ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΡΟΒΛΕΨΗ ΑΣΘΕΝΕΙΩΝ ΜΕ ΧΡΗΣΗ ΑΛΓΟΡΙΘΜΩΝ ΜΗΧΑΝΙΚΗΣ ΚΑΙ ΒΑΘΙΑΣ ΜΑΘΗΣΗΣ

Διπλωματική Εργασία

ΧΡΗΣΤΟΣ ΤΣΙΑΜΠΑΣ

Επιβλέπουσα: ΕΛΕΝΗ ΤΟΥΣΙΔΟΥ

Ιούλιος 2025

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΡΟΒΛΕΨΗ ΑΣΘΕΝΕΙΩΝ ΜΕ ΧΡΗΣΗ ΑΛΓΟΡΙΘΜΩΝ ΜΗΧΑΝΙΚΗΣ ΚΑΙ ΒΑΘΙΑΣ ΜΑΘΗΣΗΣ

Διπλωματική Εργασία

ΧΡΗΣΤΟΣ ΤΣΙΑΜΠΑΣ

Επιβλέπουσα: ΕΛΕΝΗ ΤΟΥΣΙΔΟΥ

Ιούλιος 2025

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

DISEASE PREDICTION USING MACHINE AND DEEP LEARNING ALGORITHMS

Diploma Thesis

CHRISTOS TSIAMPAS

Supervisor: ELENI TOUSIDOU

July 2025

Εγκρίνεται από την Επιτροπή Εξέτασης:

Επιβλέπουσα **ΕΛΕΝΗ ΤΟΥΣΙΔΟΥ**
ΕΔΙΠ, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος **ΜΙΧΑΗΛ ΒΑΣΙΛΑΚΟΠΟΥΛΟΣ**
Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος **ΑΣΠΑΣΙΑ ΔΑΣΚΑΛΟΠΟΥΛΟΥ**
Αναπληρώτρια Καθηγήτρια, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΠΕΡΙ ΑΚΑΔΗΜΑΪΚΗΣ ΔΕΟΝΤΟΛΟΓΙΑΣ ΚΑΙ ΠΝΕΥΜΑΤΙΚΩΝ ΔΙΚΑΙΩΜΑΤΩΝ

«Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ρητά ότι η παρούσα διπλωματική εργασία, καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας, αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Δηλώνω επίσης ότι τα αποτελέσματα της εργασίας δεν έχουν χρησιμοποιηθεί για την απόκτηση άλλου πτυχίου. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής».

Ο Δηλών

ΧΡΗΣΤΟΣ ΤΣΙΑΜΠΑΣ

Διπλωματική Εργασία

ΠΡΟΒΛΕΨΗ ΑΣΘΕΝΕΙΩΝ ΜΕ ΧΡΗΣΗ ΑΛΓΟΡΙΘΜΩΝ ΜΗΧΑΝΙΚΗΣ ΚΑΙ ΒΑΘΙΑΣ ΜΑΘΗΣΗΣ

ΧΡΗΣΤΟΣ ΤΣΙΑΜΠΑΣ

Περίληψη

Η παρούσα διπλωματική εργασία εξετάζει τη χρήση προηγμένων μεθόδων μηχανικής και βαθιάς μάθησης για την πρόβλεψη και τη διάγνωση σημαντικών ασθενειών, με έμφαση στα καρδιαγγειακά νοσήματα και τον καρκίνο του στόματος.

Για την πρόβλεψη καρδιαγγειακών παθήσεων αξιοποιήθηκαν δύο διαφορετικά σύνολα δομημένων ιατρικών δεδομένων, πάνω στα οποία υλοποιήθηκαν και συγκρίθηκαν αλγόριθμοι όπως η λογιστική παλινδρόμηση, τα τυχαία δάση (*Random Forest*), το *Gradient Boosting* και ο *Naive Bayes*. Τα αποτελέσματα κατέδειξαν ότι οι μέθοδοι *ensemble* υπερτερούν ως προς τις μετρικές ανάκλησης και *F1-score*, ενώ η λογιστική παλινδρόμηση προσφέρει σημαντική ερμηνευσιμότητα.

Επιπλέον, στο πεδίο της διάγνωσης καρκίνου του στόματος μέσω ανάλυσης ιατρικών εικόνων, εφαρμόστηκαν συνελικτικά νευρωνικά δίκτυα (*CNN*) και η αρχιτεκτονική *MobileNetV2*. Οι πειραματικές εφαρμογές έδειξαν ότι τα βαθιά δίκτυα επιτυγχάνουν υψηλά ποσοστά ακρίβειας, ευαισθησίας και βαθμολογίας *AUC*.

Τα συμπεράσματα της εργασίας υπογραμμίζουν τη δυναμική της *τεχνητής νοημοσύνης* στην υποστήριξη της πρόγνωσης και διάγνωσης, αναδεικνύοντας παράλληλα τη σημασία της σωστής αξιολόγησης, της ερμηνευσιμότητας και της συνεργασίας μεταξύ ειδικών στην υλοποίηση τέτοιων συστημάτων στην κλινική πράξη.

Λέξεις-κλειδιά:

Μηχανική Μάθηση, Βαθιά Μάθηση, Καρδιαγγειακά Νοσήματα, Καρκίνος Στόματος

Diploma Thesis

DISEASE PREDICTION USING MACHINE AND DEEP LEARNING ALGORITHMS

CHRISTOS TSIAMPAS

Abstract

This thesis explores the application of advanced *machine learning* and *deep learning* methods for the prediction and diagnosis of major diseases, focusing on cardiovascular diseases and oral cancer.

For cardiovascular risk assessment, two real-world structured medical datasets were analyzed, implementing and comparing algorithms such as *Logistic Regression*, *Random Forest*, *Gradient Boosting*, and *Naive Bayes*. The results demonstrate that *ensemble* methods (*Random Forest*, *Gradient Boosting*) achieve superior performance in recall and *F1-score*, while *Logistic Regression* provides valuable interpretability for medical decision-making. Furthermore, the detection of oral cancer was addressed through image analysis, utilizing deep Convolutional Neural Networks (*CNN*) and transfer learning with the *MobileNetV2* architecture.

Experimental findings highlight the remarkable accuracy, sensitivity, and *AUC scores* achieved by *deep learning* models in distinguishing pathological from healthy tissue in medical images. The study emphasizes the significant potential of *artificial intelligence* as a support tool for disease prognosis and diagnosis, while also discussing the importance of proper model evaluation, interpretability, and interdisciplinary collaboration for real-world clinical adoption. Future research should focus on enhancing model explainability, increasing dataset diversity, and integrating *AI* systems into clinical workflows to maximize their impact on patient care.

Keywords:

Machine Learning, Deep Learning, Cardiovascular Diseases, Oral Cancer

Πίνακας περιεχομένων

Κατάλογος σχημάτων

Κατάλογος πινάκων

Συντομογραφίες

βλ	βλέπε
π.χ.	παράδειγμα
CNN	Συνελικτικό Νευρωνικό Δίκτυο
BMI	Δείκτης Μάζας Σώματος
ROC	Καμπύλη Λειτουργίας Αποδέκτη
AUC	Εμβαδόν Καμπύλης
SGD	Τυχαία Κλίση Κατιούσα
SAM	Βελτιστοποίηση Ευαισθησίας
PSO	Βελτιστοποίηση Σμήνους Σωματιδίων
ANN	Τεχνητό Νευρωνικό Δίκτυο
SMOTE	Τεχνική Υπερδειγματοληψίας Μειοψηφίας
XGBoost	Εξαιρετική Ενίσχυση Βαθμίδας
QDA	Τετραγωνική Διακριτική Ανάλυση
GPU	Μονάδα Επεξεργασίας Γραφικών
TPU	Μονάδα Επεξεργασίας Τανυστών
EHR	Ηλεκτρονικοί Φάκελοι Υγείας
OCT	Οπτική Τομογραφία Συνοχής

Κεφάλαιο 1

Εισαγωγή

1.1 Αντικείμενο της Διπλωματικής Εργασίας

Η συνεχής πρόοδος της Τεχνητής Νοημοσύνης και της Μηχανικής Μάθησης (*Machine Learning*) τα τελευταία χρόνια έχει προσφέρει νέα εργαλεία και μεθόδους στην ιατρική έρευνα και πρακτική. Ιδιαίτερο ενδιαφέρον παρουσιάζει η εφαρμογή αυτών των τεχνολογιών στην πρόβλεψη και διάγνωση σοβαρών ασθενειών, όπως τα καρδιαγγειακά νοσήματα και ο καρκίνος του στόματος. Το αντικείμενο της παρούσας διπλωματικής εργασίας είναι η μελέτη, η υλοποίηση και η συγκριτική αξιολόγηση μοντέλων μηχανικής και βαθιάς μάθησης για την πρόβλεψη ασθενειών, με έμφαση στη χρήση πραγματικών ιατρικών δεδομένων τόσο δομημένων (*structured data*) όσο και απεικονιστικών (*image data*).

Στο πλαίσιο της εργασίας, εξετάζονται δύο διαφορετικές κατηγορίες ασθενειών και δεδομένων: (α) η πρόβλεψη καρδιαγγειακών παθήσεων με χρήση δομημένων δεδομένων ασθενών, και (β) η ανίχνευση/διάγνωση καρκίνου του στόματος μέσω ανάλυσης εικόνων μεθόδων βαθιάς μάθησης. Για κάθε περίπτωση, υλοποιούνται και αξιολογούνται διαφορετικοί αλγόριθμοι, με σκοπό την ανάδειξη των δυνατοτήτων και περιορισμών κάθε προσέγγισης. Μέσω της συγκριτικής ανάλυσης των αποτελεσμάτων, επιχειρείται να αναδειχθούν οι πιο κατάλληλες τεχνολογικές λύσεις για κάθε είδος ιατρικού προβλήματος, καθώς και οι σημαντικότεροι παράγοντες που επηρεάζουν την επίδοση των μοντέλων.

1.2 Διάρθρωση της Εργασίας

Η εργασία είναι οργανωμένη ως εξής:

- **Κεφάλαιο 2**: Παρουσιάζεται η σχετική βιβλιογραφία και αναλύονται προηγούμενες επιστημονικές εργασίες που αφορούν την εφαρμογή τεχνικών μηχανικής μάθησης στην πρόβλεψη καρδιαγγειακών παθήσεων και στη διάγνωση καρκίνου του στόματος με χρήση απεικονιστικών δεδομένων. Γίνεται συγκριτική αναφορά στα αποτελέσματα και τις προκλήσεις που έχουν αναδειχθεί από τη διεθνή έρευνα.

- **Κεφάλαιο 3**: Αναλύεται το θεωρητικό υπόβαθρο που σχετίζεται με τις τεχνικές που αξιοποιήθηκαν στα πειραματικά μέρη. Περιγράφονται οι βασικοί αλγόριθμοι μηχανικής και βαθιάς μάθησης, καθώς και οι μετρικές αξιολόγησης που χρησιμοποιήθηκαν για την αποτίμηση της απόδοσης των μοντέλων.

- **Κεφάλαια 4 & 5**: Παρουσιάζεται αναλυτικά η εφαρμογή των μεθόδων μηχανικής μάθησης για την πρόβλεψη καρδιαγγειακών παθήσεων, με χρήση δύο διαφορετικών συνόλων δομημένων ιατρικών δεδομένων. Σε κάθε κεφάλαιο αναλύονται τα δεδομένα, οι τεχνικές προεπεξεργασίας, τα επιλεγμένα μοντέλα, τα πειραματικά αποτελέσματα και ακολουθεί εκτενής συζήτηση των ευρημάτων.

- **Κεφάλαιο 6**: Περιγράφεται η εφαρμογή μεθόδων βαθιάς μάθησης για τη διάγνωση καρκίνου του στόματος μέσω ανάλυσης απεικονιστικών δεδομένων. Συγκρίνονται διαφορετικές αρχιτεκτονικές νευρωνικών δικτύων (*CNN*, *MobileNetV2*) και παρουσιάζεται η συνολική αποτίμηση των αποτελεσμάτων.

- **Κεφάλαιο 7**: Περιγράφονται τα υπολογιστικά εργαλεία, οι βασικές βιβλιοθήκες και το περιβάλλον υλοποίησης που χρησιμοποιήθηκαν για την ανάλυση, την ανάπτυξη και την αξιολόγηση των μοντέλων της εργασίας.

- **Κεφάλαιο 8**: Συνοψίζονται τα βασικά συμπεράσματα που προκύπτουν από την εργασία και διατυπώνονται προτάσεις για μελλοντική έρευνα και βελτιώσεις.

Κεφάλαιο 2

Προηγούμενες εργασίες

2.1 Πρόσφατες Μελέτες Μηχανικής Μάθησης για Πρόβλεψη Καρδιαγγειακών Παθήσεων

Μία πρόσφατη μελέτη από το Ιράν (Mehrabani-Zeinabad et al., 2023) [1] εφάρμοσε πληθώρα μοντέλων μηχανικής μάθησης σε δεδομένα 5.432 υγιών ενηλίκων από την μακροχρόνια μελέτη *Isfahan Cohort* (16ετής παρακολούθηση). Συγκρίθηκαν κλασικές στατιστικές μέθοδοι (π.χ. λογιστική παλινδρόμηση, διακριτική ανάλυση) με μοντέλα μηχανικής μάθησης όπως *kNN*, *SVM*, δέντρα απόφασης, τυχαία δάση, νευρωνικά δίκτυα και ενισχυτές βαθμίδας (*GBM*). Βρέθηκε ότι η Τετραγωνική Διακριτική Ανάλυση (*QDA*) πέτυχε τη μεγαλύτερη ακρίβεια (75,5%) αλλά με χαμηλή ευαισθησία (49,8%), ενώ αντίστροφα ένα απλό δέντρο απόφασης είχε χαμηλότερη ακρίβεια (51,9%) αλλά υψηλότερη ευαισθησία (82,5%). Ένα υβριδικό μοντέλο *Bayesian Additive Regression Trees* με ενσωματωμένο χειρισμό ελλειπουσών τιμών (*BARTm*) πέτυχε ενδιάμεση απόδοση (≈69,5% ακρίβεια, 54% ευαισθησία) χωρίς ανάγκη προεπεξεργασίας. Οι συγγραφείς κατέληξαν ότι η ανάπτυξη τοπικών μοντέλων πρόβλεψης ανά γεωγραφική περιοχή είναι χρήσιμη για την έγκαιρη ανίχνευση υψηλού κινδύνου, ενώ ο συνδυασμός κλασικών στατιστικών και τεχνικών μηχανικής μάθησης μπορεί να αξιοποιήσει τα πλεονεκτήματα και των δύο προσεγγίσεων. Συγκεκριμένα, η Τετραγωνική Διακριτική Ανάλυση αναδείχθηκε αξιόπιστη για γρήγορη και σταθερή πρόβλεψη με χαμηλό υπολογιστικό κόστος, ενώ το μοντέλο *BARTm* πρόσφερε ευελιξία χειριζόμενο αποτελεσματικά τα ελλιπή δεδομένα.

Μια άλλη μελέτη (Alwakid et al., 2025) [2] πρότεινε ένα βελτιστοποιημένο πλαίσιο μη-

χανικής μάθησης με έμφαση τόσο στην απόδοση όσο και στην ηθική διάσταση. Χρησιμοποιήθηκαν τέσσερα δημόσια σύνολα δομημένων δεδομένων καρδιοπαθειών (βάσεις *Cleveland*, *Hungarian*, *Switzerland*, *Long Beach* από την αποθήκη *UCI*) με συνολικά 918 δείγματα (μετά από ενοποίηση και απαλοιφή διπλοτύπων) και 11 χαρακτηριστικά. Εφαρμόστηκαν προηγμένες τεχνικές επιλογής χαρακτηριστικών (π.χ. *Chi-square*, *Information Gain*, *Forward/Backward Selection*) ώστε να μειωθούν σε ένα υποσύνολο 8 σημαντικών μεταβλητών. Στη συνέχεια συγκρίθηκαν μοντέλα όπως *Logistic Regression*, δέντρα απόφασης, *Random Forest*, *XGBoost* κ.ά. – με κορυφαίο τον *XGBoost*. Στο σύνολο των 8 χαρακτηριστικών, ο *XGBoost* πέτυχε εξαιρετικά αποτελέσματα (ακρίβεια 99%, ανάκληση 100%, *F1* 99%, *ευστοχία* 99%). Ακόμη και μετά από μείωση διαστασιμότητας (*PCA* σε 6 χαρακτηριστικά) ο *XGBoost* διατήρησε πολύ υψηλή επίδοση (98% ακρίβεια, 100% ανάκληση). Τα ευρήματα υπογραμμίζουν ότι ένα κατάλληλα επιλεγμένο σύνολο χαρακτηριστικών μπορεί να οδηγήσει σε σχεδόν άριστη πρόβλεψη σε κλασικά δεδομένα καρδιάς. Οι συγγραφείς όμως τονίζουν τη σημασία της ερμηνευσιμότητας και της γενίκευσης: παρά τις υψηλές μετρικές, το μικρό σχετικό μέγεθος δείγματος περιορίζει την εξαγωγή συμπερασμάτων για άλλους πληθυσμούς.

Σε δημοσίευση στο *Scientific Reports* (Rehman et al., 2025) [3], προτάθηκε ένα υβριδικό μοντέλο που συνδυάζει *Σωματιδιακή Βελτιστοποίηση Σμήνους (Particle Swarm Optimization, PSO)* με *Τεχνητό Νευρωνικό Δίκτυο (Artificial Neural Network, ANN)* για την πρόβλεψη στεφανιαίας νόσου. Χρησιμοποιήθηκαν δεδομένα κλινικών χαρακτηριστικών από ηλεκτρονικούς φακέλους υγείας (π.χ. από τη μελέτη *NHANES*) καθώς και από το σύνολο δεδομένων καρδιοπαθειών του *UCI* για εκπαίδευση και αξιολόγηση. Το προτεινόμενο πλαίσιο περιλάμβανε βήματα επιλογής μεταβλητών με αμοιβαία πληροφορία (*Mutual Information, MI*) και αντιμετώπισης ανισορροπίας κλάσεων με υπερδειγματοληψία *SMOTE*. Οι συγγραφείς συνέκριναν αρχικά συμβατικούς ταξινομητές (όπως λογιστική παλινδρόμηση, *Random Forest*, *kNN*) και βρήκαν ότι μπορούν να επιτύχουν ακρίβεια 95,8%. Ωστόσο, το προτεινόμενο μοντέλο *PSO-ANN* βελτίωσε περαιτέρω την επίδοση, επιτυγχάνοντας ακρίβεια έως 97% στην πρόβλεψη καρδιαγγειακού κινδύνου. Η βελτίωση αυτή καταδεικνύει ότι η έξυπνη βελτιστοποίηση των βαρών και η προσαρμογή του μοντέλου στα πιο σημαντικά χαρακτηριστικά μπορεί να υπερέχει των παραδοσιακών προσεγγίσεων. Το συμπέρασμα ήταν ότι η ενσωμάτωση τεχνικών προεπεξεργασίας όπως επιλογή χαρακτηριστικών (*feature selection*) και *SMOTE* με ένα υβριδικό μοντέλο μηχανικής μάθησης μπορεί να προσφέρει μεγαλύτερη ακρίβεια στην πρόγνωση σε σύγκριση με μοντέλα όπως η λογιστική παλινδρόμηση. Παράλληλα, τονίστηκε

η ανάγκη για περαιτέρω αξιολόγηση σε πραγματικά κλινικά περιβάλλοντα, ώστε να διασφαλιστεί ότι αυτά τα πιο πολύπλοκα μοντέλα πράγματι προσφέρουν ουσιαστικό όφελος έναντι των απλούστερων, ιδίως ως προς την κλινική αξιοπιστία και γενικευσιμότητα των αποτελεσμάτων.

Μια ακόμη σύγχρονη εργασία (Shah et al., 2025) [4] πρότεινε ένα υβριδικό μοντέλο συνόλου (*ensemble*) με έμφαση στην εξηγησιμότητα των αποτελεσμάτων. Οι ερευνητές συνένωσαν τρία μεγάλα δημόσια σύνολα δεδομένων (συμπεριλαμβανομένου ενός συνόλου δεδομένων 70.000 ασθενών από το *IEEE Dataport* και των κλασικών *Cleveland* και *Hungarian*) σε ένα ενιαίο σύνολο δεδομένων 70 χιλιάδων εγγραφών με 12 χαρακτηριστικά ανά ασθενή. Στο μοντέλο τους, συνδύασαν σύγχρονους αλγορίθμους ενίσχυσης και νευρωνικά δίκτυα: συγκεκριμένα, το *ensemble* χρησιμοποίησε ως βασικούς μαθητές διαφόρους ταξινομητές (*Gradient Boosting*, *CatBoost*, *LightGBM*, *SVM*, *Neural Network*) και ως μετα-εκπαιδευτή (*meta-learner*) έναν *XGBoost* που συνδύαζε τις προβλέψεις τους. Το υβριδικό αυτό μοντέλο πέτυχε *AUC-ROC* = 0,82 και ισορροπημένη απόδοση στους δύο κλάδους, με ακρίβεια 81%, ευαισθησία 83%, ειδικότητα 81% και F_1*-score* 82% στο *test set*. Οι μετρικές αυτές ήταν βελτιωμένες σε σχέση με κάθε επιμέρους μοντέλο, αναδεικνύοντας τη δύναμη των μεθόδων *ensemble*. Επιπλέον, μέσω μεθόδων *Explainable AI* (*SHAP values*, *t-SNE*, *PCA*) ανέλυσαν την επίδραση κάθε παράγοντα κινδύνου στο τελικό αποτέλεσμα. Για παράδειγμα, παράγοντες όπως η συστολική/διαστολική πίεση, ο Δείκτης Μάζας Σώματος (*BMI*) και ο λόγος χοληστερόλης-γλυκόζης αναγνωρίστηκαν οπτικά ως σημαντικοί για την ταξινόμηση. Οι συγγραφείς καταλήγουν ότι τέτοια ερμηνεύσιμα μοντέλα *ensemble* μπορούν να βελτιώσουν την ακρίβεια της πρόγνωσης σε πολύπλοκα ιατρικά δεδομένα, ενώ παράλληλα παρέχουν διαφάνεια που αυξάνει την αποδοχή τους από τους κλινικούς (σημαντικό για την αξιοπιστία της μηχανικής μάθησης στην υγεία).

Τέλος, μια συστηματική ανασκόπηση και μετα-ανάλυση (Liu et al., 2024) [5] εξέτασε συγκεντρωτικά 20 μελέτες πρόγνωσης καρδιαγγειακού κινδύνου με μηχανική μάθηση, χρησιμοποιώντας δομημένα δεδομένα ηλεκτρονικών φακέλων υγείας (*EHR*). Διαπιστώθηκε ότι τα μοντέλα μηχανικής μάθησης –ιδίως τα *Random Forest* και βαθιά νευρωνικά δίκτυα– ξεπερνούν στατιστικά σημαντικά τα παραδοσιακά σκορ κινδύνου (όπως *QRISK3*) ως προς την προγνωστική τους ικανότητα. Συγκεκριμένα, το συνολικό *AUC* των καλύτερων μοντέλων *μηχανικής μάθησης* έφτανε το 0,85–0,87, έναντι 0,76 των συμβατικών μοντέλων σε 5-10ετή πρόβλεψη κινδύνου. Παρότι αυτό υποδηλώνει ενισχυμένη διακριτική ικανότητα και καλύ-

τερη βαθμονόμηση στην πρόγνωση *CVD*, οι ερευνητές εντόπισαν υψηλή ετερογένεια μεταξύ των μελετών ($I^2 > 99\%$) και πιθανή μεροληψία δημοσίευσης. Το συμπέρασμα ήταν ότι, ενώ τα μοντέλα μηχανικής μάθησης δείχνουν σημαντικές δυνατότητες βελτίωσης στην πρόβλεψη καρδιαγγειακού κινδύνου, απαιτείται πιο αυστηρή τυποποίηση, διαφάνεια και ανεξάρτητη επικύρωση προτού ενταχθούν στην κλινική πράξη. Η μελέτη αυτή συνοψίζει εύγλωττα ότι τα προηγμένα μοντέλα μηχανικής μάθησης μπορούν να προσφέρουν ακριβέστερες προβλέψεις (π.χ. υψηλότερα *AUC*), αλλά η μεθοδολογική ποιότητα και η αξιοπιστία τους πρέπει να βελτιωθούν για ευρεία κλινική αξιοποίηση.

2.2 Πρόσφατες Έρευνες Μηχανικής Μάθησης για Διάγνωση Στοματικού Καρκίνου

Μια πρώτη μελέτη [6] αξιοποίησε κλινικές φωτογραφίες στόματος (*in-vivo* εικόνες) για αυτοματοποιημένο εντοπισμό στοματικού καρκίνου. Συγκεντρώθηκαν 700 ψηφιακές φωτογραφίες από κλινικές περιπτώσεις, εκ των οποίων οι 350 απεικόνιζαν πλακώδες καρκίνωμα του στόματος (*OSCC*) και οι υπόλοιπες 350 φυσιολογικό στοματικό βλεννογόνο. Οι ερευνητές εκπαίδευσαν ένα συνελικτικό νευρωνικό δίκτυο *DenseNet-121* για ταξινόμηση (διάκριση καρκινικών vs φυσιολογικών ιστών) και ένα δίκτυο *Faster R-CNN* για εντοπισμό περιοχών βλάβης μέσα στις φωτογραφίες. Το σύνολο δεδομένων χωρίστηκε σε 490 εικόνες για εκπαίδευση, 70 για επικύρωση και 140 για δοκιμή. Το μοντέλο *DenseNet-121* πέτυχε εξαιρετικές επιδόσεις, με ακρίβεια 99%, ευαισθησία 100%, *F1-score* 99%, ειδικότητα 100% και εμβαδόν *ROC* (*AUC*) 0.99 στο δοκιμαστικό σύνολο. Αντίστοιχα, το μοντέλο εντοπισμού *Faster R-CNN* πέτυχε *ευστοχία* 76,7%, *ευαισθησία* 82,1%, *F1-score* 79% και μέση ευστοχία περίπου 0,79 για την ανίχνευση των καρκινικών βλαβών. Οι συγγραφείς καταλήγουν στο συμπέρασμα ότι ο συνδυασμός του *DenseNet-121* και του *Faster R-CNN* παρουσιάζει υψηλή ακρίβεια και μπορεί να υποστηρίξει αποτελεσματικά τον αυτόματο έλεγχο και την έγκαιρη ανίχνευση του καρκίνου του στόματος μέσω απλών φωτογραφιών.

Μια δεύτερη μελέτη [7] εστίασε σε ιστολογικές εικόνες (ψηφιοποιημένες τομές ιστού) για τη διάγνωση του στοματικού καρκινώματος από παθολογοανατόμους. Οι ερευνητές δημιούργησαν ένα σύνολο 7.918 εικόνων υψηλής ανάλυσης από βιοψίες στοματικού ιστού, το οποίο περιλάμβανε 989 φυσιολογικές, 1.167 καρκινικές (*OSCC*) και 5.762 άλλες (π.χ.

αντιδραστικές ή προκαρκινικές) περιοχές, επιμελώς επισημασμένες από τουλάχιστον δύο ειδικούς στοματοπαθολόγους. Για την ταξινόμηση χρησιμοποιήθηκαν δύο γνωστά προεκπαιδευμένα μοντέλα *CNN* – το *VGG16* και το *ResNet-50* – με πειραματισμό σε διαφορετικές ρυθμίσεις βελτιστοποίησης (σταδιακή μείωση του ρυθμού μάθησης, αλγόριθμος *SGD* με *momentum* vs. εξελιγμένος βελτιστοποιητής *Sharpness-Aware Minimization (SAM)*). Το καλύτερο αποτέλεσμα επετεύχθη με το μοντέλο *VGG16* σε συνδυασμό με *SAM*, το οποίο πέτυχε ακρίβεια 86,2% και *ROC-AUC* 0,96 στην πολυκλασική ταξινόμηση (διάκριση μεταξύ φυσιολογικού ιστού, *OSCC* και άλλων αλλοιώσεων). Αξίζει να σημειωθεί ότι, όταν τα αποτελέσματα του καλύτερου μοντέλου παρουσιάστηκαν σε έξι στοματοπαθολόγους ως βοήθημα διάγνωσης, η διαγνωστική τους ακρίβεια βελτιώθηκε σημαντικά (p=0.031), υποδεικνύοντας ότι η συνδρομή ενός αξιόπιστου συστήματος βαθιάς μάθησης μπορεί να ενισχύσει την απόδοση των ειδικών κατά τη διάγνωση ιστολογικών δειγμάτων. Οι συγγραφείς επισημαίνουν ότι η ενσωμάτωση τέτοιων μοντέλων ως εργαλεία διπλού ελέγχου θα μπορούσε να μειώσει τα διαγνωστικά σφάλματα και να αυξήσει την ταχύτητα και αξιοπιστία στην παθολογική αξιολόγηση του στοματικού καρκίνου.

Σε μία τρίτη εργασία [8] διερευνήθηκε η χρήση της οπτικής τομογραφίας συνοχής (*OCT*) για τον εντοπισμό στοματικών καρκινωμάτων με μεθόδους βαθιάς μάθησης. Συγκεντρώθηκε ένα νέο σύνολο *OCT* δεδομένων από 19 ασθενείς του *Tianjin Stomatology Hospital* (Κίνα), που περιλάμβανε εικόνες φυσιολογικού στοματικού βλεννογόνου, δυνητικά κακοήθων βλαβών (λευκοπλακία) και στοματικού καρκίνου (*OSCC*). Για την ταξινόμηση των *OCT* τομών εφαρμόστηκαν τρία είδη *CNN* αρχιτεκτονικής – το απλό *LeNet-5*, το βαθύτερο *VGG16* και ένα *ResNet-18* – ώστε να συγκριθεί η απόδοσή τους. Οι αλγόριθμοι εκπαιδεύτηκαν με δεκαπλή διασταυρούμενη επικύρωση και αξιολογήθηκαν με δείκτες όπως ακρίβεια, ευαισθησία, ειδικότητα και θετική προγνωστική αξία. Τα αποτελέσματα έδειξαν ότι το σχετικά μικρό και λιγότερο περίπλοκο δίκτυο *LeNet-5* παρουσίασε την υψηλότερη ακρίβεια στη διάκριση των κατηγοριών: συνολική ακρίβεια 96,8%, υπερβαίνοντας τα *VGG16* (91,9%) και *ResNet-18* (90,4%). Επίσης, επιχειρήθηκε συνδυασμός *CNN* με κλασικούς ταξινομητές μηχανικής μάθησης (*SVM*, *Random Forest*), όμως ο συνδυασμός αυτός δεν ξεπέρασε την απόδοση του αυτόνομου *CNN* (η καλύτερη υβριδική προσέγγιση, *LeNet-5 + SVM*, πέτυχε 92,5% ακρίβεια). Επιπλέον, μέσω τεχνικών οπτικοποίησης (π.χ. χάρτες ενεργοποίησης τύπου *Grad-CAM*), αναδείχθηκαν οι περιοχές των *OCT* εικόνων που συνεισφέρουν περισσότερο στην απόφαση του δικτύου, βελτιώνοντας έτσι τη διαφάνεια και ερμηνευσιμότητα

του συστήματος. Συνολικά, η μελέτη καταλήγει στο ότι η αυτοματοποιημένη ανάλυση *OCT* με βαθιά μάθηση έχει τη δυναμική να προσφέρει ένα μη επεμβατικό εργαλείο υποβοήθησης για τον προληπτικό έλεγχο και τη διάγνωση του καρκίνου του στόματος, με πολύ υψηλή ακρίβεια.

Μία τέταρτη μελέτη [9] εξετάζει μια καινοτόμο προσέγγιση "πραγματικού χρόνου" για τη διάγνωση του στοματικού καρκίνου, αξιοποιώντας εικόνες *in vivo* συνεστιακής μικροσκοπίας λέιζερ υψηλής ανάλυσης από ζωντανό ιστό. Συνολικά 59 ασθενείς με στοματικές βλάβες υποβλήθηκαν σε απεικόνιση με φορητό ενδοσκόπιο συνεστιακής μικροσκοπίας, χρησιμοποιώντας δύο χρωστικές αντίθεσης (ακρυφλαβίνη και φλουορεσκεΐνη) για την ανάδειξη της μορφολογίας των πυρήνων και των ιστών. Από τους ασθενείς αυτούς συλλέχθηκαν 9.168 καρέ εικόνων σε πραγματικό χρόνο, τα οποία κατηγοριοποιήθηκαν ιστοπαθολογικά σε τέσσερις διαγνωστικές κλάσεις: χωρίς δυσπλασία, *lichenoid* αλλοίωση, δυσπλασία χαμηλού βαθμού και δυσπλασία υψηλού βαθμού/*OSCC*. Οι ερευνητές εκπαίδευσαν ένα σύστημα από τρία διαδοχικά μοντέλα *CNN* βασισμένα σε προεκπαιδευμένο *Inception-V3* (με μεταφορά μάθησης) για την αυτοματοποίηση της διαδικασίας: το πρώτο δίκτυο φιλτράρει τις θολές ή ανεπαρκούς ποιότητας εικόνες, ενώ τα επόμενα δύο δίκτυα εξειδικεύονται στην ταξινόμηση των καθαρών εικόνων ανάλογα με τη χρωστική (ξεχωριστό μοντέλο για εικόνες φλουορεσκεΐνης και για ακρυφλαβίνης). Το μοντέλο ποιότητας εικόνας απέκλεισε αποτελεσματικά τα ακατάλληλα καρέ με ακρίβεια 89,5%. Στις εικόνες με ακρυφλαβίνη, το διαγνωστικό *CNN* πέτυχε υψηλή απόδοση στην αναγνώριση *lichenoid* βλαβών (*AUC*=0,94) και δυσπλασίας χαμηλού βαθμού (*AUC*=0,91), αλλά δυσκολεύτηκε στη διάκριση των εντελώς φυσιολογικών περιοχών (*AUC*=0,44) και των σοβαρών δυσπλασιών/*καρκινωμάτων* (*AUC*=0,28). Αντίθετα, το μοντέλο για τις εικόνες με φλουορεσκεΐνη παρουσίασε άριστη διακριτική ικανότητα σε όλες τις κατηγορίες, με *AUC* περίπου 0,90–0,96. Σημαντικό είναι ότι η όλη διαδικασία ταξινόμησης είναι εξαιρετικά γρήγορη – λιγότερο από 0,1 δευτερόλεπτο ανά εικόνα – καθιστώντας την κατάλληλη για άμεση κλινική αξιοποίηση κατά την εξέταση του ασθενούς. Οι συγγραφείς καταλήγουν ότι ένας τέτοιος συνδυασμός διαδοχικών *CNN* μπορεί να λειτουργήσει ως ένα πολύ ακριβές και ταχύ σύστημα διαλογής (*triage*) των ενδοστοματικών μικροσκοπικών εικόνων, βοηθώντας τους κλινικούς να εντοπίζουν επιτόπου τις υψηλού κινδύνου βλάβες και να βελτιώσουν την έγκαιρη διάγνωση της στοματικής δυσπλασίας και του καρκίνου.

Σε μία πέμπτη μελέτη [10] διερευνάται η εφαρμογή σύγχρονων αρχιτεκτονικών *Vision Transformer* για την ανίχνευση του καρκίνου του στόματος σε κλινικές φωτογραφίες. Συλλέ-

χθηκαν 1.406 φωτογραφίες από ασθενείς με στοματικές βλάβες (αρχείο κλινικής *Oral , Maxillofacial Surgery, Charité* – Παν. Βερολίνου), οι οποίες χειροκίνητα σημειώθηκαν και ταξινομήθηκαν (ετικέτες αναφοράς από ειδικούς) ως καρκινικές ή μη. Οι ερευνητές εκπαίδευσαν ένα μοντέλο βαθιάς μάθησης βασισμένο στον *Swin Transformer*, το οποίο εκπαιδεύτηκε σε 1.265 εικόνες και αξιολογήθηκε σε 141 εικόνες ως ανεξάρτητο τεστ. Το μοντέλο *Vision Transformer* πέτυχε εξαιρετικά υψηλή ακρίβεια στην ανίχνευση του πλακώδους καρκινώματος: συγκεκριμένα, ακρίβεια 98,6% και *ROC-AUC* 0,99 στην ανάλυση του δοκιμαστικού συνόλου δεδομένων. Η επίδοση αυτή υπερέβη κατά πολύ τα ιστορικά επίπεδα διαγνωστικής ακρίβειας που αναφέρονται για μη ειδικούς κλινικούς (ευαισθησία 57,8%, ειδικότητα 31–53% χωρίς βοήθημα). Οι συγγραφείς επισημαίνουν ότι η αρωγή των κλινικών από τέτοια μοντέλα τεχνητής νοημοσύνης μπορεί να βελτιώσει σημαντικά τον ρυθμό πρώιμης ανίχνευσης του στοματικού καρκίνου, συμβάλλοντας έτσι στη βελτίωση της πρόγνωσης, της επιβίωσης και της ποιότητας ζωής των ασθενών. Επιπλέον, η χρήση αυτοματοποιημένης υποβοήθησης διάγνωσης δύναται να ενισχύσει την αξιοπιστία της εκτίμησης από γενικούς οδοντιάτρους ή ιατρούς πρωτοβάθμιας φροντίδας, οδηγώντας σε πιο έγκαιρες παραπομπές των ύποπτων περιστατικών για εξειδικευμένο έλεγχο.

Κεφάλαιο 3

Θεωρητικό υπόβαθρο

3.1 Logistic Regression

Η Λογιστική Παλινδρόμηση (*Logistic Regression*) [11] είναι μια δημοφιλής μέθοδος ταξινόμησης που χρησιμοποιεί τη λογιστική συνάρτηση (*sigmoid*) για να μοντελοποιήσει την πιθανότητα δυαδικών εκβάσεων. Συγκεκριμένα, υπολογίζει την πιθανότητα το δείγμα να ανήκει στην «θετική» κλάση (π.χ. ασθενής με νόσο) ως συνάρτηση μιας γραμμικής πρόσθεσης των χαρακτηριστικών, εφαρμοσμένης σε μια σιγμοειδή συνάρτηση ώστε η τιμή να κυμαίνεται στο [0,1].

Η πρόβλεψη γίνεται συγκρίνοντας την πιθανότητα αυτή με ένα κατώφλι (συνήθως 0.5) για να αποδοθεί ετικέτα. Η εκπαίδευση του μοντέλου βασίζεται στη μεγιστοποίηση της λογιστικής συνάρτησης πιθανοφάνειας (*log-likelihood*) ώστε να βρεθούν τα βέλτιστα βάρη β.

Η λογιστική παλινδρόμηση έχει ευρεία χρήση, μεταξύ άλλων στην ιατρική στατιστική, για την εκτίμηση του κινδύνου ασθένειας βάσει παραγόντων κινδύνου ή βιοδεικτών. Ένα βασικό προτέρημά της είναι ότι τα βάρη του μοντέλου μπορούν να ερμηνευθούν ως λογαριθμικοί λόγοι πιθανοτήτων, επιτρέποντας την κατανόηση της επίδρασης κάθε χαρακτηριστικού στην πιθανότητα εκδήλωσης του συμβάντος..

Η απλότητα και η ισχυρή θεμελίωση της μεθόδου την έχουν καταστήσει κλασσική επιλογή σε ποικίλα προβλήματα ιατρικής έρευνας, όπου έχει χρησιμοποιηθεί με επιτυχία για την περιγραφή φαινομένων και την πρόβλεψη δυαδικών εκβάσεων. Πράγματι, τα μοντέλα λογιστικής παλινδρόμησης θεωρούνται ευέλικτα, με ισχυρή ερμηνευσιμότητα, και εφαρμόζονται σε πολλά πεδία συμπεριλαμβανομένης της ιατρικής επιδημιολογίας και των κλινικών μελετών. Επιπλέον, μπορούν να επεκταθούν σε πολυκατηγοριακά προβλήματα (π.χ. πολυω-

νυμική λογιστική παλινδρόμηση) για περισσότερες από δύο κατηγορίες.

Παρά τα πλεονεκτήματά της, η λογιστική παλινδρόμηση έχει ορισμένους περιορισμούς:

- Υποθέτει ότι υπάρχει γραμμική συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών και του λογαρίθμου των *odds (log-odds)* του αποτελέσματος, κάτι που συχνά δεν ισχύει ακριβώς στον πραγματικό κόσμο.
- Δεν αποδίδει καλά σε προβλήματα όπου η σχέση χαρακτηριστικών-εκβάσεων είναι έντονα μη γραμμική, εκτός αν προηγηθούν κατάλληλοι μετασχηματισμοί ή προσθήκη πολυωνυμικών όρων.
- Κατασκευάζει έναν γραμμικό αποφαντικό κανόνα – αν τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα στον χώρο χαρακτηριστικών, η απόδοσή της μειώνεται σημαντικά.
- Η παρουσία πολλών υψηλά συσχετισμένων ή άσχετων μεταβλητών (πρόβλημα πολυσυγγραμμικότητας) μπορεί να οδηγήσει σε ασταθείς εκτιμήσεις των συντελεστών.
- Σε σύνολα δεδομένων με περισσότερες μεταβλητές από παρατηρήσεις, το μοντέλο τείνει να υπερπροσαρμόζεται (*overfitting*) αν δεν χρησιμοποιηθούν τεχνικές κανονικοποίησης *L1/L2* για περιορισμό των συντελεστών.

Παρ' όλα αυτά, με προσεκτική επιλογή χαρακτηριστικών και κανονικοποίηση, η λογιστική παλινδρόμηση παρουσιάζει καλή ακρίβεια σε πολλά ιατρικά δεδομένα και συχνά προτιμάται για την απλότητα και διαφάνειά της. Μάλιστα, σε πολλές μελέτες κλινικών προβλέψεων, έχει βρεθεί ότι πιο πολύπλοκοι αλγόριθμοι μηχανικής μάθησης δεν προσφέρουν σημαντικά υψηλότερη ακρίβεια από ένα καλά προσαρμοσμένο μοντέλο λογιστικής παλινδρόμησης, γεγονός που υπογραμμίζει την αξία της ως σημείου αναφοράς για προβλήματα ιατρικής ταξινόμησης.

3.2 Random Forest

Το μοντέλο *Random Forest* [12] είναι μια μέθοδος *ensemble* μάθησης που βασίζεται σε ένα μεγάλο πλήθος δέντρων απόφασης για την πραγματοποίηση προβλέψεων. Κάθε δέντρο απόφασης στο σύνολο εκπαιδεύεται σε διαφορετικό τυχαίο δείγμα των δεδομένων (με επαναδειγματοληψία *Bootstrap*) και σε κάθε διακλάδωση χρησιμοποιείται τυχαίο υποσύνολο χαρακτηριστικών, ώστε τα δέντρα να διαφοροποιούνται μεταξύ τους.

Κατά τη φάση πρόβλεψης, όλα τα δέντρα του δάσους ψηφίζουν και η επικρατούσα τάξη αποτελεί την τελική πρόβλεψη (πλειοψηφική απόφαση).Αυτή η προσέγγιση βελτιώνει την ακρίβεια σε σχέση με ένα μεμονωμένο δέντρο, μειώνοντας τη διασπορά και τον κίνδυνο υπερπροσαρμογής μέσω του μέσου όρου των αποτελεσμάτων πολλαπλών μοντέλων.

Βασικά πλεονεκτήματα του *Random Forest*:

- Μειωμένη τάση υπερπροσαρμογής σε σχέση με ένα μοναδικό δέντρο.
- Δυνατότητα εκτίμησης της σημασίας των χαρακτηριστικών (*feature importance*).
- Καλή απόδοση σε δεδομένα με πολλά χαρακτηριστικά και πολύπλοκες σχέσεις.
- Χειρισμός ελλιπών τιμών μέσω τεχνικών τυχαίας επαναδειγματοληψίας (*bagging*).

Περιορισμοί:

- Μεγαλύτερο υπολογιστικό κόστος σε σχέση με απλούστερα μοντέλα.
- Λειτουργεί ως μαύρο κουτί από άποψη ερμηνευσιμότητας, περιορίζοντας την αιτιολόγηση αποφάσεων.

Παρά τους περιορισμούς αυτούς, το *Random Forest* αποτελεί μια από τις πιο ευέλικτες και ισχυρές μεθόδους μηχανικής μάθησης, ιδιαίτερα δημοφιλής στον ιατρικό χώρο για την ταξινόμηση σύνθετων και πολυδιάστατων δεδομένων.

3.3 Gradient Boosting

Η μέθοδος *Gradient Boosting* [13] αποτελεί μια εξελιγμένη μέθοδο *ensemble* όπου μια σειρά από απλούς «ασθενείς» ταξινομητές (συνήθως δέντρα απόφασης μικρού βάθους) εκπαιδεύονται διαδοχικά. Κάθε νέος ταξινομητής εκπαιδεύεται ώστε να διορθώνει τα σφάλματα του συνδυαστικού μοντέλου που προηγήθηκε. Συγκεκριμένα, το μοντέλο *Gradient Boosting* όπως προτάθηκε από τον Friedman, χρησιμοποιεί την τεχνική της βαθμιαίας βελτίωσης μέσω μείωσης του βαθμολογικού σφάλματος: σε κάθε βήμα υπολογίζονται τα υπολειπόμενα σφάλματα του τρέχοντος μοντέλου και εκπαιδεύεται ένα νέο δέντρο ώστε να προβλέψει αυτά τα υπολείμματα. Το τελικό μοντέλο προκύπτει αθροίζοντας τις προβλέψεις όλων των επιμέρους δέντρων.

Με αυτό τον τρόπο, το σύστημα μαθαίνει πολύπλοκα πρότυπα στα δεδομένα και βελτιώνει προοδευτικά την ακρίβειά του ελαχιστοποιώντας μια προκαθορισμένη συνάρτηση κόστους (π.χ. το *log-loss* για ταξινόμηση). Δημοφιλείς υλοποιήσεις της μεθόδου είναι τα *XGBoost*, *LightGBM* και *CatBoost*, που έχουν βελτιστοποιηθεί για μεγάλη ταχύτητα και απόδοση. Οι αλγόριθμοι αυτοί επιδεικνύουν εξαιρετικές επιδόσεις σε δομημένα σύνολα δεδομένων, ξεπερνώντας συχνά άλλες μεθόδους σε διαγωνισμούς πρόβλεψης. Πράγματι, με κατάλληλη ρύθμιση υπερπαραμέτρων, τα μοντέλα αυτά μπορούν να συλλάβουν μη γραμμικές σχέσεις και αλληλεπιδράσεις χαρακτηριστικών πολύ αποτελεσματικά, καθιστώντας τα ικανά να πετυχαίνουν υψηλή ακρίβεια σε απαιτητικά προβλήματα ταξινόμησης. Στον χώρο της ιατρικής, έχουν αναδειχθεί ως κορυφαίες προσεγγίσεις για την πρόβλεψη κλινικών εκβάσεων από δεδομένα ασθενών, καθώς συνδυάζουν υψηλή ακρίβεια με σχετικά χαμηλό υπολογιστικό κόστος σε σχέση με βαθιά νευρωνικά δίκτυα.

Τα βασικά πλεονεκτήματα των μεθόδων *Gradient Boosting* περιλαμβάνουν:

- Πολύ υψηλή προβλεπτική ισχύ και ικανότητα μοντελοποίησης πολύπλοκων συναρτήσεων.
- Κάθε νέο μοντέλο διορθώνει τα λάθη των προηγούμενων, προσαρμόζοντας το σύνολο σε λεπτά πρότυπα.
- Σύγχρονες υλοποιήσεις με τεχνικές κανονικοποίησης (μικρός ρυθμός μάθησης, περιορισμένο βάθος δέντρων, υποδειγματοληψία χαρακτηριστικών) περιορίζουν τον κίνδυνο υπερπροσαρμογής και βελτιώνουν τη γενίκευση.
- Άριστη ισορροπία μεταξύ προκατάληψης και διασποράς, συχνά κορυφαία επιλογή σε διαγωνισμούς με δεδομένα πίνακα.
- Χρειάζεται μικρότερη υπολογιστική ισχύ σε σχέση με βαθιά δίκτυα για συγκρίσιμη ακρίβεια.
- Αποτελεσματική εκπαίδευση ακόμη και σε μεσαίου μεγέθους δείγματα.

3.4 Naive Bayes

Οι ταξινομητές *Naive Bayes* [14] στηρίζονται στο θεώρημα του Bayes για την κατά Bayes ενημέρωση της πιθανότητας μιας κλάσης δεδομένων των χαρακτηριστικών ενός δείγματος,

υπό την απλοποιητική υπόθεση ότι τα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους.

Ο χαρακτηρισμός «αφελής» προέρχεται ακριβώς από αυτή την ισχυρή υπόθεση ανεξαρτησίας: θεωρεί ότι η συμβολή κάθε χαρακτηριστικού στην πιθανότητα της κλάσης είναι ανεξάρτητη από τα άλλα χαρακτηριστικά. Παρά το γεγονός ότι αυτή η υπόθεση συχνά δεν ισχύει απολύτως (τα χαρακτηριστικά συνήθως παρουσιάζουν κάποια συσχέτιση), το μοντέλο απλοποιεί δραστικά τους υπολογισμούς της $aposteriori$ πιθανότητας, απαιτώντας μόνο τον υπολογισμό μονοδιάστατων κατανομών για κάθε χαρακτηριστικό.

Οι ταξινομητές *Naive Bayes* είναι ιδιαίτερα γρήγοροι και αποδοτικοί τόσο στην εκπαίδευση όσο και στην πρόβλεψη, με πολυπλοκότητα γραμμική ως προς το μέγεθος των δεδομένων. Δεν απαιτούν πολύ μεγάλο πλήθος δειγμάτων για να εκτιμήσουν αξιόπιστα τις πιθανότητες, χάρη στην απλότητα του μοντέλου, και μάλιστα μπορούν να αποδώσουν καλά ακόμη και με μικρά σύνολα δεδομένων. Επίσης, μπορούν εύκολα να επεκταθούν σε πολλαπλές κλάσεις (έχουν εγγενώς πολυταξινομική φύση) και κλιμακώνονται σε προβλήματα με υψηλή διαστασιμότητα (πολλά χαρακτηριστικά) – για παράδειγμα στην κατηγοριοποίηση κειμένου, όπου χρησιμοποιούνται χιλιάδες χαρακτηριστικά/λέξεις, ο *Naive Bayes* παραμένει αποτελεσματικός όταν άλλοι αλγόριθμοι δυσκολεύονται.

Τα βασικά πλεονεκτήματα των ταξινομητών *Naive Bayes* περιλαμβάνουν:

- Πολύ γρήγορη εκπαίδευση και πρόβλεψη με πολυπλοκότητα γραμμική στο μέγεθος των δεδομένων.
- Απλότητα που επιτρέπει αξιόπιστες εκτιμήσεις πιθανοτήτων ακόμη και με μικρά δείγματα.
- Εύκολη επέκταση σε προβλήματα πολλαπλών κλάσεων και υψηλής διαστασιμότητας.
- Αποτελεσματικότητα σε εφαρμογές όπως η κατηγοριοποίηση κειμένου (*text classification*).

Ωστόσο, υπάρχουν και σημαντικοί περιορισμοί:

- Η βασική υπόθεση ανεξαρτησίας μεταξύ χαρακτηριστικών σπάνια ισχύει απολύτως, με αποτέλεσμα να μειώνεται η ακρίβεια όταν υπάρχουν ισχυρές αλληλεπιδράσεις.
- Η αδυναμία μοντελοποίησης εξαρτήσεων μεταξύ χαρακτηριστικών μπορεί να περιορίσει την απόδοση σε σύνθετα ιατρικά δεδομένα. Για παράδειγμα, σε ιατρικά δεδομένα

όπου διάφορες κλινικές παράμετροι συνδέονται αιτιωδώς, ένας αφελής ταξινομητής μπορεί να αντιμετωπίσει δυσκολίες.

- Το φαινόμενο μηδενικής συχνότητας (*zero-frequency problem*): αν ένα χαρακτηριστικό δεν εμφανίζεται σε συνδυασμό με κάποια κλάση στα δεδομένα εκπαίδευσης, η πιθανότητα είναι μηδέν, επηρεάζοντας αρνητικά την τελική πρόβλεψη.

3.5 Συνελικτικά Νευρωνικά Δίκτυα (CNN)

Τα Συνελικτικά Νευρωνικά Δίκτυα (*Convolutional Neural Networks - CNN*) [15] είναι μια κατηγορία νευρωνικών δικτύων που έχουν φέρει επανάσταση στην ανάλυση οπτικών δεδομένων και εικόνων. Ένα *CNN* είναι σχεδιασμένο να μαθαίνει αυτόματα ιεραρχίες τοπικών χαρακτηριστικών από τα δεδομένα εισόδου, εκμεταλλευόμενο τη δομή τους σε μορφή πλέγματος (π.χ. εικονοστοιχεία μιας εικόνας).

Η αρχιτεκτονική ενός τυπικού *CNN* αποτελείται από:

- **Συνελικτικά επίπεδα (*convolutional layers*):** Χρησιμοποιούν μικρούς τοπικούς πυρήνες-φίλτρα που ολισθαίνουν πάνω στην εικόνα και ανιχνεύουν βασικά μοτίβα (π.χ. ακμές, γωνίες) σε κάθε περιοχή.

- **Επίπεδα υποδειγματοληψίας (*pooling layers*):** Μειώνουν τη διαστατικότητα των δεδομένων και προσδίδουν αμεταβλητότητα σε μικρές μετατοπίσεις.

- **Πλήρως συνδεδεμένα επίπεδα (*fully connected layers*):** Συνδυάζουν τα εξαγόμενα χαρακτηριστικά για την τελική ταξινόμηση.

Μέσω της διαδικασίας εκπαίδευσης με οπισθοδιάδοση (*backpropagation*), το *CNN* προσαρμόζει τα βάρη των συνελικτικών πυρήνων ώστε να αναγνωρίζει όλο και πιο σύνθετα χαρακτηριστικά στις βαθύτερες στρώσεις – από απλές ακμές στις πρώτες στρώσεις, σε σύνθετες δομές (π.χ. σχήματα οργάνων ή αλλοιώσεων) σε βαθύτερα επίπεδα.

Τα συνελικτικά δίκτυα γνώρισαν τεράστια ώθηση μετά το 2012 (*ImageNet* competition), όπου επιτεύχθηκε δραστική βελτίωση στην αναγνώριση εικόνων. Στον ιατρικό χώρο, η υιοθέτηση των *CNN* υπήρξε ραγδαία: έχουν εφαρμοστεί με επιτυχία σε:

- Ακτινολογικές εικόνες (π.χ. ανάλυση ακτινογραφιών θώρακος, ανίχνευση όγκων σε μαγνητικές (*MRI*) και αξονικές (*CT*) τομογραφίες.

- Ψηφιακές φωτογραφίες δερματολογικών βλαβών.
- Μικροσκοπικές εικόνες ιστοπαθολογίας.
- Άλλους τύπους δεδομένων όπως σήματα (π.χ. ανάλυση ηλεκτροκαρδιογραφημάτων με *1D-CNN*).

Τα κύρια πλεονεκτήματα των *CNN* είναι:

- Διαθέτουν την ικανότητα αυτόματης εκμάθησης διακριτικών χαρακτηριστικών απευθείας από τα ωμά δεδομένα, εξαλείφοντας έτσι την ανάγκη για χειροκίνητη εξαγωγή ή επιλογή χαρακτηριστικών.
- Η συνελικτική δομή εκμεταλλεύεται τις τοπικές συσχετίσεις και εισάγει αμεταβλητότητα σε μετατοπίσεις ή κλίμακα: ένα χαρακτηριστικό ανιχνεύεται ανεξαρτήτως της θέσης του στην εικόνα, χάρη στην ολίσθηση των ίδιων φίλτρων σε όλο το πεδίο.
- Μέσω της ιεραρχικής στοίβαξης πολλαπλών συνελικτικών επιπέδων, το δίκτυο συνθέτει περίπλοκες αναπαραστάσεις (π.χ. σε μια ακτινογραφία θώρακος, αρχικά μαθαίνει άκρα και γραμμές, έπειτα βασικά σχήματα όπως περιγράμματα πνευμόνων, και τελικά ανιχνεύει παθολογικά πρότυπα όπως διηθήσεις ή όζους).

Προκλήσεις και περιορισμοί:

- Απαιτούν μεγάλες ποσότητες δεδομένων για να εκπαιδευτούν αποτελεσματικά – η επίδοση τους αυξάνεται με το πλήθος των δειγμάτων, κάτι που σε ιατρικά συμφραζόμενα μπορεί να είναι δύσκολο λόγω περιορισμένου αριθμού επισημειωμένων εικόνων.
- Η έλλειψη δεδομένων μπορεί να οδηγήσει σε υπερπροσαρμογή, δηλαδή το δίκτυο μαθαίνει το θόρυβο των δεδομένων εκπαίδευσης.

Για την αντιμετώπιση αυτού του προβλήματος εφαρμόζονται τεχνικές όπως ο Τεχνητός εμπλουτισμός δεδομένων μέσω μετασχηματισμών εικόνων και η **Μεταφορά μάθησης**(***transfer learning***)[16], όπου ένα *CNN* που έχει εκπαιδευτεί σε μεγάλο γενικό σύνολο εικόνων προσαρμόζεται σε ιατρικές εικόνες με λιγότερα δεδομένα.Οι δύο βασικές στρατηγικές που εφαρμόζονται είναι:

- η χρήση προεκπαιδευμένων δικτύων ως εξαγωγείς χαρακτηριστικών (*feature extractors*), όπου το δίκτυο δεν εκπαιδεύεται εκ νέου αλλά τα εξαγόμενα χαρακτηριστικά τροφοδοτούνται σε παραδοσιακούς ταξινομητές,

- η ακριβής προσαρμογή (*fine-tuning*), δηλαδή η επανεκπαίδευση ενός προεκπαιδευμένου δικτύου στα ιατρικά δεδομένα.

Αν και οι δύο μέθοδοι χρησιμοποιούνται ευρέως, τα αποτελέσματά τους διαφέρουν ανάλογα με το πρόβλημα και το σύνολο δεδομένων. Μελέτες έχουν δείξει πως η ακριβής προσαρμογή μπορεί να υπερισχύει σε κάποιες περιπτώσεις, ενώ σε άλλες η χρήση του δικτύου ως εξαγωγέα χαρακτηριστικών δίνει καλύτερα αποτελέσματα. Σημαντικές επιτυχίες έχουν επιτευχθεί με την ακριβή προσαρμογή προεκπαιδευμένων δικτύων, όπως το *Inception v3*, το οποίο σε ιατρικά δεδομένα πέτυχε απόδοση σε επίπεδο εμπειρογνωμόνων.

3.6 MobileNetV2

Το *MobileNetV2* [17] είναι μια ειδική αρχιτεκτονική συνελικτικού νευρωνικού δικτύου, η οποία σχεδιάστηκε από την *Google* με στόχο την αποδοτική εκτέλεση αλγορίθμων όρασης σε κινητές και ενσωματωμένες συσκευές. Σε αντίθεση με τα πολύ μεγάλα και βαριά δίκτυα (όπως τα *VGG* ή *ResNet*), που επιτυγχάνουν υψηλές ακρίβειες εις βάρος πολλών παραμέτρων και υπολογισμών, το *MobileNetV2* επιδιώκει μια ισορροπία μεταξύ ακρίβειας και πολυπλοκότητας, ώστε να είναι ιδανικό για συστήματα με περιορισμένους πόρους (*επεξεργαστική ισχύ*, *μνήμη*).

Κύρια χαρακτηριστικά του *MobileNetV2* είναι:

- Η χρήση συνελικτικών πυρήνων βάθους (*depthwise separable convolutions*).

- Το καινοτόμο δομικό στοιχείο (*block*) με ανεστραμμένα υπολείμματα (*inverted residuals*) και γραμμικούς συσσωρευτές (*linear bottlenecks*).

Στην πράξη, αυτό σημαίνει ότι κάθε κλασική συνελικτική πράξη αντικαθίσταται από δύο βήματα:

- Μια συνελικτική πράξη κατά βάθος, όπου κάθε φίλτρο εφαρμόζεται ξεχωριστά σε κάθε κανάλι χαρακτηριστικών (*depthwise convolution*).

- Μια σημειακή συνελικτική 1x1 που συνδυάζει τις πληροφορίες μεταξύ των καναλιών (*pointwise convolution*).

Αυτή η τεχνική μειώνει το κόστος των συνελίξεων κατά περίπου ένα μέγεθος τάξης χωρίς σημαντική απώλεια πληροφορίας.Επιπλέον, τα ανεστραμμένα *residual blocks*:

- Πρώτα επεκτείνουν τον αριθμό των καναλιών (με φθηνή 1x1 συνελίξη).
- Στη συνέχεια εφαρμόζεται συνελικτική πράξη ως προς το βάθος (*depthwise convolution*) για την εξαγωγή χαρακτηριστικών.
- Τέλος, προβάλλουν πίσω σε χώρο χαμηλής διάστασης με γραμμική 1x1 συνελίξη (*linear bottleneck*).

Ο γραμμικός συσσωρευτής, χωρίς μη γραμμικότητα *ReLU* σε εκείνο το στάδιο, αποφεύγει την απώλεια πληροφορίας που προκαλείται από το cutoff των αρνητικών τιμών, διατηρώντας περισσότερη πληροφορία ροής.Συνδυαστικά, αυτά τα σχεδιαστικά στοιχεία επιτρέπουν στο *MobileNetV2*:

- Να έχει πολύ λιγότερες παραμέτρους και πράξεις από συγκρίσιμα δίκτυα.
- Να διατηρεί υψηλή ακρίβεια ταξινόμησης χωρίς δραστική υστέρηση.

Το *MobileNetV2* βρήκε ευρεία εφαρμογή σε περιπτώσεις όπου η ταχύτητα και η αποδοτικότητα είναι κρίσιμες:

- Χρησιμοποιείται συχνά σε εφαρμογές κινητών τηλεφώνων για αναγνώριση εικόνων, επαυξημένη πραγματικότητα και άλλες διεργασίες όρασης υπολογιστών.
- Μπορεί να τρέξει σε πραγματικό χρόνο ακόμα και σε κεντρικές μονάδες επεξεργασίας (*CPU*) κινητών συσκευών.
- Για παράδειγμα, σε εφαρμογή ιατρικής διάγνωσης μέσω έξυπνων κινητών *smartphone*, ένα προκαταρτικά εκπαιδευμένο *MobileNetV2* μπορεί να εκτελεί επιτόπου ταξινόμηση εικόνων (π.χ. φωτογραφίες δέρματος) σε κατηγορίες παθήσεων.

Παρότι ελαφρύ, το μοντέλο διατηρεί ανταγωνιστική ακρίβεια συγκριτικά με μεγαλύτερες αρχιτεκτονικές:

- Με κατάλληλη προεκπαίδευση στο *ImageNet* και ακριβή προσαρμογή σε ιατρικά δεδομένα, έχει επιτύχει υψηλά ποσοστά αναγνώρισης.
- Ενδεικτικά, σε σύστημα ταξινόμησης δερματικών βλαβών σε *Android* εφαρμογή πέτυχε ακρίβεια 93–94%, επίπεδο ικανοποιητικό για κλινική χρήση πρώτου βαθμού.

Τα πλεονεκτήματα του *MobileNetV2* συνοψίζονται σε:

- Μικρή απαίτηση μνήμης.
- Γρήγορο χρόνο απόκρισης.
- Ιδανικότητα για συστήματα επιτόπιας διάγνωσης και φορητές ιατρικές συσκευές.
- Ενσωμάτωση σε φορητά απεικονιστικά συστήματα ή εφαρμογές *τηλεϊατρικής*.
- Εκτέλεση αλγορίθμων τεχνητής νοημοσύνης χωρίς ανάγκη σύνδεσης σε ισχυρό διακομιστή ή *cloud*.

Ο βασικός συμβιβασμός είναι ότι, λόγω της επιθετικής μείωσης παραμέτρων, το *MobileNetV2* μπορεί να υστερεί ελαφρά σε ακρίβεια έναντι πολύ μεγαλύτερων δικτύων σε απαιτητικά σύνολα δεδομένων.Ωστόσο, για πολλές πρακτικές εφαρμογές όπου ο στόχος είναι ένα ελαφρύ μοντέλο με επαρκώς καλή απόδοση, το *MobileNetV2* παρέχει έναν άριστο συμβιβασμό.Τέλος, συνεχίζονται οι έρευνες βελτίωσης με επόμενες εκδόσεις και παρεμφερείς "μικρο-αρχιτεκτονικές" που στοχεύουν στη μεγιστοποίηση της ακρίβειας ανά παραμέτρους.

Στον χώρο της ιατρικής πληροφορικής, η ύπαρξη τέτοιων μοντέλων ανοίγει τον δρόμο για δημοκρατικοποίηση της τεχνητής νοημοσύνης, επιτρέποντας σε φορητές συσκευές να λειτουργούν ως έξυπνα διαγνωστικά εργαλεία σε απομακρυσμένες περιοχές ή σε περιοχές με περιορισμένους πόρους.

3.7 Μετρικές Αξιολόγησης Μοντέλων Ταξινόμησης

Για την αξιόπιστη αξιολόγηση των μοντέλων ταξινόμησης, χρησιμοποιείται ένα σύνολο μετρικών [18] που ποσοτικοποιούν την απόδοση με διαφορετικούς τρόπους.

Ακρίβεια (*accuracy*) ονομάζεται το ποσοστό των σωστών προβλέψεων του μοντέλου επί του συνόλου των περιπτώσεων (είναι δηλαδή ο λόγος των ορθών ταξινομήσεων (*True*

Positives + *True Negatives*) προς το σύνολο των δειγμάτων):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

όπου:

- *TP* (*True Positives*): αριθμός θετικών δειγμάτων που ταξινομήθηκαν σωστά ως θετικά
- *TN* (*True Negatives*): αριθμός αρνητικών δειγμάτων που ταξινομήθηκαν σωστά ως αρνητικά
- *FP* (*False Positives*): αριθμός αρνητικών δειγμάτων που ταξινομήθηκαν λανθασμένα ως θετικά
- *FN* (*False Negatives*): αριθμός θετικών δειγμάτων που ταξινομήθηκαν λανθασμένα ως αρνητικά

Η ακρίβεια δίνει μια συνολική εικόνα απόδοσης, όμως μπορεί να είναι παραπλανητική όταν υπάρχει έντονη ανισορροπία κλάσεων (π.χ. ένα μοντέλο που πάντα προβλέπει «αρνητικό» μπορεί να έχει υψηλή ακρίβεια αν η θετική τάξη είναι σπάνια).

Ευστοχία (*precision*) ή θετική προγνωστική αξία (PPV) ορίζεται ως:

$$Precision = \frac{TP}{TP + FP}$$

Δηλαδή, το ποσοστό των προγνωσμένων ως θετικών περιπτώσεων που είναι πράγματι θετικές. Απαντά στην ερώτηση «από όλες τις διαγνώσεις που έκανε το μοντέλο ως 'νόσο', πόσες ήταν σωστές;».

Ανάκληση (*recall*) ή **ευαισθησία (*sensitivity*)** είναι:

$$Recall = \frac{TP}{TP + FN}$$

Το ποσοστό των πραγματικά θετικών περιπτώσεων που ανιχνεύθηκαν σωστά από το μοντέλο. Απαντά στο «πόσους από τους ασθενείς εντόπισε το τεστ;». Η ευαισθησία είναι κρίσιμη σε περιπτώσεις που θέλουμε να μην χάνουμε θετικές περιπτώσεις (π.χ. διαγνωστικός έλεγχος για σοβαρή ασθένεια – θέλουμε υψηλή ανάκληση ώστε να μην διαφύγει κανένας ασθενής).

Η ευστοχία αντίστοιχα έχει σημασία όταν οι ψευδώς θετικές προβλέψεις έχουν σοβαρό κόστος (π.χ. ένα τεστ που δίνει πολλούς ψευδείς συναγερμούς μπορεί να προκαλεί άσκοπες

θεραπείες). Δεδομένου ότι η *ευστοχία* και η *ευαισθησία* συχνά βρίσκονται σε αντίστροφη σχέση (αύξηση της μίας μπορεί να μειώσει την άλλη), χρησιμοποιείται η μέση τιμή **F1 (*F1-score*)** ως συνοπτική μετρική, που ορίζεται ως:

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Δίνοντας μια ενιαία ένδειξη της ισορροπίας μεταξύ τους. Ένα υψηλό *F1-score* σημαίνει ότι το μοντέλο έχει τόσο υψηλή ευαισθησία όσο και υψηλή ευστοχία, άρα πετυχαίνει λίγα ψευδώς θετικά και λίγα ψευδώς αρνητικά ταυτόχρονα.

Τέλος, ένας άλλος σημαντικός δείκτης απόδοσης για δυαδικούς ταξινομητές είναι το **AUC (*Area Under the Curve*)**. Το AUC είναι το εμβαδόν κάτω από την καμπύλη *ROC* του μοντέλου και ποικίλλει από 0.5 (τυχαίος ταξινομητής) έως 1.0 (τέλειος ταξινομητής). Η καμπύλη *ROC* απεικονίζει:

- Στον άξονα των Y το *True Positive Rate*, δηλαδή την ανάκληση.
- Στον άξονα των X το *False Positive Rate*, που ορίζεται ως:

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

Μια τιμή AUC πιο κοντά στο 1 υποδηλώνει ότι το μοντέλο έχει ισχυρή ικανότητα διάκρισης μεταξύ των δύο τάξεων σε όλα τα πιθανά κατώφλια πιθανότητας – όσο υψηλότερο το AUC, τόσο μεγαλύτερη η πιθανότητα το μοντέλο να κατατάξει ένα τυχαίο θετικό δείγμα ψηλότερα από ένα τυχαίο αρνητικό.

3.8 Καμπύλες ROC (Receiver Operating Characteristic Curves)

Η καμπύλη *ROC* [19] είναι ένα εργαλείο που χρησιμοποιείται για την απεικόνιση και αξιολόγηση της απόδοσης ενός δυαδικού ταξινομητή σε όλα τα πιθανά κατώφλια ταξινόμησης. Μια καμπύλη *ROC* απεικονίζει τον συμβιβασμό μεταξύ ευαισθησίας και ειδικότητας του μοντέλου καθώς μεταβάλλεται το όριο διάκρισης.

Κάθε σημείο πάνω στην καμπύλη *ROC* αντιστοιχεί σε μια συγκεκριμένη επιλογή κατωφλίου πιθανότητας για το μοντέλο, και δείχνει:

- την **ευαισθησία** (*True Positive Rate – TPR*):

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

που είναι το ποσοστό των πραγματικά θετικών που ανιχνεύθηκαν σωστά, και

- το συμπληρωματικό της **ειδικότητας**, δηλαδή το ***False Positive Rate*** **(FPR)**:

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

που αντιστοιχεί στο ποσοστό των αρνητικών δειγμάτων που ταξινομήθηκαν λανθασμένα ως θετικά.

Μια καμπύλη *ROC* που βρίσκεται πιο κοντά στην επάνω αριστερή γωνία του διαγράμματος αντιπροσωπεύει ένα ισχυρότερο μοντέλο – δηλαδή υψηλή ευαισθησία και ταυτόχρονα υψηλή ειδικότητα.

Αντιθέτως, η διαγώνιος από το σημείο $(0, 0)$ στο $(1, 1)$ αντιστοιχεί σε έναν τυχαίο ταξινομητή χωρίς διακριτική ικανότητα, όπου:

$$TPR = FPR$$

Το εμβαδόν κάτω από την καμπύλη (*AUC – Area Under the Curve*) συνοψίζει ποσοτικά την απόδοση:

- $AUC = 1$ σημαίνει τέλεια διάκριση (η καμπύλη περνάει από το $(0, 1)$),
- $AUC = 0.5$ σημαίνει τυχαία πρόβλεψη.

Κεφάλαιο 4

Εφαρμογή μοντέλων για την πρόβλεψη καρδιαγγειακής νόσου - Α’

4.1 Εισαγωγή

Στο παρόν κεφάλαιο παρουσιάζεται το πειραματικό μέρος της διπλωματικής εργασίας, το οποίο εστιάζει στη μελέτη και ανάλυση δεδομένων που σχετίζονται με καρδιαγγειακές παθήσεις, μέσω της αξιοποίησης τεχνικών μηχανικής μάθησης. Βασικός στόχος του πειραματικού μέρους αποτελεί η διερεύνηση της δυνατότητας πρόβλεψης της εμφάνισης καρδιαγγειακής νόσου, με τη χρήση πραγματικών δεδομένων υγείας και η σύγκριση της αποδοτικότητας διαφορετικών αλγορίθμων ταξινόμησης. Ως πρώτο σύνολο δεδομένων χρησιμοποιήθηκε το *Cardiovascular Disease dataset* [20], το οποίο περιέχει δημογραφικές και κλινικές πληροφορίες από ασθενείς, καθώς και τη μεταβλητή-στόχο που αφορά στην εμφάνιση ή μη καρδιαγγειακής νόσου. Με βάση το συγκεκριμένο σύνολο δεδομένων, αναπτύχθηκε ολοκληρωμένη πειραματική διαδικασία η οποία περιλαμβάνει τα στάδια της προεπεξεργασίας των δεδομένων, της επιλογής και εκπαίδευσης κατάλληλων μοντέλων, καθώς και την αξιολόγηση της απόδοσής τους με τη χρήση κατάλληλων μετρικών.

Το πειραματικό μέρος έχει ως απώτερο σκοπό την ανάδειξη των σημαντικότερων παραγόντων που σχετίζονται με την εμφάνιση καρδιαγγειακής νόσου, αλλά και την παρουσίαση μιας μεθοδολογίας η οποία μπορεί να αξιοποιηθεί σε παρόμοιες περιπτώσεις πρόβλεψης ασθενειών μέσω ανάλυσης δεδομένων. Η ανάλυση και τα αποτελέσματα που παρουσιάζονται στα επόμενα τμήματα βασίζονται αποκλειστικά στα δεδομένα του συγκεκριμένου συνόλου και στον κώδικα που υλοποιήθηκε στο πλαίσιο της εργασίας.

4.2 Περιγραφή Δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήθηκε στην παρούσα εργασία αποτελείται από πραγματικά, ανώνυμα ιατρικά δεδομένα ασθενών και προέρχεται από δημόσια διαθέσιμη πηγή. Πιο συγκεκριμένα, το *Cardiovascular Disease dataset* περιλαμβάνει δημογραφικές, ανθρωπομετρικές και κλινικές πληροφορίες, με συνολικά 13 μεταβλητές. Κάθε εγγραφή του συνόλου δεδομένων αντιστοιχεί σε έναν ασθενή, και αποτυπώνει χαρακτηριστικά τα οποία σχετίζονται με πιθανούς παράγοντες κινδύνου για την εμφάνιση καρδιαγγειακών παθήσεων.

Οι βασικές μεταβλητές του συνόλου δεδομένων είναι οι εξής:

- **gender** (φύλο)
- **age** (ηλικία σε ημέρες)
- **height** (ύψος σε εκατοστά)
- **weight** (βάρος σε κιλά)
- **ap_hi** (συστολική αρτηριακή πίεση)
- **ap_lo** (διαστολική αρτηριακή πίεση)
- **cholesterol** (επίπεδο χοληστερόλης)
- **gluc** (επίπεδο γλυκόζης)
- **smoke** (συνήθεια καπνίσματος)
- **alco** (κατανάλωση αλκοόλ)
- **active** (σωματική δραστηριότητα)
- **cardio** (εμφάνιση καρδιαγγειακής νόσου - μεταβλητή στόχος)

Κατά την αρχική ανάλυση, το σύνολο δεδομένων περιλάμβανε 70.000 εγγραφές, ωστόσο εφαρμόστηκε καθαρισμός δεδομένων ώστε να αφαιρεθούν πιθανά διπλότυπα και ακραίες τιμές. Επιπλέον, δημιουργήθηκαν νέα παράγωγα χαρακτηριστικά, όπως η ηλικία σε έτη και ο δείκτης μάζας σώματος (BMI), τα οποία αποδείχθηκαν σημαντικά για την ταξινόμηση. Αξίζει να σημειωθεί ότι το σύνολο δεδομένων περιέχει τόσο κατηγορικές όσο και αριθμητικές

μεταβλητές, γεγονός που επιτρέπει την πολύπλευρη ανάλυση και αξιοποίηση διαφορετικών μεθόδων μηχανικής μάθησης.

Η μεταβλητή-στόχος (***cardio***) λαμβάνει διακριτές τιμές 0 ή 1, όπου η τιμή 1 δηλώνει την εμφάνιση καρδιαγγειακής νόσου. Κατά την ανάλυση της κατανομής της μεταβλητής στόχου διαπιστώθηκε σχετική ισορροπία μεταξύ των δύο κλάσεων, γεγονός που διευκολύνει τη διαδικασία εκπαίδευσης των μοντέλων και την ερμηνεία των αποτελεσμάτων. Συμπερασματικά, το συγκεκριμένο σύνολο δεδομένων παρέχει μια ρεαλιστική και αρκετά αντιπροσωπευτική εικόνα των παραγόντων που συσχετίζονται με τις καρδιαγγειακές παθήσεις, επιτρέποντας την εξαγωγή χρήσιμων συμπερασμάτων μέσω της ανάλυσης που ακολουθεί.

4.3 Προεπεξεργασία Δεδομένων

Η προεπεξεργασία των δεδομένων αποτελεί ένα ιδιαίτερα κρίσιμο στάδιο στην ανάπτυξη συστημάτων μηχανικής μάθησης, καθώς εξασφαλίζει την ποιότητα και την αξιοπιστία των αποτελεσμάτων που θα προκύψουν. Στο πλαίσιο της παρούσας εργασίας, εφαρμόστηκαν σειρά τεχνικών καθαρισμού και μετασχηματισμού των δεδομένων, ώστε να εξασφαλιστεί η καταλληλότητά τους για τα επόμενα στάδια της ανάλυσης.

Αρχικά, πραγματοποιήθηκε έλεγχος για την ύπαρξη ελλιπών ή μη έγκυρων τιμών στις εγγραφές του συνόλου δεδομένων. Από την ανάλυση διαπιστώθηκε ότι δεν υπήρχαν απουσιάζουσες τιμές, γεγονός που απλοποίησε τη διαδικασία καθαρισμού. Στη συνέχεια, εντοπίστηκαν και αφαιρέθηκαν τυχόν διπλότυπες εγγραφές, ώστε να αποφευχθεί η παραμόρφωση των αποτελεσμάτων λόγω υπερ-αναπαράστασης συγκεκριμένων περιπτώσεων.

Για την περαιτέρω βελτίωση της ποιότητας των δεδομένων, εφαρμόστηκαν τεχνικές ανίχνευσης και απομάκρυνσης ακραίων τιμών (outliers), οι οποίες εντοπίστηκαν μέσω του υπολογισμού των κατάλληλων ποσοστημορίων (2.5ο και 97.5ο εκατοστημόριο) σε βασικές αριθμητικές μεταβλητές, όπως το ύψος, το βάρος και οι τιμές της αρτηριακής πίεσης. Με τον τρόπο αυτό διατηρήθηκαν μόνο οι εγγραφές που εντάσσονται σε ρεαλιστικά πλαίσια για τον πληθυσμό που μελετάται.

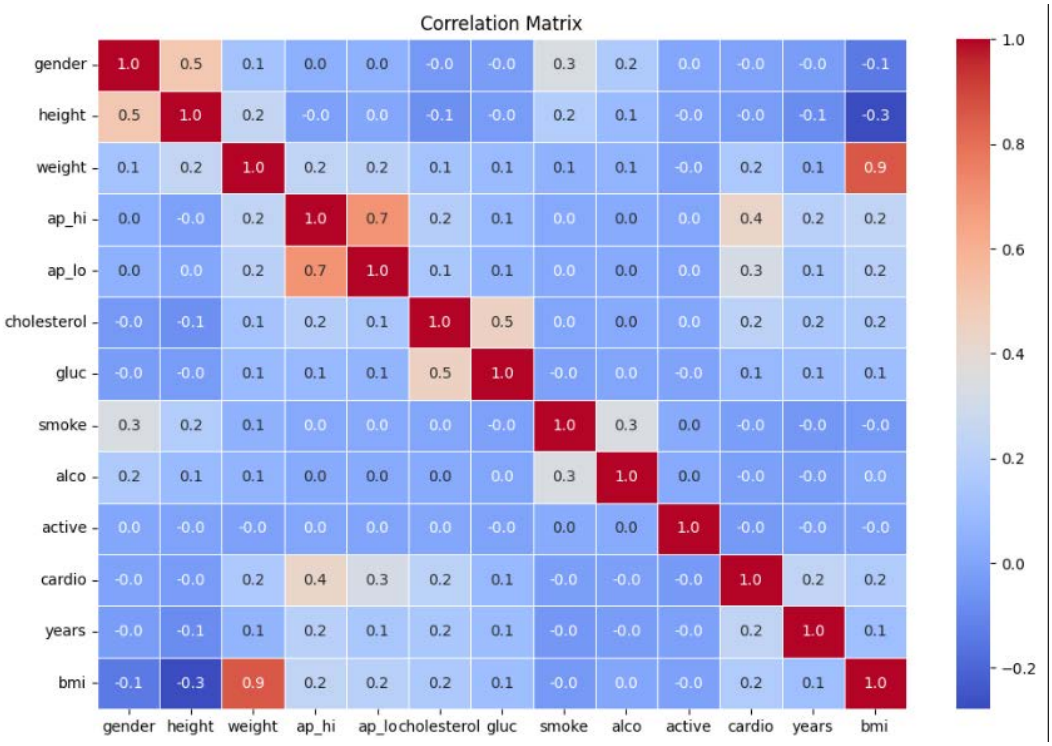
Επιπλέον, πραγματοποιήθηκε δημιουργία νέων παραγώγων χαρακτηριστικών, τα οποία κρίθηκαν σημαντικά για τη βελτίωση της απόδοσης των μοντέλων. Ειδικότερα, υπολογίστηκε η ηλικία των συμμετεχόντων σε έτη (μέσω κατάλληλου μετασχηματισμού από ημέρες), καθώς και ο δείκτης μάζας σώματος (*BMI*). Τα χαρακτηριστικά αυτά (ηλικία, δείκτης

μάζας σώματος) προστέθηκαν στο σύνολο δεδομένων και συμπεριλήφθηκαν στην εκπαίδευση των μοντέλων.

Η διαδικασία της προεπεξεργασίας ολοκληρώθηκε με την αφαίρεση μη χρήσιμων μεταβλητών, όπως το μοναδικό αναγνωριστικό κάθε εγγραφής (*id*), το οποίο δεν προσφέρει πληροφορία για την πρόβλεψη της μεταβλητής στόχου. Τέλος, πραγματοποιήθηκε οπτικοποίηση των ιδιοτήτων των μεταβλητών μέσω κατάλληλων γραφημάτων (ιστογράμματα (*histograms*), διαγράμματα box (*boxplots*), θερμικές απεικονίσεις (*heatmaps*)), ώστε να επιβεβαιωθεί η ορθότητα της επεξεργασίας και να ανιχνευθούν τυχόν περαιτέρω προβλήματα στα δεδομένα.

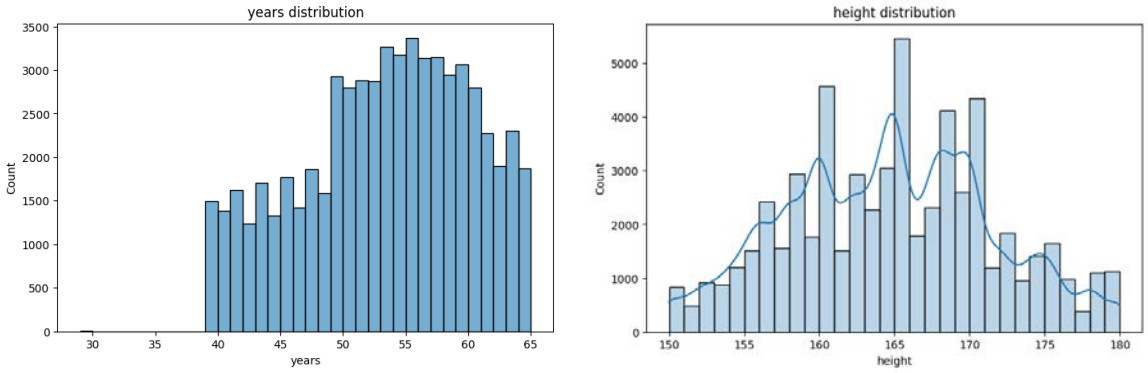
4.4 Οπτικοποίηση Δεδομένων

Ακολουθεί η οπτικοποίηση του πίνακα συσχέτισης (*correlation matrix*) στο Σχήμα 4.1 μεταξύ των αριθμητικών στηλών του συνόλου δεδομένων. Τη μεγαλύτερη θετική συσχέτιση έχει το βάρος και ο δείκτης *BMI*, κάτι που είναι αναμενόμενο καθώς ο *BMI* υπολογίζεται από το ύψος και το βάρος των ασθενών. Δοκιμάστηκε πειραματικά να αφαιρεθεί το βάρος, ωστόσο δεν παρουσίασε κάποια βελτίωση στα αποτελέσματα. Για το λόγο αυτό, αποφασίστηκε να διατηρηθεί στο τελικό σύνολο εκπαίδευσης *training set*. Διαπιστώνεται επίσης ότι δεν υπάρχει κάποια ισχυρή συσχέτιση μεταξύ της μεταβλητής στόχου (*cardio*) και των υπολοίπων χαρακτηριστικών. Αυτό μας φανερώνει ότι η πρόβλεψη της εμφάνισης καρδιαγγειακής νόσου απαιτεί το συνδυασμό πολλαπλών χαρακτηριστικών.



Σχήμα 4.1: Πίνακας συσχέτισης μεταξύ των μεταβλητών.

Η οπτικοποίηση του συνόλου δεδομένων συνεχίζεται με την παράθεση των ιστογραμμάτων συχνοτήτων για τις κύριες αριθμητικές μεταβλητές:

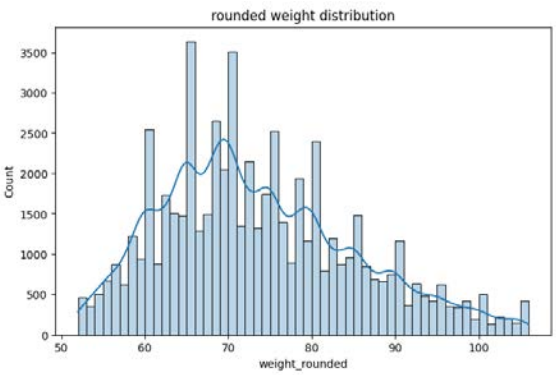


(α′) Ιστόγραμμα ηλικίας (years) (β′) Ιστόγραμμα ύψους (height)

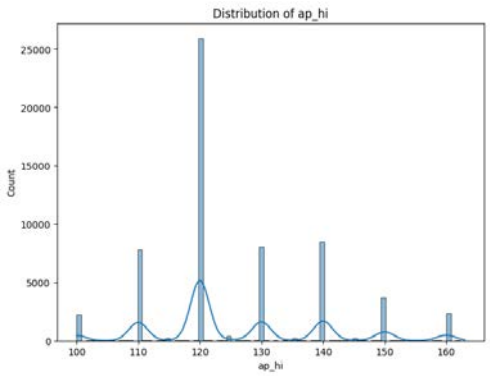
Σχήμα 4.2: Ιστογράμματα ηλικίας και ύψους.

- **Ιστόγραμμα ηλικίας** (βλ. Σχήμα 4.2α′): Το δείγμα επικεντρώνεται κυρίως σε άτομα ηλικίας 50-60 ετών, με λίγους νεότερους ή μεγαλύτερους συμμετέχοντες μέχρι την ηλικία των 65 ετών.

- **Ιστόγραμμα ύψους** (βλ. Σχήμα 4.2β′): Παρουσιάζεται μια αυξημένη συχνότητα σε τυπικά ύψη ενηλίκων ανδρών και γυναικών (160-170 εκατοστών).



(α′) Ιστόγραμμα βάρους (weight)

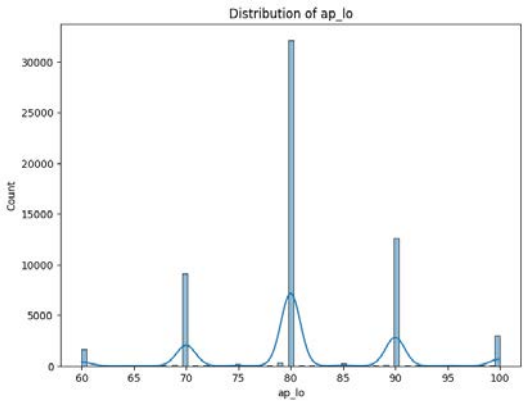


(β′) Ιστόγραμμα συστολικής πίεσης (ap_hi)

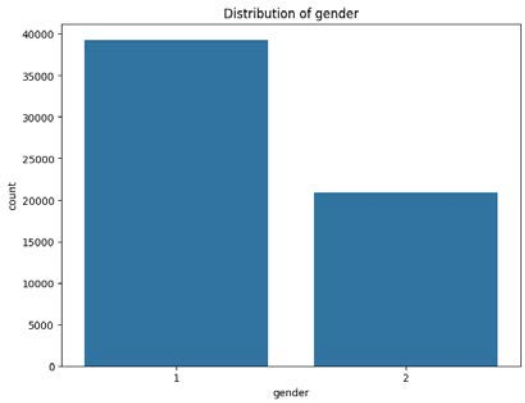
Σχήμα 4.3: Ιστογράμματα βάρους και συστολικής πίεσης.

- **Ιστόγραμμα βάρους** (βλ. Σχήμα 4.3α′): Η κατανομή του βάρους δείχνει πλειοψηφία συμμετεχόντων με βάρος 60–80 κιλών. Οι συχνότητες είναι αυξημένες σε στρογγυλές τιμές (π.χ. 65, 70, 75, 80 κιλών), ένδειξη της στρογγυλοποίησης από τους συμμετέχοντες. Οι ακραίες τιμές είναι σπάνιες, ενώ η κατανομή είναι εύλογη για ενήλικο πληθυσμό.

- **Ιστόγραμμα της συστολικής αρτηριακής πίεσης** (βλ. Σχήμα 4.3β′): Εμφανίζει έντονες αιχμές (*spikes*) σε στρογγυλές τιμές (120, 130, 140, 150, 160 mmHg), ενώ ενδιάμεσες τιμές έχουν ελάχιστες ή καμία παρατήρηση. Αυτό οφείλεται στον τρόπο συλλογής των δεδομένων, με τους περισσότερους να δηλώνουν στρογγυλοποιημένες τιμές, όπως και στο βάρος.



(α′) Ιστόγραμμα διαστολικής πίεσης (ap_lo)



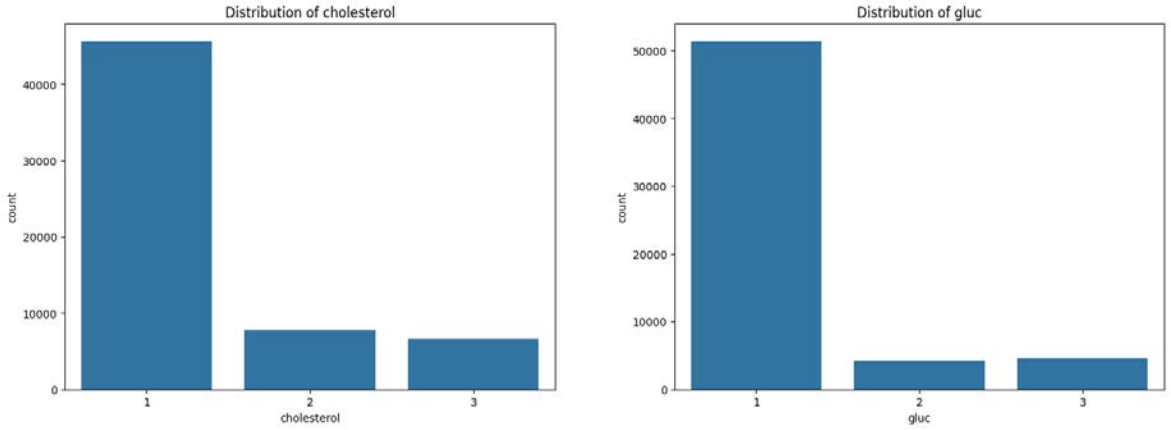
(β′) Ιστόγραμμα φύλου (gender)

Σχήμα 4.4: Ιστογράμματα διαστολικής πίεσης και φύλου.

- **Ιστόγραμμα διαστολικής πίεσης** (βλ. Σχήμα 4.4α′): Όπως και η συστολική, συγκεντρώνεται σε στρογγυλούς αριθμούς (70, 80, 90, 100 mmHg). Οι ενδιάμεσες τιμές είναι

εξαιρετικά σπάνιες, γεγονός που εξηγεί την όψη του γραφήματος. Το φαινόμενο σχετίζεται με τη διαδικασία καταγραφής/δήλωσης τιμών, όχι με σφάλμα στην ανάλυση.

- **Ιστόγραμμα φύλου** (βλ. Σχήμα 4.4β′): Η αναλογία μεταξύ ανδρών και γυναικών (*gender*) δεν είναι 1:1, με περισσότερους άνδρες στο δείγμα.



(α′) Ιστόγραμμα χοληστερόλης (cholesterol) (β′) Ιστόγραμμα γλυκόζης (gluc)

Σχήμα 4.5: Ιστογράμματα χοληστερόλης και γλυκόζης.

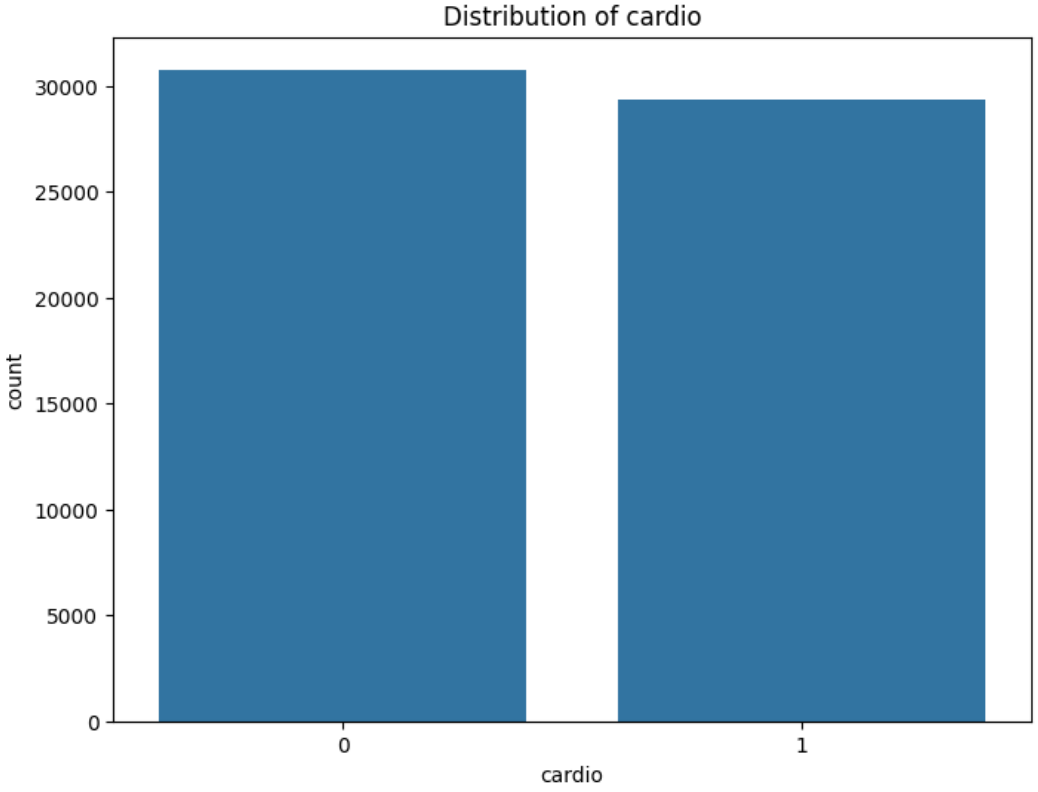
- **Ιστόγραμμα Χοληστερόλης (*cholesterol*)** (βλ. Σχήμα 4.5α′): Το μεγαλύτερο ποσοστό συμμετεχόντων έχει φυσιολογική χοληστερόλη (τιμή 1), ενώ ένα σημαντικό αλλά μικρότερο τμήμα εμφανίζει αυξημένες ή υψηλές τιμές (2 ή 3). Το εύρημα αυτό συμφωνεί με τα αναμενόμενα ποσοστά στο γενικό πληθυσμό και δείχνει ποικιλομορφία ως προς τον παράγοντα κινδύνου.
- **Ιστόγραμμα Γλυκόζης (*gluc*)** (βλ. Σχήμα 4.5β′): Αντίστοιχα, η πλειονότητα έχει φυσιολογικά επίπεδα γλυκόζης, με μικρό ποσοστό να εμφανίζει αυξημένα επίπεδα. Η διασπορά αυτή υποδηλώνει ότι στο δείγμα περιλαμβάνονται τόσο υγιή άτομα όσο και άτομα με διαταραχές μεταβολισμού.

Στο **ιστόγραμμα μεταβλητής στόχου (cardio)** (βλ. Σχήμα 4.6) η κατανομή της μεταβλητής στόχου είναι ισορροπημένη, με περίπου ίσο αριθμό ατόμων σε κάθε κατηγορία. Αυτό διευκολύνει τη διαδικασία εκπαίδευσης αλγορίθμων ταξινόμησης,καθώς μειώνει τον κίνδυνο συμπερίληψης υπέρ της πλειοψηφικής κατηγορίας.

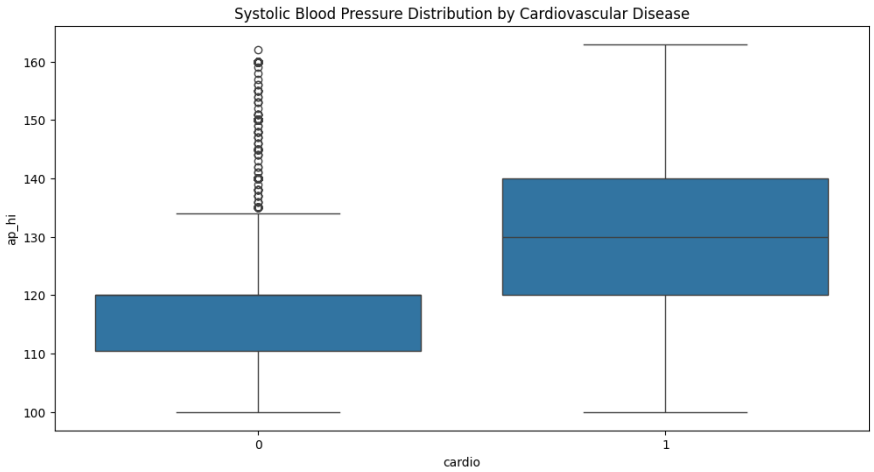
Ακολουθούν τα ***box plots* της συστολικής (*systolic*) και διαστολικής (*diastolic*) πίεσης σε σχέση με την εμφάνιση καρδιαγγειακής νόσου (*cardio*)** στο Σχήμα 4.7. Και οι δύο απεικονίσεις δείχνουν ξεκάθαρα ότι τα άτομα με καρδιαγγειακή νόσο (*cardio*=1) έχουν υψηλότερες μέσες και μέγιστες τιμές τόσο συστολικής όσο και διαστολικής πίεσης. Οι διαφορές μεταξύ των ομάδων είναι εμφανείς: τα υψηλότερα τεταρτημόρια και η διάμεσος μετατοπί-

ζονται προς μεγαλύτερες τιμές στους καρδιοπαθείς. Αυτό επιβεβαιώνει τη γνωστή συσχέτιση της υπέρτασης με την εμφάνιση καρδιαγγειακών νοσημάτων

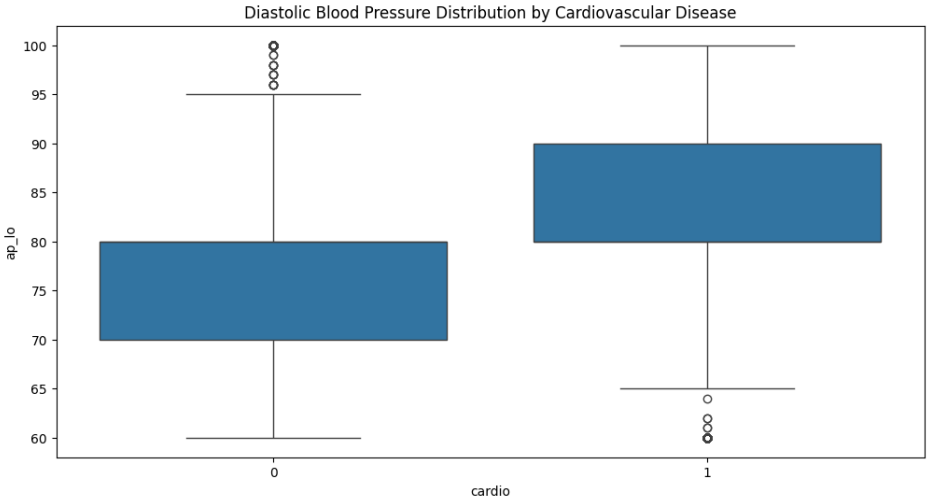
Τέλος, παρουσιάζεται το **ραβδόγραμμα της κατανομής της ηλικίας ανά κατηγορία *cardio*** (βλ. Σχήμα 4.8). Το γράφημα του πλήθους ατόμων ανά ηλικία και καρδιοπάθεια δείχνει ότι στις χαμηλές ηλικίες το ποσοστό καρδιοπαθών είναι μικρό, ενώ αυξάνει αρκετά μετά τα 55 έτη, όπου τα περιστατικά με νόσο ξεπερνούν τα υγιή. Αυτό αποτυπώνει ξεκάθαρα την επίδραση της ηλικίας ως παράγοντα κινδύνου.



Σχήμα 4.6: Ιστόγραμμα μεταβλητής στόχου

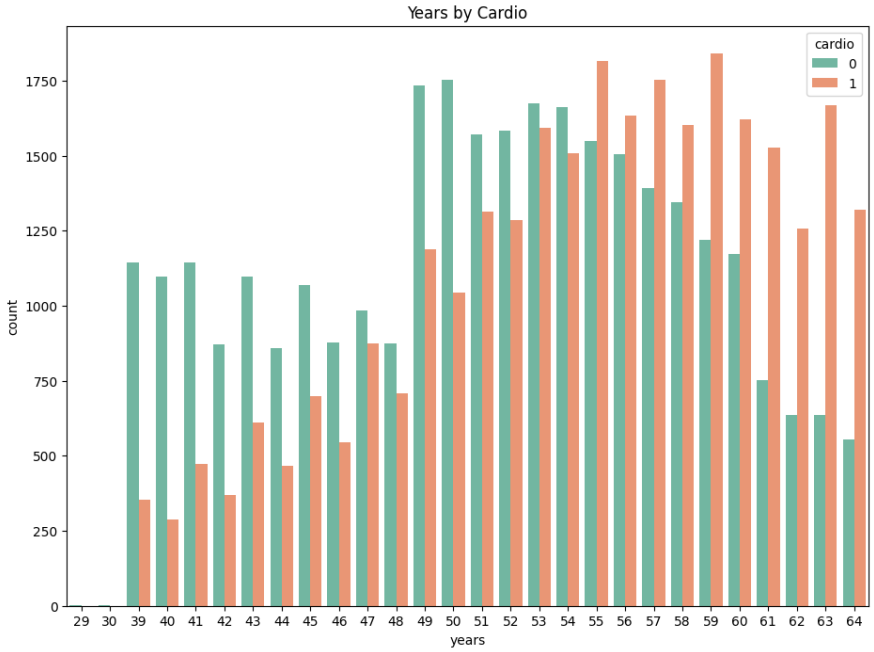


(α′) Συστολική πίεση (systolic)



(β′) Διαστολική πίεση (diastolic)

Σχήμα 4.7: Box plots συστολικής και διαστολικής πίεσης ως προς τη μεταβλητή στόχου.



Σχήμα 4.8: Κατανομή ηλικίας ως προς τη μεταβλητή στόχο.

4.5 Πειραματική Διαδικασία και Επιλογή Μεθόδων

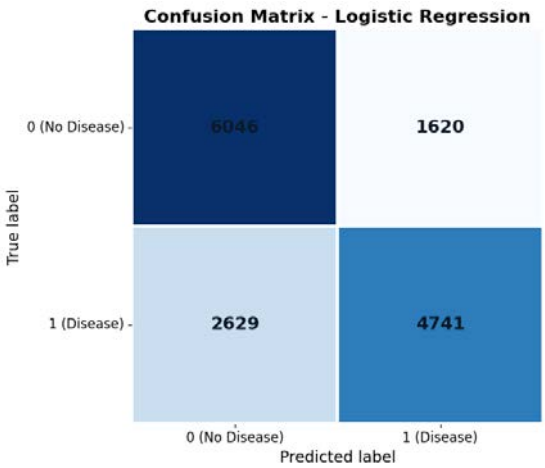
Η πειραματική διαδικασία που υιοθετήθηκε στην παρούσα εργασία περιλαμβάνει τη συστηματική εφαρμογή και αξιολόγηση διαφόρων αλγορίθμων μηχανικής μάθησης με σκοπό την πρόβλεψη της εμφάνισης καρδιαγγειακής νόσου, βάσει των διαθέσιμων δεδομένων. Αρχικά, πραγματοποιήθηκε διαχωρισμός του συνόλου δεδομένων σε σύνολο εκπαίδευσης και σύνολο ελέγχου (*test set*), με τη χρήση της μεθόδου *train/test split* σε αναλογία 75%-25%. Η επιλογή αυτή διασφαλίζει ότι η εκπαίδευση των μοντέλων πραγματοποιείται σε ανεξάρτητο υποσύνολο δεδομένων σε σχέση με εκείνο που χρησιμοποιείται για την τελική αξιολόγηση της απόδοσής τους.

Για τη διερεύνηση της προβληματικής επιλέχθηκαν και υλοποιήθηκαν διάφοροι αλγόριθμοι ταξινόμησης που χρησιμοποιούνται ευρέως στη βιβλιογραφία για ανάλογα προβλήματα. Συγκεκριμένα, εφαρμόστηκαν οι αλγόριθμοι *Logistic Regression*, *Random Forest Classifier*, *Gradient Boosting Classifier* και *Naive Bayes*. Η επιλογή των συγκεκριμένων αλγορίθμων έγινε με κριτήριο την απλότητα, την αποδοτικότητα και τη δυνατότητα ερμηνείας των αποτελεσμάτων τους. Επιπλέον, η χρήση πολλαπλών μεθόδων επιτρέπει την ευρύτερη αξιολόγηση της καταλληλότητας διαφόρων προσεγγίσεων για το συγκεκριμένο πρόβλημα.

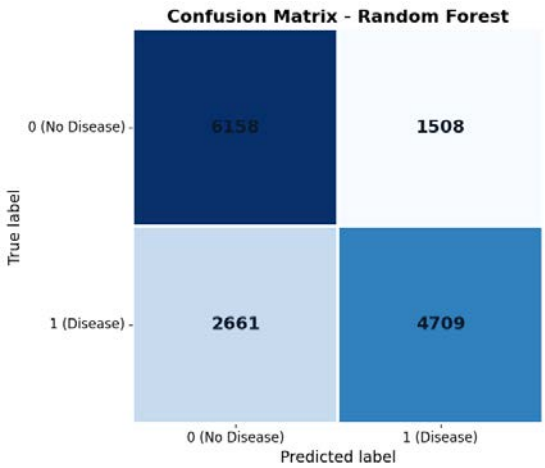
Κατά την εκπαίδευση των μοντέλων, για κάθε αλγόριθμο πραγματοποιήθηκε ρύθμιση των βασικών υπερπαραμέτρων, είτε βάσει προεπιλεγμένων τιμών είτε μέσω διαδικασιών βελτιστοποίησης (όπου απαιτήθηκε). Η εκτίμηση της απόδοσης των μοντέλων έγινε με τη χρήση διαφόρων μετρικών ταξινόμησης, όπως η ακρίβεια, η ευστοχία, η ανάκληση, η συνάρτηση *ROC-AUC*, καθώς και η ανάλυση του πίνακα σύγχυσης (*confusion matrix*) και της αναλυτικής αναφοράς ταξινόμησης (*classification report*). Οι μετρικές αυτές παρέχουν μια ολοκληρωμένη εικόνα της συμπεριφοράς των μοντέλων, τόσο ως προς τη συνολική τους απόδοση όσο και ως προς τη διακριτική τους ικανότητα ανάμεσα στις επιμέρους κλάσεις.

4.6 Πειραματικά Αποτελέσματα

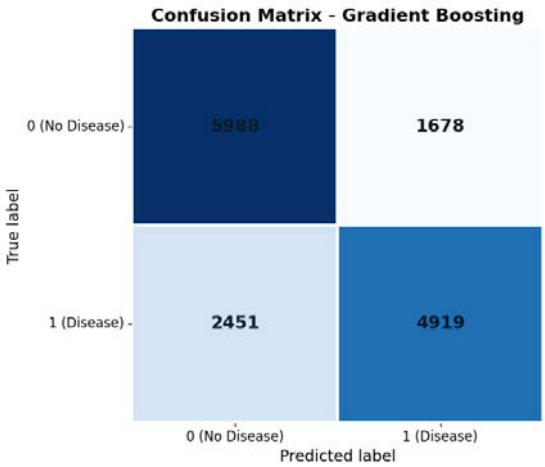
Αρχικά στο Σχήμα 4.9 παρατίθενται οι πίνακες σύγχυσης των τεσσάρων μοντέλων :



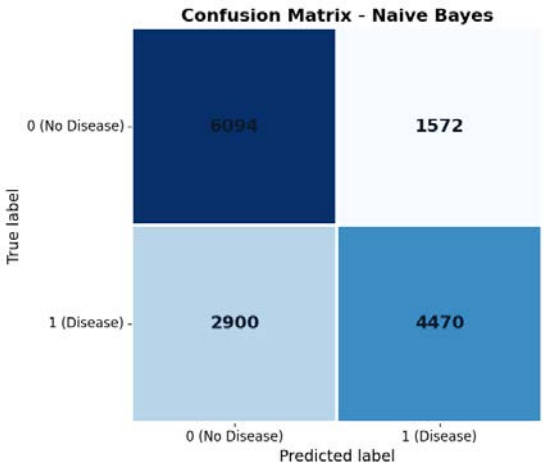
(α′) Πίνακας σύγχυσης: Logistic Regression.



(β′) Πίνακας σύγχυσης: Random Forest.



(γ′) Πίνακας σύγχυσης: Gradient Boosting.



(δ′) Πίνακας σύγχυσης: Naive Bayes.

Σχήμα 4.9: Πίνακες σύγχυσης για τα τέσσερα μοντέλα.

Τέλος, παρουσιάζεται και ο πίνακας σύγκρισης μετρικών αξιολόγησης (βλ. Πίνακα 4.1), για όλα τα μοντέλα.

Πίνακας 4.1: Αποτελέσματα απόδοσης βασικών μοντέλων ταξινόμησης

Μοντέλο	**Accuracy**	**Precision**	**Recall**	**ROC-AUC**
Logistic Regression	0.7174	0.7453	0.6433	0.7796
Random Forest	0.7227	0.7574	0.6389	0.7889
Gradient Boosting	0.7254	0.7456	0.6674	0.7911
Naive Bayes	0.7026	0.7398	0.6065	0.7603

4.7 Συζήτηση Αποτελεσμάτων

Η συγκριτική ανάλυση των αποτελεσμάτων των διαφορετικών αλγορίθμων ταξινόμησης καταδεικνύει τη σχετική αποτελεσματικότητα των επιλεγμένων μεθόδων στην πρόβλεψη της εμφάνισης καρδιαγγειακής νόσου.Από τον συγκριτικό πίνακα μετρικών (βλ. Πίνακα 4.1) παρατηρείται ότι όλοι οι αλγόριθμοι παρουσιάζουν παρόμοιες τιμές ακρίβειας.Το μοντέλο *Gradient Boosting Classifier* επιτυγχάνει την υψηλότερη τιμή ακρίβειας (*Accuracy* = 0.7254).Ακολουθεί το *Random Forest* με (*Accuracy* = 0.7227), ενώ το μοντέλο *Naive Bayes* σημείωσε τη χαμηλότερη τιμή ακρίβειας (*Accuracy* = 0.7026), επιβεβαιώνοντας τους περιορισμούς του σε πιο σύνθετα προβλήματα με αλληλεξαρτώμενες μεταβλητές.

Όσον αφορά στην **ευστοχία**,το *Random Forest* και το *Gradient Boosting* επιτυγχάνουν τις υψηλότερες τιμές, υποδηλώνοντας μεγαλύτερη αξιοπιστία στις θετικές προβλέψεις (εμφάνιση νόσου).

Η **ανάκληση**,είναι επίσης υψηλότερη για το *Gradient Boosting (0.6674)*, γεγονός που δείχνει ότι το μοντέλο έχει τη μεγαλύτερη ικανότητα να εντοπίζει τα πραγματικά περιστατικά νόσου.

Η τιμή του **ROC-AUC** για το *Gradient Boosting (0.7911)*, είναι η μεγαλύτερη μεταξύ όλων των μοντέλων, αποδεικνύοντας τη συνολική διακριτική ικανότητά του στην ταξινόμηση μεταξύ θετικών και αρνητικών περιπτώσεων.

Η εξέταση των **πινάκων σύγχυσης** επιβεβαιώνει πως τα μοντέλα *Gradient Boosting* και *Random Forest* εμφανίζουν αυξημένη ικανότητα στην ορθή αναγνώριση τόσο των ασθενών όσο και των μη ασθενών.Έχουν λιγότερα ψευδώς αρνητικά (*false negatives*) σε σχέση με τα υπόλοιπα μοντέλα.Πιο συγκεκριμένα, το *Gradient Boosting* εντόπισε 4.919 αληθώς θετικές περιπτώσεις, έναντι 4.741 του *Logistic Regression* και 4.470 του *Naive Bayes*. Ο αριθμός των ψευδώς θετικών (*false positives*) παραμένει σε διαχειρίσιμα επίπεδα, γεγονός που ενισχύει τη χρησιμότητα αυτών των μοντέλων για εφαρμογές πρόβλεψης σε πραγματικό περιβάλλον.

Κεφάλαιο 5

Εφαρμογή μοντέλων για την πρόβλεψη καρδιαγγειακής νόσου - Β'

5.1 Εισαγωγή

Στην παρούσα ενότητα παρουσιάζεται η δεύτερη πειραματική προσέγγιση, η οποία υλοποιήθηκε με τη χρήση του δημόσια διαθέσιμου συνόλου δεδομένων *Heart Disease dataset* [21]. Στόχος της ανάλυσης αυτής είναι η διερεύνηση της δυνατότητας πρόβλεψης της ύπαρξης καρδιαγγειακής νόσου, αξιοποιώντας ένα διαφορετικό σύνολο μεταβλητών και παραμέτρων σε σχέση με το προηγούμενο πείραμα, και εφαρμόζοντας σύγχρονες τεχνικές μηχανικής μάθησης.

Το συγκεκριμένο σύνολο δεδομένων περιέχει πληροφορίες για 1025 άτομα, καταγράφοντας πλήθος δημογραφικών, κλινικών και εργαστηριακών χαρακτηριστικών που σχετίζονται με την υγεία της καρδιάς, όπως η ηλικία, το φύλο, η πίεση ηρεμίας, τα επίπεδα χοληστερόλης, καθώς και άλλοι δείκτες σχετικοί με την καρδιακή λειτουργία. Η μεταβλητή-στόχος (*target*) αντιστοιχεί στην ύπαρξη ή μη καρδιαγγειακής νόσου, επιτρέποντας τη διατύπωση του προβλήματος ως δυαδική ταξινόμηση.

Η πειραματική διαδικασία που ακολουθήθηκε περιλαμβάνει όλα τα βασικά στάδια της ανάλυσης δεδομένων, από την αρχική εξερεύνηση και προεπεξεργασία, έως την εκπαίδευση και αξιολόγηση αλγορίθμων ταξινόμησης. Κεντρικός σκοπός της ενότητας αυτής είναι αφενός η σύγκριση της αποδοτικότητας διαφορετικών μοντέλων και αφετέρου η ανάδειξη κρίσιμων παραγόντων που επηρεάζουν την πρόβλεψη της καρδιαγγειακής νόσου, μέσα από τη μελέτη ενός εναλλακτικού συνόλου δεδομένων.

5.2 Περιγραφή Δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήθηκε στη δεύτερη πειραματική προσέγγιση αποτελείται από 1025 εγγραφές και 14 συνολικά χαρακτηριστικά, τα οποία σχετίζονται με παράγοντες κινδύνου και ενδείξεις καρδιαγγειακής νόσου. Πρόκειται για το *Heart Disease dataset*, το οποίο έχει χρησιμοποιηθεί ευρέως στη διεθνή βιβλιογραφία για την ανάπτυξη και αξιολόγηση μοντέλων πρόβλεψης καρδιακών παθήσεων.

Κάθε εγγραφή στο σύνολο δεδομένων αντιστοιχεί σε έναν ασθενή και περιλαμβάνει τόσο δημογραφικά όσο και κλινικά δεδομένα. Ενδεικτικά, καταγράφονται:

- **age** (ηλικία),
- **sex** (φύλο),
- **cp** (τύπος πόνου στο στήθος),
- **trestbps** (συστολική αρτηριακή πίεση ηρεμίας),
- **chol** (επίπεδα χοληστερόλης),
- **fbs** (γλυκόζη νηστείας),
- **restecg** (αποτελέσματα ηλεκτροκαρδιογραφήματος ηρεμίας),
- **thalach** (μέγιστη καταγραφείσα καρδιακή συχνότητα),
- **exang** (εμφάνιση στηθάγχης κατά την άσκηση),
- **oldpeak** (τιμή *oldpeak* - πτώση ST),
- **slope** (κλίση ST κατά την άσκηση),
- **ca** (αριθμός αγγείων που χρωματίστηκαν),
- **thal** (τύπος θαλλίου).

Η μεταβλητή-στόχος (`target`) λαμβάνει δυαδικές τιμές (0 ή 1), όπου το 1 υποδηλώνει την παρουσία καρδιαγγειακής νόσου και το 0 την απουσία της. Το συγκεκριμένο σύνολο δεδομένων δεν περιέχει ελλιπείς τιμές, γεγονός που απλοποιεί τη διαδικασία προεπεξεργασίας και ενισχύει την αξιοπιστία των πειραματικών αποτελεσμάτων. Επιπλέον, περιλαμβάνει τόσο αριθμητικά όσο και κατηγορικά χαρακτηριστικά, επιτρέποντας τη διερεύνηση πολύπλευρων παραγόντων που σχετίζονται με την υγεία της καρδιάς.

5.3 Προεπεξεργασία Δεδομένων

Για το *Heart Disease dataset*, η διαδικασία προεπεξεργασίας ξεκίνησε με τη διεξοδική εξερεύνηση των χαρακτηριστικών του συνόλου δεδομένων, προκειμένου να εντοπιστούν τυχόν σφάλματα, ελλιπείς τιμές ή ακραίες παρατηρήσεις που θα μπορούσαν να επηρεάσουν αρνητικά τα τελικά αποτελέσματα.

Αρχικά, εντοπίστηκαν και αφαιρέθηκαν οι εγγραφές με ελλιπείς τιμές (π.χ. *thal* = 0, *ca* = 4), ώστε να διασφαλιστεί η ποιότητα και πληρότητα του συνόλου δεδομένων. Στη συνέχεια, πραγματοποιήθηκε έλεγχος για ακραίες τιμές με στατιστικές μεθόδους και γραφήματα (όπως boxplots), χωρίς να προκύψει η ανάγκη μαζικής αφαίρεσης περαιτέρω εγγραφών.. Στη συνέχεια, πραγματοποιήθηκε έλεγχος για την κατανομή των μεταβλητών, καθώς και για την παρουσία πιθανών ακραίων τιμών σε βασικά κλινικά χαρακτηριστικά. Ο εντοπισμός ακραίων τιμών έγινε τόσο με στατιστικές μεθόδους (π.χ. περιγραφικά μέτρα, *boxplots*) όσο και με οπτικοποιήσεις, χωρίς να διαπιστωθεί η ανάγκη μαζικής αφαίρεσης εγγραφών.

Επιπρόσθετα, για τη βέλτιστη απόδοση των αλγορίθμων μηχανικής μάθησης που χρησιμοποιήθηκαν, εφαρμόστηκε κανονικοποίηση (*standardization*) στα αριθμητικά χαρακτηριστικά, μέσω της μεθόδου `StandardScaler` της βιβλιοθήκης *scikit-learn*. Η κανονικοποίηση κρίθηκε αναγκαία προκειμένου να αντιμετωπιστούν οι διαφορετικές κλίμακες τιμών ανάμεσα στα χαρακτηριστικά και να βελτιωθεί η σύγκλιση των αλγορίθμων κατά την εκπαίδευση.

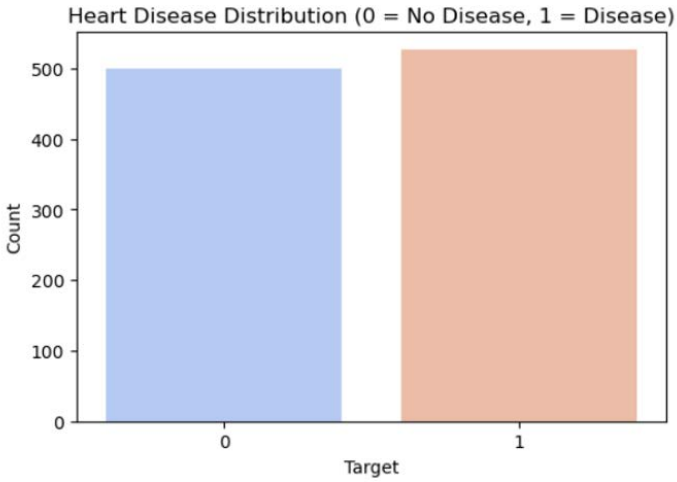
Η προεπεξεργασία ολοκληρώθηκε με τη μετατροπή των κατάλληλων χαρακτηριστικών σε αριθμητική μορφή, όπου απαιτούνταν, και τον διαχωρισμό του συνόλου δεδομένων σε σύνολο εκπαίδευσης και ελέγχου, προκειμένου να διασφαλιστεί η αντικειμενική αξιολόγηση της απόδοσης των μοντέλων.

5.4 Οπτικοποίηση Δεδομένων

Αρχικά παρουσιάζεται το διάγραμμα κατανομής του πλήθους των εγγραφών ως προς τη στήλη-στόχο στο Σχήμα 5.1:

Ακολουθούν τα ιστογράμματα συχνοτήτων (βλ. Σχήμα 5.2) για ενδεικτικές στήλες του συνόλου δεδομένων συνοδευόμενα από την ανάλυσή τους:

- **Ιστόγραμμα ηλικίας (*age*)** (βλ. Σχήμα 5.2α′): Η ηλικία των ασθενών παρουσιάζει περίπου κανονική κατανομή με κεντρική τάση στα 55–60 έτη. Δεν υπάρχουν ακραίες



Σχήμα 5.1: Κατανομή εγγραφών ως προς τη μεταβλητή-στόχο.

τιμές ή σημαντικές ανωμαλίες.

- **Ιστόγραμμα τύπου πόνου στο στήθος (*cp*)** (βλ. Σχήμα 5.2β′): Η μεταβλητή cp παίρνει τιμές 0 (τυπική στηθάγχη), 1 (άτυπη στηθάγχη), 2 (μη στηθαγχικός πόνος), και 3 (ασύμπτωματικός), οι οποίες κωδικοποιούν τον τύπο του πόνου στο στήθος κάθε ασθενούς, με την τιμή 0 να εμφανίζεται συχνότερα.

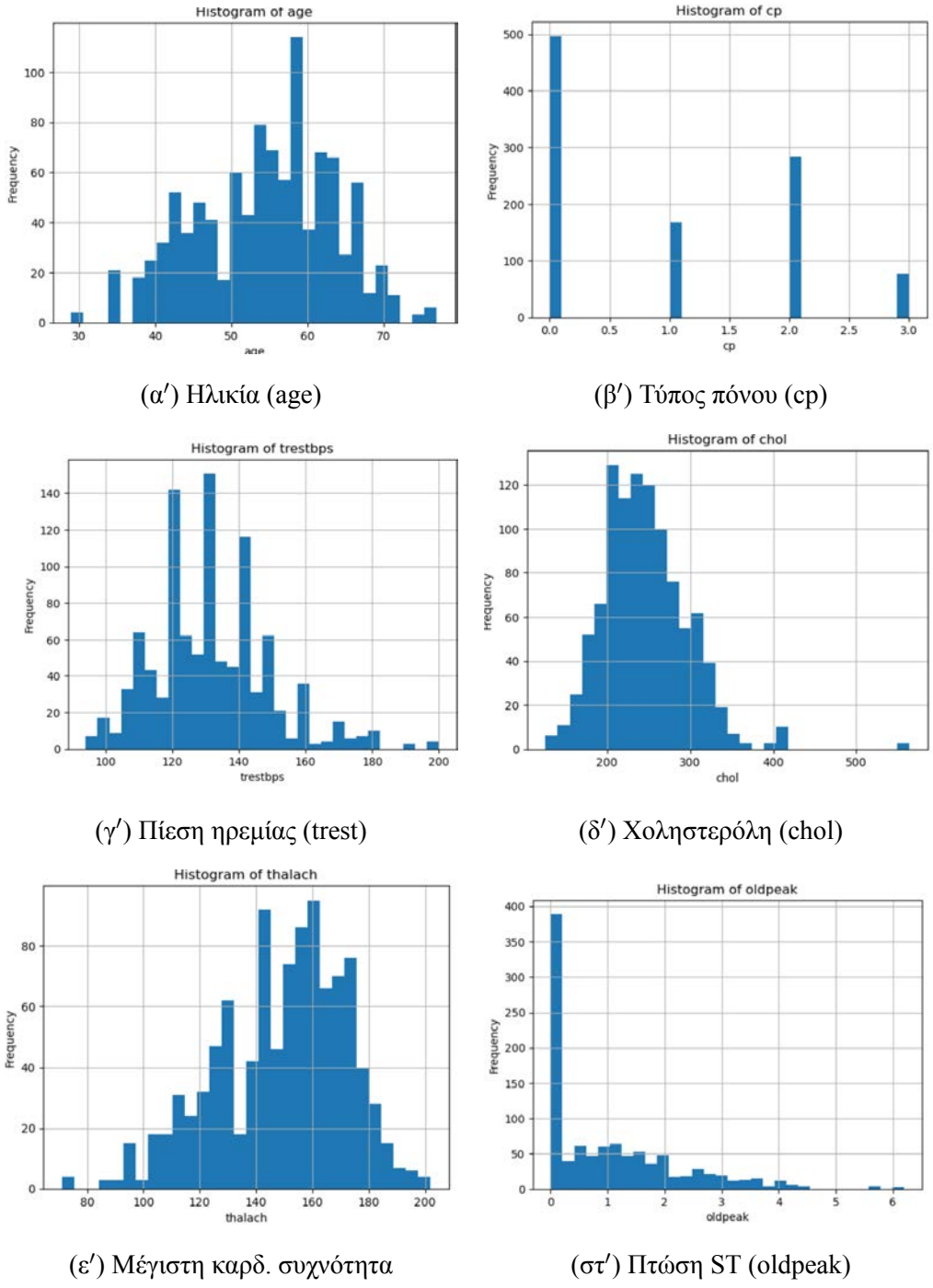
- **Ιστόγραμμα πίεσης ηρεμίας (*trest*)** (βλ. Σχήμα 5.2γ′): Η πίεση ηρεμίας συγκεντρώνεται κυρίως μεταξύ 110 και 140 mmHg, ενώ παρατηρούνται κάποιες υψηλές τιμές (>180 mmHg).

- **Ιστόγραμμα χοληστερόλης (*chol*)** (βλ. Σχήμα 5.2δ′): Παρουσιάζεται ασύμμετρη κατανομή με αρκετές ακραίες τιμές οι οποίες ενδέχεται να έχουν κλινική σημασία.Οι περιπτώσεις που ξεπερνούν τα 240 mg/dL συνδέονται με αυξημένο κίνδυνο καρδιαγγειακών παθήσεων [22].

- **Ιστόγραμμα μέγιστης καρδιακής συχνότητας (*thalac*)** (βλ. Σχήμα 5.2ε′): Η κατανομή είναι κανονική με μέση τιμή περίπου 140–160 bpm και χωρίς σημαντικές ακραίες τιμές.

- **Ιστόγραμμα πτώσης ST (*oldpeak*)** (βλ. Σχήμα 5.2στ′): Η μεταβλητή *oldpeak* [23] αντιστοιχεί στο βάθος της κατάθλιψης (πτώσης) του διαστήματος ST στο ηλεκτροκαρδιογράφημα κατά τη δοκιμασία κοπώσεως. Χαμηλές τιμές (κοντά στο 0) υποδηλώνουν φυσιολογική ανταπόκριση ή ήπια ισχαιμία, ενώ υψηλότερες τιμές (>1 mm) σχετίζονται με σημαντική ισχαιμία του μυοκαρδίου και σοβαρότερη καρδιακή δυσλειτουργία. Στο

συγκεκριμένο ιστόγραμμα, η πλειονότητα των τιμών βρίσκεται κοντά στο 0, ωστόσο παρατηρούνται και υψηλότερες τιμές, που υποδηλώνουν αυξημένο κίνδυνο για καρδιαγγειακά προβλήματα σε ορισμένους ασθενείς.



(α′) Ηλικία (age)

(β′) Τύπος πόνου (cp)

(γ′) Πίεση ηρεμίας (trest)

(δ′) Χοληστερόλη (chol)

(ε′) Μέγιστη καρδ. συχνότητα

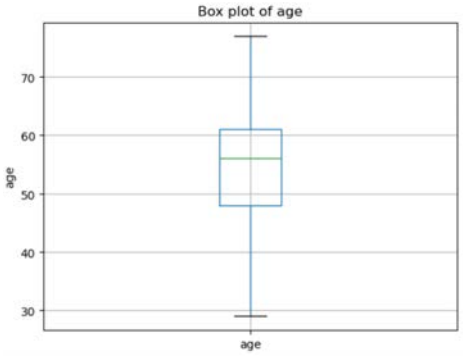
(στ′) Πτώση ST (oldpeak)

Σχήμα 5.2: Ιστογράμματα συχνοτήτων για βασικές στήλες του συνόλου δεδομένων.

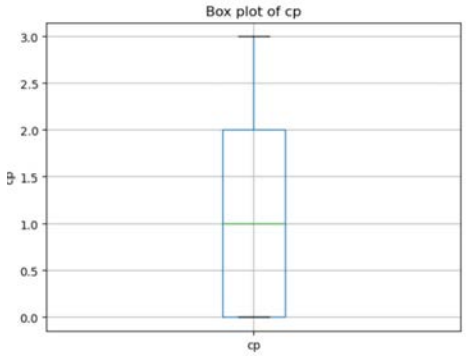
Η οπτικοποίηση συνεχίζεται με την παράθεση των *boxplot* διαγραμμάτων (βλ. Σχήμα 5.3) για τις αντίστοιχες στήλες συνοδευόμενων από την ανάλυσή τους:

- **Boxplot ηλικίας (*age*)** (βλ. Σχήμα 5.3α′): Η κατανομή εμφανίζει σχετικά ομαλή διασπορά με λίγες μικρές τιμές κάτω από το πρώτο τεταρτημόριο, χωρίς σημαντικές ακραίες τιμές ή ανωμαλίες.

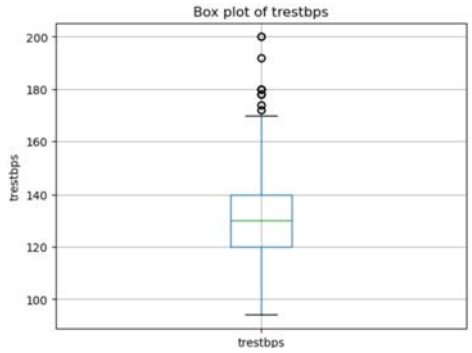
- **Boxplot τύπου πόνου στο στήθος (*cp*)** (βλ. Σχήμα 5.3β′): Δεδομένου ότι πρόκειται για κατηγορική μεταβλητή, τα δεδομένα απεικονίζονται ως διακριτές τιμές χωρίς ακραίες αποκλίσεις.



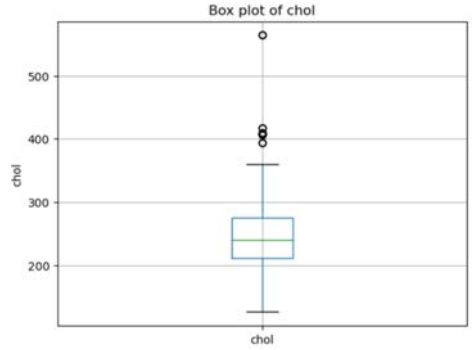
(α′) Ηλικία (age)



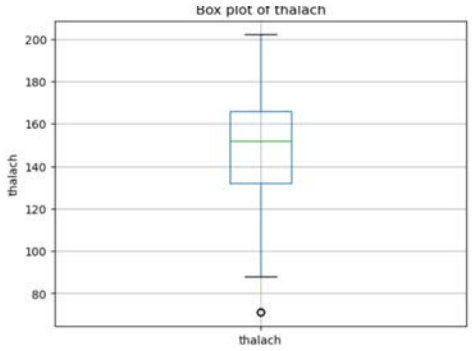
(β′) Τύπος πόνου (cp)



(γ′) Πίεση ηρεμίας (trest)



(δ′) Χοληστερόλη (chol)



(ε′) Μέγιστη καρδ. συχνότητα

Σχήμα 5.3: Boxplots για βασικές στήλες του συνόλου δεδομένων.

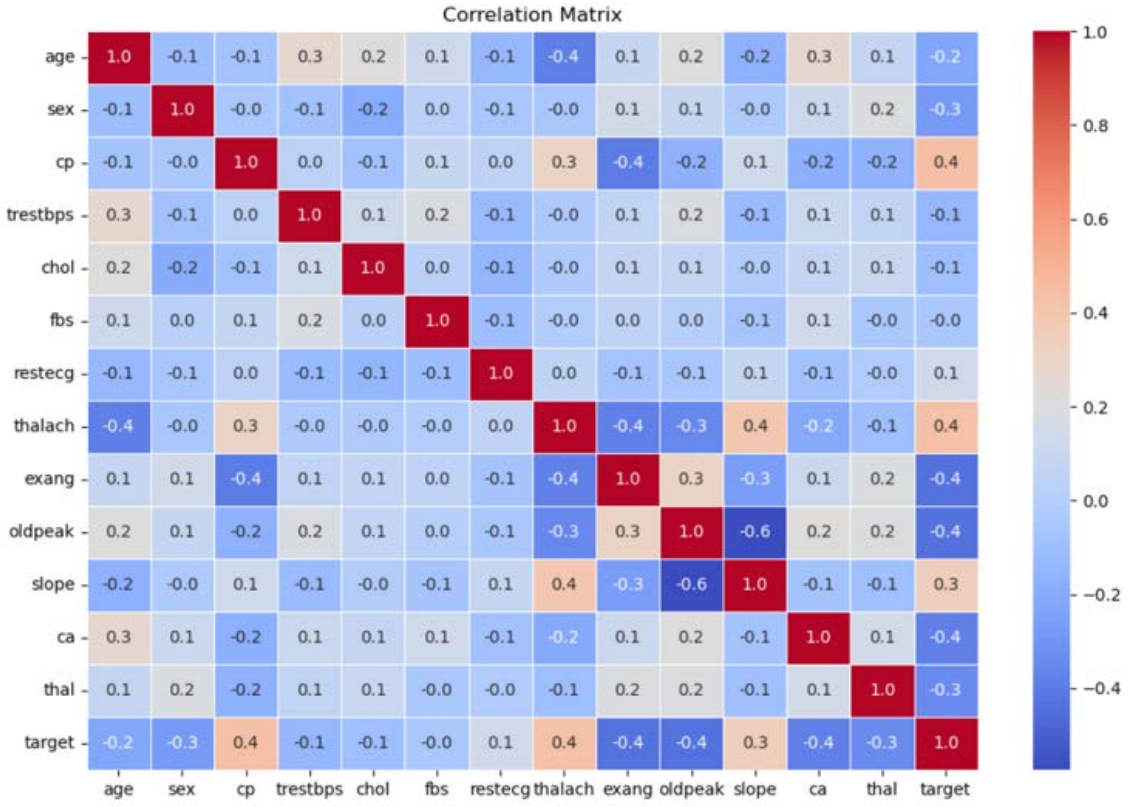
- **Boxplot πίεσης ηρεμίας (*trest*)** (βλ. Σχήμα 5.3γ′): Η πλειονότητα των τιμών εντοπί-

ζεται εντός φυσιολογικού εύρους (περίπου 90–120 mmHg), ωστόσο παρατηρούνται και υψηλότερες τιμές που ξεπερνούν τα 140 mmHg — τιμή που συνήθως υποδεικνύει υπέρταση. Οι ακραίες αυτές παρατηρήσεις αντανακλούν την κλινική ποικιλία του δείγματος, καθώς το σύνολο δεδομένων περιλαμβάνει και ασθενείς με αυξημένο καρδιαγγειακό κίνδυνο.

- **Boxplot χοληστερόλης (*chol*)** (βλ. Σχήμα 5.3δ′): Διακρίνονται αρκετές υψηλές τιμές εκτός του τυπικού εύρους, γεγονός που αντανακλά πιθανές περιπτώσεις υπερχοληστερολαιμίας στο δείγμα, χωρίς αυτό να αποτελεί απαραίτητα απόκλιση από την κλινική πραγματικότητα.

- **Boxplot μέγιστης καρδιακής συχνότητας (*thalac*)** (βλ. Σχήμα 5.3ε′): Η κατανομή είναι ομαλή με λίγες χαμηλές ακραίες τιμές.

Τέλος, παρατίθεται ο πίνακας συσχέτισης για το σύνολο δεδομένων (βλ. Σχήμα 5.4).



Σχήμα 5.4: Πίνακας συσχέτισης για τις αριθμητικές μεταβλητές του συνόλου δεδομένων.

5.5 Πειραματική Διαδικασία και Επιλογή Μεθόδων

Η πειραματική διαδικασία που ακολουθήθηκε για το Heart Disease dataset σχεδιάστηκε με στόχο τη συστηματική αξιολόγηση της αποτελεσματικότητας διαφόρων αλγορίθμων ταξινόμησης στην πρόβλεψη της ύπαρξης καρδιαγγειακής νόσου. Αρχικά, το σύνολο δεδομένων διαχωρίστηκε σε δύο τμήματα: σύνολο εκπαίδευσης και σύνολο ελέγχου, με αναλογία 75% και 25% αντίστοιχα, διασφαλίζοντας έτσι την ανεξαρτησία των δεδομένων κατά την εκπαίδευση και αξιολόγηση των μοντέλων.

Για τη διερεύνηση του προβλήματος επιλέχθηκαν και εφαρμόστηκαν ευρέως χρησιμοποιούμενοι αλγόριθμοι μηχανικής μάθησης, συγκεκριμένα ο ***Random Forest Classifier***, ο ***Logistic Regression***, ο ***Gradient Boosting Classifier*** και ο ***Naive Bayes***. Η επιλογή αυτών των αλγορίθμων βασίστηκε τόσο στη βιβλιογραφία όσο και στη σχετική τους απόδοση σε προβλήματα ταξινόμησης ιατρικών δεδομένων, επιτρέποντας τη συγκριτική αξιολόγηση διαφορετικών προσεγγίσεων ως προς την ακρίβεια, τη γενίκευση και την ερμηνευσιμότητα.

Η εκπαίδευση των μοντέλων πραγματοποιήθηκε με χρήση των προεπεξεργασμένων δεδομένων, ακολουθώντας τις βέλτιστες πρακτικές επιλογής και ρύθμισης των παραμέτρων τους. Για την εκτίμηση της απόδοσης κάθε μοντέλου αξιοποιήθηκαν μετρικές όπως η *ακρίβεια*, η *ευστοχία*, η *ανάκληση* και η *ROC-AUC*, ενώ έγινε και ανάλυση των *πινάκων σύγχυσης* για την πλήρη κατανόηση των επιδόσεων κάθε αλγορίθμου ως προς τα πραγματικά θετικά και αρνητικά περιστατικά.

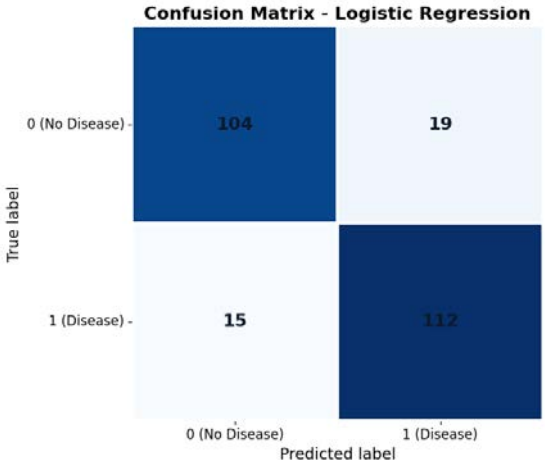
5.6 Πειραματικά Αποτελέσματα

Αρχικά παρατίθεται ο συγκριτικός Πίνακας (βλ. Πίνακα 5.1) με τις μετρικές αξιολόγησης των μοντέλων.

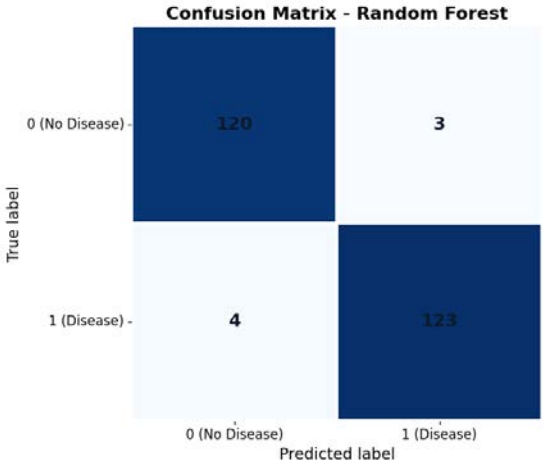
Μοντέλο	Accuracy	Precision	Recall	ROC-AUC
Logistic Regression	0.8640	0.8550	0.8819	0.9381
Random Forest	0.9720	0.9762	0.9685	0.9949
Gradient Boosting	0.9520	0.9752	0.9291	0.9930
Naive Bayes	0.8280	0.8333	0.8268	0.9159

Πίνακας 5.1: Σύγκριση Βασικών Μετρικών ανά Μοντέλο

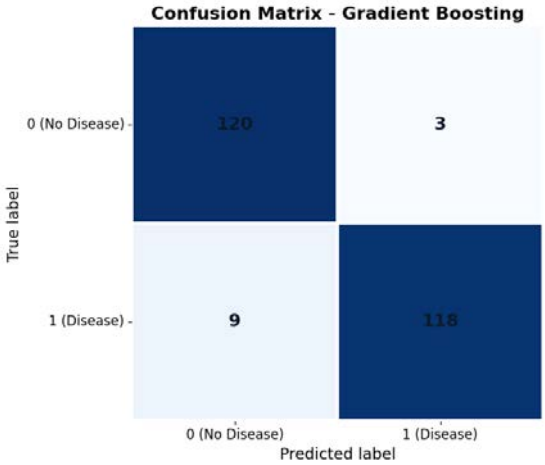
Ακολοθούν, οι πίνακες σύγχυσης των τεσσάρων μοντέλων (βλ. Σχήμα 5.5)



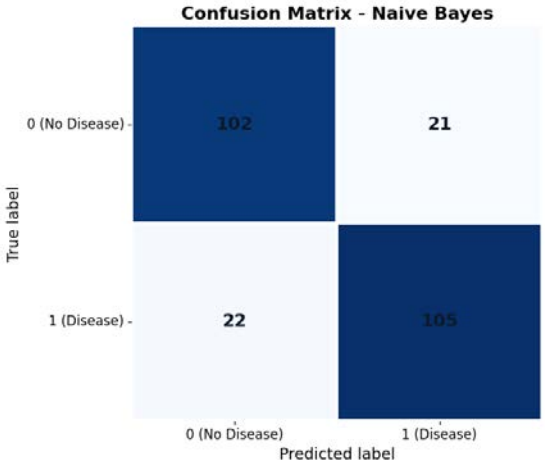
(α′) Πίνακας σύγχυσης: Logistic Regression.



(β′) Πίνακας σύγχυσης: Random Forest.



(γ′) Πίνακας σύγχυσης: Gradient Boosting.



(δ′) Πίνακας σύγχυσης: Naive Bayes.

Σχήμα 5.5: Πίνακες σύγχυσης για τα τέσσερα μοντέλα.

5.7 Συζήτηση Αποτελεσμάτων

Η ανάλυση των αποτελεσμάτων από την εφαρμογή των αλγορίθμων μηχανικής μάθησης στο *Heart Disease dataset* καταδεικνύει σημαντικές διαφορές στην απόδοση και τη διακριτική ικανότητα κάθε μοντέλου. Σύμφωνα με τον συγκριτικό Πίνακα 5.1. Τα *ensemble* μοντέλα, και ειδικότερα ο ***Random Forest Classifier*** και ο ***Gradient Boosting Classifier***, εμφάνισαν τις υψηλότερες επιδόσεις σε σχεδόν όλες τις μετρικές αξιολόγησης. Ο ***Random Forest Classifier*** σημείωσε την υψηλότερη τιμή ακρίβειας (*ακρίβεια* = 0.9720), υψηλή ευστοχία (*ευστοχία* = 0.9762) και σχεδόν τέλεια ανάκληση (*ανάκληση* = 0.9949). Η ανάλυση του *πίνακα σύγχυσης* επιβεβαιώνει την εξαιρετική του απόδοση, εντοπίζοντας σχεδόν το σύνολο των πε-

ριστατικών εμφάνισης νόσου και ελαχιστοποιώντας ψευδώς θετικά και αρνητικά. Ο ***Gradient Boosting Classifier*** παρουσίασε παρόμοια επίπεδα επιδόσεων, γεγονός που επιβεβαιώνει τη σταθερά υψηλή αποδοτικότητα των μεθόδων *ensemble* στην πρόβλεψη πολύπλοκων ιατρικών φαινομένων.

Αντίθετα, το μοντέλο ***Logistic Regression*** εμφάνισε χαμηλότερες επιδόσεις με τιμή ακρίβειας (*ακρίβεια* = 0.8640) και τιμή ευστοχίας (*ευστοχία* = 0.8550). Η χαμηλότερη επίδοση του μοντέλου αποδίδεται κυρίως στους περιορισμούς που προκύπτουν από την υπόθεση γραμμικότητας μεταξύ των χαρακτηριστικών. Παρά ταύτα, παρουσιάζει ικανοποιητική απόδοση λόγω της απλότητας και ταχύτητάς του.

Το ***Naive Bayes*** κατέγραψε αξιοσημείωτη συμπεριφορά με τιμή ακρίβειας (*ακρίβεια* = 0.8280) και τιμή ευστοχίας (*ευστοχία* = 0.8333). Ο πίνακας σύγχυσης αναδεικνύει υψηλή ικανότητα εντοπισμού ατόμων με καρδιαγγειακή νόσο *(true positives)*. Ωστόσο εμφάνισε αυξημένα ψευδώς θετικά (*false positives*) σε σχέση με τα *ensemble* μοντέλα.

Η αξιολόγηση μέσω ***ROC-AUC*** επιβεβαιώνει τη συνολική υπεροχή των *ensemble* αλγορίθμων, καθώς ο ***Gradient Boosting*** και ο ***Random Forest*** ξεπερνούν το 0.99, δηλώνοντας εξαιρετική διακριτική ικανότητα ανάμεσα στις δύο κλάσεις.

Κεφάλαιο 6

Εφαρμογή μοντέλων για την πρόβλεψη καρκίνου του στόματος

Στην παρούσα ενότητα παρουσιάζεται η τρίτη πειραματική προσέγγιση, η οποία αφορά τη διάγνωση καρκίνου του στόματος με τη χρήση τεχνικών βαθιάς μάθησης σε ιατρικές εικόνες. Η ανάλυση αυτή βασίζεται σε ένα ευρύ και ισορροπημένο σύνολο εικόνων δύο κατηγοριών [24], με στόχο την αυτόματη διάκριση μεταξύ φυσιολογικών περιπτώσεων και περιπτώσεων ακανθοκυτταρικού καρκινώματος (*Squamous Cell Carcinoma*).

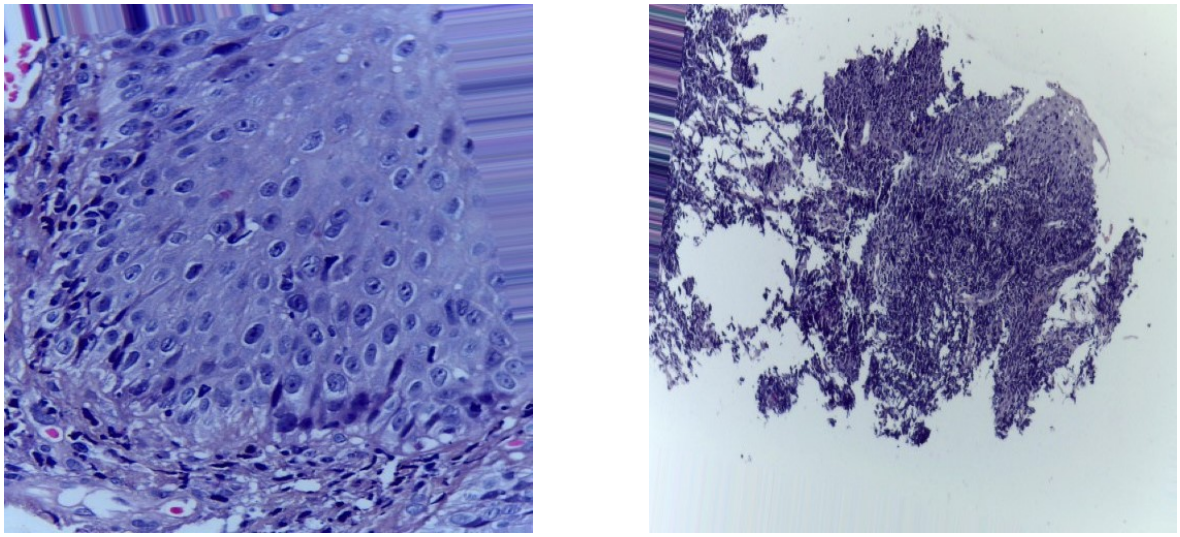
Η βασική καινοτομία του πειραματικού αυτού μέρους είναι η αξιολόγηση και σύγκριση δύο διαφορετικών προσεγγίσεων μηχανικής όρασης: της εκπαίδευσης ενός συνελικτικού νευρωνικού δικτύου (*CNN*) από την αρχή και της αξιοποίησης μεταφοράς μάθησης (*transfer learning*) μέσω του προεκπαιδευμένου μοντέλου *MobileNetV2*. Μέσω αυτών των μεθοδολογιών επιχειρείται να διερευνηθεί ο βαθμός αποτελεσματικότητας κάθε προσέγγισης στην έγκαιρη και αξιόπιστη ανίχνευση του στοματικού καρκίνου, συμβάλλοντας στην προαγωγή της αυτοματοποιημένης ιατρικής διάγνωσης.

6.1 Περιγραφή Δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήθηκε σε αυτή την πειραματική διαδικασία αποτελείται από 10.002 ιατρικές εικόνες του στοματικού βλεννογόνου, χωρισμένες σε δύο ισοδύναμες κατηγορίες: *Normal* (φυσιολογικές περιπτώσεις) και *Squamous Cell Carcinoma* (περιπτώσεις ακανθοκυτταρικού καρκινώματος). Κάθε κατηγορία περιλαμβάνει 5.001 εικόνες, διασφαλίζοντας έτσι ισορροπία κλάσεων και αποφεύγοντας το πρόβλημα της μεροληψίας προς

κάποια κατηγορία.

Οι εικόνες προέρχονται από διάφορες πηγές ιατρικής έρευνας και διαγνωστικών κέντρων και έχουν μετατραπεί σε ενιαία μορφή και μέγεθος, κατάλληλη για είσοδο σε νευρωνικά δίκτυα. Η υψηλή ανάλυση και η ποικιλία των εικόνων επιτρέπουν τη μελέτη λεπτομερειών που σχετίζονται με την παθολογία του στοματικού καρκίνου, καθιστώντας το συγκεκριμένο σύνολο δεδομένων ιδανικό για την ανάπτυξη και αξιολόγηση συστημάτων βαθιάς μάθησης.



(α') τύπου α – φυσιολογικό (β') τύπου β – ακανθοκυτταρικό καρκίνωμα

Σχήμα 6.1: Ενδεικτικό παράδειγμα εικόνας από το σύνολο δεδομένων στοματικού βλεννογόνου.

Στη συνέχεια, παρουσιάζονται αναλυτικά οι δύο διαφορετικές υλοποιήσεις που εφαρμόστηκαν στο παραπάνω σύνολο δεδομένων: (1) εκπαίδευση *CNN* από την αρχή και (2) μεταφορά μάθησης με *MobileNetV2*.

6.2 Πρόβλεψη με Συνελικτικό Νευρωνικό Δίκτυο (CNN)

Η πρώτη παραλλαγή υλοποιήθηκε μέσω της κατασκευής και εκπαίδευσης ενός συνελικτικού νευρωνικού δικτύου (*CNN*) από το μηδέν. Η προεπεξεργασία (*preprocessing*) των δεδομένων περιλάμβανε το χωρίσμο (*splitting*) των εικόνων σε σύνολα εκπαίδευσης (*training*), επικύρωσης (*validation*) και ελέγχου (*testing*), τη μετατροπή των εικόνων σε ενιαίο μέγεθος και την κανονικοποίησή τους (*normalization*). Για την ενίσχυση της γενικευσιμότητας του μοντέλου, εφαρμόστηκαν τεχνικές εμπλουτισμού δεδομένων, όπως τυχαίες αναστροφές, περιστροφές, μεταβολές στην αντίθεση και τη φωτεινότητα.

Επιπλέον, το δίκτυο αποτελείται από τρία συνελικτικά επίπεδα με αυξανόμενο αριθμό φίλτρων και επίπεδα υποδειγματοληψίας για μείωση των διαστάσεων. Πριν την έξοδο, εφαρμόζεται απόρριψη νευρώνων (*dropout*) 50% για αποφυγή υπερπροσαρμογής. Η εκπαίδευση πραγματοποιήθηκε με το βελτιστοποιήτη (*optimizer*) *Adam* (ρυθμός μάθησης 0.0005) και binary crossentropy, ενώ εφαρμόστηκε και πρώιμος τερματισμός (*early stopping*) με υπομονή (*patience*) 6 και επαναφορά των καλύτερων βαρών.

Στο Σχήμα 6.2 απεικονίζεται η αρχιτεκτονική του συνελικτικού νευρωνικού δικτύου:

Layer (type)	Output Shape	Param #
sequential (Sequential)	(None, 128, 128, 3)	0
rescaling (Rescaling)	(None, 128, 128, 3)	0
conv2d (Conv2D)	(None, 128, 128, 32)	896
max_pooling2d (MaxPooling2D)	(None, 64, 64, 32)	0
conv2d_1 (Conv2D)	(None, 64, 64, 64)	18,496
max_pooling2d_1 (MaxPooling2D)	(None, 32, 32, 64)	0
conv2d_2 (Conv2D)	(None, 32, 32, 128)	73,856
max_pooling2d_2 (MaxPooling2D)	(None, 16, 16, 128)	0
flatten (Flatten)	(None, 32768)	0
dense (Dense)	(None, 128)	4,194,432
dropout (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 1)	129

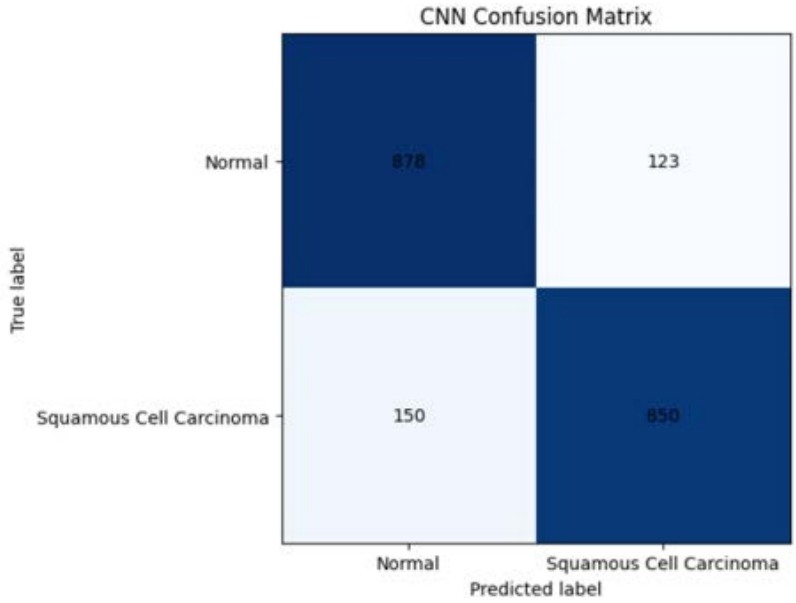
Total params: 4,287,809 (16.36 MB)
Trainable params: 4,287,809 (16.36 MB)
Non-trainable params: 0 (0.00 B)

Σχήμα 6.2: Πειραματική διαδικασία CNN

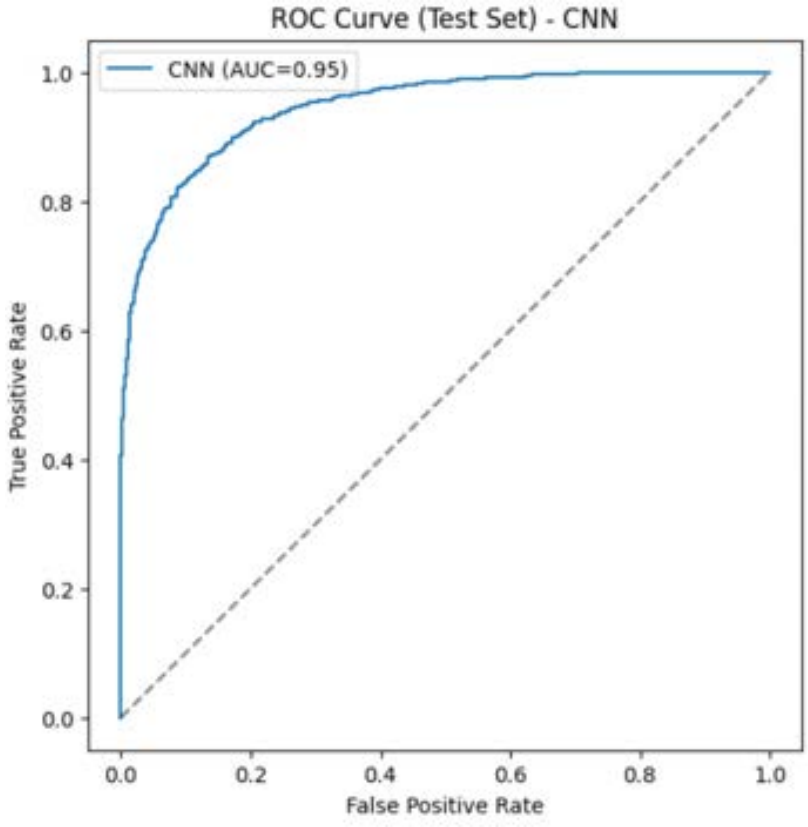
Ακολουθεί ο Πίνακας 6.1 με τα αποτελέσματα των μετρικών για το συγκεκριμένο μοντέλο. Στο Σχήμα 6.3 παρατίθεται ο πίνακας σύγχυσης. Στη συνέχεια εμφανίζεται η καμπύλη *ROC* στο Σχήμα 6.4. Τέλος, στο Σχήμα 6.5 απεικονίζονται οι καμπύλες εκπαίδευσης και επικύρωσης του μοντέλου :

Μοντέλο	Accuracy	Precision	Recall	F1-score	ROC-AUC
CNN	0.8636	0.8736	0.85	0.8616	0.9462

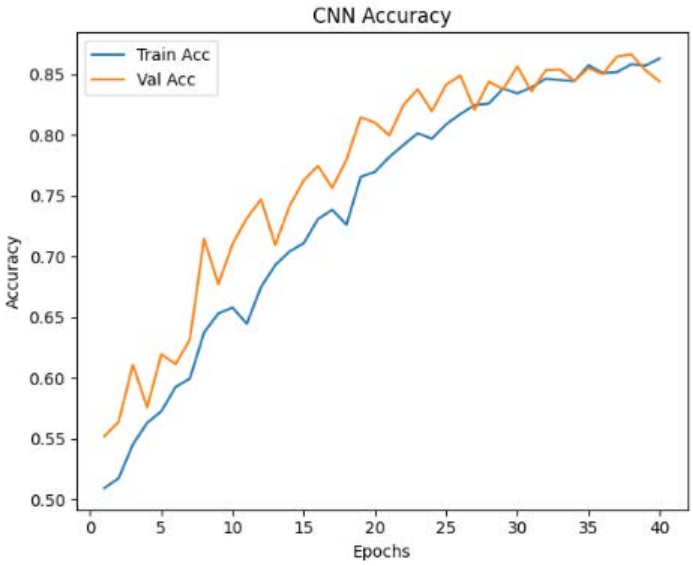
Πίνακας 6.1: Αποτελέσματα απόδοσης του μοντέλου CNN



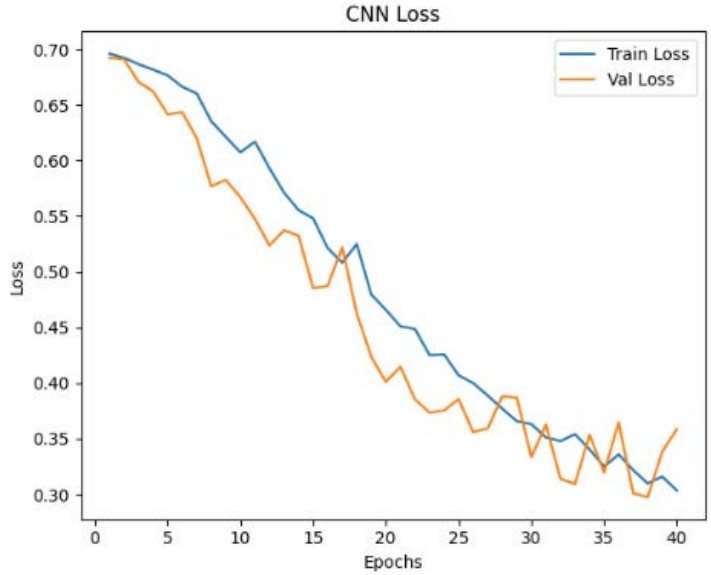
Σχήμα 6.3: Πίνακας σύγχυσης CNN



Σχήμα 6.4: Καμπύλη ROC CNN



(α′) Καμπύλη εκπαίδευσης για το CNN



(β′) Καμπύλη επικύρωσης για το CNN

Σχήμα 6.5: Καμπύλες εκπαίδευσης και επικύρωσης του μοντέλου CNN.

6.3 Πρόβλεψη με Μεταφορά Μάθησης (MobileNetV2)

Η δεύτερη παραλλαγή υλοποιήθηκε με χρήση μεταφοράς μάθησης βασισμένη στο προεκπαιδευμένο συνελικτικό νευρωνικό δίκτυο *MobileNetV2*, το οποίο έχει εκπαιδευτεί στο μεγάλο και ποικιλόμορφο σύνολο *ImageNet*, παρέχοντας μια ισχυρή βάση γνώσης για βασικά οπτικά χαρακτηριστικά όπως ακμές, υφές και σχήματα.

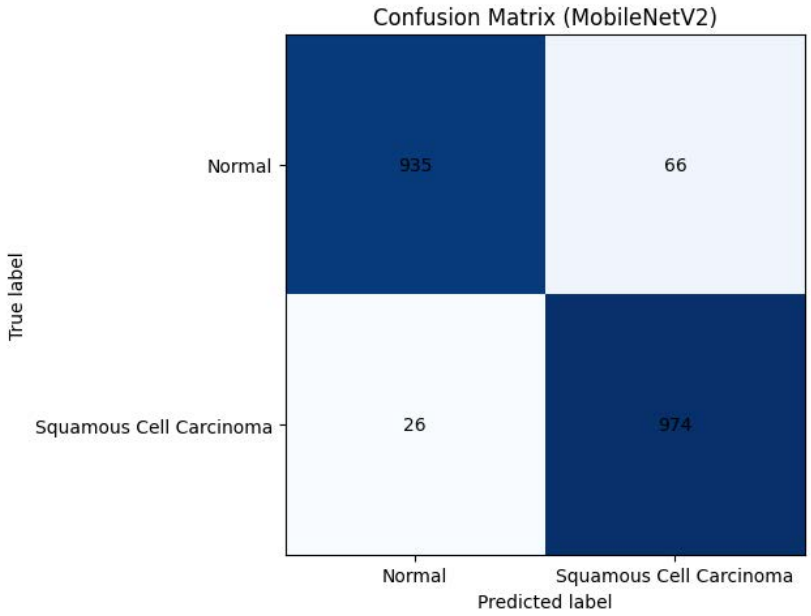
Στην υλοποίηση, το *MobileNetV2* φορτώθηκε χωρίς το τελικό επίπεδο ταξινόμησης, ώστε να χρησιμοποιηθούν μόνο τα προεκπαιδευμένα χαρακτηριστικά που εξάγει. Για να προσαρμοστεί το μοντέλο στα δεδομένα του *Oral Cancer Dataset*,χρησιμοποιήθηκε ακριβής προσαρμογή. Ξεπαγώθηκαν τα τελευταία 30% των στρωμάτων, ενώ τα πρώτα 70% παρέμειναν παγωμένα, διατηρώντας σταθερά τα ήδη εκμαθημένα βάρη. Η είσοδος του μοντέλου περιλαμβάνει εφαρμογή τεχνικών εμπλουτισμού δεδομένων. Δοκιμάστηκαν, όπως και στη υλοποίηση με το *CNN*, τυχαίες αναστροφές, περιστροφές, μεταβολές στην αντίθεση και τη φωτεινότητα, ώστε να αυξηθεί η γενικευσιμότητα. Στη συνέχεια, γίνεται κανονικοποίηση (*rescaling* στο διάστημα [0,1]) των εικονοστοιχείων (*pixel*) πριν την είσοδο στο βασικό μοντέλο *MobileNetV2*. Τα χαρακτηριστικά που προκύπτουν από το *MobileNetV2* συμπυκνώνονται με μέσο όρο σε κάθε κανάλι με *global average pooling*, ενώ εφαρμόζονται στρώματα απόρριψης νευρώνων (*dropout layers*) για την αποφυγή υπερπροσαρμογής. Ακολουθεί ένα πλήρως συνδεδεμένο επίπεδο με 128 νευρώνες και ενεργοποίηση *ReLU*, και άλλο ένα στρώμα απόρριψης. Τέλος, το τελικό επίπεδο εξόδου αποτελείται από έναν νευρώνα με *sigmoid* ενεργοποίηση για δυαδική ταξινόμηση. Η εκπαίδευση του μοντέλου πραγματοποιείται με τον βελτιστοποιητή *Adam* και *binary crossentropy loss*. Χρησιμοποιούνται μηχανισμοί επανάκλησης όπως το πρώιμο σταμάτημα, που σταματάει την εκπαίδευση αν η απώλεια επικύρωσης *validation loss* δεν βελτιώνεται για 6 συνεχόμενες επαναλήψεις εκπαίδευσης(*epochs*) και η μείωση του ρυθμού μάθησης (*ReduceLROnPlateau*), που μειώνει το ρυθμό μάθησης αν το λάθος επικύρωσης σταθεροποιηθεί, βελτιώνοντας τη σταθερότητα της εκπαίδευσης.

Τέλος, η αξιολόγηση του μοντέλου γίνεται σε ξεχωριστό σύνολο δοκιμής, όπου υπολογίζονται μετρικές όπως ακρίβεια, ευστοχία, ανάκληση, *F1-score* και *ROC-AUC*, παρέχοντας ολοκληρωμένη εικόνα της απόδοσης όπως φαίνεται και στον Πίνακα 6.2.

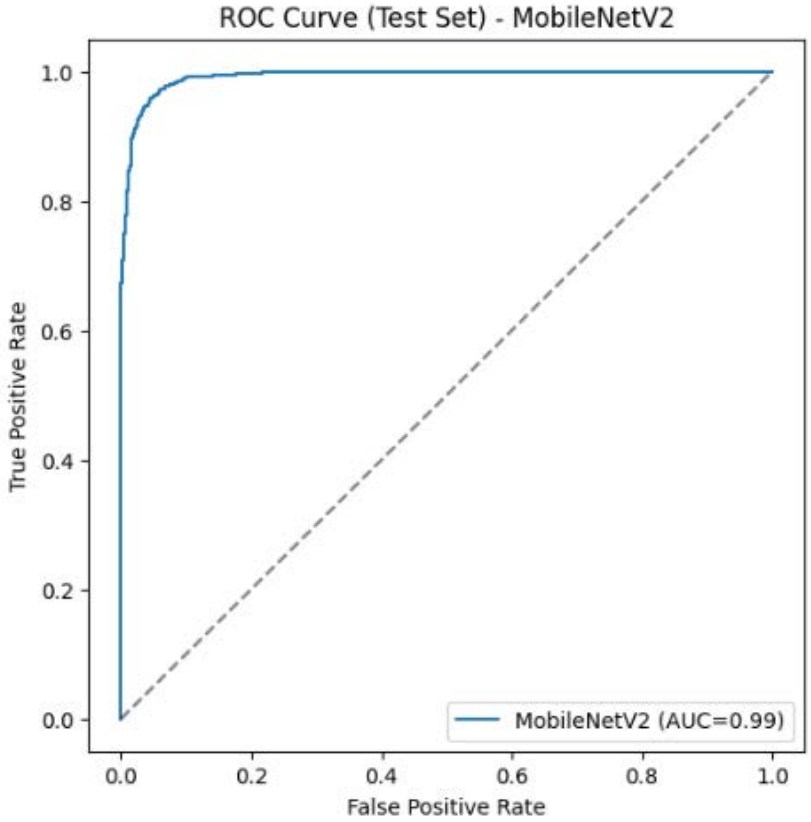
Μοντέλο	Accuracy	Precision	Recall	F1-score	ROC-AUC
MobileNetV2	0.9540	0.9365	0.974	0.9549	0.9926

Πίνακας 6.2: Αποτελέσματα απόδοσης του μοντέλου MobileNetV2

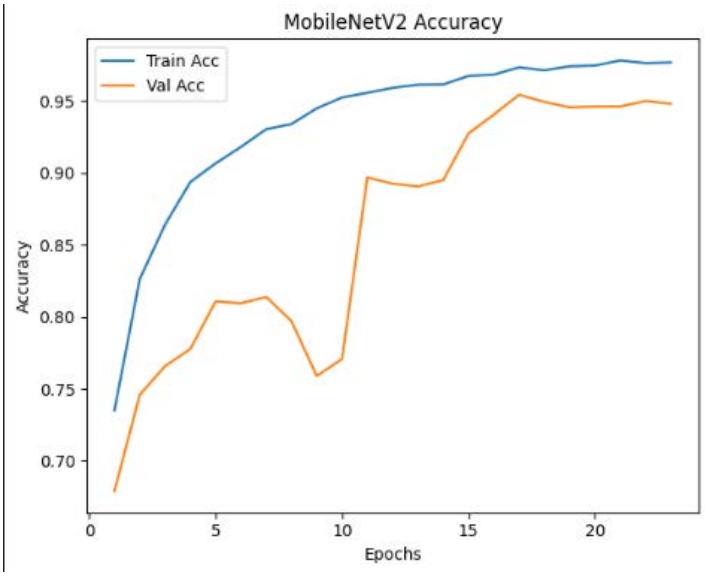
Ακολουθεί ο πίνακας σύγχυσης για το *MobileNetV2* στο Σχήμα 6.6, ενώ στο Σχήμα 6.7 παρουσιάζεται η καμπύλη *ROC*. Τέλος στο Σχήμα 6.8 απεικονίζονται οι καμπύλες εκπαίδευσης και επικύρωσης:



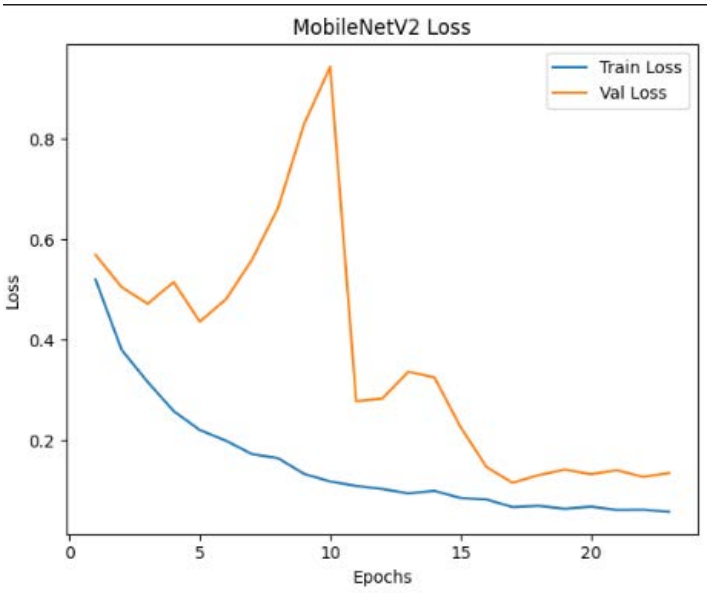
Σχήμα 6.6: Πίνακας σύγχυσης: MobileNetV2



Σχήμα 6.7: Καμπύλη ROC MobileNetV2



(α′) Καμπύλη εκπαίδευσης: MobileNetV2



(β′) Καμπύλη επικύρωσης: MobileNetV2

Σχήμα 6.8: Καμπύλες εκπαίδευσης και επικύρωσης του μοντέλου MobileNetV2.

6.4 Συγκριτική Ανάλυση CNN και MobileNetV2

Η σύγκριση των δύο διαφορετικών προσεγγίσεων στο Πίνακα 6.3, δηλαδή της εκπαίδευσης ενός κλασικού συνελικτικού νευρωνικού δικτύου (*CNN*) από την αρχή και της μεταφοράς μάθησης με τη χρήση του προεκπαιδευμένου μοντέλου *MobileNetV2*, αναδεικνύει σαφώς την ανωτερότητα της δεύτερης μεθόδου σε όλα τα επίπεδα αξιολόγησης.

Μοντέλο	Accuracy	Precision	Recall	F1-score	ROC-AUC
CNN	0.8636	0.8736	0.85	0.8616	0.9462
MobileNetV2	0.9540	0.9365	0.974	0.9549	0.9926

Πίνακας 6.3: Σύγκριτικός πίνακας απόδοσης των μοντέλων CNN και MobileNetV2

Η ανάλυση των πινάκων σύγχυσης του *CNN* (βλ. Σχήμα 6.3) και του *MobileNetV2* (βλ. Σχήμα 6.6) φανερώνει μείωση των ψευδώς αρνητικών περιπτώσεων (26 για *MobileNetV2* έναντι 150 για *CNN*), στοιχείο κρίσιμο για ιατρικές εφαρμογές όπου απαιτείται ελαχιστοποίηση των διαφυγόντων διαγνώσεων, καθώς και μείωση των ψευδώς θετικών περιπτώσεων, συμβάλλοντας στη μείωση άσκοπων εξετάσεων και ψυχολογικής επιβάρυνσης στους ασθενείς.

Η σαφής υπεροχή της *MobileNetV2* αποδίδεται στην αξιοποίηση προεκπαιδευμένων χαρακτηριστικών από το μεγάλο σύνολο δεδομένων *ImageNet*, καθώς η μεταφορά μάθησης επιτρέπει την ανίχνευση πολύπλοκων μοτίβων σε ιατρικές εικόνες που είναι δύσκολα ανιχνεύσιμα από δίκτυα με τυχαία αρχικοποίηση όπως το *CNN*, ενώ παράλληλα μειώνει την ανάγκη για τεράστια ποσότητα εξειδικευμένων δεδομένων, διευκολύνοντας τη χρήση της μεθόδου και σε περιπτώσεις περιορισμένων επισημασμένων εικόνων. Οι καμπύλες μάθησης του *MobileNetV2* (βλ. Σχήμα 6.8) παρουσιάζουν σταθερή βελτίωση και ταχεία σύγκλιση, με πολύ μικρή απόκλιση μεταξύ του συνόλου εκπαίδευσης και του συνόλου επικύρωσης, γεγονός που υποδεικνύει καλή γενίκευση και περιορισμό της υπερπροσαρμογής. Να σημειωθεί ότι η τιμή του *AUC* (= 0.99) αποτελεί ισχυρό δείκτη της διαχωριστικής ισχύος του μοντέλου, σημαντικό για την αποτελεσματική διαχείριση ιατρικών κινδύνων.

Συνοψίζοντας, η μεταφορά μάθησης με σύγχρονες αρχιτεκτονικές όπως η *MobileNetV2* προσφέρει σαφή πλεονεκτήματα έναντι των δικτύων που εκπαιδεύονται από το μηδέν και έχει ουσιαστικές προεκτάσεις για την κλινική πράξη, επιτρέποντας την ανάπτυξη ισχυρών διαγνωστικών εργαλείων με περιορισμένους υπολογιστικούς πόρους.

Κεφάλαιο 7

Τεχνικές λεπτομέρειες

Στο κεφάλαιο αυτό παρουσιάζονται συνοπτικά τα βασικά εργαλεία και περιβάλλοντα που χρησιμοποιήθηκαν για την υλοποίηση του πειραματικού μέρους της διπλωματικής εργασίας. Ο κώδικας αναπτύχθηκε στη γλώσσα προγραμματισμού *Python* και εκτελέστηκε στο διαδικτυακό περιβάλλον *Google Colab*.

7.1 Η γλώσσα προγραμματισμού Python

Η *Python* αποτελεί μια από τις πιο δημοφιλείς και ευέλικτες γλώσσες προγραμματισμού, ιδανική για εφαρμογές *μηχανικής μάθησης* και *ανάλυσης δεδομένων*. Η *Python* υποστηρίζει *αντικειμενοστραφή* αλλά και *δομημένο προγραμματισμό*, δίνοντας τη δυνατότητα ανάπτυξης κώδικα σε διάφορα επίπεδα πολυπλοκότητας. Στην παρούσα εργασία χρησιμοποιήθηκαν αποκλειστικά οι εξής βασικές βιβλιοθήκες:

- ***pandas***: Βιβλιοθήκη για εύκολο χειρισμό και ανάλυση δεδομένων, με έμφαση στους πίνακες δεδομένων (*DataFrames*). Υποστηρίζει εισαγωγή και εξαγωγή σε μορφές όπως *CSV* και *Excel*.

- ***numpy***: Θεμελιώδης βιβλιοθήκη για αριθμητικούς υπολογισμούς και διαχείριση πολυδιάστατων πινάκων (*arrays*).

- ***scikit-learn (sklearn)***: Βιβλιοθήκη *μηχανικής μάθησης* με υλοποιήσεις πλήθους αλγορίθμων *ταξινόμησης*, *παλινδρόμησης* και *ομαδοποίησης*, καθώς και εργαλεία *προεπεξεργασίας* και *αξιολόγησης μοντέλων*.

- ***TensorFlow***: Πλατφόρμα *ανοιχτού κώδικα* για την κατασκευή και εκπαίδευση *βαθιών νευρωνικών δικτύων*. Χρησιμοποιήθηκε για την υλοποίηση και εκπαίδευση των *συνελικτικών νευρωνικών δικτύων*.

- ***matplotlib*** και ***seaborn***: Βιβλιοθήκες για *οπτικοποίηση δεδομένων* και αποτελεσμάτων μέσω γραφημάτων και διαγραμμάτων.

7.2 Google Colab

Το *Google Colaboratory (Colab)* είναι ένα δωρεάν *online* περιβάλλον προγραμματισμού *Python* που βασίζεται σε *Jupyter notebooks*. Επιτρέπει την εκτέλεση κώδικα σε απομακρυσμένους *server* της *Google*, προσφέροντας πρόσβαση σε *GPU* και *TPU* για επιτάχυνση υπολογισμών. Τα βασικά πλεονεκτήματα του περιβάλλοντος είναι:

- Απλή πρόσβαση με λογαριασμό *Google* και σύνδεση στο διαδίκτυο, χωρίς τοπική εγκατάσταση.

- Υποστήριξη εκτέλεσης απαιτητικών εφαρμογών ακόμα και σε συσκευές με περιορισμένους πόρους.

- Δυνατότητα ομαδικής συνεργασίας μέσω κοινής χρήσης των *notebooks*.

Κεφάλαιο 8

Επίλογος

Η παρούσα διπλωματική εργασία ασχολήθηκε με τη μελέτη, υλοποίηση και συγκριτική αξιολόγηση σύγχρονων μεθόδων μηχανικής και βαθιάς μάθησης για την πρόβλεψη και τη διάγνωση σημαντικών ασθενειών, όπως τα καρδιαγγειακά νοσήματα και ο καρκίνος του στόματος. Μέσα από την ανάλυση διαφορετικών συνόλων δεδομένων – τόσο δομημένων όσο και απεικονιστικών – υλοποιήθηκαν, εκπαιδεύτηκαν και αξιολογήθηκαν διάφορα μοντέλα ταξινόμησης, με στόχο την ανάδειξη των δυνατοτήτων και των περιορισμών κάθε προσέγγισης.

8.1 Σύνοψη και συμπεράσματα

Στην περίπτωση της πρόβλεψης καρδιαγγειακών παθήσεων, η συγκριτική ανάλυση μοντέλων όπως το *Logistic Regression* , το *Random Forest*, το *Gradient Boosting* και ο *Naive Bayes*, έδειξε ότι οι μέθοδοι *ensemble* (όπως το *Random Forest* και το *Gradient Boosting*) υπερέχουν συνήθως σε απόδοση, κυρίως ως προς τις μετρικές της ανάκλησης και του *F1-score*. Ωστόσο, η λογιστική παλινδρόμηση διατηρεί πλεονεκτήματα ως προς την ερμηνευσιμότητα και την ευκολία εφαρμογής σε ιατρικά δεδομένα, ενώ ο *Naive Bayes* μπορεί να χρησιμοποιηθεί ως απλή γραμμή βάσης (*baseline*). Οι πειραματικές εφαρμογές σε δύο διαφορετικά ιατρικά σύνολα δεδομένων ανέδειξαν τη σημασία της σωστής προεπεξεργασίας, της επιλογής κατάλληλων μετρικών αξιολόγησης και της βελτιστοποίησης των υπερπαραμέτρων για τη μέγιστη απόδοση των μοντέλων

Αναφορικά με τη διάγνωση του καρκίνου του στόματος μέσω ανάλυσης εικόνας, τα αποτελέσματα έδειξαν ότι τα συνελικτικά νευρωνικά δίκτυα (*CNN*) και οι ελαφριές αρχιτεκτονικές μεταφοράς μάθησης (όπως το *MobileNetV2*) προσφέρουν εξαιρετικές επιδόσεις στην

αναγνώριση παθολογικών ευρημάτων. Τα βαθιά δίκτυα κατάφεραν να διακρίνουν με υψηλή ακρίβεια καρκινικές από μη καρκινικές εικόνες, επιβεβαιώνοντας τη δυναμική των μεθόδων βαθιάς μάθησης στην ανάλυση ιατρικών εικόνων. Ιδιαίτερη έμφαση δόθηκε στη σωστή αξιολόγηση με μετρικές όπως η *βαθμολογία AUC*, το *F1-score* και η καμπύλη *ROC*, που προσφέρουν πιο αξιόπιστη αποτίμηση σε κλινικά σενάρια.

8.2 Μελλοντικές επεκτάσεις

Η παρούσα διπλωματική εργασία μπορεί να αποτελέσει αφετηρία για περαιτέρω μελέτη και ανάπτυξη τόσο στον τομέα της πρόβλεψης καρδιαγγειακών παθήσεων όσο και στη διάγνωση του καρκίνου του στόματος μέσω ανάλυσης εικόνας. Παρά τα αξιόλογα αποτελέσματα που επιτεύχθηκαν, διαφαίνονται ορισμένες κατευθύνσεις που θα μπορούσαν να ενισχύσουν ακόμη περισσότερο την πρακτική αξία και την επιστημονική συνεισφορά της εργασίας.

Αρχικά, η περαιτέρω βελτίωση της διαδικασίας εξόρυξης και επιλογής χαρακτηριστικών (*feature engineering*) στα δομημένα ιατρικά δεδομένα θα μπορούσε να οδηγήσει σε αύξηση της διακριτικής ικανότητας των μοντέλων. Στην κατεύθυνση της ανάλυσης ιατρικών εικόνων, σημαντική προοπτική αποτελεί η δοκιμή και άλλων προηγμένων αρχιτεκτονικών βαθιάς μάθησης, όπως τα *EfficientNet*, *Vision Transformers*. Η σύγκριση των αποτελεσμάτων τους με αυτά που επιτεύχθηκαν με το *MobileNetV2* μπορεί να οδηγήσει σε περαιτέρω βελτίωση της ακρίβειας και της αξιοπιστίας στη διάγνωση παθολογικών καταστάσεων του στόματος.

Τέλος, πολύ σημαντική είναι και η αξιολόγηση των αναπτυγμένων μοντέλων σε νέα, ανεξάρτητα ή και πραγματικά κλινικά δεδομένα. Η δοκιμή τους σε εξωτερικά σύνολα ή σε συνθήκες καθημερινής κλινικής πρακτικής θα επιτρέψει τη διερεύνηση της γενικευσιμότητας και της πρακτικής τους χρησιμότητας.

Βιβλιογραφία

[1] K. Mehrabani-Zeinabad, A. Feizi, M. Sadeghi, H. Roohafza, M. Talaei, and N. Sarrafzadegan. Cardiovascular disease incidence prediction by machine learning and statistical techniques: a 16-year cohort study from eastern mediterranean region. *BMC Medical Informatics and Decision Making*, 23(1):72, 2023.

[2] G. Alwakid, F. Ul Haq, N. Tariq, M. Humayun, M. Shaheen, and M. Alsadun. Optimized machine learning framework for cardiovascular disease diagnosis: a novel ethical perspective. *BMC Cardiovascular Disorders*, 25(1):123, 2025.

[3] M. U. Rehman, S. Naseem, A. Butt, A. Javed, M. Asif, and H. T. Mirza. Predicting coronary heart disease with advanced machine learning classifiers for improved cardiovascular risk assessment. *Scientific Reports*, 15:13361, 2025.

[4] P. Shah, M. Shukla, N. H. Dholakia, and H. Gupta. Predicting cardiovascular risk with hybrid ensemble learning and explainable ai. *Scientific Reports*, 15:17927, 2025.

[5] T. Liu, A. J. Krentz, L. Lu, and V. Ćurčin. Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis. *European Heart Journal – Digital Health*, 6(1):7–22, 2024.

[6] K. Warin, W. Limprasert, S. Suebnukarn, S. Jinaporntham, and P. Jantana. Automatic classification and detection of oral cancer in photographic images using deep learning algorithms. *Journal of Oral Pathology Medicine*, 50(9):911–918, 2021.

[7] S. Sukegawa, S. Ono, F. Tanaka, Y. Inoue, T. Hara, K. Yoshii, and M. Miyake. Effectiveness of deep learning classifiers in histopathological diagnosis of oral squamous cell carcinoma by pathologists. *Scientific Reports*, 13(1):11676, 2023.

[8] Z. Yang, H. Pan, J. Shang, J. Zhang, and Y. Liang. Deep-learning-based automated identification and visualization of oral cancer in optical coherence tomography images. *Biomedicines*, 11(3):802, 2023.

[9] R. S. Ramani, I. Tan, L. Bussau, L. A. O'Reilly, J. Silke, C. Angel, and T. Yap. Convolutional neural networks for accurate real-time diagnosis of oral epithelial dysplasia and oral squamous cell carcinoma using high-resolution in vivo confocal microscopy. *Scientific Reports*, 15(1):2555, 2025.

[10] T. Flügge, R. Gaudin, A. Sabatakakis, D. Tröltzsch, M. Heiland, N. van Nistelrooij, and S. Vinayahalingam. Detection of oral squamous cell carcinoma in clinical photographs using a vision transformer. *Scientific Reports*, 13(1):2296, 2023.

[11] D. G. Kleinbaum, K. Dietz, M. Gail, and M. Klein. *Logistic regression*. Springer-Verlag, New York, 2002.

[12] Y. Liu, Y. Wang, and J. Zhang. New machine learning algorithm: Random forest. In *International conference on information computing and applications*, pages 246–252, Berlin, Heidelberg, September 2012. Springer Berlin Heidelberg.

[13] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54:1937–1967, 2021.

[14] I. Rish. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, August 2001.

[15] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 33(12):6999–7019, 2021.

[16] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[17] K. Dong, C. Zhou, Y. Ruan, and Y. Li. Mobilenetv2 model for image classification. In *2020 2nd International Conference on Information Technology and Computer Application (ITCA)*, pages 476–480. IEEE, December 2020.

[18] M. Hossin and M. N. Sulaiman. A review on evaluation metrics for data classification evaluations. *International journal of data mining knowledge management process*, 5(2):1, 2015.

[19] J. Fan, S. Upadhye, and A. Worster. Understanding receiver operating characteristic (roc) curves. *Canadian Journal of Emergency Medicine*, 8(1):19–20, 2006.

[20] Kaggle Dataset. Cardiovascular disease dataset. `https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset`. Πρόσβαση: 2025.

[21] Kaggle Dataset. Heart disease dataset. `https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset`. Πρόσβαση: 2025.

[22] American Heart Association. Χοληστερόλη και Κίνδυνος Καρδιαγγειακών. `https://www.heart.org/en/health-topics/cholesterol`. Πρόσβαση: 2025.

[23] Peter M. Okin, Mary J. Roman, Elisa T. Lee, James M. Galloway, Barbara V. Howard, and Richard B. Devereux. Combined echocardiographic left ventricular hypertrophy and electrocardiographic st depression improve prediction of mortality in american indians: The strong heart study. *Circulation*, 109(6):714–719, 2004.

[24] Kaggle Dataset. Medical scan classification dataset - oral cancer. `https://www.kaggle.com/datasets/arjunbasandrai/medical-scan-classification-dataset?select=Oral+Cancer`. Πρόσβαση: 2025.

ΠΑΡΑΡΤΗΜΑ

Παράρτημα Α

ΚΩΔΙΚΑΣ ΚΑΙ ΔΕΔΟΜΕΝΑ

Οι κώδικες (Jupyter notebooks) που χρησιμοποιήθηκαν στο προγραμματιστικό μέρος της εργασίας, βρίσκονται διαθέσιμοι και τα αντίστοιχα σύνολα δεδομένων βρίσκονται εδώ. Ο υπερσύνδεσμος οδηγεί στο αποθετήριο Github.