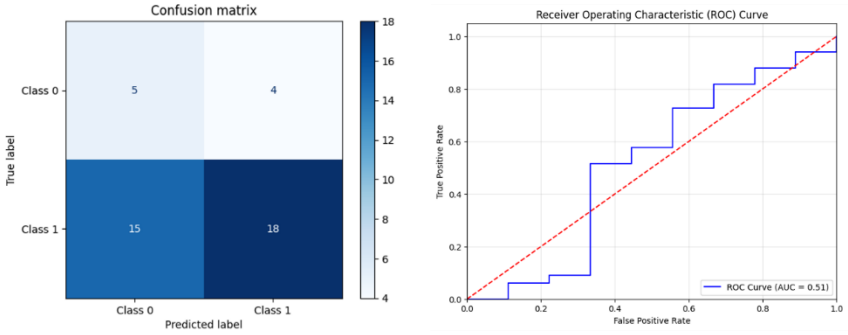
Εικόνα 4.51: Confusion matrix & ROC curve (CD326-APC+) – Random Forest



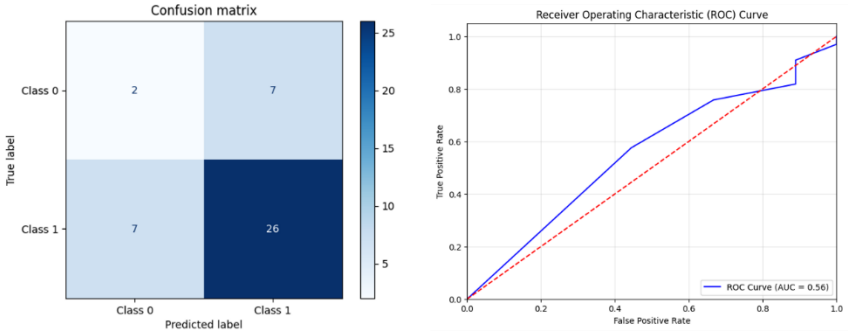
Εικόνα 4.52: Confusion matrix & ROC curve (CD326-APC+) – KNN



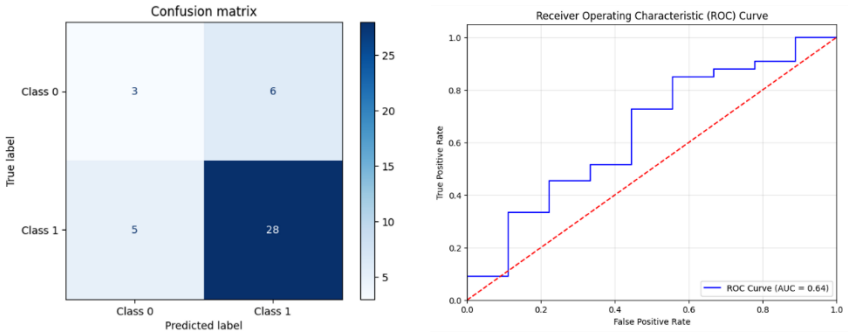
Εικόνα 4.53: Confusion matrix & ROC curve (CD326-APC+) – SVM



Εικόνα 4.54: Confusion matrix & ROC curve (CD326-APC+) – Naive Bayes

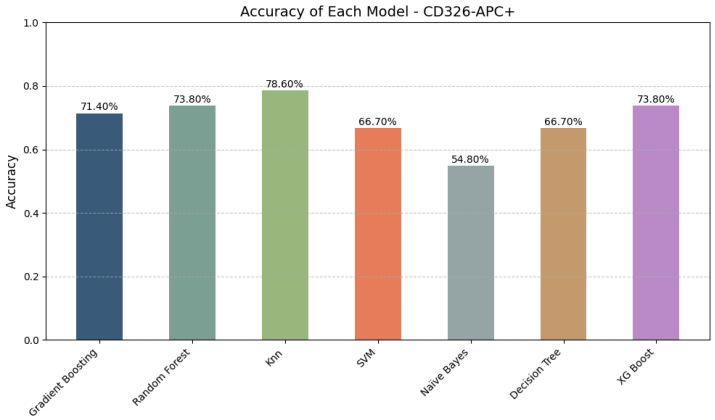


Εικόνα 4.55: Confusion matrix & ROC curve (CD326-APC+) – Decision Tree

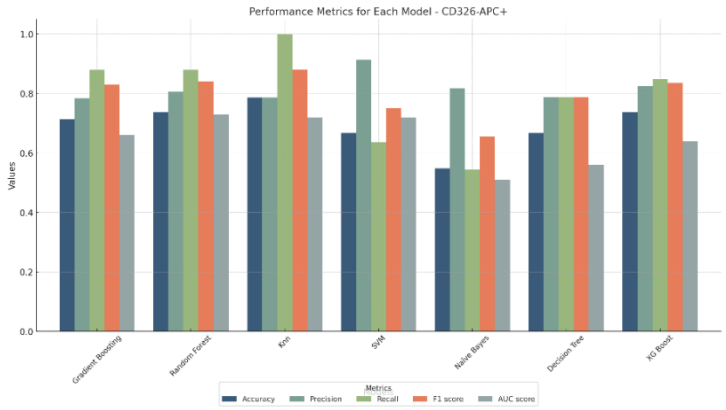


Εικόνα 4.56: Confusion matrix & ROC curve (CD326-APC+) – XG Boost

Για τα μοντέλα στο σύνολο δεδομένων CD326-APC+, παρατηρούμε ότι τα δύο πιο αποδοτικά μοντέλα είναι το KNN και το Random Forest. Συγκεκριμένα, το KNN ξεχωρίζει με την υψηλότερη επίδοση στο Recall (100%), το οποίο εξασφαλίζει την ανίχνευση όλων των θετικών κλάσεων. Ωστόσο, η εξαιρετική αυτή απόδοση στο Recall μπορεί να υποδεικνύει υπερεκπαίδευση, κάτι που ίσως επηρεάσει την ικανότητά του να γενικεύει καλά σε νέα δεδομένα. Από την άλλη, το Random Forest προσφέρει σταθερές και υψηλές επιδόσεις σε όλους τους δείκτες, με ακρίβεια 73.8%, Precision 80.6%, και Recall 87.9%, ενώ η AUC score (0.73) είναι η υψηλότερη μεταξύ των μοντέλων, κάνοντάς το ένα από τα πιο αξιόπιστα για το σύνολο δεδομένων. Το Gradient Boosting και το XG Boost επίσης παρουσιάζουν καλές επιδόσεις, με F1-scores 82.9% και 83.6%, αντίστοιχα, αλλά υστερούν ελαφρώς σε AUC scores (0.66 και 0.64). Αντίθετα, τα SVM, Decision Tree και Naïve Bayes εμφανίζουν χαμηλότερη συνολική απόδοση, με χαμηλότερα AUC scores και λιγότερο ισορροπημένα αποτελέσματα σε Precision και Recall. Συνολικά, τα KNN και Random Forest ξεχωρίζουν ως τα δύο πιο αποδοτικά μοντέλα για τη συγκεκριμένη περίπτωση.



Εικόνα 4.57: Ακρίβεια για όλα τα μοντέλα - CD326-APC+

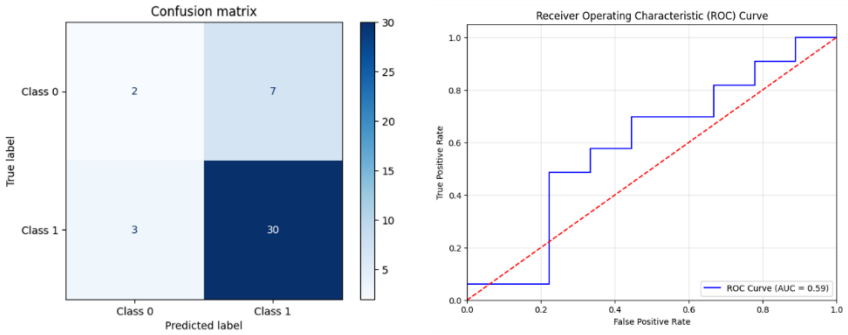


Εικόνα 4.58: Απόδοση όλων των μετρικών - CD326-APC+

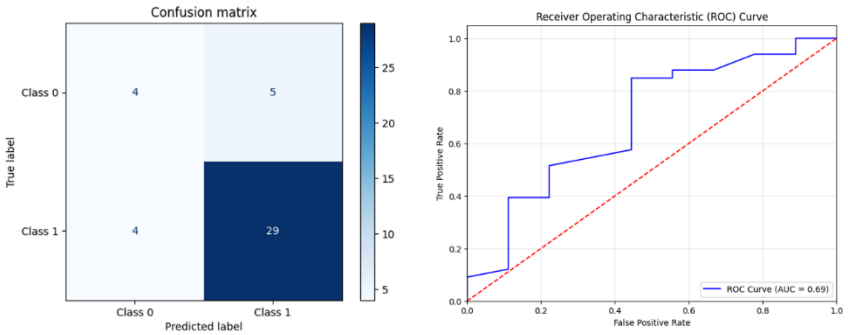
Lact-FITC+

Πίνακας 4.8: Απόδοση των μοντέλων για τον βιοδείκτη Lact-FITC+

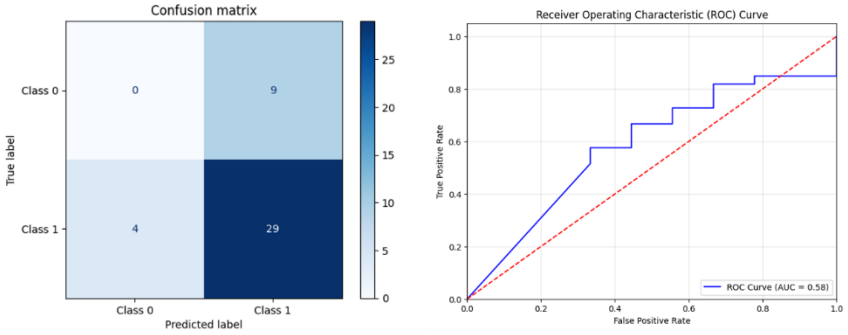
	Accuracy	Precision	Recall	F1-score	AUC score
Gradient Boosting	76.2%	81.1%	90.9%	85.7%	59%
Random Forest	78.6%	85.3%	87.9%	86.6%	69%
KNN	69%	76.3%	87.9%	81.7%	58%
SVM	73.8%	84.4%	81.8%	83.1%	71%
Naïve Bayes	71.4%	80%	84.8%	82.4%	54%
Decision Tree	66.7%	78.8%	78.8%	78.8%	60%
XG Boost	76.8%	81.6%	93.9%	87.3%	63%



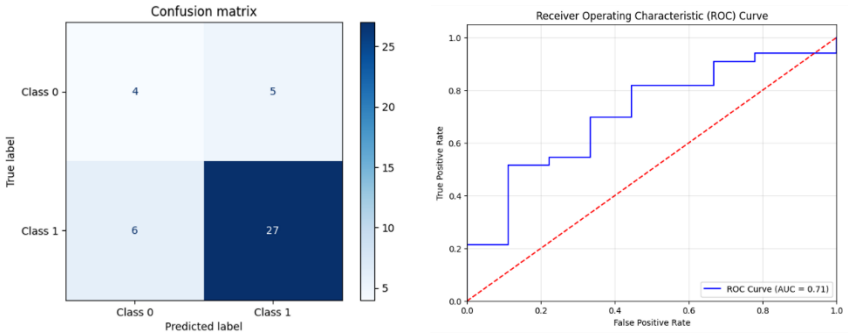
Εικόνα 4.59: Confusion matrix & ROC curve (Lact-FITC+) – Gradient Boosting



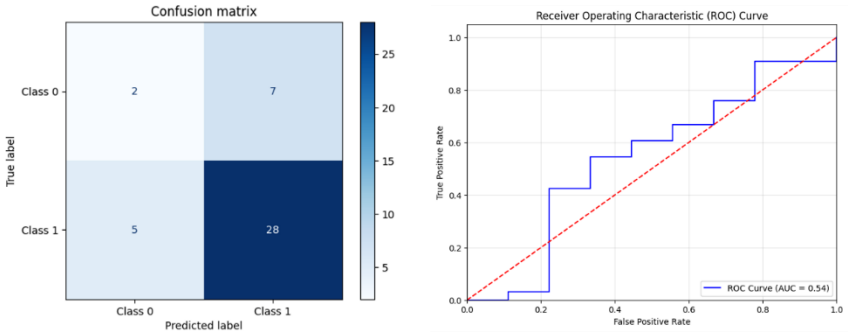
Εικόνα 4.60: Confusion matrix & ROC curve (Lact-FITC+) – Random Forest



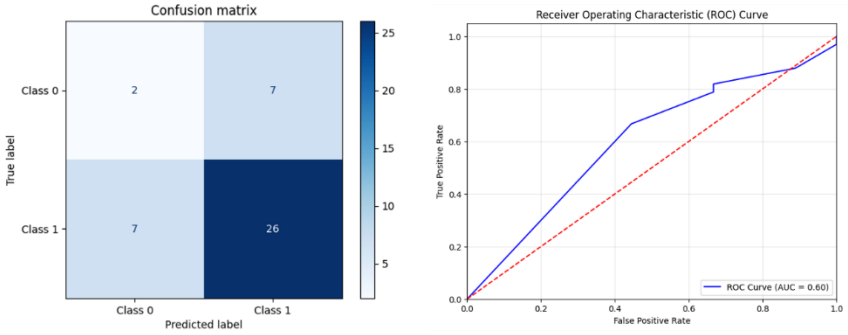
Εικόνα 4.61: Confusion matrix & ROC curve (Lact-FITC+) – KNN



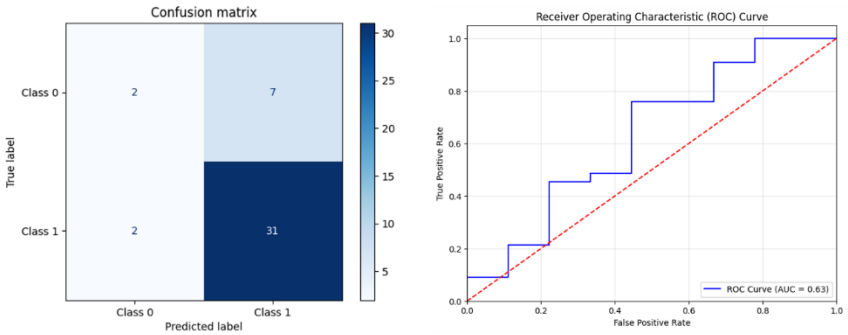
Εικόνα 4.62: Confusion matrix & ROC curve (Lact-FITC+) – SVM



Εικόνα 4.63: Confusion matrix & ROC curve (Lact-FITC+) – Naive Bayes

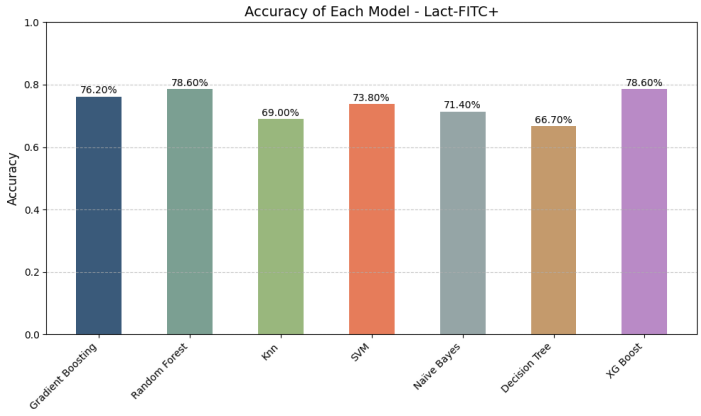


Εικόνα 4.64: Confusion matrix & ROC curve (Lact-FITC+) – Decision Tree

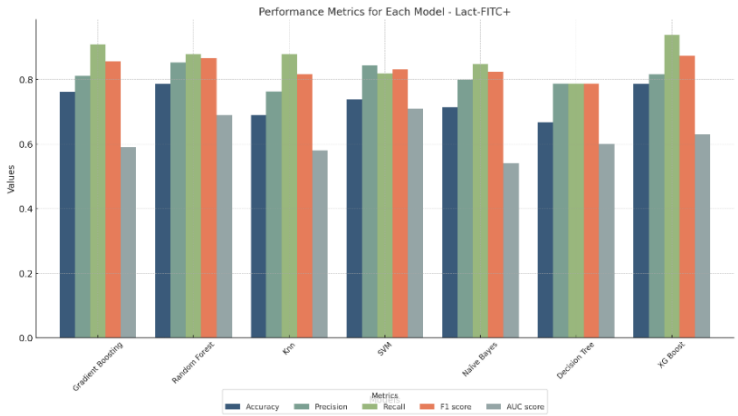


Εικόνα 4.65: Confusion matrix & ROC curve (Lact-FITC+) – XG Boost

Για το Lact-FITC+, μπορούμε να παρατηρήσουμε ότι τα δύο καλύτερα μοντέλα είναι το Random Forest και το XG Boost. Το Random Forest διακρίνεται με ακρίβεια 78.6%, υψηλή Precision (85.3%) και καλή AUC score (0.69), υποδεικνύοντας την ικανότητά του να εντοπίζει σωστά τις θετικές περιπτώσεις με συνολικά ισορροπημένη απόδοση. Το XG Boost, με παρόμοια ακρίβεια (76.8%), ξεχωρίζει για το υψηλότερο Recall (93.9%) και το καλύτερο F1 score (87.3%), καθιστώντας το ιδανικό για εφαρμογές όπου η ανίχνευση των θετικών περιπτώσεων είναι κρίσιμη. Αντίθετα, το KNN παρουσιάζει χαμηλότερη ακρίβεια (69%) και AUC score (0.58), ενώ το SVM, παρά την υψηλή Precision (84.4%), έχει μικρότερο Recall (81.8%) και F1 score (83.1%), που το κατατάσσουν πιο χαμηλά. Τα υπόλοιπα μοντέλα, όπως το Gradient Boosting και το Naïve Bayes, εμφανίζουν ικανοποιητική απόδοση, αλλά υστερούν ελαφρώς σε ισορροπία μεταξύ των μετρικών. Συνολικά, τα μοντέλα Random Forest και XG Boost υπερέχουν λόγω της γενικά ισχυρής και ισορροπημένης απόδοσής τους στους βασικούς δείκτες.



Εικόνα 4.66: Ακρίβεια για κάθε μοντέλο – Lact-FITC+



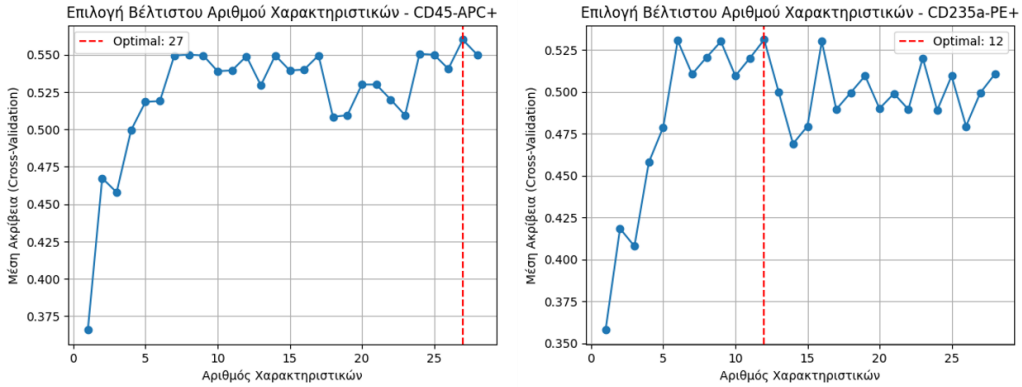
Εικόνα 4.67: Απόδοση όλων των μετρικών – Lact-FITC+

4.2 Μέρος 2ο: Διάκριση του είδους του εγκεφαλικό (4 κλάσεις)

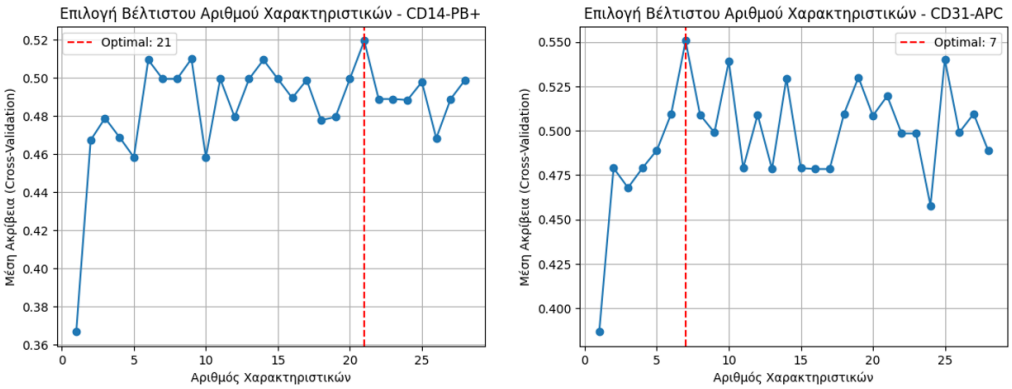
4.2.1 Αποτελέσματα από την επιλογή χαρακτηριστικών

Στο 2ο μέρος της εργασίας, εξετάζεται η απόδοση των μοντέλων ως προς τη ικανότητά τους να ταξινομούν με ακρίβεια τους ασθενείς στις αντίστοιχες κατηγορίες εγκεφαλικών επεισοδίων. Οι κατηγορίες αυτές περιλαμβάνουν: Control, Vessel Occlusion, Bleeding, και LVO οι οποίες έχουν κωδικοποιηθεί με αριθμούς από το 0 έως το 3, αντίστοιχα. Η κατανομή των κατηγοριών είναι ως εξής: η κατηγορία 0 εμφανίζεται 33 φορές, η κατηγορία 1 εμφανίζεται 67 φορές, η κατηγορία 2 εμφανίζεται 21 φορές, ενώ η κατηγορία 3 εμφανίζεται 19 φορές. Η μεθοδολογία που ακολουθήθηκε για το δεύτερο μέρος της ανάλυσης ήταν αντίστοιχη με εκείνη του πρώτου μέρους. Αρχικά, πραγματοποιήθηκε επιλογή χαρακτηριστικών με στόχο την πρόβλεψη της στήλης «General Type», η οποία υποδεικνύει την κατηγορία στην οποία ανήκει το άτομο που εξετάστηκε. Τα δεδομένα διαχωρίστηκαν σε σύνολα εκπαίδευσης και ελέγχου σε αναλογία 70%-30%.

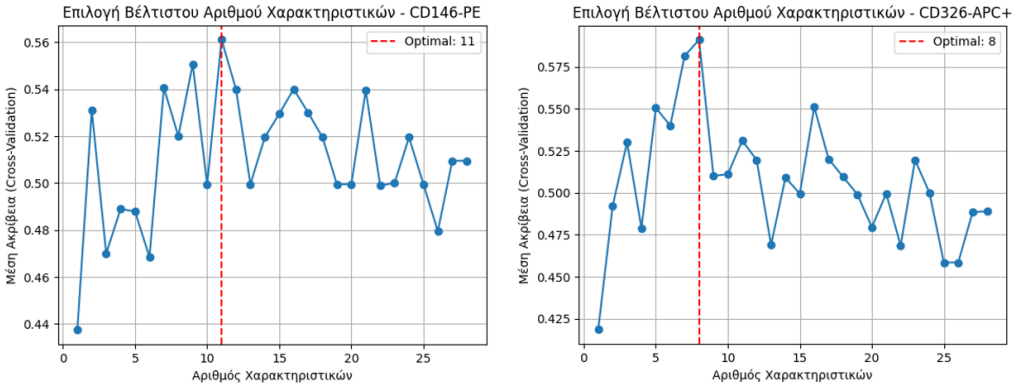
Όπως και στο πρώτο μέρος, για την επιλογή των σημαντικότερων χαρακτηριστικών χρησιμοποιήθηκε ο αλγόριθμος Recursive Feature Elimination (RFE) σε συνδυασμό με τον Random Forest Classifier. Η ανάλυση έδειξε ότι ο βέλτιστος αριθμός χαρακτηριστικών ποικίλλει ανάλογα με τον βιοδείκτη, όπως απεικονίζεται στα διαγράμματα που ακολουθούν. Ωστόσο, παρατηρήθηκε ότι η επιλογή 10 χαρακτηριστικών παρείχε υψηλή και σταθερή απόδοση, σύμφωνα με τα αποτελέσματα του αλγορίθμου RFE. Επιπλέον, στο πρώτο μέρος της ανάλυσης, επιλέχθηκε ο ίδιος αριθμός χαρακτηριστικών, γεγονός που εξασφαλίζει ομοιομορφία στα αποτελέσματα της έρευνας. Συνεπώς, αποφασίστηκε η χρήση των 10 καλύτερων χαρακτηριστικών για τη συνέχεια της ανάλυσης, διατηρώντας την ισορροπία μεταξύ της ακρίβειας του μοντέλου και της απλότητας της διαδικασίας, ενώ παράλληλα διασφαλίζεται η συνέπεια και η συγκρισιμότητα των αποτελεσμάτων.



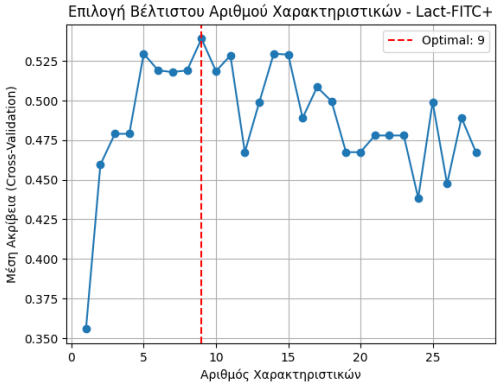
Εικόνα 4.68: Επιλογή βέλτιστου αριθμού χαρακτηριστικών μέρος 2 (CD45-APC+ - CD235a-PE+)



Εικόνα 4.69: Επιλογή βέλτιστου αριθμού χαρακτηριστικών μέρος 2 (CD14-PB+ - CD31-APC)



Εικόνα 4.70: Επιλογή βέλτιστου αριθμού χαρακτηριστικών μέρος 2 (CD146-PE - CD326-APC+)



Εικόνα 4.71: Επιλογή βέλτιστου αριθμού χαρακτηριστικών μέρος 2 (Lact-FITC+)

Στη συνέχεια, τα δύο πιο αποδοτικά μοντέλα, όπως αυτά αναδείχθηκαν στο πρώτο μέρος της ανάλυσης, επιλέχθηκαν για την αξιολόγηση της απόδοσής τους στο συγκεκριμένο πρόβλημα. Η επιλογή αυτών των μοντέλων βασίστηκε στη λογική ότι η αποδοτικότητά τους στις αρχικές κλάσεις (στο πρώτο μέρος της ταξινόμησης) τα καθιστά αξιόπιστες και αποτελεσματικές επιλογές. Συγκεκριμένα, κρίθηκαν τα καταλληλότερα για να χρησιμοποιηθούν ως βασικά μοντέλα σε μια πιο σύνθετη ταξινόμηση με περισσότερες κλάσεις, εξασφαλίζοντας έτσι τη συνέπεια και την αποδοτικότητα στη διαδικασία ανάλυσης.

Πίνακας 4.9: Τα 2 πιο αποδοτικά μοντέλα για κάθε βιοδείκτη

Βιοδείκτες	**2 καλύτερα μοντέλα**
CD45-APC+	Gradient Boosting, Random Forest
CD235a-PE+	Random Forest, XG Boost
CD14-PB+	Random Forest, Naïve Bayes
CD31-APC	Gradient Boosting, XG Boost
CD146-PE	Gradient Boosting, Random Forest
CD326-APC+	Random Forest, KNN
Lact-FITC+	Random Forest, XG Boost

Στο δεύτερο μέρος της ανάλυσης, υπήρξαν δυο διαφοροποιήσεις στον τρόπο απεικόνισης και αξιολόγησης των μοντέλων και των αποτελεσμάτων. Η πρώτη διαφοροποίηση αφορούσε την αξιολόγηση των μετρικών, καθώς χρησιμοποιήθηκαν οι macro μέσες τιμές τους. Αυτή η επιλογή κρίθηκε καταλληλότερη για προβλήματα με πολλές κατηγορίες, καθώς οι

macro μέσες τιμές λαμβάνουν υπόψη τη μέση απόδοση των μοντέλων σε όλες τις κατηγορίες, δίνοντας ίσο βάρος σε κάθε κλάση, ανεξαρτήτως μεγέθους του δείγματος. Αυτό εξασφαλίζει μια πιο ισορροπημένη εκτίμηση της απόδοσης, ιδιαίτερα σε δεδομένα με ανισοκατανομή μεταξύ των κλάσεων όπως είναι στην συγκεκριμένη περίπτωση. Επιπλέον, για την απεικόνιση της απόδοσης των μοντέλων μέσω καμπυλών ROC, δημιουργήθηκαν οι καμπύλες one-vs-one καθώς και οι one-vs-rest. Οι καμπύλες one-vs-one συγκρίνουν κάθε δυνατή ζεύξη κλάσεων, δίνοντας λεπτομερή εικόνα για την ικανότητα διάκρισης μεταξύ δύο συγκεκριμένων κατηγοριών. Αντίστοιχα, οι καμπύλες one-vs-rest αξιολογούν κάθε κλάση ξεχωριστά έναντι όλων των υπολοίπων, παρέχοντας μια πιο γενική εικόνα για την απόδοση του μοντέλου ως προς τη διάκριση μιας κατηγορίας από το σύνολο. Συνδυάζοντας αυτές τις προσεγγίσεις, επιτεύχθηκε πιο πλήρης και ακριβής αξιολόγηση της ικανότητας ταξινόμησης των μοντέλων.

Πίνακας 4.10: Επιλεγμένα χαρακτηριστικά για κάθε βιοδείκτη για το 2ο μέρος

Βιοδείκτες	**Επιλεγμένα χαρακτηριστικά για το 2ο μέρος**
CD45-APC+	Age, RR syst., Thrombocytes, wbc, glucose, Intensity, Std, entropy, 500-600 nm, 900-1000 nm
CD235a-PE+	Age, RR syst., Thrombocytes, wbc, glucose, Mean, Std, kurtosis, skewness, 200-300 nm
CD14-PB+	Age, RR syst., Thrombocytes, wbc, glucose, Mean, Median, skewness, entropy, 600-700 nm
CD31-APC	Age, RR syst., Thrombocytes, wbc, glucose, entropy, 100-200 nm, 500-600 nm, 600-700 nm, 800-900 nm
CD146-PE	Age, RR syst., Thrombocytes, wbc, glucose, Median, skewness, 100-200 nm, 200-300 nm, 300-400 nm
CD326-APC+	Age, RR syst., Thrombocytes, wbc, glucose, Std, skewness, entropy, 100-200 nm, 300-400 nm
Lact-FITC+	Age, RR syst., Thrombocytes, wbc, glucose, Mean, kurtosis, skewness, 400-500 nm, 700-800 nm

4.2.2 Αποτελέσματα Απόδοσης Μοντέλων Μηχανικής Μάθησης

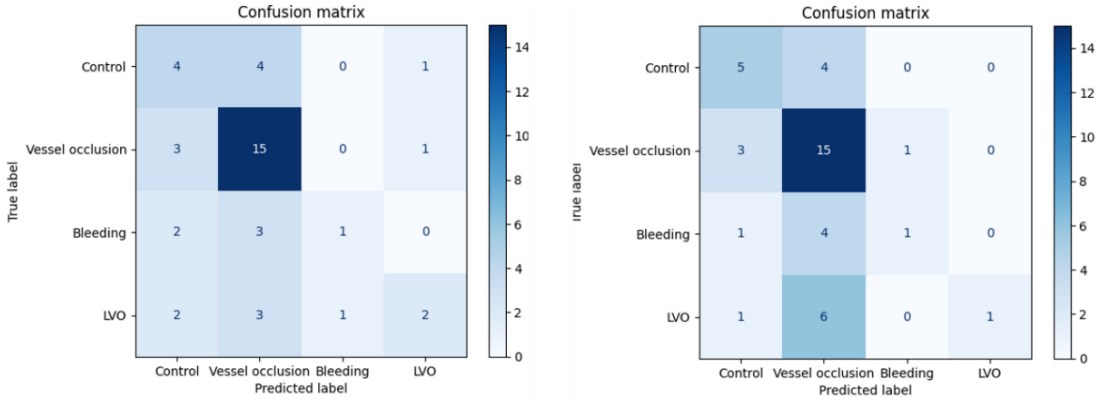
CD45-APC+

Για την ταξινόμηση των δεδομένων CD45-APC+, χρησιμοποιήθηκαν τα μοντέλα Gradient Boosting και Random Forest. Τα βέλτιστα αποτελέσματα των υπερπαραμέτρων που προέκυψαν για το Gradient Boosting παρατίθενται παρακάτω. Η ακρίβεια (macro) ήταν ίδια για τα δύο μοντέλα, αγγίζοντας το 52,40%, υποδεικνύοντας παρόμοια συνολική απόδοση. Ωστόσο, το Random Forest υπερέχει σημαντικά σε Precision (macro) με 62,90%, συγκριτικά με το 49,10% του Gradient Boosting, υποδεικνύοντας ότι το Random Forest διαχειρίζεται καλύτερα τα ψευδώς θετικά. Από την άλλη πλευρά, το Gradient Boosting είχε ελαφρώς υψηλότερο Recall (macro) (41,30% έναντι 40,90%) και F1-score (macro) (41,60% έναντι 40,60%), καθιστώντας το ελαφρώς καλύτερο στην ισορροπία μεταξύ Precision και Recall. Συνολικά, το Random Forest δείχνει καλύτερη ακρίβεια στις θετικές προβλέψεις, ενώ το Gradient Boosting προσφέρει πιο ισορροπημένη απόδοση.

Πίνακας 4.11: Απόδοση των μοντέλων για τον βιοδείκτη CD45-APC+

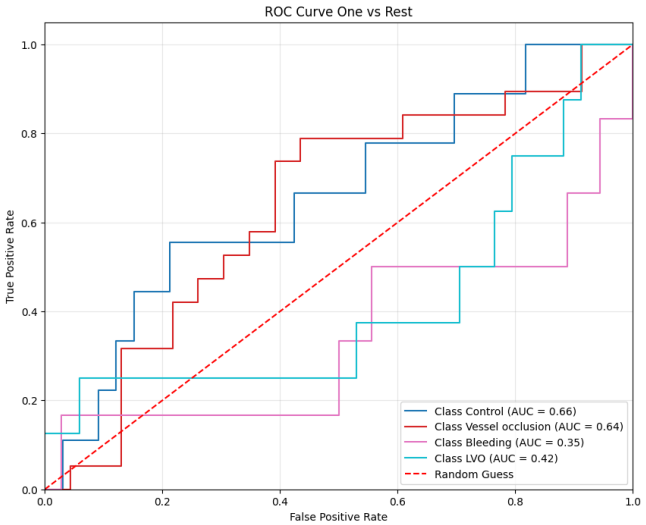
Model	Gradient Boosting	Random Forest

Accuracy (macro)	52.40%	52.40%
Precision (macro)	49.10%	62.90%
Recall (macro)	41.30%	40.90%
F1 – score (macro)	41.60%	40.60%

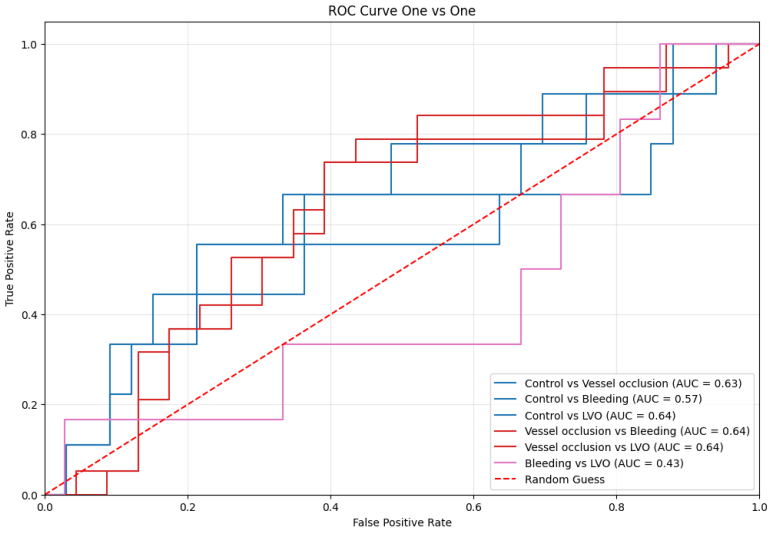


Εικόνα 4.72: Πίνακες σύγχυσης για τα μοντέλα Gradient Boosting (αριστερά) και Random Forest (δεξιά)

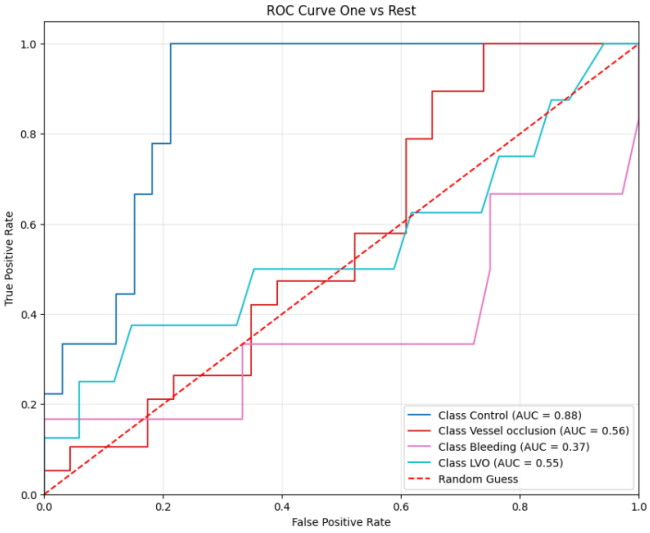
Για το μοντέλο Gradient Boosting, οι καμπύλες ROC δείχνουν ότι οι τιμές AUC για τις διάφορες κατηγορίες και συγκρίσεις είναι γενικά μέτριες έως χαμηλές, με την κατηγορία "Control" να επιτυγχάνει το υψηλότερο AUC (0.66) στην ανάλυση One vs Rest. Αντίθετα, η διάκριση μεταξύ κατηγοριών όπως "Bleeding" και "LVO" είναι ιδιαίτερα δύσκολη, με πολύ χαμηλά AUC (0.35 και 0.42 αντίστοιχα). Για το μοντέλο Random Forest, παρατηρείται ότι η κατηγορία "Control" έχει την καλύτερη απόδοση, με υψηλό AUC (0.88 στο One vs Rest και 0.91 στο One vs One). Αντίθετα, οι κατηγορίες "Bleeding" και "LVO" παρουσιάζουν χαμηλές τιμές AUC, ιδιαίτερα στις συγκρίσεις μεταξύ τους (0.41 και 0.55), υποδεικνύοντας δυσκολία στη διάκριση.



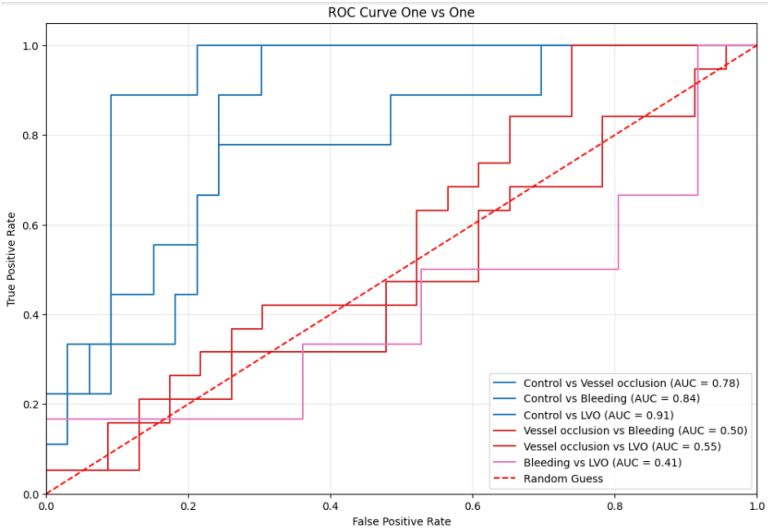
Εικόνα 4.73: Καμπύλη ROC one vs rest - Gradient Boosting



Εικόνα 4.74: : Καμπύλη ROC one vs one - Gradient Boosting



Εικόνα 4.75: Καμπύλη ROC one vs rest – Random Forest



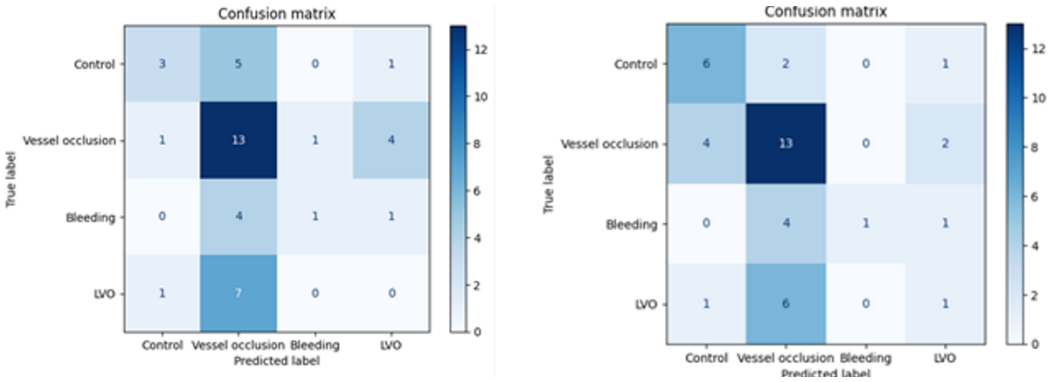
Εικόνα 4.76: Καμπύλη ROC one vs one – Random Forest CD45-APC+

CD235a-PE+

Από το 1ο μέρος της ανάλυσης προέκυψε ότι τα 2 πιο αποδοτικά μοντέλα είναι το Random Forest και το XG Boost. Στην συγκεκριμένη ταξινόμηση μεταξύ 4 κλάσεων, το XG Boost πετυχαίνει υψηλότερη ακρίβεια και Precision, γεγονός που υποδεικνύει καλύτερη ικανότητα αποφυγής ψευδώς θετικών προβλέψεων. Παράλληλα, παρουσιάζει καλύτερο Recall, και υψηλότερο F1-score, καταδεικνύοντας καλύτερη συνολική ισορροπία μεταξύ Precision και Recall.

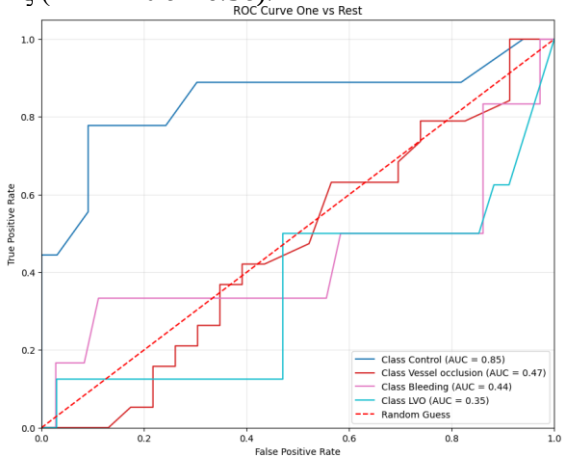
Πίνακας 4.12: Απόδοση των μοντέλων για τον βιοδείκτη CD235a-PE+

Model	Random Forest	XG Boost
Accuracy (macro)	40.5%	50%
Precision (macro)	38.7%	56.6%
Recall (macro)	29.6%	41.1%
F1 – score (macro)	30.5%	40.8%

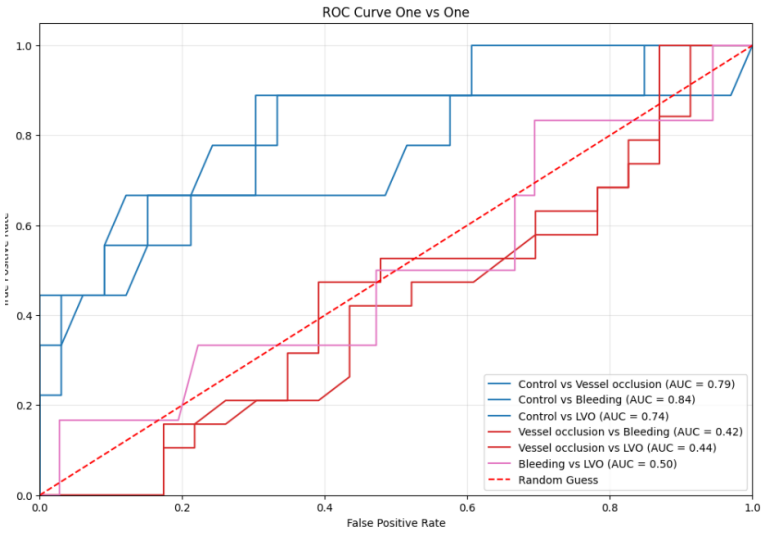


Εικόνα 4.77: Πίνακες σύγχυσης για τα μοντέλα Random Forest (αριστερά) και XG Boost (δεξιά)

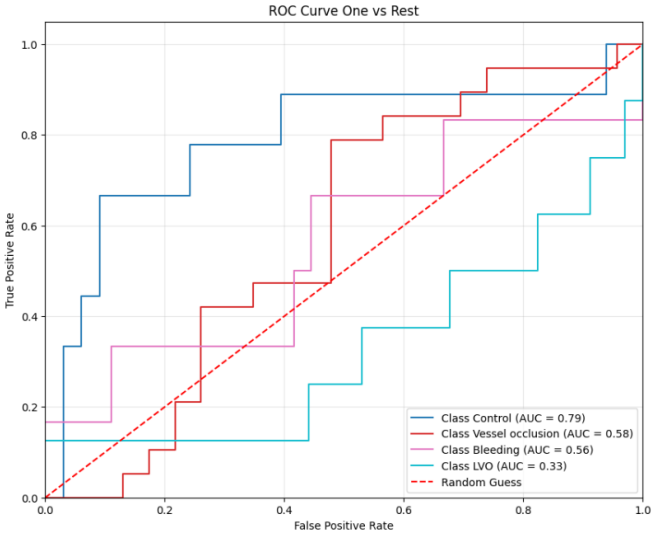
Για το Random Forest, στην ανάλυση One vs Rest, η κατηγορία "Control" επιτυγχάνει το υψηλότερο AUC (0.85), ενώ οι κατηγορίες "Vessel Occlusion", "Bleeding" και "LVO" έχουν χαμηλά AUC (0.47, 0.44, 0.35 αντίστοιχα), δείχνοντας δυσκολία διάκρισης. Στην ανάλυση One vs One, οι καλύτερες τιμές AUC παρατηρούνται μεταξύ "Control vs Bleeding" (0.84) και "Control vs Vessel Occlusion" (0.79), ενώ οι συγκρίσεις που περιλαμβάνουν "LVO" και "Bleeding" έχουν τις χαμηλότερες τιμές (0.42–0.50). Συνολικά, η απόδοση είναι καλή για την κατηγορία "Control", αλλά περιορισμένη στις υπόλοιπες. Από την άλλη, για το μοντέλο XG Boost στην σύγκριση One vs Rest, η κατηγορία "Control" έχει την καλύτερη απόδοση με AUC 0.79, ενώ οι κατηγορίες "Vessel Occlusion" (0.58) και "Bleeding" (0.56) εμφανίζουν μέτρια αποτελέσματα, με την "LVO" να έχει τη χαμηλότερη απόδοση (0.33). Στην ανάλυση One vs One, οι καλύτερες διακρίσεις είναι μεταξύ "Control vs Bleeding" (AUC = 0.77) και "Control vs LVO" (AUC = 0.76), ενώ οι κατηγορίες "Vessel Occlusion" και "Bleeding" παραμένουν δυσδιάκριτες (AUC ~ 0.54-0.56).



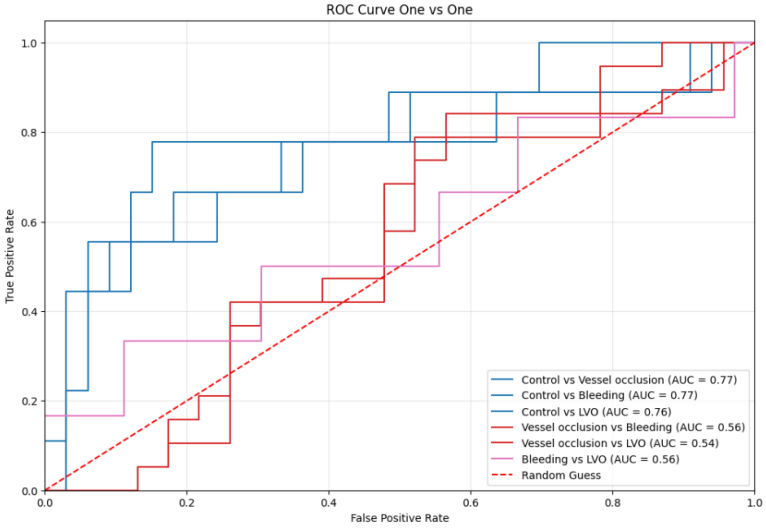
Εικόνα 4.78: : Καμπύλη ROC one vs rest – Random Forest



Εικόνα 4.79: : Καμπύλη ROC one vs one – Random Forest



Εικόνα 4.80: Καμπύλη ROC one vs rest – XG Boost



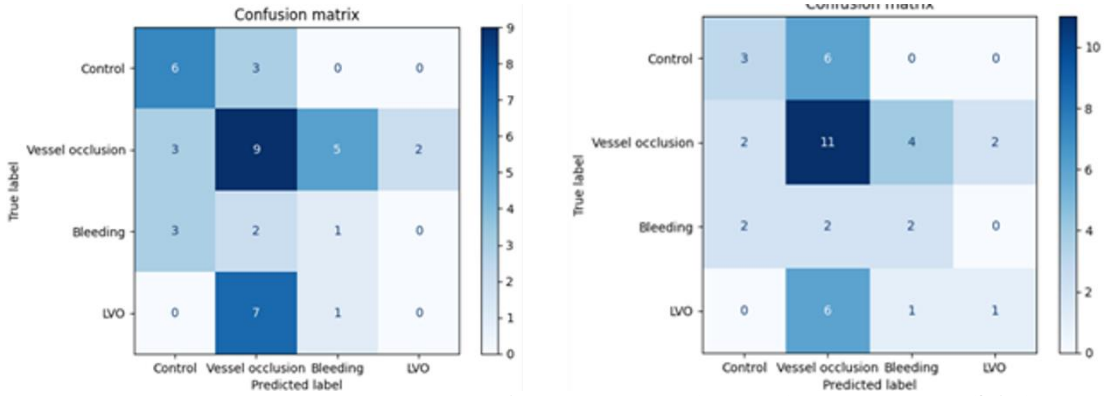
Εικόνα 4.81: Καμπύλη ROC one vs one - XG Boost

CD14-PB+:

Για το CD14-PB+, έχουν χρησιμοποιηθεί τα μοντέλα Naïve Bayes και το Random Forest, με το πρώτο να υπερέχει ελαφρώς σε όλες τις μετρικές. Συγκεκριμένα, το Naïve Bayes επιτυγχάνει υψηλότερη ακρίβεια, Precision καθώς και F1-score, υποδεικνύοντας καλύτερη συνολική απόδοση.

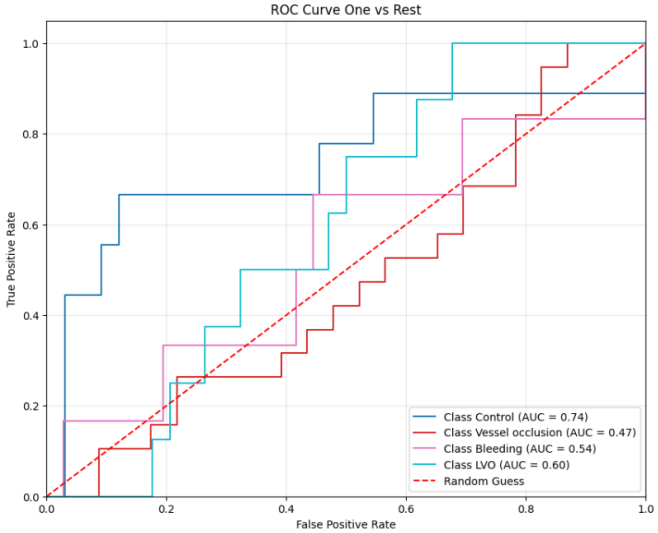
Πίνακας 4.13: Απόδοση των μοντέλων για τον βιοδείκτη CD14-PB+

Model	Random Forest	Naïve Bayes
Accuracy (macro)	38.1%	40.5%
Precision (macro)	26.8%	37.2%
Recall (macro)	32.7%	34.3%
F1 – score (macro)	29.4%	34.1%

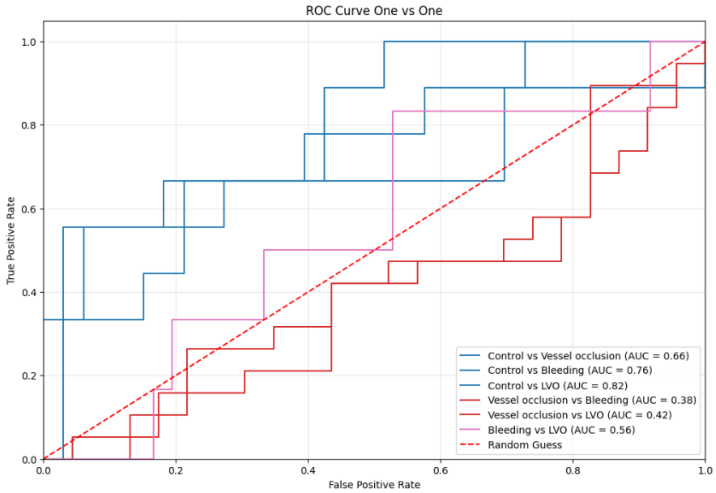


Εικόνα 4.82: Πίνακες σύγχυσης για τα μοντέλα Random Forest (αριστερά) και Naïve Bayes (δεξιά)

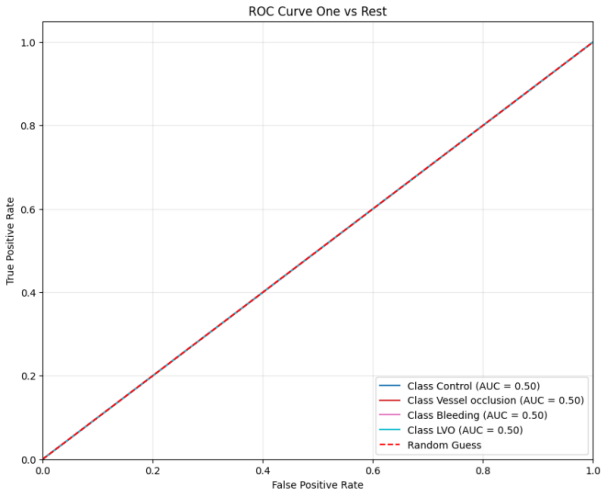
Όσον αφορά τις καμπύλες ROC, το μοντέλο Random Forest δείχνει την καλύτερη απόδοση για την κατηγορία "Control" (AUC = 0.74 στο One vs Rest και 0.82 στο One vs One), ενώ οι κατηγορίες "Vessel Occlusion" και "Bleeding" έχουν χαμηλότερες τιμές AUC (0.38–0.56). Για το μοντέλο Naïve Bayes, στη σύγκριση One vs Rest, όλες οι κατηγορίες παρουσιάζουν AUC 0.50, γεγονός που δείχνει ότι το μοντέλο δεν διακρίνει καλύτερα από μια τυχαία πρόβλεψη. Παράλληλα, για One vs One, οι καλύτερες τιμές AUC παρατηρούνται στις συγκρίσεις "Control vs LVO" (AUC = 0.82) και "Control vs Bleeding" (AUC = 0.76), ενώ οι συγκρίσεις που περιλαμβάνουν τις κατηγορίες "Vessel Occlusion" και "Bleeding" έχουν χαμηλές τιμές AUC (0.38–0.56).



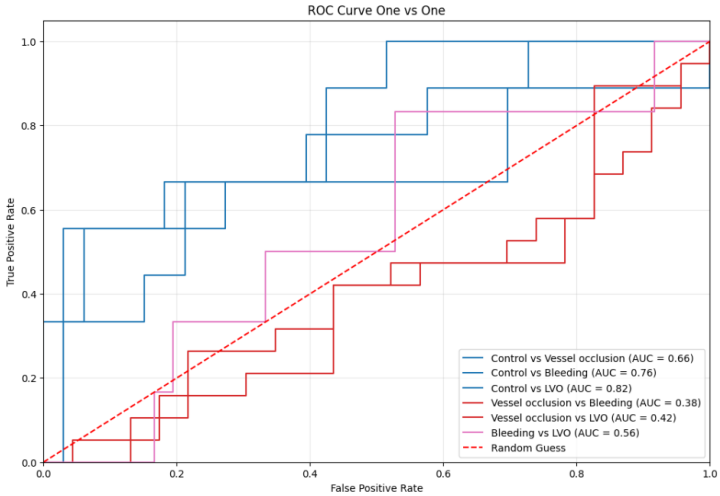
Εικόνα 4.83: Καμπύλη ROC one vs rest – Random Forest



Εικόνα 4.84: Καμπύλη ROC one vs one – Random Forest



Εικόνα 4.85: Καμπύλη ROC one vs rest - Naive Bayes



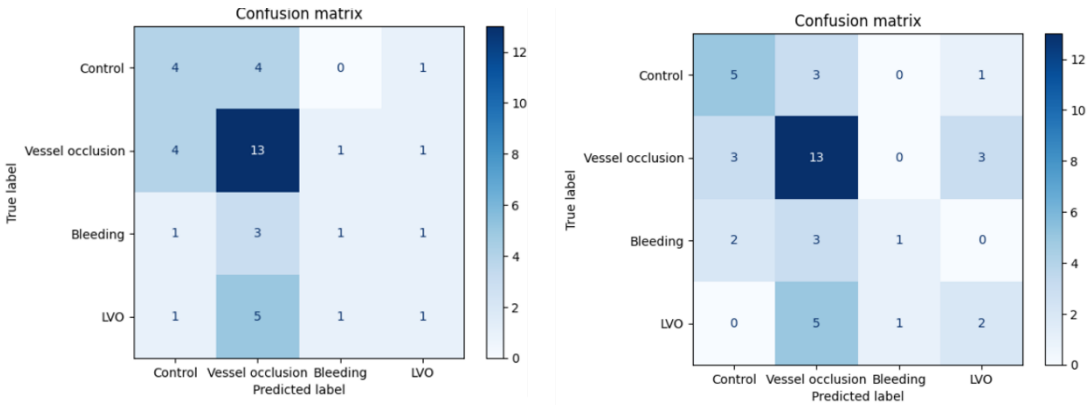
Εικόνα 4.86: Καμπύλη ROC one vs one - Naive Bayes

CD31-APC

Για το CD31-APC, επιλέχθηκαν τα μοντέλα Gradient Boosting και XG Boost. Παρατηρούμε ότι το XG Boost υπερέχει σε όλες τις μετρικές. Το Gradient Boosting, αν και έχει βελτιστοποιηθεί μέσω επιλογής υπερπαραμέτρων, υστερεί στη συνολική απόδοση. Παρατίθενται στον παρακάτω πίνακα συνολικά τα αποτελέσματα για την απόδοση των μετρικών.

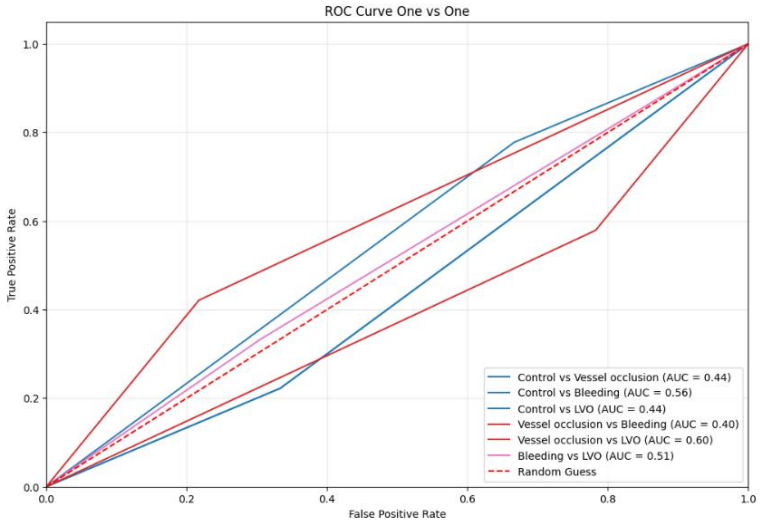
Πίνακας 4.14: Απόδοση των μοντέλων για τον βιοδείκτη CD31-APC

Model	Gradient Boosting	XG Boost
Accuracy (macro)	45.2%	50%
Precision (macro)	37.6%	46.9%
Recall (macro)	35.5%	41.4%
F1 – score (macro)	35%	41.7%

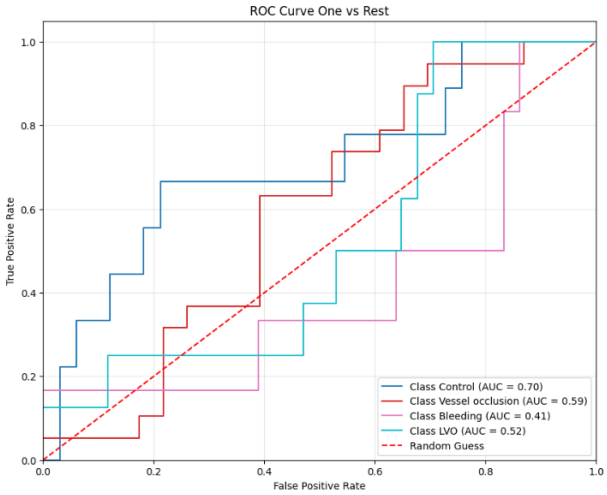


Εικόνα 4.87: Πίνακες σύγχυσης για τα μοντέλα Gradient Boosting (αριστερά) και XG Boost (δεξιά)

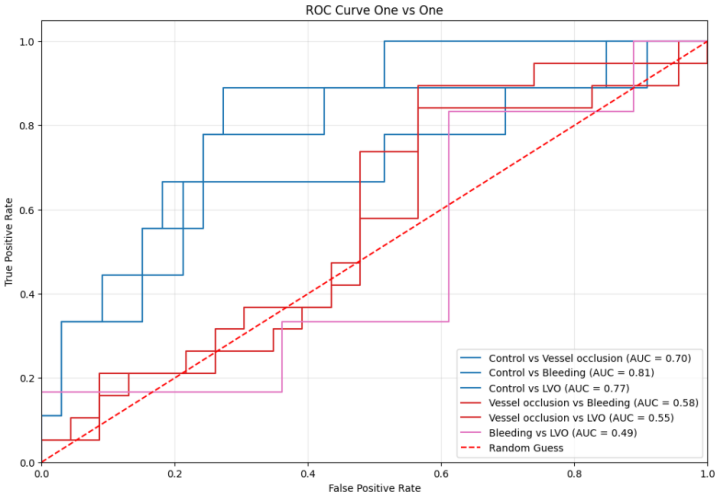
Στις καμπύλες ROC για το Gradient Boosting, η ανάλυση One vs Rest δείχνει ότι η κατηγορία "Control" έχει την καλύτερη απόδοση με AUC 0.70, ενώ οι κατηγορίες "Vessel Occlusion" (0.59) και "LVO" (0.52) έχουν μέτρια αποτελέσματα. Η κατηγορία "Bleeding" παρουσιάζει τη χαμηλότερη απόδοση με AUC 0.41, δείχνοντας δυσκολία στη διάκριση. Στην ανάλυση One vs One, οι τιμές AUC είναι γενικά χαμηλές, με τις περισσότερες συγκρίσεις να κυμαίνονται μεταξύ 0.40 και 0.60, υποδεικνύοντας σημαντικές δυσκολίες στη διάκριση των κατηγοριών. Για το μοντέλο XG Boost, η κατηγορία "Control" παρουσιάζει την καλύτερη απόδοση με AUC 0.79, ενώ οι κατηγορίες "LVO" (0.61) και "Vessel Occlusion" (0.59) επιδεικνύουν μέτρια αποτελέσματα. Από την άλλη, η κατηγορία "Bleeding" έχει τη χαμηλότερη τιμή AUC (0.41), γεγονός που υποδεικνύει δυσκολία στη διάκριση από τις υπόλοιπες κατηγορίες. Στις συγκρίσεις μεταξύ δύο κατηγοριών, οι τιμές AUC είναι υψηλότερες για τα ζεύγη "Control vs Bleeding" (0.81) και "Control vs LVO" (0.77), ενώ οι κατηγορίες "Vessel Occlusion" και "Bleeding" εμφανίζουν τις χαμηλότερες αποδόσεις (AUC = 0.49–0.58).



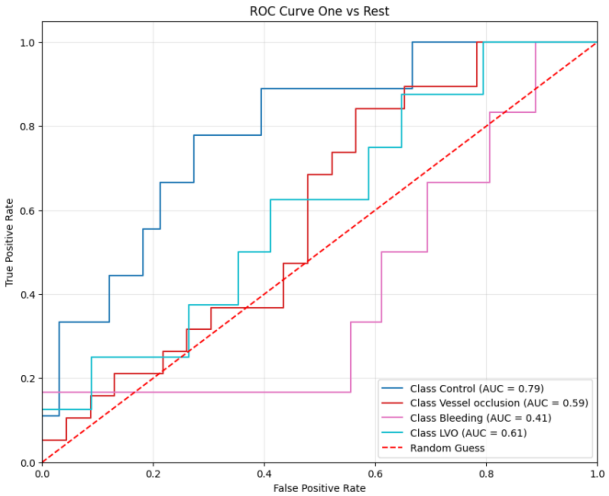
Εικόνα 4.88: Καμπύλη ROC one vs one – Gradient Boosting



Εικόνα 4.89: Καμπύλη ROC one vs rest – Gradient Boosting



Εικόνα 4.90: Καμπύλη ROC one vs one – XG Boost



Εικόνα 4.91: Καμπύλη ROC one vs rest – XG Boost

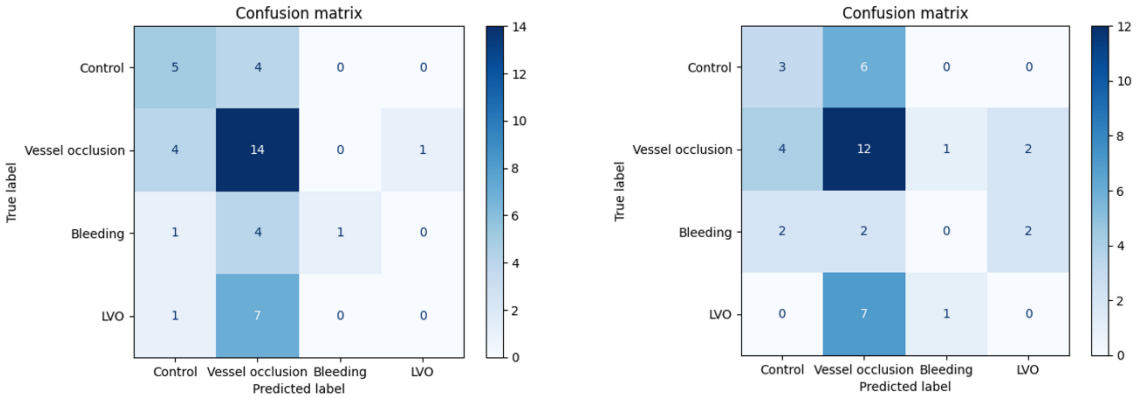
CD146-PE

Παρακάτω παρατίθενται ο πίνακας με τα αποτελέσματα για την απόδοση των 2 μοντέλων. Τα αποτελέσματα δείχνουν ότι το Gradient Boosting υπερέχει σε όλες τις μετρικές, επιτυγχάνοντας υψηλότερη ακρίβεια (Accuracy 47.6%), καλύτερη ισορροπία μεταξύ Precision και Recall, καθώς και αυξημένο F1-score. Αντίθετα, το Random Forest εμφανίζει σαφώς

χαμηλότερη απόδοση, με Precision μόλις 19.4% και F1-score 21.4%, γεγονός που υποδηλώνει μειωμένη ικανότητα του μοντέλου να προβλέψει σωστά τις κατηγορίες.

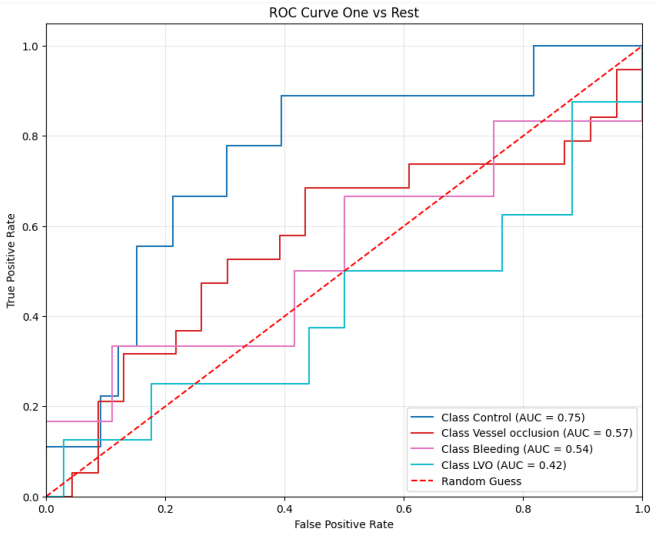
Πίνακας 4.15: Απόδοση των μοντέλων για τον βιοδείκτη CD146-PE

Model	Gradient Boosting	Random Forest
Accuracy (macro)	47.6%	35.7%
Precision (macro)	48.4%	19.4%
Recall (macro)	36.5%	24.1%
F1 – score (macro)	34.2%	21.4%

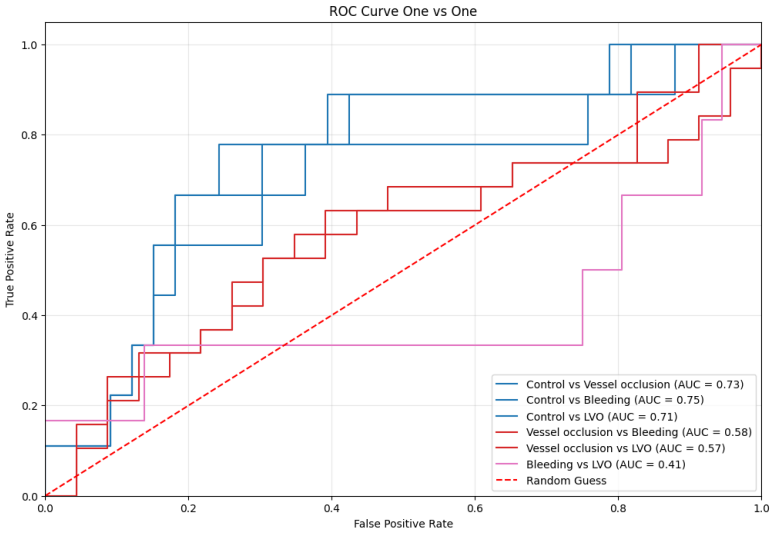


Εικόνα 4.92: Πίνακες σύγχυσης για τα μοντέλα Gradient Boosting (αριστερά) και Random Forest (δεξιά)

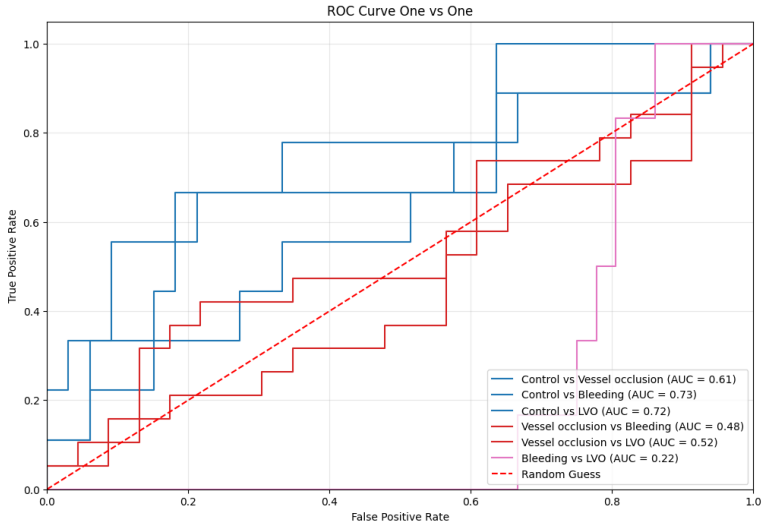
Όπως παρατηρείται για το μοντέλο Gradient Boosting, στις παρακάτω καμπύλες ROC, η κατηγορία "Control" εμφανίζει την καλύτερη απόδοση με AUC 0.75 τόσο στο One vs Rest όσο και σε συγκρίσεις όπως "Control vs Bleeding". Οι κατηγορίες "Vessel Occlusion" (AUC = 0.57–0.73) και "Bleeding" (AUC = 0.54–0.58) έχουν μέτρια αποτελέσματα, ενώ η κατηγορία "LVO" παρουσιάζει τη χαμηλότερη απόδοση με AUC 0.42–0.57. Παράλληλα, για το Random Forest, η κατηγορία "Control" έχει την καλύτερη απόδοση, με AUC 0.71 στο One vs Rest και 0.73 στη σύγκριση με την κατηγορία "Bleeding". Οι κατηγορίες "Vessel Occlusion" και "Bleeding" παρουσιάζουν μέτριες τιμές AUC (0.48–0.61), ενώ η κατηγορία "LVO" έχει τη χαμηλότερη απόδοση, με AUC 0.43 στο One vs Rest και μόλις 0.22 στη σύγκριση με την κατηγορία "Bleeding".



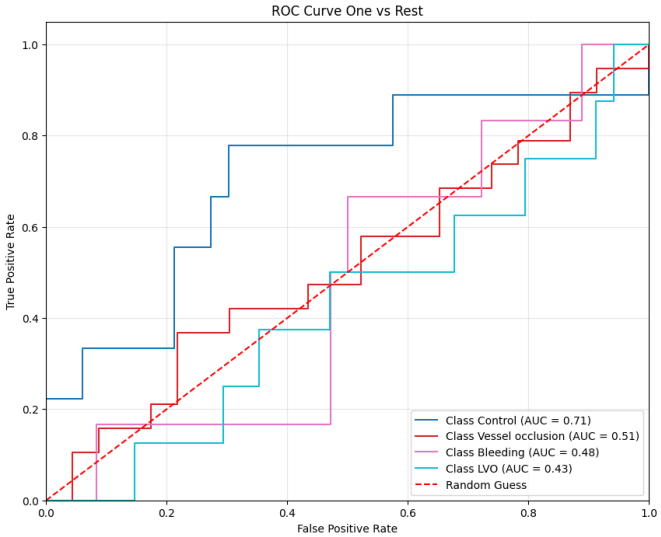
Εικόνα 4.93: Καμπύλη ROC one vs rest - Gradient Boosting



Εικόνα 4.94: Καμπύλη ROC one vs one - Gradient Boosting



Εικόνα 4.95: Καμπύλη Roc one vs one - Random Forest



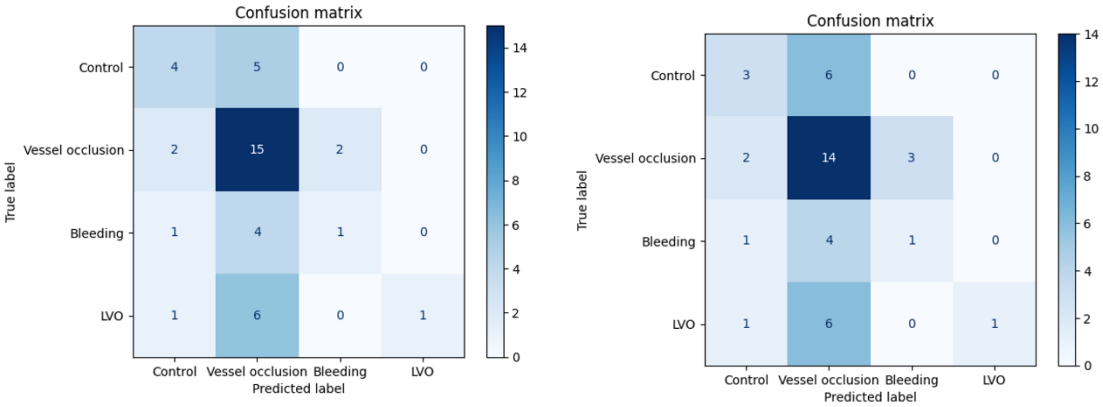
Εικόνα 4.96: Καμπύλη Roc one vs rest - Random Forest

CD326-APC+

Αντίστοιχα, για τον δείκτη CD326-APC+, εξετάστηκαν τα μοντέλα Random Forest και K-Nearest Neighbors. Από τα αποτελέσματα που παρουσιάζονται παρακάτω, παρατηρείται ότι το το Random Forest υπερέχει του KNN σε όλες τις μετρικές.

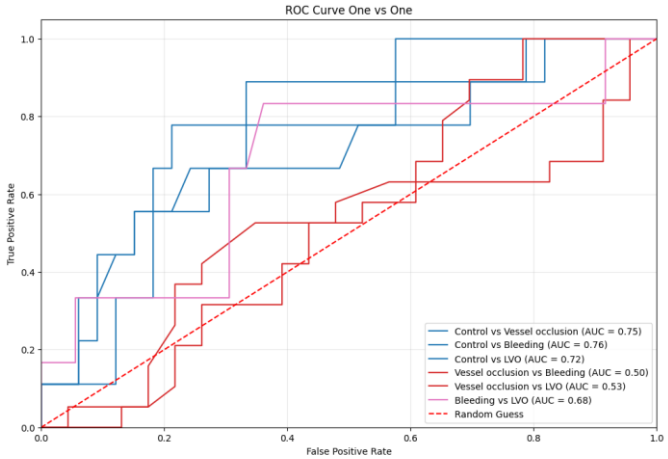
Πίνακας 4.16: Απόδοση των μοντέλων για τον βιοδείκτη CD326-APC+

Model	Random Forest	KNN
Accuracy (macro)	50%	45.2%
Precision (macro)	58.3%	53.6%
Recall (macro)	38.1%	34%
F1 – score (macro)	38.2%	34.2%

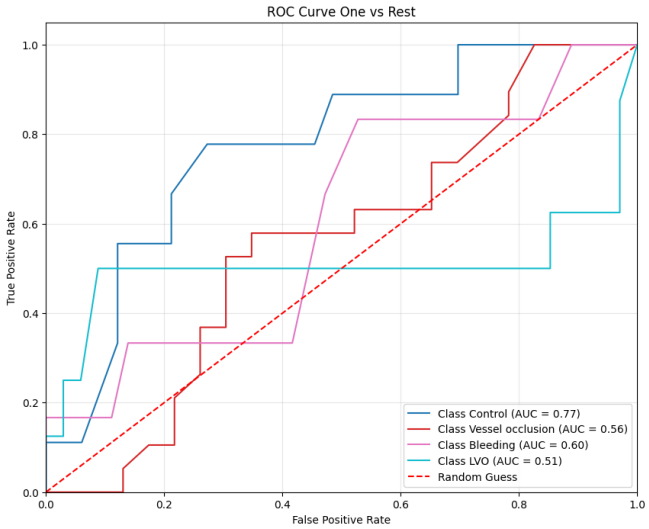


Εικόνα 4.97: Πίνακες σύγχυσης για τα μοντέλα Random Forest (αριστερά) και KNN (δεξιά)

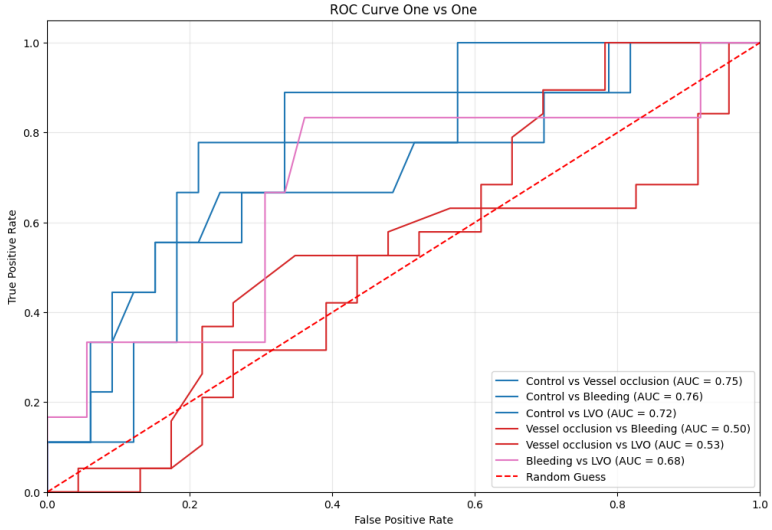
Για το Random Forest, οι καμπύλες ROC δείχνουν ότι η κατηγορία "Control" έχει την καλύτερη απόδοση, με AUC 0.77 στο One vs Rest και 0.76 στη σύγκριση με την κατηγορία "Bleeding". Οι κατηγορίες "Vessel Occlusion" και "Bleeding" παρουσιάζουν μέτριες αποδόσεις (AUC 0.56–0.60 στο One vs Rest και 0.50–0.53 στο One vs One). Η κατηγορία "LVO" εμφανίζει χαμηλότερη διάκριση, με AUC 0.51 στο One vs Rest, αλλά βελτιώνεται στο One vs One με AUC 0.68 στη σύγκριση με την κατηγορία "Bleeding". Όσον αφορά το Knn, στην σύγκριση One vs Rest, οι κατηγορίες "Control" και "Bleeding" επιτυγχάνουν AUC 0.60, ενώ οι "Vessel Occlusion" και "LVO" παρουσιάζουν χαμηλότερες τιμές με AUC 0.52 και 0.46 αντίστοιχα, υποδεικνύοντας δυσκολία στη διάκριση. Στο One vs One, οι καλύτερες επιδόσεις παρατηρούνται στις συγκρίσεις "Control vs Vessel Occlusion" (AUC = 0.75) και "Control vs Bleeding" (AUC = 0.76). Αντίθετα, οι συγκρίσεις που περιλαμβάνουν τις κατηγορίες "Vessel Occlusion" και "LVO" έχουν χαμηλότερες αποδόσεις (AUC 0.50–0.53), με την κατηγορία "Bleeding vs LVO" να σημειώνει μέτρια απόδοση (AUC 0.68).



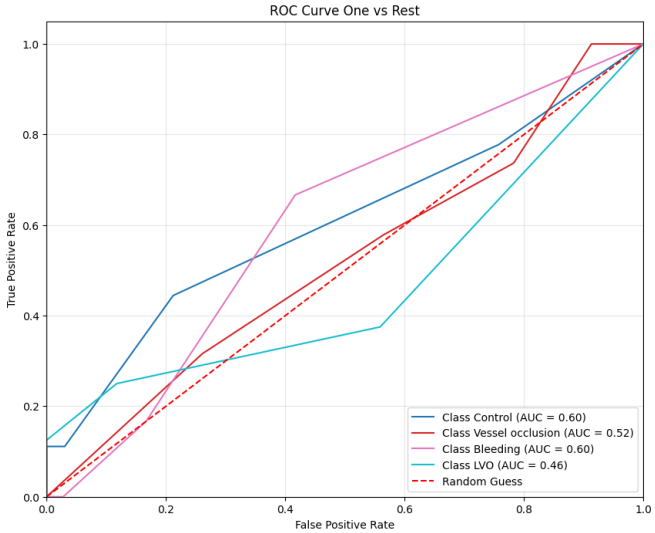
Εικόνα 4.98: Καμπύλη Roc one vs one - Random Forest



Εικόνα 4.99: Καμπύλη Roc one vs rest - Random Forest



Εικόνα 4.100: Καμπύλη Roc one vs one – KNN



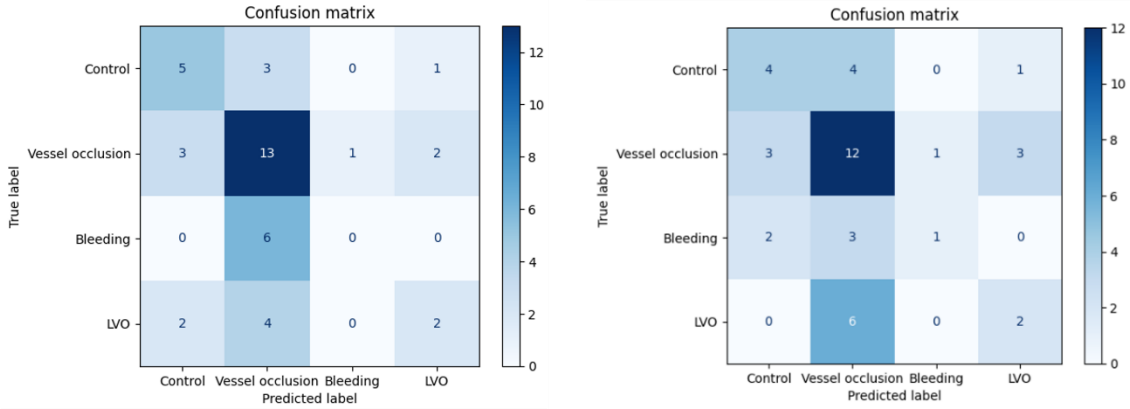
Εικόνα 4.101: Καμπύλη Roc one vs rest - KNN

Lact-FITC+

Για το Lact-FITC+, χρησιμοποιήθηκαν τα μοντέλα Random Forest και XG Boost, τα οποία παρουσιάζουν παρόμοιες επιδόσεις με μικρές διαφορές στις μετρικές οι οποίες παρουσιάζονται στον παρακάτω πίνακα.

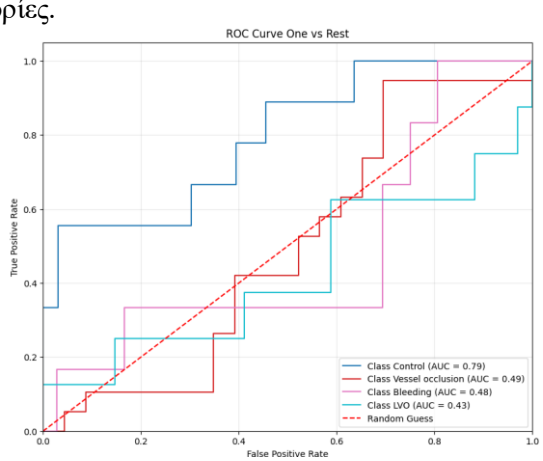
Πίνακας 4.17: Απόδοση των μοντέλων για τον βιοδείκτη Lact-FITC+4

Model	Random Forest	XG Boost
Accuracy (macro)	47.6%	45.2%
Precision (macro)	35%	43.9%
Recall (macro)	37.2%	37.3%
F1 – score (macro)	35.3%	38.1%

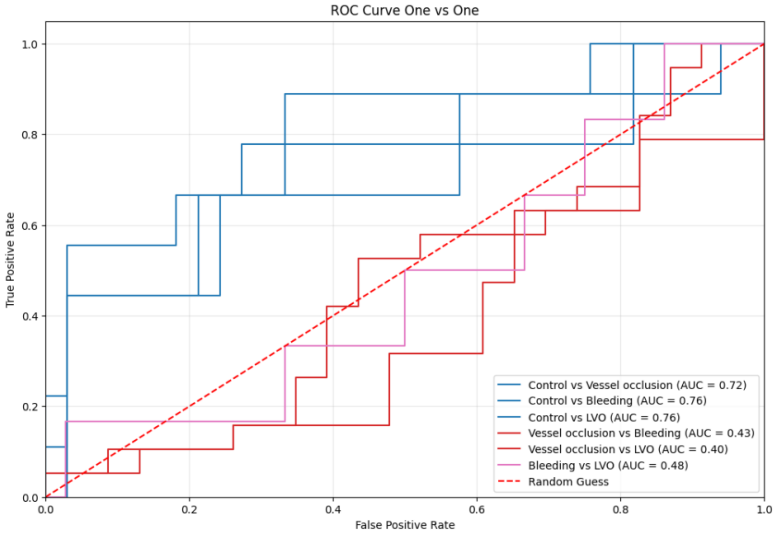


Εικόνα 4.102: Πίνακες σύγχυσης για το Random Forest (αριστερά) και για το XG Boost (δεξιά)

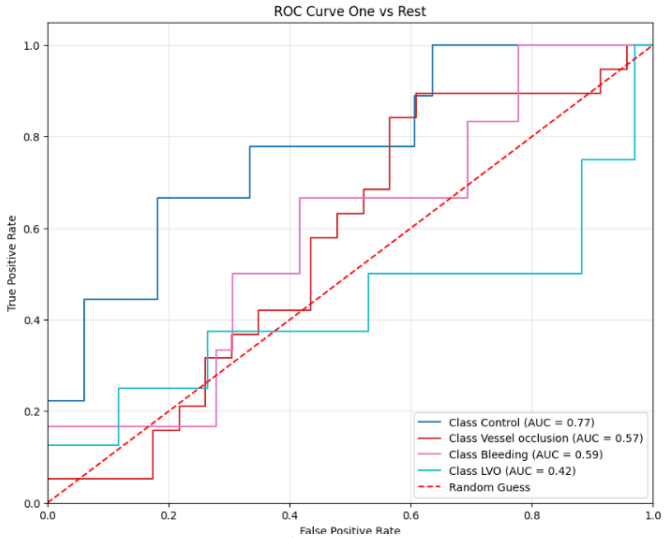
Για το Random Forest, στην ανάλυση One vs Rest, η κατηγορία "Control" παρουσιάζει την καλύτερη απόδοση με AUC 0.79, ενώ οι κατηγορίες "Vessel Occlusion" (0.49), "Bleeding" (0.48) και "LVO" (0.43) έχουν χαμηλές επιδόσεις, υποδεικνύοντας περιορισμένη διακριτική ικανότητα. Παράλληλα, στο One vs One, οι καλύτερες επιδόσεις εμφανίζονται στις συγκρίσεις "Control vs Bleeding" (AUC = 0.76) και "Control vs LVO" (AUC = 0.76), ενώ οι κατηγορίες "Vessel Occlusion" και "Bleeding" εμφανίζουν τις χαμηλότερες τιμές AUC (0.40–0.48). Από την άλλη, για το μοντέλο XG Boost, στο One vs Rest, η κατηγορία "Control" παρουσιάζει την καλύτερη απόδοση με AUC 0.77, ενώ οι κατηγορίες "Bleeding" και "Vessel Occlusion" εμφανίζουν μέτριες τιμές, με την "LVO" να σημειώνει τη χαμηλότερη διάκριση. Στη σύγκριση One vs One, οι συγκρίσεις "Control vs Bleeding" (AUC = 0.72) και "Control vs LVO" (AUC=0.76) έχουν την καλύτερη απόδοση, ενώ οι κατηγορίες που περιλαμβάνουν "Vessel Occlusion" και "LVO" εμφανίζουν χαμηλότερες τιμές (AUC 0.50–0.64). Συνολικά, το XG Boost φαίνεται να διακρίνει αποδοτικά την κατηγορία "Control", ενώ αντιμετωπίζει δυσκολίες με τις υπόλοιπες κατηγορίες.



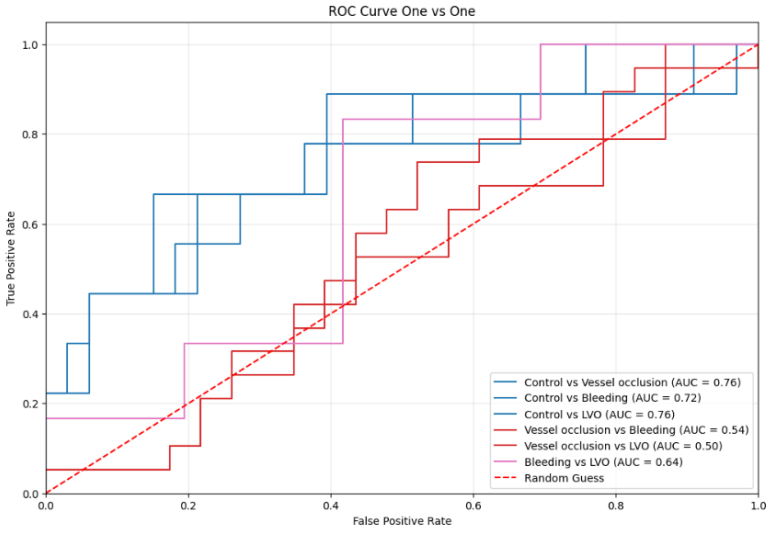
Εικόνα 4.103: Καμπύλη Roc one vs rest – Random Forest



Εικόνα 4.104: Καμπύλη Roc one vs one – Random Forest



Εικόνα 4.105: Καμπύλη Roc one vs rest – XG Boost



Εικόνα 4.106: Καμπύλη Roc one vs one – XG Boost

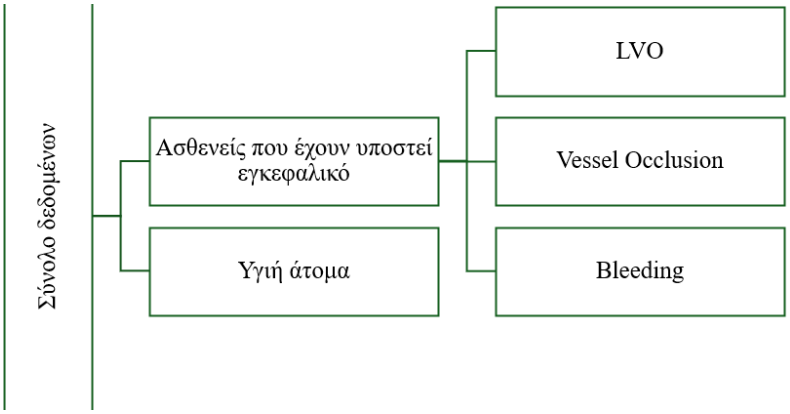
4.3 Μέρος 3: Διάγνωση σε 2 στάδια

Στο τρίτο μέρος της έρευνας, εφαρμόζεται η προσέγγιση "Two-Stage Hybrid Data Classifiers", δηλαδή γίνεται διάγνωση σε 2 στάδια. Η θεμελιώδης λογική αυτής της μεθόδου βασίζεται στη διάκριση των δεδομένων σε πλειοψηφικές και μειοψηφικές κατηγορίες, με την εφαρμογή διαφορετικών ταξινομητών για την ανίχνευση κάθε κατηγορίας ξεχωριστά. Αυτή η στρατηγική στοχεύει στη βελτίωση της απόδοσης του μοντέλου, ιδίως σε περιπτώσεις ανισόρροπων δεδομένων, όπως συμβαίνει όταν οι κατηγορίες δεν εκπροσωπούνται ισομερώς στο σύνολο δεδομένων.

Στο πρώτο στάδιο, τα δεδομένα κατηγοριοποιούνται σε πλειοψηφικές και μειοψηφικές κατηγορίες, με τον ταξινομητή να εστιάζει στην ανίχνευση των πλειοψηφικών κατηγοριών. Η χρήση αυτής της μεθόδου συμβάλλει στη μείωση των λαθών στις συχνότερες κατηγορίες, επιτρέποντας παράλληλα μια πιο ακριβή γενίκευση του μοντέλου. Στην συγκεκριμένη περίπτωση, το 1ο στάδιο περιλαμβάνει την ταξινόμηση των ασθενών που έχουν υποστεί εγκεφαλικό επεισόδιο. Στο 1ο μέρος της εργασίας, υλοποιήθηκε η ταξινόμηση των ασθενών και των υγιών ατόμων, για το 1ο στάδιο θα χρησιμοποιηθεί το μοντέλο που εμφανίστηκε ως το πιο αποδοτικό στο 1ο μέρος της εργασίας.

Στο δεύτερο στάδιο, αφού οι πλειοψηφικές κατηγορίες έχουν κατανεμηθεί, ο ταξινομητής εστιάζει στις μειοψηφικές κατηγορίες, οι οποίες στην περίπτωση μας περιλαμβάνουν τις κατηγορίες εγκεφαλικού επεισοδίου Control, Vessel Occlusion, Bleeding, και LVO. Αξίζει να σημειωθεί ότι, παρόλο που η κατηγορία Control αναφέρεται σε υγιή άτομα, το 1ο μοντέλο ταξινόμησης δεν έχει 100% ακρίβεια. Αυτό σημαίνει ότι μπορεί να ταξινομηθούν λανθασμένα κάποια άτομα ως υγιή, ενώ στην πραγματικότητα να έχουν υποστεί εγκεφαλικό επεισόδιο. Αυτά τα λάθη ταξινόμησης από το 1ο μοντέλο ενδέχεται να περάσουν στο 2ο στάδιο.

Για την ανίχνευση αυτών των μειοψηφικών κατηγοριών, εφαρμόζονται συχνά τεχνικές oversampling, όπως η μέθοδος SMOTE, με στόχο την εξισορρόπηση των δεδομένων και την αύξηση της σημασίας των μειοψηφικών κατηγοριών κατά τη διάρκεια της εκπαίδευσης του μοντέλου.



Εικόνα 4.107: Διάγραμμα ροής για το μέρος 3 (διάγνωση σε 2 στάδια)

Η προσέγγιση αυτή έχει εφαρμοστεί με επιτυχία σε πλήθος μελετών ([97-100]) και έχει αποδείξει την αποτελεσματικότητά της στην ενίσχυση της ακρίβειας ανίχνευσης, ιδίως για τις μειοψηφικές κατηγορίες. Έχει, επίσης, αποφέρει βελτιωμένα αποτελέσματα σε σύγκριση με άλλες μεθόδους, καθώς μειώνει το ποσοστό των ψευδών θετικών και ενισχύει τη γενική ακρίβεια του μοντέλου, καθιστώντας την κατάλληλη για εφαρμογές όπου η σωστή ανίχνευση σπάνιων περιστατικών είναι κρίσιμη. Επιπλέον, η προσέγγιση αυτή ενδυναμώνει τη γενική απόδοση του συστήματος ταξινόμησης, προσφέροντας μια ευέλικτη και αποδοτική λύση σε προβλήματα ανισοκατανεμημένων δεδομένων, όπως εκείνα που αφορούν ιατρικές διαγνώσεις και ανίχνευση σπάνιων περιστατικών.

4.3.1 Αποτελέσματα από την Εφαρμογή Επιλογής Χαρακτηριστικών στο 1ο Στάδιο

Βιοδείκτες	Μοντέλο 1ου σταδίου	Μοντέλο 2ου σταδίου	Accuracy	Precision	Recall	F1-score
CD45-APC+	**Random Forest**	**Random Forest**	50%	18%	25%	20%
CD235a-PE+	**Random Forest**	**Random Forest**	73%[1]	57%	50%	50%
CD14-PB+	**Random Forest**	**KNN**	67%	17%	25%	20%
CD31-APC	**Gradient Boosting**	**KNN**	55%	14%	25%	18%
CD146-PE	**Random Forest**	**XG Boost**	46%	46%	50%	40%
CD326-APC+	**Random Forest**	**KNN**	37%	10%	25%	14%
Lact-FITC+	**Random Forest**	**SVM**	40%	13%	20%	16%

4.3.2 Αποτελέσματα από την Εφαρμογή Επιλογής Χαρακτηριστικών και στα 2 Στάδια

Στο δεύτερο μέρος του κεφαλαίου 4.3, η μεθοδολογία του Two-Stage Hybrid Classifier παραμένει η ίδια, ωστόσο, παρουσιάζεται μια σημαντική τροποποίηση στην προσέγγιση λόγω της εφαρμογής μεθόδων επιλογής χαρακτηριστικών. Συγκεκριμένα, πριν από το 1ο στάδιο, εφαρμόζεται μια διαδικασία επιλογής χαρακτηριστικών, η οποία αποσκοπεί στην επιλογή των πλέον σημαντικών χαρακτηριστικών για τη διάκριση μεταξύ των ασθενών που έχουν υποστεί εγκεφαλικό επεισόδιο και των υγιών ατόμων. Στο επόμενο 2ο στάδιο, εφαρμόζεται ξανά μια διαδικασία επιλογής χαρακτηριστικών για να εντοπιστούν τα κατάλληλα χαρακτηριστικά που θα επιτρέψουν τη διάγνωση του τύπου του εγκεφαλικού επεισοδίου.

Αυτή η διαφοροποίηση στην επιλογή χαρακτηριστικών για κάθε στάδιο προκύπτει από την παρατήρηση ότι τα χαρακτηριστικά που είναι σημαντικά για τη διάγνωση του εγκεφαλικού επεισοδίου γενικά (δηλαδή αν ο ασθενής έχει ή όχι εγκεφαλικό) διαφέρουν από εκείνα που είναι κρίσιμα για την κατηγοριοποίηση του τύπου εγκεφαλικού επεισοδίου. Δηλαδή, τα χαρακτηριστικά που επηρεάζουν τη διάγνωση της παρουσίας ενός εγκεφαλικού επεισοδίου μπορεί να μην είναι τα ίδια με εκείνα που επισημαίνουν τη διαφοροποίηση μεταξύ των διαφόρων τύπων εγκεφαλικών επεισοδίων (π.χ., Vessel Occlusion, LVO, Bleeding).

Βιοδείκτες	Επιλεγμένα χαρακτηριστικά για το 3ο μέρος (διάγνωση τύπου εγκεφαλικού με διπλή διάγνωση)
CD45-APC+	RR syst., RR diast., Thrombocytes, crp, kurtosis, 100-200 nm, 600-700 nm, 700-800 nm, 800-900 nm, 900-1000 nm
CD235a-PE+	RR diast., Thrombocytes, Mean, Median, Std, skewness, 100-200 nm, 300-400 nm, 400-500 nm, 500-600 nm

[1] Η ακρίβεια του μοντέλου έχει υπολογιστεί στο 73%, η οποία δείχνει το ποσοστό των σωστών προβλέψεων σε σχέση με το συνολικό αριθμό των προβλέψεων. Ωστόσο, είναι σημαντικό να σημειωθεί ότι το μοντέλο δεν έχει προβλέψει καθόλου την κατηγορία LVO. Η συγκεκριμένη διαπίστωση αναλύεται στο κεφάλαιο 5 «Συμπεράσματα».

CD14-PB+	Age, RR diast., Thrombocytes, wbc, crp, Total, skewness, entropy, 500-600 nm, 900-1000 nm
CD31-APC	Thrombocytes, wbc, crp, Mean, kurtosis, skewness, 400-500 nm, 500-600 nm, 600-700 nm, 800-900 nm
CD146-PE	RR diast., Thrombocytes, glucose, Total, Median, entropy, 200-300 nm, 500-600 nm, 700-800 nm, 900-1000 nm
CD326-APC+	Thrombocytes, wbc, glucose, Total, Intensity, entropy, 300-400 nm, 400-500 nm, 500-600 nm, 600-700 nm
Lact-FITC+	Age, RR syst., Thrombocytes, glucose, Total, Mean, kurtosis, skewness, entropy, 500-600 nm

Βιοδείκτες	**Μοντέλο 1ου σταδίου**	**Μοντέλο 2ου σταδίου**	**Accuracy**	**Precision**	**Recall**	**F1-score**
CD45-APC+	**Random Forest**	**Random Forest**	59%	34%	40%	35%
CD235a-PE+	**Random Forest**	KNN	55%	19%	34%	24%
CD14-PB+	**Random Forest**	**KNN**	46%	15%	33%	20%
CD31-APC	**Gradient Boosting**	**SVM**	46%	44%	53%	40%
CD146-PE	**Random Forest**	**Random Forest**	36%	22%	27%	23%
CD326-APC+	**Random Forest**	**Random Forest**	37%	21%	40%	25%
Lact-FITC+	**Random Forest**	**Gradient Boosting**	55%	48%	48%	46%

4.4 Μέρος 4: Διάκριση του είδους του εγκεφαλικού μεταξύ 3 κλάσεων

Από τα παραπάνω αποτελέσματα, παρατηρείται ότι ορισμένες κατηγορίες διαχωρίζονται με μεγαλύτερη ευκολία σε σύγκριση με άλλες. Η συγκεκριμένη παρατήρηση οδηγεί στην εξέταση της συγχώνευσης των κατηγοριών Vessel Occlusion και LVO σε μία ενιαία κατηγορία Ισχαιμικού Εγκεφαλικού, απόφαση η οποία παρουσιάζει ενδιαφέρον τόσο κλινικά όσο και μεθοδολογικά.

Από κλινικής σκοπιάς, τα ισχαιμικά εγκεφαλικά επεισόδια, ανεξάρτητα από το αν προκαλούνται από απόφραξη μεγάλου αγγείου (LVO) ή μικρότερου αγγείου (Vessel Occlusion), μοιράζονται έναν κοινό παθοφυσιολογικό μηχανισμό ο οποίος σχετίζεται με τη διακοπή της ροής του αίματος στον εγκέφαλο λόγω θρόμβωσης ή εμβολής, οδηγώντας σε παρόμοιες κλινικές επιπτώσεις. Από μεθοδολογική σκοπιά, η ανάλυση των ROC καμπυλών κατέδειξε ότι το μοντέλο εμφανίζει σημαντικές δυσκολίες στη διάκριση αυτών των δύο κατηγοριών, γεγονός που υποδηλώνει ότι τα χαρακτηριστικά τους είναι ιδιαίτερα όμοια και αλληλεπικαλυπτόμενα. Αντίθετα, η διάκριση μεταξύ Ισχαιμικού και Αιμορραγικού εγκεφαλικού (Bleeding) είναι πιο σαφής, καθώς τα δύο αυτά είδη εγκεφαλικού προκύπτουν

από διαφορετικούς παθογενετικούς μηχανισμούς και εκδηλώνονται με διακριτές κλινικές και βιολογικές παραμέτρους.

Επομένως, η συγχώνευση των κατηγοριών Vessel Occlusion και LVO σε μία ενιαία κλάση Ισχαιμικού εγκεφαλικού συμβάλλει στην απλοποίηση της ταξινόμησης, επιτρέποντας στο μοντέλο να αποφύγει την περιττή σύγχυση μεταξύ παρόμοιων υποκατηγοριών και να επιτύχει πιο ακριβή και σταθερή ταξινόμηση μεταξύ των τριών βασικών κατηγοριών: Control, Ischemic stroke και Hemorrhage stroke. Επιπλέον, η συγκεκριμένη προσέγγιση αναμένεται να βελτιώσει την απόδοση των μοντέλων ταξινόμησης, οδηγώντας σε αυξημένα ποσοστά Accuracy, Recall και AUC scores, ενώ παράλληλα καθιστά τη διαφοροποίηση μεταξύ των τύπων εγκεφαλικού πιο ρεαλιστική και κλινικά χρήσιμη.

4.4.1 Αποτελέσματα από την επιλογή χαρακτηριστικών

Τα βέλτιστα αποτελέσματα που προέκυψαν για κάθε βιοδείκτη καταγράφονται στον ακόλουθο πίνακα. Επιπλέον, παρουσιάζεται η τιμή AUC, η οποία αποτυπώνει την ικανότητα διάκρισης του μοντέλου μεταξύ των κατηγοριών-στόχων, συγκεκριμένα μεταξύ Ισχαιμικού Εγκεφαλικού και Αιμορραγικού Εγκεφαλικού.

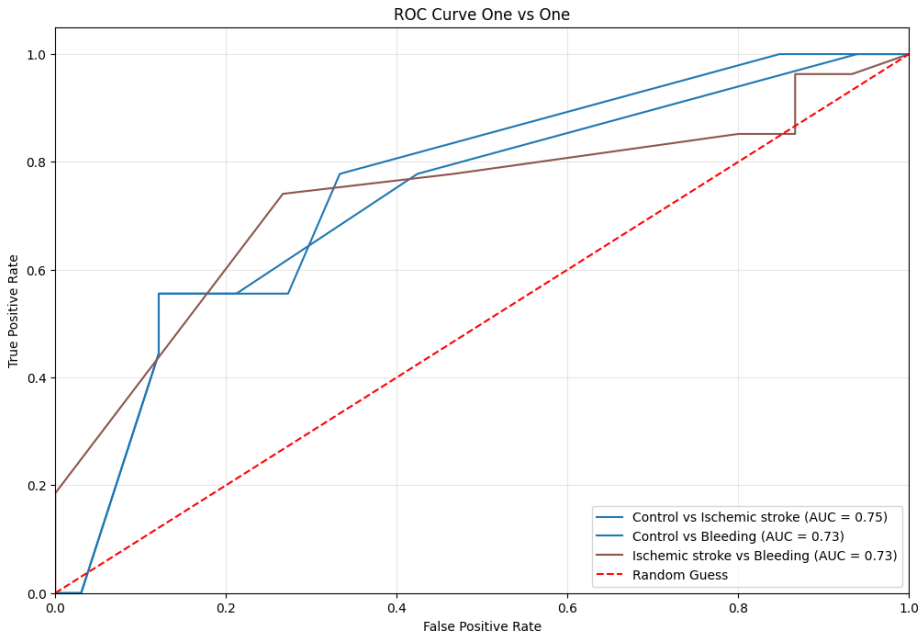
Πίνακας 4.18: Επιλεγμένα χαρακτηριστικά για το μέρος 2.1

Βιοδείκτες	**Επιλεγμένα χαρακτηριστικά**
CD45-APC+	Age, RR syst., RR diast., wbc, glucose, Intensity, Std, entropy, 500-600 nm, 900-1000 nm
CD235a-PE+	Age, RR syst., Thrombocytes, wbc, glucose, Total Concentration, Median, Std, kurtosis, 200-300 nm
CD14-PB+	Age, RR syst., RR diast., Oxygen saturation, Thrombocytes, wbc, glucose, Median, Std, skewness
CD31-APC	Age, RR syst., Thrombocytes, wbc, glucose, entropy, 100-200 nm, 500-600 nm, 600-700 nm, 800-900 nm
CD146-PE	Age, RR syst., RR diast., Thrombocytes, wbc, glucose, Median, Std, skewness, 100-200 nm
CD326-APC+	Age, RR syst., Thrombocytes, wbc, glucose, Std, skewness, entropy, 100-200 nm, 300-400 nm
Lact-FITC+	Age, RR syst., RR diast., Thrombocytes, wbc, glucose, kurtosis, 400-500 nm, 500-600 nm, 900-1000 nm

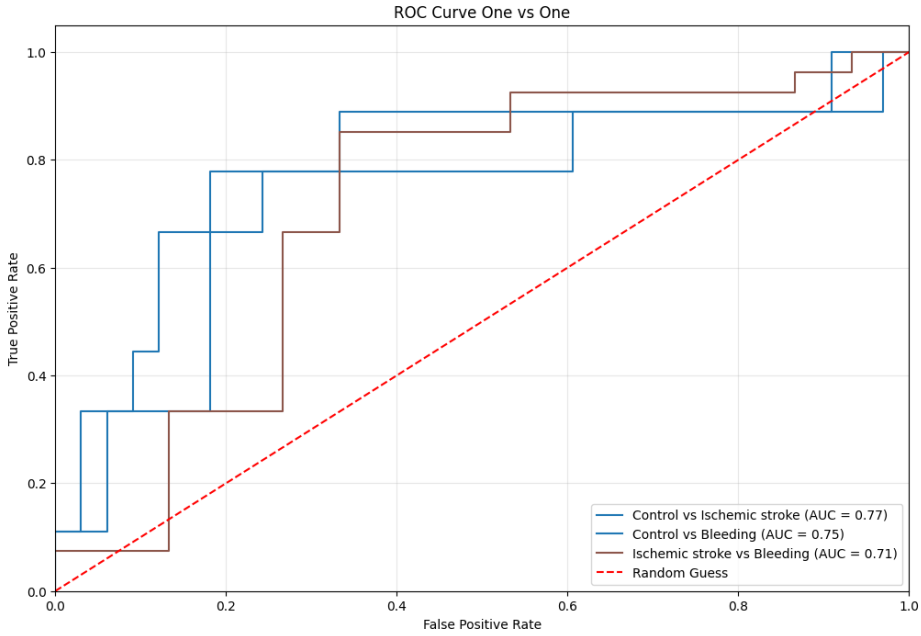
4.4.2 Αποτελέσματα Απόδοσης Μοντέλων Μηχανικής Μάθησης

Βιοδείκτες	**Μοντέλο**	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC score (Ischemic stroke vs Bleeding)**	**AUC score (Control vs Ischemic stroke)**	**AUC score (Control vs Bleeding)**
CD45-APC+	**Random Forest**	66.7%	71.1%	46.3%	48.3%	54%	81%	78%
CD235a-PE+	**Gradient Boosting**	73.8%	77.6%	56.8%	61.1%	66%	75%	78%

CD14-PB+	**Decision Tree**	64.3%	55.6%	50%	49.7%	73%	75%	73%
CD31-APC	**XG Boost**	66.7%	56.2%	48.8%	49.6%	61%	68%	76%
CD146-PE	**XG Boost**	64.3%	38%	43.2%	40.4%	64%	71%	70%
CD326-APC+	**Gradient Boosting**	69%	72.1%	47.5%	49.3%	71%	77%	75%
Lact-FITC+	**Gradient Boosting**	61.9%	67.4%	43.8%	45.5%	67%	74%	78%



Εικόνα 4.108: ROC curve one vs one - CD14-PB+ (Decision Tree)



Εικόνα 4.109: ROC curve one vs one - CD326-APC+ (Gradient Boosting)

5. Κεφάλαιο: Συμπεράσματα

5.1 Ανάλυση Αποτελεσμάτων

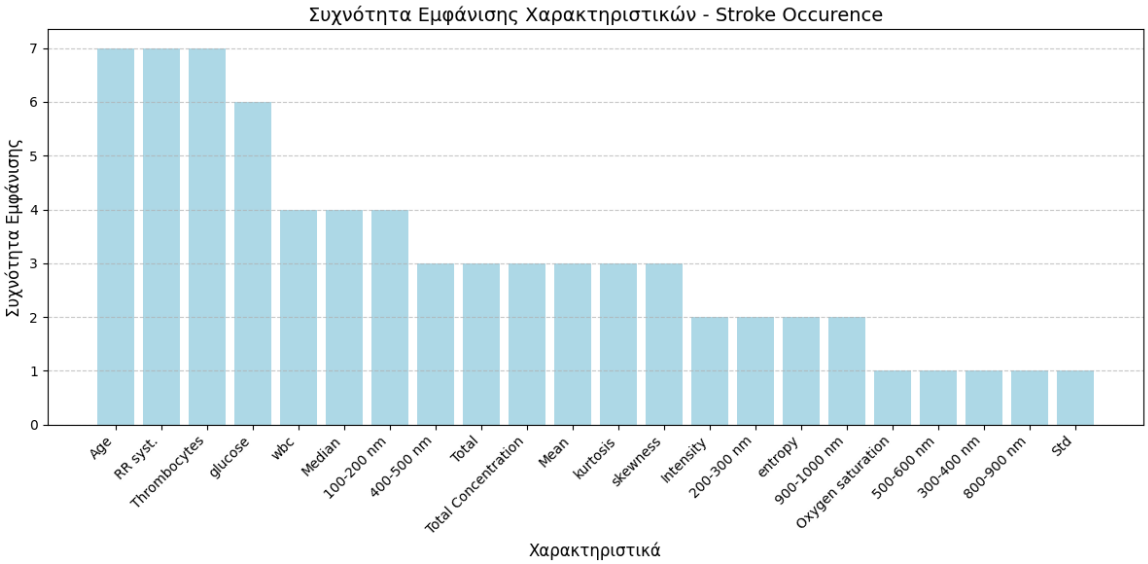
5.1.1 Συμπεράσματα από την επιλογή χαρακτηριστικών

Όσον αφορά την επιλογή των χαρακτηριστικών, τα καταλληλότερα χαρακτηριστικά που αναδείχθηκαν μέσα από τη διαδικασία ανάλυσης παρουσιάζονται συνολικά στον παρακάτω πίνακα, καλύπτοντας τα δύο πρώτα μέρη της εργασίας (Μέρος 1^{ο}: διάγνωση εγκεφαλικού – Μέρος 2^{ο}: Διάγνωση τύπου εγκεφαλικού (4 κλάσεις)).

Πίνακας 5.1: Συγκεντρωτικός πίνακας από την επιλογή χαρακτηριστικών

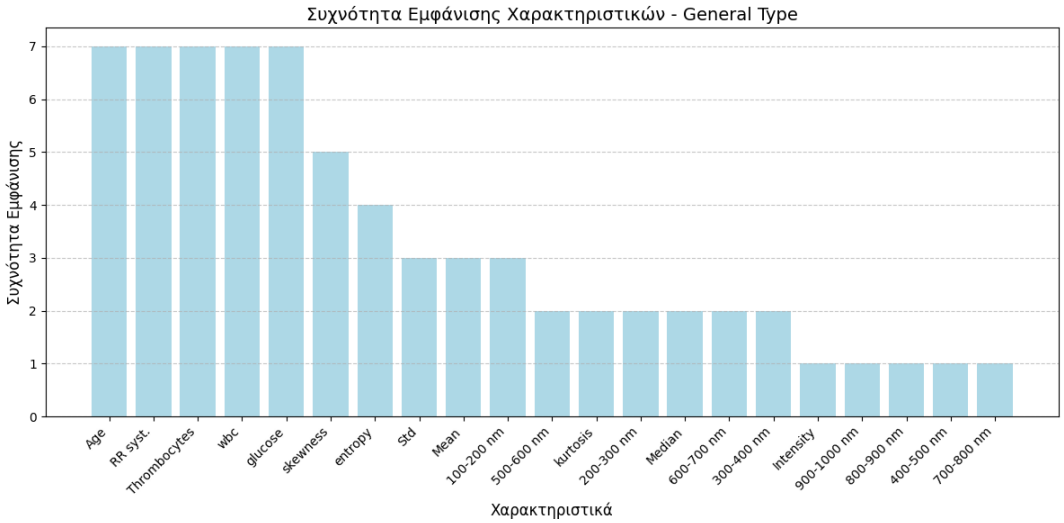
Βιοδείκτες	**Επιλεγμένα χαρακτηριστικά για το 1^{ο} μέρος (διάγνωση εγκεφαλικού)**
CD45-APC+	Age, RR syst., Oxygen saturation, Thrombocytes, wbc, glucose, Intensity, Median, 100-200 nm, 400-500 nm
CD235a-PE+	Age, RR syst., Thrombocytes, wbc, glucose, Total, Total Concentration, Mean, Median, kurtosis
CD14-PB+	Age, RR syst., Thrombocytes, Total, Mean, kurtosis, skewness, 200-300 nm, 400-500 nm, 500-600 nm
CD31-APC	Age, RR syst., Thrombocytes, wbc, glucose, Total, entropy, 100-200 nm, 300-400 nm, 800-900 nm
CD146-PE	Age, RR syst., Thrombocytes, glucose, Median, Std, kurtosis, skewness, 100-200 nm, 900-1000 nm
CD326-APC+	Age, RR syst., Thrombocytes, glucose, Total Concentration, Median, skewness, entropy, 100-200 nm, 200-300 nm
Lact-FITC+	Age, RR syst., Thrombocytes, wbc, glucose, Intensity, Total Concentration, Mean, 400-500 nm, 900-1000 nm
	Επιλεγμένα χαρακτηριστικά για το 2ο μέρος (διάγνωση τύπου εγκεφαλικού)
CD45-APC+	Age, RR syst., Thrombocytes, wbc, glucose, Intensity, Std, entropy, 500-600 nm, 900-1000 nm
CD235a-PE+	Age, RR syst., Thrombocytes, wbc, glucose, Mean, Std, kurtosis, skewness, 200-300 nm
CD14-PB+	Age, RR syst., Thrombocytes, wbc, glucose, Mean, Median, skewness, entropy, 600-700 nm
CD31-APC	Age, RR syst., Thrombocytes, wbc, glucose, entropy, 100-200 nm, 500-600 nm, 600-700 nm, 800-900 nm
CD146-PE	Age, RR syst., Thrombocytes, wbc, glucose, Median, skewness, 100-200 nm, 200-300 nm, 300-400 nm
CD326-APC+	Age, RR syst., Thrombocytes, wbc, glucose, Std, skewness, entropy, 100-200 nm, 300-400 nm
Lact-FITC+	Age, RR syst., Thrombocytes, wbc, glucose, Mean, kurtosis, skewness, 400-500 nm, 700-800 nm

Από το 1ο μέρος της εργασίας, η ανάλυση της συχνότητας εμφάνισης των χαρακτηριστικών έδειξε ότι τα χαρακτηριστικά Age, RR syst. και Thrombocytes εμφανίζονται 7 φορές, καθιστώντας τα ως τα πιο συχνά επιλεγμένα. Ακολουθεί το glucose με 6 εμφανίσεις, ενώ τα wbc και 100-200 nm εμφανίζονται 4 φορές. Τα Intensity, Median, 400-500 nm, Total, Total Concentration, Mean, kurtosis, και skewness καταγράφηκαν 3 φορές. Τέλος, τα 200-300 nm, 900-1000 nm, και entropy εμφανίζονται 2 φορές, ενώ τα Oxygen saturation, 300-400 nm, 800-900 nm, και Std καταγράφονται μόνο μία φορά.

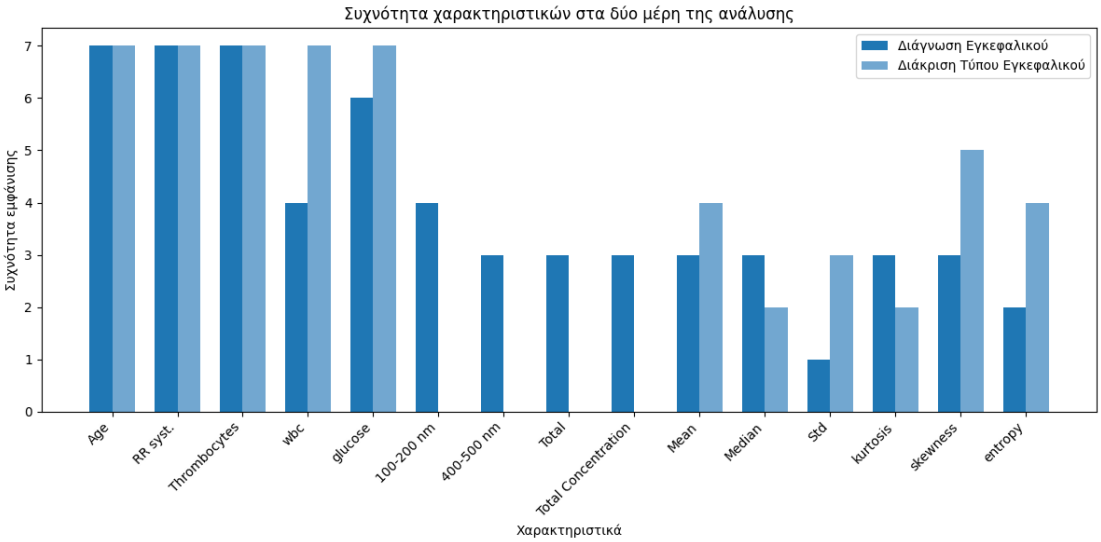


Εικόνα 5.1: Διάγραμμα συχνότητας εμφάνισης χαρακτηριστικών για το 1ο μέρος

Για το 2ο μέρος, η ανάλυση δείχνει ότι τα χαρακτηριστικά Age, RR syst., Thrombocytes, wbc, και glucose εμφανίζονται σταθερά σε όλους τους βιοδείκτες. Επιπλέον, χαρακτηριστικά όπως το skewness εμφανίζεται 5 φορές, το Mean και το entropy εμφανίζεται 4 φορές, το Std εμφανίζεται 3 φορές, το kurtosis καθώς και το Median εμφανίζονται 2 φορές. Παρατηρούνται πολλές ομοιότητες στην επιλογή χαρακτηριστικών ανάμεσα στους βιοδείκτες, κάτι το οποίο υποδηλώνει ότι υπάρχουν κοινοί βιολογικοί μηχανισμοί ή χαρακτηριστικά που είναι εξίσου σημαντικά για την ανάλυση όλων των βιοδεικτών. Παρόλα αυτά, η διαφοροποίηση σε ορισμένα χαρακτηριστικά δείχνει ότι κάθε βιοδείκτης μπορεί να προσθέτει μια μοναδική πληροφορία στην ερμηνεία της κατάστασης λόγω της εξειδίκευσης στη βιολογική του λειτουργία.



Εικόνα 5.2: Διάγραμμα συχνότητας εμφάνισης χαρακτηριστικών για το 2ο μέρος



Εικόνα 5.3: Συχνότητα εμφάνισης χαρακτηριστικών και στα 2 μέρη της ανάλυσης

Η ανάλυση της συχνότητας εμφάνισης των χαρακτηριστικών στα δύο μέρη της μελέτης δείχνει ότι ορισμένα χαρακτηριστικά είναι σταθερά σημαντικά τόσο για τη διάγνωση του εγκεφαλικού όσο και για τη διαφοροποίηση του τύπου του. Τα Age, RR syst., Thrombocytes, wbc και glucose εμφανίζονται σε όλες τις περιπτώσεις, υποδηλώνοντας ότι είναι βασικά χαρακτηριστικά τα οποία σχετίζονται άμεσα με την εκδήλωση του εγκεφαλικού επεισοδίου. Στη διάγνωση του εγκεφαλικού, κρίσιμο ρόλο φαίνεται να διαδραματίζουν τα φασματικά χαρακτηριστικά και τα δεδομένα των εξωκυτταρικών κυστιδίων, όπως το 100-200 nm, 400-500 nm, Total και Total Concentration, τα οποία σχετίζονται με τη συγκέντρωση και τις ιδιότητες των EVs. Αντίθετα, για τη διαφοροποίηση του τύπου του εγκεφαλικού (ισχαιμικό ή αιμορραγικό), μεγαλύτερη σημασία έχουν τα στατιστικά χαρακτηριστικά της κατανομής των δεδομένων, όπως το skewness, entropy, Std και Mean, γεγονός που δείχνει ότι η μορφολογία και η κατανομή των EVs διαφοροποιούνται ανάλογα με τον τύπο του εγκεφαλικού επεισοδίου.

Η παραπάνω επιλογή των χαρακτηριστικών μπορεί να εξηγηθεί από ιατρικές μελέτες όπως αναφέρεται και στο θεωρητικό κομμάτι της εργασίας. Αρχικά, η ηλικία αποτελεί έναν από τους πιο καθοριστικούς παράγοντες κινδύνου για το εγκεφαλικό επεισόδιο. Καθώς τα αγγεία «γερνούν», γίνονται λιγότερο ελαστικά και πιο επιρρεπή σε αθηροσκλήρωση, αυξάνοντας τον κίνδυνο τόσο για ισχαιμικά όσο και για αιμορραγικά εγκεφαλικά. Το RR syst. (συστολική πίεση) είναι ένας από τους πιο κρίσιμους δείκτες καρδιαγγειακής υγείας, ειδικά για ισχαιμικά εγκεφαλικά επεισόδια. Η υπέρταση μπορεί να προκαλέσει βλάβες στα αιμοφόρα αγγεία του εγκεφάλου, καθιστώντας τα πιο ευάλωτα σε ρήξη ή θρόισμα.

Επιπλέον, τα αιμοπετάλια (Thrombocytes) είναι υπεύθυνα για την πήξη του αίματος και τη δημιουργία θρόμβων. Υψηλά επίπεδα αιμοπεταλίων μπορούν να συμβάλλουν στη δημιουργία θρόμβου, το οποίο μπορεί να προκαλέσει ισχαιμικό εγκεφαλικό επεισόδιο. Αντίθετα, πολύ χαμηλά επίπεδα αιμοπεταλίων μπορεί να σχετίζονται με αιμορραγικό εγκεφαλικό επεισόδιο.

Όσον αφορά, τα λευκά αιμοσφαίρια (wbc) είναι δείκτες φλεγμονής και η αύξησή τους έχει συσχετιστεί με τη σοβαρότητα του εγκεφαλικού επεισοδίου. Η φλεγμονώδης απόκριση που προκαλείται μετά από ένα εγκεφαλικό μπορεί να επηρεάσει τη διαδικασία ανάρρωσης, ενώ υψηλά WBC σχετίζονται με αυξημένο κίνδυνο εγκεφαλοαγγειακής νόσου.

Παράλληλα, η υψηλή γλυκόζη στο αίμα (συνήθως σε περιπτώσεις διαβήτη) αυξάνει τον κίνδυνο εγκεφαλικού επεισοδίου, καθώς μπορεί να προκαλέσει βλάβες στα αιμοφόρα αγγεία και να αυξήσει την πιθανότητα θρόισμα και αθηρωμάτωσης.

Τα φασματικά χαρακτηριστικά των EVs (100-200 nm, 400-500 nm) σχετίζονται με την απορρόφηση και εκπομπή φωτός από εξωκυτταρικά κυστίδια, τα οποία μπορεί να παρέχουν ενδείξεις για την ενδοθηλιακή βλάβη, τη φλεγμονή και τη θρόμβωση. Οι αλλαγές σε αυτά τα μήκη κύματος μπορεί να αντικατοπτρίζουν βιολογικές μεταβολές που σχετίζονται με την παθοφυσιολογία του εγκεφαλικού επεισοδίου.

Η συγκέντρωση των EVs (Total & Total Concentration) είναι σημαντική, καθώς αυξημένος αριθμός εξωκυτταρικών κυστιδίων μπορεί να αποτελεί ένδειξη εγκεφαλικής βλάβης. Τα EVs εμπλέκονται σε μηχανισμούς όπως η φλεγμονή, η πήξη του αίματος και η νευρωνική επικοινωνία, καθιστώντας τα κρίσιμα στη διάγνωση του εγκεφαλικού. Παράλληλα, τα στατιστικά χαρακτηριστικά της κατανομής των EVs (Mean, Median, Std, Kurtosis, Skewness, Entropy) παίζουν σημαντικό ρόλο στην ταξινόμηση του τύπου εγκεφαλικού επεισοδίου. Η τυπική απόκλιση (Std) υποδηλώνει την ποικιλομορφία της κατανομής των EVs, ενώ οι δείκτες ασυμμετρίας (Skewness) και κυρτότητας (Kurtosis) μπορούν να αποκαλύψουν αν η κατανομή έχει περισσότερες ακραίες τιμές, οι οποίες σχετίζονται με παθολογικές καταστάσεις. Η εντροπία (Entropy), ως μέτρο της ποικιλομορφίας του δείγματος, μπορεί να διακρίνει διαφορετικούς τύπους εγκεφαλικών επεισοδίων.

5.2.2 Συμπεράσματα για το 1ο Μέρος

Για το 1ο μέρος της ανάλυσης, αφού ολοκληρώθηκε η επιλογή των χαρακτηριστικών, αξιολογήθηκαν διάφορα μοντέλα για κάθε βιοδείκτη. Στον παρακάτω πίνακα, απεικονίζονται τα δύο πιο αποδοτικά μοντέλα για κάθε περίπτωση, βασισμένα στην απόδοσή τους σε διάφορους δείκτες. Ειδικότερα, Random Forest φαίνεται να είναι το πιο συχνά εμφανιζόμενο μοντέλο σε αυτή την ανάλυση (σε 6 από τις 7 περιπτώσεις), ακολουθούμενο από το Gradient Boosting και το XG Boost, τα οποία εμφανίζονται ως τα βέλτιστα σε 3 περιπτώσεις. Παράλληλα, τα μοντέλα KNN και Naïve Bayes εμφανίζονται από 1 φορά.

Η συχνότητα εμφάνισης των συγκεκριμένων μοντέλων ως τα πιο αποδοτικά μπορεί να εξηγηθεί βάσει των παρακάτω χαρακτηριστικών τους. Συγκεκριμένα, το Random Forest είναι συνήθως πολύ ανθεκτικό σε μικρότερα σύνολα δεδομένων, καθώς δημιουργεί πλήθος δέντρων από τυχαία δείγματα και με τυχαία χαρακτηριστικά, μειώνοντας έτσι την υπερπροσαρμογή. Από την άλλη, το Gradient Boosting, ως τεχνική ενίσχυσης, επικεντρώνεται στην επίλυση των λαθών των προηγούμενων μοντέλων. Αυτή η προσέγγιση έχει αποδειχθεί πολύ αποδοτική σε προβλήματα όπου τα δεδομένα είναι περιορισμένα, καθώς ενσωματώνει τη δυνατότητα να προσαρμόσει τα μοντέλα για να καλύψει τα λάθη τους. Παράλληλα, η συνεχής εκπαίδευση των μοντέλων στο Gradient Boosting επιτρέπει καλύτερη απόδοση, ακόμα και όταν οι πόροι είναι περιορισμένοι, κάτι που είναι ιδιαίτερα χρήσιμο σε μικρά ιατρικά σύνολα δεδομένων.

Αντίστοιχα, το XG Boost συνδυάζει τα πλεονεκτήματα του Gradient Boosting με τη δυνατότητα κανονικοποίησης και προσαρμογής παραμέτρων για να ελαχιστοποιήσει την υπερπροσαρμογή και να μεγιστοποιήσει την απόδοση.

Συνολικά, τα αποτελέσματα της μελέτης έδειξαν ότι η χρήση αλγορίθμων Μηχανικής Μάθησης μπορεί να συμβάλει σημαντικά στη διάγνωση εγκεφαλικών επεισοδίων, προσφέροντας υψηλή ακρίβεια και αξιοπιστία σε σύγκριση με τις παραδοσιακές μεθόδους. Οι αλγόριθμοι Gradient Boosting, XGBoost και Random Forest εμφάνισαν την καλύτερη απόδοση, αποδεικνύοντας ότι οι μέθοδοι βασισμένες σε δέντρα απόφασης είναι ιδιαίτερα αποτελεσματικές στη συγκεκριμένη εφαρμογή.

Πίνακας 5.2: Συνολικός πίνακας για τα 2 πιο αποδοτικά μοντέλα για κάθε βιοδείκτη για το 1ο μέρος

Βιοδείκτες	**2 πιο αποδοτικά μοντέλα για κάθε βιοδείκτη για το 1ο μέρος**					
		Accuracy	**Precision**	**Recall**	**F1-score**	**AUC score**
CD45-APC+	**Gradient Boosting**	78.6%	85.3%	87.9%	86.6%	74%
	Random Forest	78.7%	86.3%	87.9%	86.6%	81%
CD235a-PE+	**Random Forest**	88.1%	91.2%	93.9%	92.5%	82%

	XG Boost	88.1%	88.9%	97%	92.8%	80%
CD14-PB+	**Random Forest**	85.7%	86.5%	97%	91.4%	74%
	Naïve Bayes	85.7%	88.6%	93.9%	91.2%	84%
CD31-APC	**Gradient Boosting**	81%	85.7%	90.9%	88.2%	75%
	XG Boost	81%	83.8%	93.9%	88.6%	72%
CD146-PE	**Gradient Boosting**	76.2%	81.1%	90.9%	85.7%	60%
	Random Forest	78.6%	83.3%	90.9%	87%	66%
CD326-APC+	**Random Forest**	73.8%	80.6%	87.9%	84.1%	73%
	KNN	78.6%	78.6%	100%	88%	72%
Lact-FITC+	**Random Forest**	78.6%	85.3%	87.9%	86.6%	69%
	XG Boost	76.8%	81.6%	93.9%	87.3%	63%

Με βάση τα παραπάνω αποτελέσματα, παρατηρείται ότι ο βιοδείκτης CD235a-PE+ είναι ο πιο αποδοτικός βιοδείκτης, με το μοντέλο XG Boost να επιτυγχάνει Recall 97% και F1-score 92.8%, ενώ το Random Forest εμφανίζει Precision 91.2% και συνολική ακρίβεια 88.1%, γεγονός που υποδηλώνει υψηλή αξιοπιστία στην ανίχνευση του εγκεφαλικού. Πολύ καλά αποτελέσματα παρατηρούνται επίσης για τους CD14-PB+ και CD31-APC, οι οποίοι εμφανίζουν Accuracy άνω του 85%, καθώς και υψηλές τιμές Recall και F1-score. Αντίθετα, ο CD326-APC+ εμφανίζει χαμηλότερη απόδοση, καθώς η ακρίβειά του κυμαίνεται μεταξύ 73.8%-78.6%, αν και το μοντέλο KNN πέτυχε Recall 100%, γεγονός που ενδέχεται να οφείλεται σε overfitting. Συνολικά, τα αποτελέσματα υποδηλώνουν ότι οι βιοδείκτες CD235a-PE+, CD14-PB+ και CD31-APC είναι οι πιο αξιόπιστοι για τη διάγνωση του εγκεφαλικού, ενώ η επιλογή του κατάλληλου αλγορίθμου ταξινόμησης μπορεί να βελτιώσει σημαντικά την ακρίβεια των προβλέψεων.

Η υψηλή αποδοτικότητα των βιοδεικτών CD235a-PE+, CD14-PB+ και CD31-APC στη διάγνωση του εγκεφαλικού εξηγείται από τη στενή τους σχέση με βασικούς παθοφυσιολογικούς μηχανισμούς που εμπλέκονται στην εμφάνιση του εγκεφαλικού επεισοδίου, όπως αναφέρεται και στο θεωρητικό μέρος της παρούσας εργασίας. Συγκεκριμένα, ο CD235a-PE+, ως δείκτης των ερυθρών αιμοσφαιρίων και των EVs που προέρχονται από αυτά, σχετίζεται με μικροαγγειακές βλάβες και ενδοθηλιακή δυσλειτουργία, οι οποίες μπορούν να διαταράξουν την αιματική ροή και να αυξήσουν τον κίνδυνο ισχαιμικού εγκεφαλικού. Ο CD14-PB+, που εκφράζεται στα μονοκύτταρα και τα μακροφάγα, συνδέεται με φλεγμονώδεις αποκρίσεις, οι οποίες είναι κρίσιμες τόσο στην πρόκληση όσο και στην εξέλιξη του εγκεφαλικού, καθώς η αυξημένη παραγωγή EVs με CD14+ υποδηλώνει αγγειακή φλεγμονή και δυσλειτουργία του αιματοεγκεφαλικού φραγμού. Παράλληλα, ο CD31-APC, ως δείκτης ενδοθηλιακών κυττάρων και αιμοπεταλίων, αντικατοπτρίζει διαταραχές στην αγγειακή λειτουργία και την υπερπηκτικότητα, παράγοντες που συμβάλλουν σημαντικά στην εμφάνιση εγκεφαλικών θρομβώσεων.

Αξίζει να σημειωθεί ότι οι τρεις αυτοί βιοδείκτες μοιράζονται 6 από τα 10 επιλεγμένα χαρακτηριστικά. Η κοινή παρουσία των Age, RR syst., Thrombocytes, Total, Mean και

Kurtosis υποδεικνύει ότι αυτοί οι δείκτες πιθανώς αντικατοπτρίζουν σημαντικές αλλαγές στη μικροκυκλοφορία, την πήξη του αίματος και τη φλεγμονώδη απόκριση, μηχανισμούς που είναι κεντρικοί στην παθοφυσιολογία του εγκεφαλικού. Επομένως, η επαναλαμβανόμενη παρουσία αυτών των χαρακτηριστικών στους πιο αποδοτικούς βιοδείκτες υποδηλώνει ότι η ανάλυση συγκεκριμένων αιματολογικών, αγγειακών και φλεγμονωδών παραμέτρων, σε συνδυασμό με τα δεδομένα των EVs, μπορεί να προσφέρει έναν αξιόπιστο τρόπο διάγνωσης του εγκεφαλικού. Συνολικά, τα ευρήματα αυτά ενισχύουν την υπόθεση ότι η χρήση αυτών των βιοδεικτών σε συνδυασμό με προχωρημένους αλγορίθμους μηχανικής μάθησης μπορεί να οδηγήσει σε βελτιωμένες στρατηγικές ανίχνευσης του εγκεφαλικού και πιθανώς στην ανάπτυξη προγνωστικών μοντέλων για έγκαιρη παρέμβαση.

5.2.3 Συμπεράσματα για το 2ο Μέρος

Στο 2ο μέρος της εργασίας, εξετάζεται η απόδοση των μοντέλων ως προς τη ικανότητά τους να ταξινομούν με ακρίβεια τους ασθενείς στις αντίστοιχες κατηγορίες εγκεφαλικών επεισοδίων. Οι κατηγορίες αυτές περιλαμβάνουν: Control, Vessel Occlusion, Bleeding, και LVO οι οποίες έχουν κωδικοποιηθεί με αριθμούς από το 0 έως το 3, αντίστοιχα. Στη συνέχεια, τα δύο πιο αποδοτικά μοντέλα, όπως αυτά αναδείχθηκαν στο πρώτο μέρος της ανάλυσης, επιλέχθηκαν για την αξιολόγηση της απόδοσής τους στο συγκεκριμένο πρόβλημα. Η επιλογή αυτών των μοντέλων βασίστηκε στη λογική ότι η αποδοτικότητά τους στις αρχικές κλάσεις (στο πρώτο μέρος της ταξινόμησης) τα καθιστά αξιόπιστες και αποτελεσματικές επιλογές. Συγκεκριμένα, κρίθηκαν τα καταλληλότερα για να χρησιμοποιηθούν ως βασικά μοντέλα σε μια πιο σύνθετη ταξινόμηση με περισσότερες κλάσεις, εξασφαλίζοντας έτσι τη συνέπεια και την αποδοτικότητα στη διαδικασία ανάλυσης. Στον παρακάτω μοντέλα, εμφανίζονται τα πιο αποδοτικά μοντέλα για κάθε έναν από τους βιοδείκτες.

Πίνακας 5.3: Απόδοση των μοντέλων για το μέρος 2.1

Βιοδείκτες	**Μοντέλα**	**Accuracy**	**Precision**	**Recall**	**F1-score**
CD45-APC+	Gradient Boosting	52.40%	62.90%	40.90%	40.60%
CD235a-PE+	XG Boost	50%	56.6%	41.1%	40.8%
CD14-PB+	Naïve Bayes	40.5%	37.2%	34.3%	34.1%
CD31-APC	XG Boost	50%	46.9%	41.45	41.7%
CD146-PE	Gradient Boosting	47.6%	48.4%	36.5%	34.2%
CD326-APC+	Random Forest	50%	58.3%	38.1%	38.2%
Lact-FITC+	XG Boost	45.2%	43.9%	37.3%	38.1%

Τα αποτελέσματα του 2ου μέρους της ανάλυσης δείχνουν ότι τα επιλεγμένα μοντέλα παρουσιάζουν χαμηλές επιδόσεις σε όλους τους βιοδείκτες. Συγκεκριμένα, τα ποσοστά Accuracy κυμαίνονται από 40.5% έως 52.4%, γεγονός που υποδηλώνει ότι τα μοντέλα δυσκολεύονται να διαχωρίσουν με ακρίβεια τους διαφορετικούς τύπους εγκεφαλικού επεισοδίου (Control, Vessel Occlusion, Bleeding, LVO).

Επίσης, παρατηρείται ότι ο CD45-APC+ με το Gradient Boosting πέτυχε το υψηλότερο Accuracy (52.4%), αλλά το Recall του είναι μόλις 40.9%, υποδηλώνοντας ότι πολλά δείγματα ταξινομούνται λανθασμένα. Παρομοίως, ο CD235a-PE+ με το XG Boost εμφάνισε Accuracy 50%, με Recall 41.1%, ενώ ο CD31-APC είχε αντίστοιχα 50% Accuracy και Recall 41.45%. Αυτό σημαίνει ότι η ικανότητα ανίχνευσης των διαφόρων τύπων εγκεφαλικού είναι περιορισμένη, κάτι που μπορεί να οφείλεται είτε σε έλλειψη διακριτών προτύπων μεταξύ των κατηγοριών είτε σε αλληλεπικάλυψη χαρακτηριστικών μεταξύ των διαφορετικών τύπων εγκεφαλικού. Τέλος, η απόδοση των βιοδεικτών είναι σημαντικά χαμηλότερη σε σχέση με το 1ο μέρος.

5.2.3 Συμπεράσματα για το 3ο Μέρος

Στο 1ο μέρος της εργασίας, επιλέξαμε τα μοντέλα που παρουσίασαν την καλύτερη ακρίβεια για το 1ο στάδιο ταξινόμησης. Αυτός ο στόχος καθορίστηκε βάσει της ανάγκης για την πιο αξιόπιστη ταξινόμηση των δεδομένων, επιλέγοντας τα μοντέλα που φάνηκαν να αποδίδουν καλύτερα στην αρχική φάση της εργασίας. Ωστόσο, πρέπει να σημειωθεί ότι η απόδοση του μοντέλου επηρεάστηκε σημαντικά από τον περιορισμένο αριθμό των δεδομένων. Το σύνολο δεδομένων ήταν πολύ μικρό, γεγονός που είχε αντίκτυπο στα αποτελέσματα των ταξινομητών. Με λίγα δεδομένα, τα μοντέλα αντιμετωπίζουν δυσκολίες στην εκμάθηση των υποκείμενων σχέσεων ανάμεσα στα χαρακτηριστικά και τις κατηγορίες, με αποτέλεσμα οι προβλέψεις να είναι λιγότερο ακριβείς και να καταγράφονται υψηλότερες τιμές λάθους.

Ο περιορισμένος αριθμός δειγμάτων επηρεάζει επίσης την ικανότητα των μοντέλων να αναγνωρίσουν τις λιγότερο εκπροσωπούμενες κατηγορίες, κάτι που ενδέχεται να είναι υπεύθυνο για τις χαμηλότερες επιδόσεις σε μετρικές όπως η precision, recall και F1-score για ορισμένες κατηγορίες.

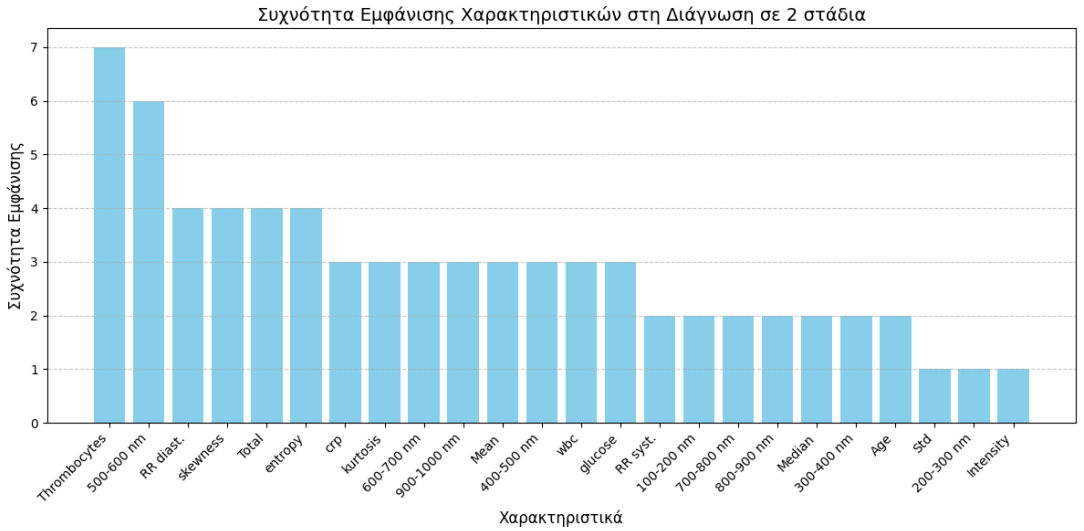
Συγκεκριμένα, τα αποτελέσματα της παρούσας ανάλυσης υποδεικνύουν ότι τα χρησιμοποιούμενα μοντέλα μηχανικής μάθησης εμφανίζουν γενικά χαμηλή απόδοση, με τον βιοδείκτη CD235a-PE+ να παρουσιάζει την υψηλότερη ακρίβεια (73%), αν και με μέτρια τιμή Precision (57%) και Recall (50%). Οι υπόλοιποι βιοδείκτες εμφανίζουν σημαντικά χαμηλότερες επιδόσεις, με ορισμένους, όπως CD326-APC+ (Accuracy 37%) και Lact-FITC+ (Accuracy 40%), να καταγράφουν ιδιαίτερα χαμηλές τιμές Precision και F1-score, γεγονός που υποδηλώνει αδυναμία του μοντέλου να διακρίνει επαρκώς τις θετικές περιπτώσεις.

Όπως προαναφέρθηκε, στο δεύτερο μέρος του κεφαλαίου 4.3, η μεθοδολογία του Two-Stage Hybrid Classifier παραμένει η ίδια, ωστόσο, παρουσιάζεται μια σημαντική τροποποίηση στην προσέγγιση λόγω της εφαρμογής μεθόδων επιλογής χαρακτηριστικών. Πριν από την έναρξη του 1ου σταδίου, πραγματοποιείται μια διαδικασία επιλογής χαρακτηριστικών, με στόχο τον εντοπισμό των πλέον σημαντικών μεταβλητών που διαφοροποιούν τους ασθενείς με εγκεφαλικό επεισόδιο από τα υγιή άτομα. Στη συνέχεια, κατά το 2ο στάδιο, εφαρμόζεται εκ νέου διαδικασία επιλογής χαρακτηριστικών, προκειμένου να προσδιοριστούν οι βέλτιστες μεταβλητές που θα συμβάλουν στην ακριβή διάγνωση του τύπου του εγκεφαλικού επεισοδίου.

Πίνακας 5.4: Σύγκριση μεθόδων (απευθείας διάγνωση - διάγνωση σε 2 στάδια)

Βιοδείκτες	**Επιλεγμένα χαρακτηριστικά για το 2ο μέρος (διάγνωση τύπου εγκεφαλικού)**
CD45-APC+	Age, RR syst., Thrombocytes, wbc, glucose, Intensity, Std, entropy, 500-600 nm, 900-1000 nm
CD235a-PE+	Age, RR syst., Thrombocytes, wbc, glucose, Mean, Std, kurtosis, skewness, 200-300 nm
CD14-PB+	Age, RR syst., Thrombocytes, wbc, glucose, Mean, Median, skewness, entropy, 600-700 nm
CD31-APC	Age, RR syst., Thrombocytes, wbc, glucose, entropy, 100-200 nm, 500-600 nm, 600-700 nm, 800-900 nm
CD146-PE	Age, RR syst., Thrombocytes, wbc, glucose, Median, skewness, 100-200 nm, 200-300 nm, 300-400 nm
CD326-APC+	Age, RR syst., Thrombocytes, wbc, glucose, Std, skewness, entropy, 100-200 nm, 300-400 nm
Lact-FITC+	Age, RR syst., Thrombocytes, wbc, glucose, Mean, kurtosis, skewness, 400-500 nm, 700-800 nm

	Επιλεγμένα χαρακτηριστικά για το 3ο μέρος (διάγνωση τύπου εγκεφαλικού με διπλή διάγνωση)
CD45-APC+	RR syst., RR diast., Thrombocytes, crp, kurtosis, 100-200 nm, 600-700 nm, 700-800 nm, 800-900 nm, 900-1000 nm
CD235a-PE+	RR diast., Thrombocytes, Mean, Median, Std, skewness, 100-200 nm, 300-400 nm, 400-500 nm, 500-600 nm
CD14-PB+	Age, RR diast., Thrombocytes, wbc, crp, Total, skewness, entropy, 500-600 nm, 900-1000 nm
CD31-APC	Thrombocytes, wbc, crp, Mean, kurtosis, skewness, 400-500 nm, 500-600 nm, 600-700 nm, 800-900 nm
CD146-PE	RR diast., Thrombocytes, glucose, Total, Median, entropy, 200-300 nm, 500-600 nm, 700-800 nm, 900-1000 nm
CD326-APC+	Thrombocytes, wbc, glucose, Total, Intensity, entropy, 300-400 nm, 400-500 nm, 500-600 nm, 600-700 nm
Lact-FITC+	Age, RR syst., Thrombocytes, glucose, Total, Mean, kurtosis, skewness, entropy, 500-600 nm



Εικόνα 5.4: Συχνότητα εμφάνισης χαρακτηριστικών για την διάγνωση σε 2 στάδια

Από την ανάλυση της συχνότητας εμφάνισης των χαρακτηριστικών στη διάγνωση σε 2 στάδια, παρατηρείται ότι τα Thrombocytes (αιμοπετάλια), το φασματικό εύρος 500-600 nm και η διαστολική πίεση (RR diastolic) είναι τα πιο συχνά χαρακτηριστικά, γεγονός που υποδηλώνει τη διαγνωστική τους σημασία στη διαφοροποίηση των τύπων εγκεφαλικού επεισοδίου. Ακολουθούν μεσαίας συχνότητας χαρακτηριστικά, όπως CRP (C-reactive protein), entropy, skewness και διάφορες φασματικές περιοχές (600-700 nm, 900-1000 nm, 400-500 nm). Τα χαρακτηριστικά με τη χαμηλότερη συχνότητα εμφάνισης, όπως Age, Std, Intensity και 200-300 nm, εμφανίζονται μόνο 1-3 φορές, γεγονός που υποδηλώνει ότι η διαγνωστική τους αξία είναι λιγότερο καθοριστική. Συνολικά, η παρουσία πολλαπλών φασματικών χαρακτηριστικών σε υψηλή συχνότητα υποδηλώνει τη σημασία της φασματικής ανάλυσης στη διαφοροποίηση των τύπων εγκεφαλικού, ενώ η επαναλαμβανόμενη εμφάνιση των Thrombocytes και CRP ενισχύει τη σημασία αιματολογικών και φλεγμονωδών παραμέτρων στη διάγνωση.

Στον παραπάνω πίνακα, αποτυπώνονται τα επιλεγμένα χαρακτηριστικά για τις 2 μεθόδους διάγνωσης τύπου εγκεφαλικού, της άμεσης διάκρισης του τύπου εγκεφαλικού επεισοδίου και της διπλής διάγνωσης μέσω 2-stage hybrid classification, Η σύγκριση των δύο

μεθόδων ταξινόμησης αναδεικνύει σημαντικές διαφορές ως προς την επιλογή χαρακτηριστικών. Στην απευθείας ταξινόμηση, τα επιλεγμένα χαρακτηριστικά είναι γενικότερα, με κύριες μεταβλητές τις Age, RR systolic, Thrombocytes, wbc και glucose, οι οποίες φαίνεται να έχουν κεντρικό ρόλο στη διάκριση μεταξύ των τύπων εγκεφαλικού. Αντιθέτως, στη διπλή διάγνωση, εντοπίζονται επιπλέον σημαντικοί δείκτες, όπως οι RR diastolic, CRP, Total και Intensity, οι οποίοι υποδηλώνουν ότι μια πιο εξειδικευμένη προσέγγιση μπορεί να βελτιώσει την ακριβέστερη ταξινόμηση. Επιπλέον, η ανάλυση φασματικών δεδομένων (nm ranges) παρουσιάζει μεγαλύτερη ποικιλία στη δεύτερη μέθοδο, γεγονός που ενισχύει την υπόθεση ότι η υβριδική προσέγγιση μπορεί να ανιχνεύσει επιπλέον διακριτικά στοιχεία μεταξύ των τύπων εγκεφαλικού επεισοδίου.

Η ενίσχυση της παρουσίας φασματικών χαρακτηριστικών και η προσθήκη νέων κλινικών δεικτών στη 2-stage hybrid classification υποδηλώνει ότι η πολυεπίπεδη ανάλυση των δεδομένων επιτρέπει την ανίχνευση περισσότερων διαγνωστικών μοτίβων. Αυτό μπορεί να συμβάλει στη βελτίωση της απόδοσης του ταξινομητή, ειδικά σε περιπτώσεις όπου τα χαρακτηριστικά των διαφορετικών τύπων εγκεφαλικού παρουσιάζουν μικρές διαφοροποιήσεις. Η αξιοποίηση των επιπλέον χαρακτηριστικών που προκύπτουν από τη δεύτερη μέθοδο ενδέχεται να ενισχύσει την ακρίβεια της αρχικής ταξινόμησης. Μελλοντικές βελτιώσεις θα μπορούσαν να περιλαμβάνουν βελτιστοποίηση της επιλογής χαρακτηριστικών (feature selection) μέσω ensemble models, προκειμένου να διερευνηθεί η συνεισφορά των νέων μεταβλητών στη συνολική απόδοση του ταξινομητή.

Όσον αφορά την απόδοση των μοντέλων, η ακρίβεια κυμαίνεται από 36% έως 59%, με τον βιοδείκτη CD235a-PE+ να παρουσιάζει τη μεγαλύτερη ακρίβεια (59%), ενώ οι υπόλοιποι βιοδείκτες έχουν σχετικά χαμηλότερες επιδόσεις. Η τιμή της precision παραμένει χαμηλή για τους περισσότερους βιοδείκτες, με εξαιρέσεις όπως ο βιοδείκτης Lact-FITC+, ο οποίος παρουσιάζει precision 48%. Παράλληλα, η recall και η F1-score είναι επίσης χαμηλές για τους περισσότερους βιοδείκτες, επιβεβαιώνοντας ότι το μοντέλο δεν καταφέρνει να εντοπίσει σωστά τις θετικές περιπτώσεις και, συνεπώς, η απόδοσή του στη διάκριση των διαγνωστικών κατηγοριών είναι περιορισμένη. Αν και το μοντέλο εντοπίζει κάποιες θετικές περιπτώσεις, κυρίως στον βιοδείκτη Lact-FITC+, η απόδοση παραμένει κάτω από το επιθυμητό επίπεδο, καθώς η F1-score δεν υπερβαίνει το 50% σε κάποια κατηγορία.

Ωστόσο, παρατηρείται ότι όταν εφαρμόζεται η επιλογή χαρακτηριστικών και στο 2ο στάδιο, τα αποτελέσματα εμφανίζουν βελτίωση. Η χρήση της επιλογής χαρακτηριστικών στο δεύτερο στάδιο επιτρέπει στο μοντέλο να εστιάσει στα πιο διακριτικά χαρακτηριστικά για τη διάγνωση του τύπου του εγκεφαλικού επεισοδίου, γεγονός που ενδέχεται να βελτιώνει την ακρίβεια και τις υπόλοιπες μετρικές απόδοση σε σχέση με το 1ο στάδιο.

Αυτή η βελτίωση προκύπτει επειδή τα χαρακτηριστικά που είναι πιο σημαντικά για την αναγνώριση της παρουσίας εγκεφαλικού επεισοδίου μπορεί να διαφέρουν από εκείνα που είναι πιο χρήσιμα για την ταξινόμηση του τύπου εγκεφαλικού. Έτσι, η εφαρμογή της επιλογής χαρακτηριστικών και στα δύο στάδια ενισχύει τη διακριτικότητα του μοντέλου, επιτρέποντας του να επιλέξει πιο κατάλληλα χαρακτηριστικά για κάθε στάδιο και να βελτιώσει τη διάκριση μεταξύ των κατηγοριών.

Τέλος, είναι σημαντικό να σημειωθεί ότι το συγκεκριμένο κομμάτι της έρευνας παραμένει σε ερευνητικό επίπεδο. Ως εκ τούτου, υπάρχει δυνατότητα για περαιτέρω επέκταση της μεθόδου και για τη χρήση περισσότερων δεδομένων, προκειμένου να βελτιωθούν οι επιδόσεις και να επιτευχθεί μια πιο αξιόπιστη και ακριβής διάγνωση.

5.2.4 Συμπεράσματα για το 4ο Μέρος

Η μετάβαση από το 2ο στο 4ο μέρος της εργασίας βασίζεται στην ανάλυση των καμπυλών ROC one-vs-one, οι οποίες επιτρέπουν τη σύγκριση κάθε δυνατού ζεύγους κατηγοριών, παρέχοντας λεπτομερή εικόνα για την ικανότητα των μοντέλων να διακρίνουν μεταξύ συγκεκριμένων κατηγοριών. Από την εξέταση των αποτελεσμάτων, διαπιστώθηκε ότι ορισμένες κατηγορίες διαχωρίζονται ευκολότερα σε σχέση με άλλες, γεγονός που ανέδειξε την ανάγκη επαναπροσδιορισμού της ταξινόμησης. Επομένως, η συγκεκριμένη παρατήρηση

οδηγεί στην εξέταση της συγχώνευσης των κατηγοριών Vessel Occlusion και LVO σε μία ενιαία κατηγορία Ισχαιμικού Εγκεφαλικού, απόφαση που παρουσιάζει ενδιαφέρον τόσο σε κλινικό όσο και σε μεθοδολογικό επίπεδο.

Από κλινικής σκοπιάς, η συγχώνευση αυτή παρουσιάζει ιδιαίτερο ενδιαφέρον για τη μελέτη της συσχέτισης της φυσιολογίας αυτών των καταστάσεων με EVs, καθώς τόσο οι Vessel Occlusion όσο και τα LVO μοιράζονται έναν κοινό παθοφυσιολογικό μηχανισμό που αφορά τη διακοπή της ροής του αίματος στον εγκέφαλο, λόγω θρόισμα ή εμβολής. Αυτές οι καταστάσεις οδηγούν σε παρόμοιες κλινικές συνέπειες και έχουν κοινά χαρακτηριστικά ως προς την αιτιολογία τους, γεγονός που ενισχύει την αναγκαιότητα συγχώνευσης αυτών των κατηγοριών σε μία ενιαία κλάση Ισχαιμικού Εγκεφαλικού. Αυτή η συγχώνευση καθιστά τη μελέτη της φυσιολογίας τους σε σχέση με τα Embolic Events πιο σαφή, διευκολύνοντας την αναγνώριση και τη θεραπευτική προσέγγιση.

Από μεθοδολογικής σκοπιάς, η ανάλυση των καμπυλών ROC έδειξε ότι το μοντέλο εμφανίζει σημαντικές δυσκολίες στη διάκριση αυτών των δύο κατηγοριών (Vessel Occlusion και LVO), γεγονός που υποδηλώνει ότι τα χαρακτηριστικά τους είναι ιδιαίτερα όμοια και αλληλεπικαλυπτόμενα. Η συγχώνευση τους σε μία κατηγορία Ισχαιμικού Εγκεφαλικού μειώνει την σύγχυση μεταξύ των παρόμοιων υποκατηγοριών και επιτρέπει στο μοντέλο να εστιάσει σε πιο διακριτές κατηγορίες, όπως το Control και το Αιμορραγικό Εγκεφαλικό (Bleeding), το οποίο προκύπτει από διαφορετικό παθογενετικό μηχανισμό και εκδηλώνεται με διακριτές κλινικές και βιολογικές παραμέτρους.

Η συγχώνευση αυτών των κατηγοριών σε μία ενιαία κλάση Ισχαιμικού Εγκεφαλικού έχει ως αποτέλεσμα την απλοποίηση της ταξινόμησης και τη βελτίωση της απόδοσης των μοντέλων ταξινόμησης, οδηγώντας σε αυξημένα ποσοστά Accuracy, Recall και AUC scores, ενώ παράλληλα καθιστά τη διαφοροποίηση μεταξύ των τύπων εγκεφαλικού πιο ρεαλιστική και κλινικά χρήσιμη. Αυτή η προσέγγιση ενισχύει την κλινική σημασία των αποτελεσμάτων, διευκολύνοντας την αξιολόγηση και την εφαρμογή θεραπευτικών στρατηγικών για τους ασθενείς με ισχαιμικό εγκεφαλικό επεισόδιο.

Η ανάλυση των αποτελεσμάτων για την συγχώνευση των κατηγοριών δείχνει ότι ο CD235a-PE+ αποτελεί τον πιο αξιόπιστο βιοδείκτη συνολικά, με υψηλή ακρίβεια (73.8%) και F1-score (61.1%), καθιστώντας τον ιδιαίτερα αποτελεσματικό για την ταξινόμηση εγκεφαλικού επεισοδίου. Ωστόσο, ο CD14-PB+ εμφανίζει το υψηλότερο AUC (73%) για τη διάκριση μεταξύ Ισχαιμικού και Αιμορραγικού εγκεφαλικού, γεγονός που τον καθιστά τον βέλτιστο βιοδείκτη για αυτή την κρίσιμη διαφοροποίηση. Ο CD326-APC+ και ο Lact-FITC+ έχουν επίσης καλή διακριτική ικανότητα μεταξύ των δύο τύπων εγκεφαλικού (AUC = 71% και 67% αντίστοιχα).

Όσον αφορά τους αλγορίθμους ταξινόμησης, το Gradient Boosting χρησιμοποιείται στους περισσότερους βιοδείκτες και προσφέρει σταθερή απόδοση, ενώ το Decision Tree υπερέχει στην ταξινόμηση του CD14-PB+, καθιστώντας το ιδανικό για τη διάκριση Ισχαιμικού vs Αιμορραγικού εγκεφαλικού. Από την άλλη, η διάκριση Control vs Ischemic Stroke και Control vs Hemorrhage Stroke είναι σαφώς πιο εύκολη, με τον CD45-APC+ και τον CD235a-PE+ να εμφανίζουν υψηλές τιμές AUC (81% και 78% αντίστοιχα).

Επομένως, οι βιοδείκτες CD14-PB+, CD326-APC+ και Lact-FITC+ αναδεικνύονται ως οι πιο αξιόπιστοι για τη διάκριση μεταξύ Ισχαιμικού και Αιμορραγικού εγκεφαλικού, επιτυγχάνοντας τα υψηλότερα AUC scores. Ακολουθεί συγκριτικός πίνακας που παρουσιάζει τα επιλεγμένα χαρακτηριστικά για τη διάγνωση με 4 κλάσεις (Vessel Occlusion, LVO, Control, Bleeding) και 3 κλάσεις (Ischemic Stroke, Control, Bleeding) για τους τρεις πιο αποδοτικούς βιοδείκτες.

Πίνακας 5.5: Σύγκριση των επιλεγμένων χαρακτηριστικών για την διάκριση 4 κατηγοριών και 3 κατηγοριών

	Διάκριση μεταξύ 4 κλάσεων (Vessel Occlusion, LVO, Control, Bleeding)	**Διάκριση μεταξύ 3 κλάσεων (Ischemic Stroke, Control, Bleeding)**

CD14-PB+	Age, RR syst., Thrombocytes, wbc, glucose, Mean, Median, skewness, entropy, 600-700 nm	'Age', 'RR syst.', 'RR diast.', 'Oxygen saturation', 'Thrombocytes', 'wbc', 'glucose', 'Median', 'Std', 'skewness'
CD326-APC+	Age, RR syst., Thrombocytes, wbc, glucose, Std, skewness, entropy, 100-200 nm, 300-400 nm	'Age', 'RR syst.', 'Thrombocytes', 'wbc', 'glucose', 'Std', 'skewness', 'entropy', '100-200 nm', '300-400 nm'
Lact-FITC+	Age, RR syst., Thrombocytes, wbc, glucose, Mean, kurtosis, skewness, 400-500 nm, 700-800 nm	'Age', 'RR syst.', 'RR diast.', 'Thrombocytes', 'wbc', 'glucose', 'kurtosis', '400-500 nm', '500-600 nm', '900-1000 nm'

Η ανάλυση των επιλεγμένων χαρακτηριστικών ανάμεσα στις δύο διαφορετικές ταξινομήσεις (4 κλάσεων: Vessel Occlusion, LVO, Control, Bleeding και 3 κλάσεων: Ischemic Stroke, Control, Bleeding) αποκαλύπτει σημαντικές κλινικές διαφορές.

Τα θεμελιώδη αιματολογικά και μεταβολικά χαρακτηριστικά, όπως η ηλικία, η συστολική πίεση (RR syst.), τα αιμοπετάλια (Thrombocytes), τα λευκά αιμοσφαίρια (wbc) και η γλυκόζη (glucose), παραμένουν σταθεροί προγνωστικοί δείκτες, επιβεβαιώνοντας την κρίσιμη σημασία τους στη φυσιοπαθολογία του εγκεφαλικού επεισοδίου. Η υπεργλυκαιμία, για παράδειγμα, έχει συσχετιστεί με αυξημένη νευρωνική βλάβη μετά από ισχαιμικό επεισόδιο, ενώ τα αυξημένα επίπεδα λευκών αιμοσφαιρίων υποδηλώνουν φλεγμονώδη απόκριση, η οποία παίζει κεντρικό ρόλο στην εξέλιξη τόσο του ισχαιμικού όσο και του αιμορραγικού εγκεφαλικού.

Στη διάκριση μεταξύ 4 κλάσεων, τα φασματικά χαρακτηριστικά των εξωκυτταρικών κυστιδίων, όπως οι απορροφήσεις σε μήκη κύματος 600-700 nm και 700-800 nm, φαίνεται να είναι κρίσιμα για τη λεπτομερή διάκριση μεταξύ των δύο τύπων ισχαιμικού εγκεφαλικού (Vessel Occlusion και LVO), πιθανώς επειδή αντανακλούν μικροαγγειακές ή ενδοθηλιακές διαταραχές που διαφοροποιούν τις δύο υποκατηγορίες. Ωστόσο, για τις 3 κλάσεις, όπου τα ισχαιμικά επεισόδια ενοποιούνται σε μία κατηγορία, αυτά τα χαρακτηριστικά καθίστανται λιγότερο κρίσιμα. Αντίθετα, αιμοδυναμικοί δείκτες όπως η διαστολική πίεση (RR diast.) και ο κορεσμός οξυγόνου (Oxygen Saturation) επιλέγονται, γεγονός που υποδηλώνει ότι η συνολική αιμοδυναμική απορρύθμιση αποτελεί σημαντικότερο προγνωστικό παράγοντα όταν η ταξινόμηση επικεντρώνεται στη διάκριση Ισχαιμικού από Αιμορραγικό εγκεφαλικό. Η υπόταση ή η σοβαρή διακύμανση της πίεσης μπορεί να επιδεινώσει την ισχαιμική βλάβη, ενώ η υποξία σχετίζεται με αυξημένη πιθανότητα αιμορραγικής μετατροπής.

Επιπλέον, η ανάλυση της στατιστικής κατανομής των δεδομένων των EVs φαίνεται να προσαρμόζεται στη δομή της ταξινόμησης. Στη διάκριση 4 κλάσεων, χρησιμοποιούνται Mean, Median, Skewness και Entropy, υποδηλώνοντας ότι η συνολική ποσότητα και η κατανομή των EVs μπορεί να διαφέρει μεταξύ των υποκατηγοριών του Ισχαιμικού εγκεφαλικού. Αντίθετα, στη διάκριση 3 κλάσεων, δίνεται περισσότερη έμφαση στη μεταβλητότητα και τις ακραίες τιμές των δεδομένων, καθώς χαρακτηριστικά όπως το Standard Deviation (Std) και το Kurtosis αντικαθιστούν τα Mean και Median, πιθανώς επειδή η παθοφυσιολογία του αιμορραγικού εγκεφαλικού χαρακτηρίζεται από μεγαλύτερη ετερογένεια και υψηλότερη μεταβλητότητα στα προφίλ των EVs.

Συνολικά, η μετάβαση από 4 σε 3 κλάσεις οδηγεί σε αλλαγή στη βαρύτητα των χαρακτηριστικών, καθώς η λεπτομερής διάκριση μεταξύ των υποκατηγοριών του Ισχαιμικού εγκεφαλικού καθίσταται λιγότερο σημαντική και δίνεται μεγαλύτερη έμφαση στη γενική παθοφυσιολογική διαφοροποίηση μεταξύ ισχαιμικού και αιμορραγικού εγκεφαλικού.

Αυτή η προσαρμογή στη σημασία των χαρακτηριστικών φαίνεται να βελτιστοποιεί την απόδοση των μοντέλων ταξινόμησης, παρέχοντας πιο αξιόπιστες διαγνωστικές πληροφορίες για την κλινική διαφοροποίηση των τύπων εγκεφαλικού επεισοδίου. Συγκεκριμένα, η τιμή για την Accuracy αυξάνεται από ~40-52% σε έως και 73.8%, ενώ η Precision βελτιώνεται σημαντικά, ιδιαίτερα στον CD235a-PE+ (από 56.6% σε 77.6%) και τον CD326-APC+ (από 58.3% σε 72.1%), μειώνοντας τις ψευδώς θετικές ταξινομήσεις. Το Recall παραμένει χαμηλότερο από την Precision, αλλά εμφανίζει βελτίωση στον CD14-PB+ (από 34.3% σε

50%), υποδηλώνοντας καλύτερη ανίχνευση των θετικών περιπτώσεων. Το F1-score αυξάνεται συνολικά, με μέγιστη τιμή 61.1% (CD235a-PE+), δείχνοντας καλύτερη ισορροπία μεταξύ Precision και Recall. Τέλος, το AUC (Ischemic Stroke vs Bleeding) βελτιώνεται από 40-50% σε έως και 73% (CD14-PB+), υποδηλώνοντας σαφέστερο διαχωρισμό μεταξύ των τύπων εγκεφαλικού. Συνολικά, η μετάβαση στις 3 κλάσεις ενισχύει την ακρίβεια και τη διαγνωστική αξία των μοντέλων, καθιστώντας τα πιο αξιόπιστα στη διαφοροποίηση μεταξύ Ισχαιμικού και Αιμορραγικού εγκεφαλικού.

5.2 Σύγκριση με State of Art Αποτελέσματα

Παρόμοιες μελέτες έχουν αναπτυχθεί, όπως αναφέρεται και στα υποκεφάλαια "1.2 Βιβλιογραφική Ανασκόπηση" και "3.1.6 Σημασία έγκαιρης διάγνωσης – Ρόλος των Βιοδεικτών".

Συγκεκριμένα, η χρήση αλγορίθμων μηχανικής μάθησης, όπως οι SVM, ANFIS, KNN, DT και SVM, για τη διάκριση ισχαιμικού και αιμορραγικού εγκεφαλικού επεισοδίου, με ακρίβεια 99,1%, έχει μελετηθεί σε αιματολογικούς δείκτες. Ειδικότερα, η σύγκριση με την υπάρχουσα βιβλιογραφία δείχνει ότι, ενώ οι αιματολογικοί δείκτες (ANC, CPK, WBC) είναι σημαντικοί, η εφαρμογή μηχανικής μάθησης σε συνδυασμό με αυτούς μπορεί να ενισχύσει σημαντικά την ακριβή διάγνωση.

Παράλληλα, η χρήση microRNAs για τη διάγνωση του ισχαιμικού εγκεφαλικού με την εφαρμογή του αλγορίθμου SVM και την επιτυχία AUC=0.958 αποδεικνύει τη δύναμη της γενετικής προσέγγισης στη διάγνωση εγκεφαλικού επεισοδίου. Αυτή η μέθοδος ενσωματώνει μία από τις πιο πρόσφατες τεχνικές για την ανάλυση του ισχαιμικού εγκεφαλικού επεισοδίου, διαφοροποιούμενη από τα δεδομένα που χρησιμοποιούν κυρίως απεικονιστικά μέσα (CT, MRI), προσφέροντας δυνατότητες για πιο αξιόπιστες και προηγμένες διαγνωστικές προσεγγίσεις.

Μια από τις πιο καινοτόμες προσεγγίσεις είναι η χρήση εξωκυτταρικών κυστιδίων (EVs) για τη διάγνωση του ισχαιμικού εγκεφαλικού με τη χρήση next-generation sequencing (NGS), η οποία προσφέρει υψηλή ακρίβεια με AUC=92%. Αυτή η μέθοδος διαφοροποιείται από τις παραδοσιακές διαγνωστικές προσεγγίσεις που χρησιμοποιούν μόνο απεικονιστικά δεδομένα ή ιατρικά αρχεία, δίνοντας τη δυνατότητα για μη επεμβατική και πρώιμη διάγνωση του εγκεφαλικού επεισοδίου.

Επιπλέον, έχουν αναλυθεί συνολικά 518 βιοδείκτες (βλ. 3.1.6), με την πλειονότητα των μελετών να επικεντρώνεται στους πρωτεϊνικούς βιοδείκτες, ενώ σε μικρότερο ποσοστό εξετάστηκαν RNA και γονιδιακή έκφραση. Δύο βιοδείκτες που εμφανίστηκαν συχνά στις μελέτες είναι το S100B και η MMP-9, οι οποίοι απέδειξαν τη δυνατότητά τους να διακρίνουν το ισχαιμικό εγκεφαλικό από άλλες καταστάσεις. Παρά ταύτα, το 89% των βιοδεικτών αξιολογήθηκε μόνο μία φορά, με ένα μικρό ποσοστό να μελετήθηκε περισσότερες από μία φορές. Οι βιοδείκτες GFAP και S100B αναδείχθηκαν ως οι συχνότερα αναφερόμενοι, με την αυξημένη συγκέντρωσή τους να χρησιμεύει ως δείκτης για τη διαφοροποίηση του ισχαιμικού από το αιμορραγικό εγκεφαλικό.

Σύμφωνα με τις παραπάνω μελέτες, παρά τις υποσχέσεις των βιοδεικτών για τη διάγνωση του εγκεφαλικού, τα αποτελέσματα υποδεικνύουν ότι δεν είναι ακόμη επαρκώς αξιόπιστοι ώστε να αντικαταστήσουν τις παραδοσιακές απεικονιστικές τεχνικές. Η ανάγκη για την περαιτέρω βελτίωση των διαγνωστικών μεθόδων με τη χρήση νέων βιοδεικτών και την ολοκληρωμένη ανάλυση δεδομένων παραμένει επιτακτική.

Στην παρούσα εργασία, αναδείχθηκε μια σημαντική προσέγγιση που αφορά την επιλογή των χαρακτηριστικών που είναι πιο σχετικά για τη διάγνωση του εγκεφαλικού επεισοδίου καθώς και για τη διάκριση μεταξύ των διαφορετικών κατηγοριών. Τα αποτελέσματα του πρώτου μέρους είναι ιδιαίτερα ικανοποιητικά και αποδεικνύουν ότι η χρήση των EVs σε συνδυασμό με τις μεθόδους μηχανικής μάθησης, μπορεί να αποτελέσει πολύτιμο εργαλείο υποβοήθησης της λήψης αποφάσεων από τους γιατρούς. Από την άλλη, τα αποτελέσματα των υπόλοιπων σταδίων δεν παρουσίασαν εξαιρετικά ικανοποιητικά δεδομένα, γεγονός που μπορεί να αποδοθεί εν μέρει στον περιορισμένο αριθμό δειγμάτων στο σύνολο δεδομένων. Παρά ταύτα, η παρούσα εργασία εισάγει μια διαφορετική προσέγγιση, η οποία επικεντρώνεται στην επέκταση της χρήσης των EVs ως βιοδεικτών και εξετάζει συστηματικά τις δυνατότητές τους να ταξινομούν τους ασθενείς, χωρίς να περιορίζεται σε μία μόνο κατηγορία.

5.3 Περιορισμοί

Στην παρούσα εργασία υπήρξαν ορισμένοι περιορισμοί. Ο κυριότερος ήταν το περιορισμένο μέγεθος του συνόλου δεδομένων, καθώς το δείγμα αποτελούνταν μόλις από 140 παρατηρήσεις. Οι αλγόριθμοι μηχανικής μάθησης αποδίδουν βέλτιστα όταν εκπαιδεύονται σε μεγάλες και ισορροπημένες βάσεις δεδομένων, επιτρέποντας την αξιόπιστη αναγνώριση προτύπων και τη μείωση της πιθανότητας υπερπροσαρμογής. Η περιορισμένη διαθεσιμότητα δεδομένων δεν επέτρεψε την εφαρμογή πιο σύνθετων τεχνικών, όπως νευρωνικά δίκτυα (Deep Learning), τα οποία απαιτούν μεγάλο όγκο δεδομένων για αποτελεσματική εκπαίδευση. Συνεπώς, η χρήση τέτοιων μεθόδων θα μπορούσε ενδεχομένως να βελτιώσει την ακρίβεια της ταξινόμησης, υπό την προϋπόθεση ότι θα είναι διαθέσιμο ένα μεγαλύτερο και πιο αντιπροσωπευτικό σύνολο δεδομένων.

5.4 Μελλοντικές Επεκτάσεις

Η παρούσα μελέτη ανέδειξε τη δυνατότητα διάγνωσης του εγκεφαλικού επεισοδίου μέσω μοντέλων μηχανικής μάθησης, τα οποία, παρά το περιορισμένο μέγεθος του δείγματος, παρουσίασαν αρκετά ικανοποιητικά αποτελέσματα. Η περαιτέρω αύξηση του συνόλου δεδομένων θα μπορούσε να ενισχύσει σημαντικά την ακρίβεια των μοντέλων και να επιτρέψει τη χρήση πιο προηγμένων μεθόδων τεχνητής νοημοσύνης, όπως τα συνελικτικά νευρωνικά δίκτυα (CNNs). Επιπλέον, μια από τις σημαντικότερες προοπτικές για μελλοντική έρευνα είναι η διεύρυνση της μελέτης των εξωκυτταρικών κυστιδίων (EVs), με ενσωμάτωση επιπλέον βιοδεικτών ή συνδυαστικών προσεγγίσεων με γενετικούς δείκτες και άλλες αιματολογικές παραμέτρους. Η περαιτέρω διερεύνηση των EVs, καθώς και η ανάπτυξη πολυπαραγοντικών μοντέλων που συνδυάζουν κλινικά, αιματολογικά και μοριακά δεδομένα, θα μπορούσε να οδηγήσει σε ακόμα μεγαλύτερη ακρίβεια στην έγκαιρη διάγνωση του εγκεφαλικού επεισοδίου.

Παράλληλα, η εφαρμογή των αναπτυγμένων μοντέλων σε πολλαπλά κλινικά περιβάλλοντα αποτελεί ένα κρίσιμο βήμα προς την επικύρωση των ευρημάτων. Αν και τα δεδομένα που χρησιμοποιήθηκαν στην παρούσα μελέτη ήταν πραγματικά κλινικά δεδομένα, προερχόμενα από δείγματα ασθενών, το μικρό τους μέγεθος και η ενδεχόμενη ομοιογένεια του δείγματος ενδέχεται να περιορίζουν τη γενίκευση των αποτελεσμάτων. Επομένως, η διεξαγωγή μελετών σε μεγαλύτερα και πιο ετερογενή δείγματα πληθυσμού, καθώς και η δοκιμή των μοντέλων σε πολυκεντρικές μελέτες ή διαφορετικά κλινικά περιβάλλοντα, θα επιτρέψει μια πιο αξιόπιστη αξιολόγηση της αποτελεσματικότητάς τους στην καθημερινή ιατρική πράξη. Η ενσωμάτωση αυτών των τεχνολογιών σε πρωτόκολλα επείγουσας διάγνωσης θα μπορούσε να βελτιώσει τη διαχείριση των περιστατικών εγκεφαλικού, επιτρέποντας ταχύτερη και πιο αξιόπιστη αναγνώριση των ασθενών που χρειάζονται άμεση παρέμβαση.

Τέλος, μια ακόμα σημαντική προοπτική είναι η ανάπτυξη αυτοματοποιημένων διαγνωστικών εργαλείων, τα οποία θα ενσωματώνουν τα μοντέλα μηχανικής μάθησης σε πλατφόρμες ανάλυσης δεδομένων σε πραγματικό χρόνο. Τέτοιες πλατφόρμες θα μπορούσαν να εφαρμοστούν σε νοσοκομεία, κινητές ιατρικές μονάδες ή ακόμα και σε εφαρμογές τηλεϊατρικής, επιτρέποντας τη γρήγορη και ακριβή αναγνώριση ενός εγκεφαλικού επεισοδίου και τη λήψη θεραπευτικών αποφάσεων με ελάχιστη καθυστέρηση.

6. Κεφάλαιο 6: Βιβλιογραφία

[1] Sherif, Fayroz F., and Khaled S. Ahmed. "A Machine Learning Approach for Stroke Differential Diagnosis by Blood Biomarkers." *Journal of Advances in Information Technology* 15.1 (2024).

[2] Zhao, Xinyi, et al. "Machine learning analysis of MicroRNA expression data reveals novel diagnostic biomarker for ischemic stroke." *Journal of Stroke and Cerebrovascular Diseases* 30.8 (2021): 105825.

[3] Sirsat, Manisha Sanjay, Eduardo Fermé, and Joana Camara. "Machine learning for brain stroke: a review." *Journal of Stroke and Cerebrovascular Diseases* 29.10 (2020): 105162.

[4] Bang, Ji Hoon, et al. "Machine Learning-Based Etiologic Subtyping of Ischemic Stroke Using Circulating Exosomal microRNAs." *International Journal of Molecular Sciences* 25.12 (2024): 6761.

[5] Park, Jae Hyun, and Jisook Moon. "Early Diagnosis of Brain Diseases Using Artificial Intelligence and EV Molecular Data: A Proposed Noninvasive Repeated Diagnosis Approach." *Cells* 12.1 (2022): 102.

[6] Rout, Madhusmita, et al. "Improving Stroke Outcome Prediction Using Molecular and Machine Learning Approaches in Large Vessel Occlusion." *Journal of Clinical Medicine* 13.19 (2024): 5917.

[7] Manwani, Bharti, et al. "Small RNA signatures of acute ischemic stroke in L1CAM positive extracellular vesicles." *Scientific Reports* 14.1 (2024): 13560.

[8] Awad, Mariette, and Rahul Khanna. *Efficient learning machines: theories, concepts, and* applications for engineers and system designers. Springer nature, 2015.

[9] Samuel, Arthur L. "Some studies in machine learning using the game of checkers." IBM Journal of research and development 3.3 (1959): 210-229.

[10] Muditomo, Arianto, and Syamsari Syamsari. "POSSIBILITY OF USING MACHINE LEARNING ALGORITHMS TO MODERNIZE AGRICULTURAL RESEARCH IN INDONESIA." Proceedings of the Business Innovation and Engineering Conference (BIEC 2022). 2023.

[11] Ahmad, Ghulab Nabi, et al. "Efficient medical diagnosis of human heart diseases using machine learning techniques with and without GridSearchCV." *IEEE Access* 10 (2022): 80151-80173.

[12] Nasteski, Vladimir. "An overview of the supervised machine learning methods." *Horizons. b* 4.51-62 (2017): 56.

[13] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. "Unsupervised learning. The elements of statistical learning." Springer New York, 2009. 485-585.

[14] Barbara, Daniel, and Sushil Jajodia. "Applications of data mining in computer security." Vol. 6. Springer Science & Business Media, 2002.`

[15] Kaelbling, Leslie Pack, Michael L. Littman, and Andrew W. Moore. "Reinforcement learning: A survey." *Journal of artificial intelligence research* 4 (1996): 237-285.

[16] Gradient boosting machines, a tutorial Alexey Natekin1 * and Alois Knoll 2 1 fortiss GmbH, Munich, Germany 2 Department of Informatics, Technical University Munich, Garching, Munich, Germany

[17] 1999 REITZ LECTURE GREEDY FUNCTION APPROXIMATION: A GRADIENT BOOSTING MACHINE1 By Jerome H. Friedman Stanford University

[18] Application of XGBoost algorithm in the optimization of pollutant concentration Jiangtao Li a , Xingqin An a,* , Qingyong Li b,**, Chao Wang a , Haomin Yu b , Xinyuan Zhou b , Yangli-ao Geng

[19] Experimenting XGBoost Algorithm for Prediction and Classifi cation of Different Datasets Ramraj Sa Nishant Uzirb Sunil Rb and Shatadeep Banerjeeb

[20] Omarzai, Fraidoon. “XGBoost Regression in Depth.” *Medium*, 25 Mar. 2020, https://medium.com/@fraidoonomarzai99/xgboost-regression-in-depth-cb2b3f623281. Accessed 16 Feb. 2025.
[21] Chen, Tianqi and Guestrin, Carlos, «XGBoost: A Scalable Tree Boosting System,» June 2016.
[22] “XGBoost.” *Wikipedia*, Wikimedia Foundation, Feb. 2025, https://en.wikipedia.org/wiki/XGBoost#cite_ref-paper_16-0. Accessed Feb. 2025.
[23] “XGBoost.” *GitHub*, https://github.com/dmlc/xgboost. Accessed 16 Feb. 2025.
[24] Ghosh, Siddhartha. “An End-to-End Guide to Understand the Math Behind XGBoost.” *Analytics Vidhya*, 6 Sept. 2018, https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/. Accessed Feb. 2025.
[25] “XGBoost vs Gradient Boost: Differences and Use Cases.” *ML Journey*, 1 Dec. 2021, https://mljourney.com/xgboost-vs-gradient-boost-differences-and-use-cases/. Accessed 16 Feb. 2025.
[26] Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer’s Disease: A Systematic Review Alessia Sarica1 *, Antonio Cerasa1 and Aldo Quattrone1, 2 1 Institute of Bioimaging and Molecular Physiology, National Research Council, Catanzaro, Italy, 2 Institute of Neurology, University Magna Graecia, Catanzaro, Italy
[27] Rigatti, Steven J. "Random forest." *Journal of Insurance Medicine* 47.1 (2017): 31-39.
[28] Liu, Yingchun. "Random forest algorithm in big data environment." *Computer modelling & new technologies* 18.12A (2014): 147-151.
[29] Vembandasamy, K., R. Sasipriya, and E. Deepa. "Heart diseases detection using Naive Bayes algorithm." *International Journal of Innovative Science, Engineering & Technology* 2.9 (2015): 441-444.
[30] International Journal of Science and Research (IJSR) ISSN: 2319-7064 ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426 Volume 9 Issue 1, January 2020 www.ijsr.net Licensed Under Creative Commons Attribution CC BY Machine Learning Algorithms - A Review Batta Mahesh
[31] Fan,M., Wei, L., He, Z., Wei, W., & Lu, X. (2016). Defect inspection of solder bumps using the scanning acoustic microscopy and fuzzy SVM algorithm. Microelectronics Reliability, 65, 192-197.
[32] Long, S., Huang, X., Chen, Z., Pardhan, S., & Zheng, D. (2019). Automatic detection of hard exudates in color retinal images using dynamic threshold and SVM classification: algorithm development and evaluation.BioMed research international,2019.
[33] Abd Elkarim, I. S., & Agbinya, J. (2019). A Review of Parallel Support Vector Machines (PSVMs) for Big Data classification. Australian Journal of Basic and Applied Sciences, 13(12), 61-71.
[34] Awad, Mariette, et al. "Support vector machines for classification." *Efficient learning machines: Theories, concepts, and applications for engineers and system designers* (2015): 39-66.
[35] Xing, Wenchao, and Yilin Bei. "Medical health big data classification based on KNN classification algorithm." *Ieee Access* 8 (2019): 28808-28819.
[36] Kataria, Aman, and M. D. Singh. "A review of data classification using k-nearest neighbour algorithm." *International Journal of Emerging Technology and Advanced Engineering* 3.6 (2013): 354-360.
[37] Charbuty, Bahzad, and Adnan Abdulazeez. "Classification based on decision tree algorithm for machine learning." *Journal of Applied Science and Technology Trends* 2.01 (2021): 20-28.

[38] Priyam, Anuja, et al. "Comparative analysis of decision tree classification algorithms." *International Journal of current engineering and technology* 3.2 (2013): 334-337.
[39] Antoniadis A, Lambert-Lacroix S, Leblanc F. Effective dimension reduction methods for tumor classification using gene expression data. Bioinformatics 2003;19(5):563–70.
[40] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn 2002;46:389–422.
[41] Yu J-S, Ongarello S, Fiedler R, Chen X-W, Toffolo G, Cobelli C, et al. Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data. Bioinformatics 2005;21:2200–9.
[42] Liu H, Motoda H. Feature selection for knowledge discovery and data mining. Kluwer Academic; 1998.
[43] Shirbani, F., and H. Soltanian Zadeh. "Fast SFFS-based algorithm for feature selection in biomedical datasets." *AUT Journal of Electrical Engineering* 45.2 (2013): 43-56.
[44] Pudil, Pavel, Jana Novovičová, and Josef Kittler. "Floating search methods in feature selection." *Pattern recognition letters* 15.11 (1994): 1119-1125.
[45] Park, Cheolsoo, Clive Cheong Took, and Joon-Kyung Seong. "Machine learning in biomedical engineering." *Biomedical Engineering Letters* 8 (2018): 1-3.
[46] A novel feature selection approach for biomedical data classification Yonghong Peng *, Zhiqing Wu, Jianmin Jiang School of Informatics, University of Bradford, Bradford BD7 1DP, UK
[47] Ramezan, Christopher A. "Transferability of Recursive Feature Elimination (RFE)-Derived Feature Sets for Support Vector Machine Land Cover Classification." *Remote Sensing* 14.24 (2022): 6218.
[48] https://scott.fortmann-roe.com/docs/BiasVariance.html
[49] Conceptual complexity and the bias/variance tradeoff Erica Briscoe a , Jacob Feldman b,⇑ a Aerospace, Transportation & Advanced Systems Laboratory, Georgia Tech Research Institute, United States bDepartment of Psychology, Center for Cognitive Science, Rutgers University – New Brunswick, United States
[50] https://www.kaggle.com/code/sid321axn/regularization-techniques-in-deep-learning
[51] Anguita, Davide, et al. "K-Fold Cross Validation for Error Rate Estimate in Support Vector Machines." *DMIN*. 2009.
[52] M.R. Wayahdi, D. Syahputra, S. Hafiz, N. Ginting, Evaluation of the K-Nearest Neighbor Model With K-Fold Cross Validation on Image Classification, J. Infokum. 9 (2020) 1–6.
[53] Nti, Isaac Kofi, Owusu Nyarko-Boateng, and Justice Aning. "Performance of machine learning algorithms with different K values in K-fold CrossValidation." International Journal of Information Technology and Computer Science 13.6 (2021): 61-71.
[54] Schratz, Patrick, et al. "Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data." *Ecological Modelling* 406 (2019): 109-120.
[55] Ahsan, Md Manjurul, et al. "Effect of data scaling methods on machine learning algorithms and model performance." *Technologies* 9.3 (2021): 52.
[56] Haixiang, Guo, et al. "Learning from class-imbalanced data: Review of methods and applications." *Expert systems with applications* 73 (2017): 220-239.
[57] Viloria, Amelec, Omar Bonerge Pineda Lezama, and Nohora Mercado-Caruzo. "Unbalanced data processing using oversampling: machine learning." Procedia Computer Science 175 (2020): 108-113.
[58] An Overview of Overfitting and its Solutions Xue Ying Building 1, Huizhong Tower, NO.1 Shangdi Seven Street, Haidian District Beijing 10 0085 China

[59] Zhang, Tong, and Bin Yu. "Boosting with early stopping: Convergence and consistency." (2005): 1538-1579.
[60] Shorten, Connor, and Taghi M. Khoshgoftaar. "A survey on image data augmentation for deep learning." *Journal of big data* 6.1 (2019): 1-48.
[61] Reddy, G. Thippa, et al. "Analysis of dimensionality reduction techniques on big data." *Ieee Access* 8 (2020): 54776-54788.
[62] Tian, Yingjie, and Yuqi Zhang. "A comprehensive survey on regularization strategies in machine learning." *Information Fusion* 80 (2022): 146-166.
[63] "Stroke." *MedlinePlus*, U.S. National Library of Medicine, https://medlineplus.gov/ency/article/000726.htm. Accessed 5 Feb. 2025.
[64] "Stroke." *National Heart, Lung, and Blood Institute*, U.S. Department of Health and Human Services, https://www.nhlbi.nih.gov/health/stroke. Accessed 16 Feb. 2025.
[65] Coupland, A.P., Thapar, A., Qureshi, M.I., Jenkins, H. and Davies, A.H., 2017. The definition of stroke. Journal of the Royal Society of Medicine, 110(1), pp.9-12.
[66] Patil, Smita, et al. "Detection, diagnosis and treatment of acute ischemic stroke: current and future perspectives." Frontiers in medical technology 4 (2022): 748949.
[67] Feske, Steven K. "Ischemic stroke." *The American journal of medicine* 134.12 (2021): 1457-1464.
[68] Hanley, Daniel F., et al. "Hemorrhagic stroke: introduction." *Stroke* 44.6_suppl_1 (2013): S65-S66.
[69] https://watchlearnlive.heart.org/index.php?moduleSelect=tisatk
[70] Murphy, S.J. and Werring, D.J., 2020. Stroke: causes and clinical features. Medicine, 48(9), pp.561-566.
[71] Khan, S.N. and Vohra, E.A., 2007. Risk factors for stroke: A hospital based study. Pak J Med Sci, 23(1), pp.17-22.
[72] Boehme, Amelia K., Charles Esenwa, and Mitchell SV Elkind. "Stroke risk factors, genetics, and prevention." *Circulation research* 120.3 (2017): 472-495.
[73] Jones, Abbeygail, et al. "Functional stroke symptoms: a narrative review and conceptual model." The Journal of neuropsychiatry and clinical neurosciences 32.1 (2020): 14-23.
[74] Musuka, Tapuwa D., et al. "Diagnosis and management of acute ischemic stroke: speed is critical." Cmaj 187.12 (2015): 887-893.
[75] Lee, E. C., Ha, T. W., Lee, D. H., Hong, D. Y., Park, S. W., Lee, J. Y., Lee, M. R., and Oh, J. S. "Utility of Exosomes in Ischemic and Hemorrhagic Stroke Diagnosis and Treatment." International Journal of Molecular Sciences, vol. 23, no. 15, 28 July 2022, p. 8367,
[76] Li, Qianyun, et al. *"Multi-Level Biomarkers for Early Diagnosis of Ischaemic Stroke: A Systematic Review and Meta-Analysis."* **International Journal of Molecular Sciences**
[77] Bustamante, Alejandro, et al. "Blood biomarkers for the early diagnosis of stroke: the stroke-chip study." Stroke 48.9 (2017): 2419-2425.
[78] Ollen-Bittle, Nikita, et al. "Mechanisms and biomarker potential of extracellular vesicles in stroke." *Biology* 11.8 (2022): 1231.
[79] Tian, Ye, and Xiaomei Yan. "Flow Cytometry for Single Extracellular Vesicle Analysis." *Extracellular Vesicles: From Bench to Bedside*. Singapore: Springer Nature Singapore, 2024. 111-124.
[80] Welsh, Joshua A., et al. "A compendium of single extracellular vesicle flow cytometry." *Journal of extracellular vesicles* 12.2 (2023): e12299.
[81] Simeone, Pasquale, et al. "Diameters and fluorescence calibration for extracellular vesicle analyses by flow cytometry." *International Journal of Molecular Sciences* 21.21 (2020): 7885.

[82] Hossin, Mohammad, and Md Nasir Sulaiman. "A review on evaluation metrics for data classification evaluations." *International journal of data mining & knowledge management process* 5.2 (2015): 1.

[83] N.V. Chawla, N. Japkowicz and A. Kolcz, "Editorial: Special issue on learning from imbalanced data sets", SIGKDD Explorations, 6 (2004) 1-6.

[84] Q. Gu, L. Zhu and Z. Cai, "Evaluation Measures of the Classification Performance of Imbalanced Datasets", in Z. Cai et al. (Eds.) ISICA 2009, CCIS 51. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 461-471.

[85] J. Huang and C. X. Ling, "Constructing new and better evaluation measures for machine learning", in R. Sangal, H. Mehta and R. K. Bagga (Eds.) Proc. of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007), San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007, pp.859-864.

[86] A. Rakotomamonyj, "Optimizing area under ROC with SVMs", in J. Hernandez-Orallo, C. Ferri, N. Lachiche and P. A. Flach (Eds.) 1st Int. Workshop on ROC Analysis in Artificial Intelligence (ROCAI 2004), Valencia, Spain, 2004, pp. 71-80.

[87] M. V. Joshi, "On evaluating performance of classifiers for rare classes", in Proceedings of the 2002 IEEE Int. Conference on Data Mining (ICDN 2002) ICDM'02, Washington, D. C., USA: IEEE Computer Society, 2002, pp. 641-644.

[88] K. Hajian-Tilaki, ―Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation.‖ Casp. J. Intern. Med., vol. 4, no. 2, pp. 627– 635, 2013.

[89] S. Safari, A. Baratloo, M. Elfil, and A. Negida, ―Evidence Based Emergency Medicine; Part 5 Receiver Operating Curve and Area under the Curve.‖ Emerg.

[90] Yu, Lei, and Huan Liu. "Feature selection for high-dimensional data: A fast correlation-based filter solution." *Proceedings of the 20th international conference on machine learning (ICML-03)*. 2003.

[91] Stenz, Katrine Tang, et al. "Extracellular vesicles in acute stroke diagnostics." *Biomedicines* 8.8 (2020): 248.

[92] Ziegler-Heitbrock HW, Ulevitch RJ. CD14: cell surface receptor and differentiation marker. Immunol Today. 1993 Mar;14(3):121-5. doi: 10.1016/0167-5699(93)90212-4. PMID: 7682078.

[93] Kim SW, Kim H, Cho HJ, Lee JU, Levit R, Yoon YS. Human peripheral blood-derived CD31+ cells have robust angiogenic and vasculogenic properties and are effective for treating ischemic vascular disease. J Am Coll Cardiol. 2010 Aug 10;56(7):593-607. doi: 10.1016/j.jacc.2010.01.070. PMID: 20688215; PMCID: PMC2917842.

[94] Patriarca C, Macchi RM, Marschner AK, Mellstedt H. Epithelial cell adhesion molecule expression (CD326) in cancer: a short review. Cancer Treat Rev. 2012 Feb;38(1):68-75. doi: 10.1016/j.ctrv.2011.04.002. Epub 2011 May 14. PMID: 21576002.

[95] Zhang Y, Lu C, Zhang J. Lactoferrin and Its Detection Methods: A Review. Nutrients. 2021 Jul 22;13(8):2492. doi: 10.3390/nu13082492. PMID: 34444652; PMCID: PMC8398339.

[96] Malhotra, Konark, Jeffrey Gornbein, and Jeffrey L. Saver. "Ischemic strokes due to large-vessel occlusions contribute disproportionately to stroke-related dependence and death: a review." *Frontiers in neurology* 8 (2017): 651.

[97] Demidova, L.A. Two-Stage Hybrid Data Classifiers Based on SVM and kNN Algorithms. *Symmetry* **2021**, *13*, 615. https://doi.org/10.3390/sym13040615

[98] Hussain, Jamal, Samuel Lalmuanawma, and Lalrinfela Chhakchhuak. "A two-stage hybrid classification technique for network intrusion detection system." *International journal of computational intelligence systems* 9.5 (2016): 863-875.

[99] Hussain, Jamal, Samuel Lalmuanawma, and Lalrinfela Chhakchhuak. "A novel network intrusion detection system using two-stage hybrid classification

technique." *International Journal of Computer & Communication Engineering Research (IJCCER) Volume* (2015).

[100] Jansi Rani, M., and Durairaj Devaraj. "Two-stage hybrid gene selection using mutual information and genetic algorithm for cancer data classification." *Journal of medical systems* 43.8 (2019): 235.

7. Παράρτημα – Βέλτιστες παράμετροι για τα μοντέλα

Παράρτημα Α – Μέρος 1ο

Στο παρακάτω παράρτημα παρατίθενται οι βέλτιστες τιμές των υπερπαραμέτρων που προέκυψαν για καθένα από τα μοντέλα μηχανικής μάθησης στα οποία εφαρμόστηκε η μέθοδος Grid Search CV όπως αναφέρθηκε εκτενώς παραπάνω, η οποία εξασφαλίζει τη συστηματική αναζήτηση του βέλτιστου συνδυασμού παραμέτρων. Η διαδικασία αυτή αποσκοπούσε στη βέλτιστη απόδοση των μοντέλων για κάθε βιοδείκτη, εξασφαλίζοντας τη μέγιστη αξιοπιστία.

Πίνακας 7.1: Βέλτιστες τιμές υπερπαραμέτρων για το μοντέλο Gradient Boosting

	learning_rate	max_depth	n_estimators
CD45-APC+	1	3	100
CD235a-PE+	0.01	5	200
CD14-PB+	0.01	5	200
CD31-APC	0.01	5	200
CD146-PE	1	3	100
CD326-APC+	0.1	3	100
Lact-FITC+	0.01	5	200

Πίνακας 7.2: Βέλτιστες τιμές υπερπαραμέτρων για το μοντέλο Random Forest

	max_depth	max_features	min_samples_leaf	min_samples_split	n_estimatos
CD45-APC+	None	Sqrt	1	2	200
CD235a-PE+	None	Sqrt	1	2	200
CD14-PB+	None	Sqrt	1	2	200
CD31-APC	None	Sqrt	1	2	200
CD146-PE	None	Sqrt	1	2	100
CD326-APC+	None	Sqrt	1	2	200
Lact-FITC+	None	Sqrt	1	2	50

Πίνακας 7.3: Βέλτιστες τιμές υπερπαραμέτρων για το μοντέλο K-Nearest Neighbors

	n_neighbors	weights
CD45-APC+	20	uniform
CD235a-PE+	10	uniform
CD14-PB+	10	uniform
CD31-APC	3	uniform
CD146-PE	5	uniform
CD326-APC+	20	uniform
Lact-FITC+	3	distance

Πίνακας 7.4: Βέλτιστες τιμές υπερπαραμέτρων για το μοντέλο SVM

	C	kernel	Gamma
CD45-APC+	100	poly	scale
CD235a-PE+	100	rbf	gamma
CD14-PB+	100	rbf	scale
CD31-APC	100	rbf	scale
CD146-PE	100	rbf	scale

CD326-APC+	100	poly	scale
Lact-FITC+	100	rbf	scale

Πίνακας 7.5: Βέλτιστες τιμές για το μοντέλο Gaussian Naïve Bayes

	priors	var_smoothing
CD45-APC+	default=None	default=1e-9
CD235a-PE+	default=None	default=1e-9
CD14-PB+	default=None	default=1e-9
CD31-APC	default=None	default=1e-9
CD146-PE	default=None	default=1e-9
CD326-APC+	default=None	default=1e-9
Lact-FITC+	default=None	default=1e-9

Πίνακας 7.6: Βέλτιστες τιμές υπερπαραμέτρων για το μοντέλο Decision Tree

	max_depth	min_samples_split	min_samples_leaf
CD45-APC+	3	2	1
CD235a-PE+	3	2	4
CD14-PB+	10	2	5
CD31-APC	3	2	1
CD146-PE	10	2	1
CD326-APC+	3	2	10
Lact-FITC+	3	2	4

Πίνακας 7.7: Βέλτιστες τιμές υπερπαραμέτρων για το μοντέλο XG Boost

	max_depth	n_estimators	learning_rate	subsample
CD45-APC+	default = 6	default = 100	default = 0.3	default = 1
CD235a-PE+	default = 6	default = 100	default = 0.3	default = 1
CD14-PB+	default = 6	default = 100	default = 0.3	default = 1
CD31-APC	default = 6	default = 100	default = 0.3	default = 1
CD146-PE	default = 6	default = 100	default = 0.3	default = 1
CD326-APC+	default = 6	default = 100	default = 0.3	default = 1
Lact-FITC+	default = 6	default = 100	default = 0.3	default = 1

Παράρτημα Β – Μέρος 2ο

Πίνακας 7.8: Συγκεντρωτικός πίνακας με τις βέλτιστες παραμέτρους (2ο μέρος)

Βέλτιστες παράμετροι	**Μοντέλα που χρησιμοποιήθηκαν**	
CD45-APC+	**Gradient Boosting**	**Random Forest**
	learning_rate 0.1, max_depth: 3, n_estimators: 100	max_depth: 10, max_features: 'sqrt', min_samples_leaf: 1, min_samples_split: 2, n_estimators: 100
CD235a-PE+	**Random Forest**	**XG Boost**
	max_depth: None max_features: None, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 100	max_depth: default = 6, n_estimator: default = 100, learning_rate: default = 0.3, subsample: default = 1
CD14-PB+	**Random Forest**	**Naïve Bayes**
	max_depth: None, max_features: 'sqrt', min_samples_leaf: 2, min_samples_split: 10, n_estimators: 10	priors: default=None, var_smoothing:default=1e-9
CD31-APC	**Gradient Boosting**	**XG Boost**
	learning_rate: 0.1, max_depth: 4, n_estimators: 50, subsample: 0.8	max_depth: default = 6, n_estimator: default = 100, learning_rate: default = 0.3, subsample: default = 1
CD146-PE	**Gradient Boosting**	**Random Forest**
	learning_rate: 0.1, max_depth: 3, n_estimators: 100	max_depth: None, max_features: None, min_samples_leaf: 1, min_samples_split: 5, n_estimators: 200
CD326-APC+	**Random Forest**	**KNN**
	max_depth: None, max_features: 'sqrt', min_samples_leaf: 1, min_samples_split: 2, n_estimators: 100	knn__n_neighbors: 5, knn__weights: 'uniform'
Lact-FITC+	**Random Forest**	**XG Boost**
	max_depth: None, max_features: None, min_samples_leaf: 2, min_samples_split: 5, n_estimators: 200	max_depth: default = 6, n_estimator: default = 100 learning_rate: default = 0.3, subsample: default = 1

Παράρτημα Γ – Μέρος 4ο

Βέλτιστες παράμετροι	Μοντέλα που χρησιμοποιήθηκαν
CD45-APC+	**Random Forest**
	max_depth: None, max_features: 'sqrt', min_samples_leaf: 2, min_samples_split: 5, n_estimators: 100
CD235a-PE+	**Gradient Boosting**
	n_estimators: 100, learning rate: 0.1, max_depth: 3, min_samples_leaf: 1, max_features: 1
CD14-PB+	**Decision Tree**
	max_depth=5, min_samples_leaf=4, random_state=42
CD31-APC	**XG Boost**
	max_depth: default = 6, n_estimator: default = 100, learning_rate: default = 0.3, subsample: default = 1
CD146-PE	**XG Boost**
	max_depth: default = 6, n_estimator: default = 100, learning_rate: default = 0.3, subsample: default = 1
CD326-APC+	**Gradient Boosting**
	'learning_rate': 0.2, 'max_depth': 4, 'n_estimators': 100, 'subsample': 0.8
Lact-FITC+	**Gradient Boosting**
	'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 150, 'subsample': 1.0