Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων Και Μηχανικων Υπολογιστων
Τομεας Συστηματων Μεταδοσης Πληροφοριας Και Τεχνολογιας Υλικων

Διπλωματικη Εργασια

Συγκριτική Ανάλυση Μοντέλων Μηχανικής Μάθησης για την Πρόβλεψη Θνησιμότητας σε Εγκεφαλικούς Όγκους

Σπουδαστής:
Προκοπίου Στυλιανός

Επιβλέπων:
Γεώργιος Ματσόπουλος
Καθηγητής ΕΜΠ

Αθήνα, Νοέμβριος 2024

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Συγκριτική Ανάλυση Μοντέλων Μηχανικής Μάθησης για την Πρόβλεψη Θνησιμότητας σε Εγκεφαλικούς Όγκους

Σπουδαστής:
Στυλιανός Προκοπίου

Επιβλέπων:
Γεώργιος Ματσόπουλος
Καθηγητής ΕΜΠ

Εγκρίθηκε από την κάτωθι τριμελή επιτροπή την 29η Νοεμβρίου 2024.

Γεώργιος Ματσόπουλος Καθηγητής ΕΜΠ	Παναγιώτης Τσανάκας Καθηγητής ΕΜΠ	Αθανάσιος Δ. Παναγόπουλος Καθηγητής ΕΜΠ

Αθήνα, Νοέμβριος 2024

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

Copyright © Στυλιανός Προκοπίου, 2024
Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρε- ται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.
Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Περίληψη

Οι όγκοι εγκεφάλου, ιδιαίτερα οι μεταστατικοί όγκοι που προέρχονται από άλλους καρκίνους, αποτελούν σημαντική πρόκληση στην πρόγνωση και τον σχεδιασμό θεραπείας των ασθενών. Αυτή η διατριβή εξερευνά την εφαρμογή της μηχανικής μάθησης (Machine Learning, ML) για την ταξινόμηση του θανάτου σε ασθενείς με όγκους εγκεφάλου, χρησιμοποιώντας δεδομένα πολλαπλών τρόπων.
Το αρχικό τμήμα παρέχει μια επισκόπηση των όγκων εγκεφάλου, των μεταστάσεων και της κλινικής τους σημασίας. Στη συνέχεια, εμβαθύνει στα θεμελιώδη στοιχεία της μηχανικής μάθησης, εξηγώντας διάφορες μεθόδους και τις δυνατότητές τους για την ανάλυση ιατρικών προβλημάτων.
Ακολούθως, η διατριβή περιγράφει λεπτομερώς την εξέταση ενός δείγματος δεδομένων (n = 200) που αποκτήθηκε από το Αρχείο Εικόνων Καρκίνου (TCIA). Τα δεδομένα περιλαμβάνουν εικόνες MRI, διατμηματώσεις και αντίστοιχες κλινικές πληροφορίες (ηλικία, φύλο, κατάσταση θανάτου κ.λπ.).
Για τον ποσοτικό προσδιορισμό των χαρακτηριστικών του όγκου από τις εικόνες MRI χρησιμοποιούνται τεχνικές εξαγωγής χαρακτηριστικών radiomics . Στη συνέχεια, τα εξαχθέντα χαρακτηριστικά υποβάλλονται σε προεπεξεργασία για να διασφαλιστεί η καταλληλότητά τους για αλγορίθμους μηχανικής μάθησης.
Το βασικότερο κομμάτι της διατριβής περιλαμβάνει την εξαγωγή χαρακτηριστικών radiomics και την μελέτη της προβλεπτικής τους ικανότητας και κατ' επέκταση τη συγκριτική ανάλυση και βελτιστοποίηση διαφορετικών μεθόδων μηχανικής μάθησης για την ταξινόμηση του θανάτου των ασθενών. Αυτές οι μέθοδοι περιλαμβάνουν μια σειρά από αλγορίθμους επιβλεπόμενης μάθησης που χρησιμοποιούνται συνήθως σε ιατρικές εφαρμογές.
Το τμήμα των αποτελεσμάτων παρουσιάζει μια ολοκληρωμένη σύγκριση των μετρήσεων επιδόσεων που επιτεύχθηκαν από κάθε μέθοδο μηχανικής μάθησης. Περιλαμβάνει μια λεπτομερή ανάλυση των παραγόντων που επηρεάζουν την ακρίβεια ταξινόμησης και τον εντοπισμό του βέλτιστου αλγορίθμου για την πρόβλεψη θανάτου σε αυτήν την ειδική ομάδα ασθενών.
Τέλος, η διατριβή ολοκληρώνεται με τη συζήτηση των συνεπειών των ευρημάτων, των πιθανών περιορισμών της μελέτης και των κατευθύνσεων για μελλοντική έρευνα στην ταξινόμηση όγκων εγκεφάλου χρησιμοποιώντας προηγμένες προσεγγίσεις μηχανικής μάθησης. Αυτό το έργο συμβάλλει στην ανάπτυξη πιο ισχυρών και ακριβών μεθόδων για την πρόβλεψη των αποτελεσμάτων των ασθενών σε περιπτώσεις όγκων εγκεφάλου, συμβάλλοντας τελικά σε εξατομικευμένες στρατηγικές θεραπείας και βελτιωμένη φροντίδα των ασθενών.

Λέξεις κλειδιά: Όγκος εγκεφάλου, Μετάσταση, Μηχανική Μάθηση, Επιβλεπόμενη Μάθηση, Ακτινομική, Εξαγωγή Χαρακτηριστικών, Απεικόνιση Μαγνητικού Συντονισμού (MRI), Ταξινόμηση Θανάτου, Πρόγνωση, TCIA (Αρχείο Εικόνων Καρκίνου), Ιατρική Απεικόνιση, Έκβαση Ασθενή

Abstract

Brain tumors, particularly metastatic tumors originating from other cancers, pose a significant challenge in patient prognosis and treatment planning. This thesis explores the application of machine learning (ML) for death classification in brain tumor patients using multimodal data.
The initial section provides an overview of brain tumors, metastasis, and their clinical significance. It then delves into the fundamentals of machine learning, explaining various methods and their potential for medical problem analysis.
Following this, the thesis details the exploration of a sample dataset (n=200) acquired from The Cancer Imaging Archive (TCIA). The data includes MRI images, segmentations, and corresponding clinical information (age, sex, death status, etc.).
Radiomic feature extraction techniques are employed to quantify tumor characteristics from the MRI images. Subsequently, the extracted features undergo preprocessing to ensure their suitability for machine learning algorithms.
A core component of the thesis involves the extraction of radiomics features and the study of their predictive ability which extends to the comparative analysis and optimization of many distinct machine learning methods for patient death classification. These methods encompass a range of supervised learning algorithms commonly used in medical applications.
The results section presents a comprehensive comparison of the performance metrics achieved by each machine learning method. It includes a detailed analysis of factors influencing the classification accuracy and the identification of the optimal algorithm for death prediction in this specific patient cohort. Finally, the thesis concludes by discussing the implications of the findings, potential limitations of the study, and avenues for future research in brain tumor classification utilizing advanced machine learning approaches.
This work contributes to the development of more robust and accurate methods for predicting patient outcomes in brain tumor cases, ultimately aiding in personalized treatment strategies and improved patient care.

Keywords: Brain tumor, Metastasis, Machine Learning, Supervised Learning, Radiomics, Feature Extraction, MRI, Death Classification, Prognosis, TCIA, Medical Imaging, Patient Outcome

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον κ. Γεώργιο Ματσόπουλο, καθηγητή και επιβλέπων της παρούσας εργασίας για την άψογη συνεργασία και την εμπιστοσύνη που μου έδειξε έτσι ώστε να μπορώ να αναλάβω αυτήν την εργασία. Θα ήθελα επίσης να ευχαριστήσω και τους κυρίους Παναγιώτη Τσανάκα και Αθανάσιο Δ. Παναγόπουλο, οι οποίοι σαν μέλη της υπόλοιπης τριμελής επιτροπής συνέβαλαν επίσης στο πέρας αυτού του έργου. Δεν θα έπρεπε να παραλείψω τους υποψήφιους διδάκτορες, Κωνσταντίνο Γεώργα και Θεόδωρο Π. Βαγενά, οι οποίοι βοήθησαν πολύ ενεργά σε αυτήν την εργασία, δίνοντας κατευθύνσεις, ιδέες και προτάσεις για την επίλυση πολλών προβλημάτων που προέκυψαν, πάντα με υπομονή και ευγένεια. Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου, πηγή της δύναμης και έμπνευσης μου, για την συνεχή τους στήριξη προς εμένα κατά την διάρκεια των ακαδημαϊκών μου χρόνων και όχι μόνο.

Πίνακας περιεχομένων

Περίληψη ***5***

Abstract ***7***

Ευχαριστίες ***9***

Πίνακας περιεχομένων ***10***

Ευρετήριο Εικόνων ***12***

Κεφάλαιο 1: Εισαγωγή ***14***

1.1: Ορισμός του Προβλήματος **14**

1.2: Κίνητρο της Εργασίας και Στόχοι **15**

1.3: Οργάνωση της Εργασίας **15**

Κεφάλαιο 2: Ορισμός και Περιγραφή του Προβλήματος ***17***

2.1: Καρκινικοί Όγκοι στον Εγκέφαλο **17**

2.1.1: Ανατομία Εγκεφάλου 18

2.1.2: Καρκίνος Εγκεφάλου 20

2.1.3: Τύποι Καρκίνου στον Εγκέφαλο 21

2.1.4: Συμπτώματα Όγκων Εγκεφάλου 24

2.1.5: Παράγοντες Κινδύνου για Όγκο Εγκεφάλου 24

2.1.6: Μετάσταση Εγκεφάλου 24

2.2: Έλεγχος και Απεικόνιση Καρκίνου Εγκεφάλου **28**

2.2.1: Μαγνητική Τομογραφία (MRI / functional MRI) 29

2.2.2: Μαγνητική Τομογραφία Όγκων Εγκεφάλου 30

2.2.2: Επεξεργασία Ιατρικών Εικόνων 32

Κεφάλαιο 3: Μηχανική Μάθηση και Κατάταξη ***35***

3.1: Μηχανική Μάθηση **35**

3.1.1: Εισαγωγή στην Μηχανική Μάθηση 35

3.1.2: Μηχανική Μάθηση 35

3.1.3: Νευρωνικά Δίκτυα 37

3.2: Αλγόριθμοι Μηχανικής Μάθησης και Ταξινόμηση **44**

3.2.1: k-Nearest Neighbors (k-NN) 44

3.2.2: Support Vector Machines (SVM) 47

3.2.3: Λογιστική Παλινδρόμηση (Logistic Regression, LR) 50

3.2.4 Random Forest (RF) 54

3.2.5 XGBoost (XGB) 58

3.3: Μηχανική Μάθηση σε Ιατρικές Εφαρμογές.....62
3.3.1: Επεξεργασία Ιατρικών Εικόνων 64
3.3.2: Τμηματοποίηση και Καταχώρηση 67

Κεφάλαιο 4: Ανάλυση Δεδομένων και Επιλογή Μοντέλου 71

4.1: Δεδομένα Μοντέλου 71
4.1.1: Συλλογή Δεδομένων 71
4.1.2: Προετοιμασία Δεδομένων..... 73
4.1.3: Απεικόνιση και Επεξεργασία Δεδομένων 75
4.1.4: Εξαγωγή Χαρακτηριστικών..... 78

4.2: Διερευνητική Ανάλυση και Προεπεξεργασία Χαρακτηριστικών.....83
4.2.1: Χωρισμός σε Training και Test Set..... 83
4.2.2: Αφαίρεση Ακραίων Τιμών 84
4.2.3: Κανονικοποίηση Δεδομένων 86
4.2.4: Εκκαθάριση Συγγραμικότητας..... 87
4.2.5: Μείωση Διαστατικότητας..... 88
4.2.6: Διόρθωση της Ανισορροπίας με Συνθετικά Σεδομένα..... 90
4.2.7: Επιλογή Χαρακτηριστικών με την Μέθοδο LGBM 91

4.3: Ανάλυση Μοντέλων Ταξινόμησης.....93
4.3.1: Εκπαίδευση Μοντέλων..... 93
4.3.2: Μετρικές Αξιολόγησής Μοντέλων 94
4.3.3: Βελτιστοποίηση Μοντέλων 98

Κεφάλαιο 5: Αποτελέσματα, Παρατηρήσεις και Συμπεράσματα 101

5.1: Αποτελέσματα.....102
5.1.1: Πείραμα 1: Βασική Προεπεξεργασία (Robust Normalization, Stratified K-Fold Cross-Validation) 104
5.1.2: Πείραμα 2: Αφαίρεση Ακραίων Τιμών 106
5.1.3: Πείραμα 3: Αφαίρεση της Πολυσυγγραμμικότητας 107
5.1.4: Πείραμα 4: Διόρθωση της Ανισορροπίας με Συνθετικά Δεδομένα (SMOTE)..... 109
5.1.5: Πείραμα 5: Μείωση Διαστάσεων με Χρήση PCA 111
5.1.6: Πείραμα 6: Επιλογή Χαρακτηριστικών με τον Ταξινομητή LGBM 113

5.2: Βελτιστοποίηση Αποτελεσμάτων 115
5.2.1: Ενσωμάτωση Δημογραφικών δεδομένων..... 116
5.2.2: Συνδυασμοί Μοντέλων: Ensemble, Blend και Stack 120
5.2.3: Αναζήτηση Βέλτιστων Υπερπαραμέτρων 123

5.3: Συμπεράσματα και Μελλοντικές Επεκτάσεις.....128

Βιβλιογραφία.....131

Ευρετήριο Εικόνων

Εικόνα 1- Ανατομία του Ανθρώπινου Εγκεφάλου από [56] 19
Εικόνα 2 - Υγιής Εγκέφαλος και Εγκέφαλος με Όγκο από [57] 21
Εικόνα 3 - Διαφορετικοί Τύποι Όγκων Εγκεφάλου από [58] 22
Εικόνα 4 - Κύκλος Μετάστασης στον Εγκέφαλο από [59] 25
Εικόνα 5 Μαγνητικές Τομογραφίες για διάφορα είδη όγκων από [60] 30
Εικόνα 6 - Διαφορετικά είδη MRI.....32
Εικόνα 7 Βασική Δομή Perceptron 38
Εικόνα 8 - Συναρτήσεις Ενεργοποίησης.....39
Εικόνα 9 – Νευρωνικό δίκτυο πολλών επιπέδων (Multi Layer Perceptron).....40
Εικόνα 10 - Απεικόνιση K-NN Ταξινόμησης 46
Εικόνα 11 - Απεικόνιση Δυαδικής Ταξινόμησης SVM από [61] 49
Εικόνα 12 - Ταξινόμηση με Λογιστική Παλινδρόμηση 52
Εικόνα 13 Random Forest Classifier [62] 56
Εικόνα 14 - Διάγραμμα ροής αλγορίθμου XGBoost.....59
Εικόνα 15 Εφαρμογές της Μηχανική Μάθησης στην Ιατρική από [63] 63
Εικόνα 16 Τμηματοποίηση Εικόνας (Image Segmentation) από [64].....66
Εικόνα 17 Δομή συνελικτικού δικτύου U-net από [65] 68
Εικόνα 18 - Κλινικά δεδομένα 74
Εικόνα 19 - Περιγραφές των δημογραφικών δεδομένων.....74
Εικόνα 20 – 3D Απεικόνιση NifTI αρχείου 76
Εικόνα 21 - Απεικόνιση στο λογισμικό Slicer 76
Εικόνα 22 - Οπτικοποίηση με το εργαλείο SimpleITK 78
Εικόνα 23 - Ρύθμιση Εξαγωγής Χαρακτηριστικών - Image Type 80
Εικόνα 24 – Κατηγορίες χαρακτηριστικών και πλήθος των επιμέρους τάξεων 81
Εικόνα 25 - Ρύθμιση Εξαγωγής Χαρακτηριστικών - Feature Class.....81
Εικόνα 26 - Ρύθμιση Εξαγωγής Χαρακτηριστικών - Setting.....82
Εικόνα 27 - Χαρακτηριστικά Ραδιομικής για Όλα τα Δείγματα 83
Εικόνα 28 - Οπτικοποίηση Ακραίων Τιμών με Boxplot & Histogram.....85
Εικόνα 29 - Μοντέλα Εκπαίδευσης για την Πειραματική Διαδικασία.....94
Εικόνα 30 - Πίνακας Σύγχυσης (Confusion Matrix) 95
Εικόνα 31 - K-Fold Cross Validation (10 folds).....99
Εικόνα 32 - Grid Search Cross Validation 100
Εικόνα 33 - Κατανομή των Κατηγοριών των Δειγμάτων.....102
Εικόνα 34 - Σετ Δοκιμής - Κατανομή των Κατηγοριών των Δειγμάτων 103
Εικόνα 35 - Σετ Εκπαίδευσης - Κατανομή των Κατηγοριών των Δειγμάτων 103
Εικόνα 36 - Ρυμθίσεις Πειράματος 1 - Basic 105
Εικόνα 37 - Επιδόσεις Καλύτερων Μοντέλων Πειράματος 1 - Basic 105
Εικόνα 38 - Ρυμθίσεις Πειράματος 2 - Remove Outliers 106
Εικόνα 39 - Επιδόσεις Καλύτερων Μοντέλων Πειράματος 2 - Remove Outliers 107
Εικόνα 40 - Ρυμθίσεις Πειράματος 3 - Remove Multicollinearity.....108

Εικόνα 41 - Επιδόσεις Καλύτερων Μοντέλων Πειράματος 3 - Remove Multicollinearity.....109
Εικόνα 42 - Ρυμθίσεις Πειράματος 4 - Fix Imbalance with Synthetic Data.....110
Εικόνα 43 - Επιδόσεις Καλύτερων Μοντέλων Πειράματος 4 - Fix Imbalance with Synthetic Data 111
Εικόνα 44 - Πίνακας Συσχέτισης Πειράματος 4 - PCA.....111
Εικόνα 45 - Ρυμθίσεις Πειράματος 5 - PCA.....112
Εικόνα 46 - Επιδόσεις Καλύτερων Μοντέλων Πειράματος 5 - PCA.....113
Εικόνα 47 - Ρυμθίσεις Πειράματος 6 - Feature Selection 114
Εικόνα 48 - Επιδόσεις Καλύτερων Μοντέλων Πειράματος 6 - Feature Selection.....115
Εικόνα 49 - Ενσωμάτωση Δημογραφικών Δεδομένων στο Σύνολο Χαρακτηριστικών 116
Εικόνα 50 - Ρυθμίσεις Εκπαίδευσης Μοντέλων για Σύνολο με Κλινικά Δεδομένα.....117
Εικόνα 51 - Πίνακας Συσχέτισης – Σύνολο με Κλινικά Δεδομένα 117
Εικόνα 52 - Επιδόσεις Καλύτερων Μοντέλων - Σύνολο με Κλινικά Δεδομένα.....118
Εικόνα 53 - SVM Μετρικές ανά κλάση - Σύνολο με Κλινικά Δεδομένα 119
Εικόνα 54 - SVM Confusion Matrix - Σύνολο με Κλινικά Δεδομένα.....119
Εικόνα 55 - SVM Σημαντικότητα Χαρακτηριστικών Σύνολο με Κλινικά Δεδομένα 120
Εικόνα 56 – Ρυθμίσεις μοντέλου blender_v1 121
Εικόνα 57 - Μετρικές ανά κατηγορία για το μοντέλο blender_v1.....121
Εικόνα 58 - Επιδόσεις Μετρικών κατα την εκπαίδευση του μοντέλου blender_v1.....122
Εικόνα 59 - Ρυθμίσεις μοντέλου blender_v2 122
Εικόνα 60 - Μετρικές ανά κατηγορία για το μοντέλο blender_v2.....123
Εικόνα 61 - Επιδόσεις Μοντέλων - Grid Search 124
Εικόνα 62 - Συγκριτικοί Πίνακες Σύγχυσης - Grid Search.....125
Εικόνα 63 - Επιδόσεις Μοντέλων - Bayesian Search.....125
Εικόνα 64 - Συγκριτικοί Πίνακες Σύγχυσης - Bayesian Search 126
Εικόνα 65 - Σύγκριση Υπερπαραμέτρων Επιλεγμένων Μοντέλων - Βελτιστοποίηση 127
Εικόνα 66 - Σύγκριση Μετρικών ανα Μοντέλο και Πείραμα 129

Κεφάλαιο 1: Εισαγωγή

1.1: Ορισμός του Προβλήματος

Οι όγκοι εγκεφάλου αποτελούν μια σοβαρή ιατρική κατάσταση με σημαντική ετερογένεια - πράγμα που σημαίνει ότι μπορούν να διαφοροποιηθούν σημαντικά ως προς τον τύπο, τη σοβαρότητα και την πρόγνωση. Η ακριβής διάγνωση και η έγκαιρη ανίχνευση των μεταστάσεων (εξάπλωση) είναι καθοριστικές για τον προσδιορισμό της καλύτερης θεραπευτικής αγωγής και τη βελτίωση των αποτελεσμάτων των ασθενών. Ωστόσο, οι παραδοσιακές μέθοδοι, όπως η οπτική ανάλυση των εξετάσεων MRI, μπορεί να είναι υποκειμενικές και περιορισμένες. Αυτό μπορεί να οδηγήσει σε λανθασμένη διάγνωση ή καθυστερημένη διάγνωση, επηρεάζοντας την αποτελεσματικότητα της θεραπείας και την επιβίωση των ασθενών.

Η παρούσα διατριβή επικεντρώνεται στην αξιοποίηση της ισχύος της μηχανικής μάθησης για την αντιμετώπιση αυτών των προκλήσεων. Οι αλγόριθμοι μηχανικής μάθησης έχουν τη δυνατότητα να αναλύουν τεράστια ποσά ιατρικών δεδομένων, συμπεριλαμβανομένων των εξετάσεων MRI, και να εντοπίζουν διακριτικά μοτίβα που μπορεί να διαφύγουν της ανθρώπινης παρατήρησης. Με την εκπαίδευση αυτών των αλγορίθμων σε ένα μεγάλο σύνολο δεδομένων με επισήμανση περιστατικών όγκων εγκεφάλου, ο στόχος είναι να αναπτυχθεί ένα σύστημα που μπορεί να ταξινομεί αυτόματα τους τύπους όγκων με υψηλή ακρίβεια και να προβλέπει την πιθανότητα μετάστασης.

Αυτή η πρόοδος θα είχε σημαντικές επιπτώσεις στη βελτίωση της διάγνωσης και του σχεδιασμού θεραπείας των όγκων εγκεφάλου. Ο πιο ακριβής προσδιορισμός μπορεί να οδηγήσει σε στοχευμένες θεραπείες ειδικές για τον τύπο του όγκου, ενώ η έγκαιρη πρόβλεψη της μετάστασης επιτρέπει ταχύτερη παρέμβαση και πιθανώς βελτιώνει τα ποσοστά επιβίωσης των ασθενών. Συνολικά, αυτή η έρευνα συμβάλλει σε έναν αναπτυσσόμενο τομέα της ιατρικής μηχανικής μάθησης που αποσκοπεί στη μεταμόρφωση του τρόπου με τον οποίο διαγιγνώσκουμε και διαχειριζόμαστε σύνθετες ιατρικές καταστάσεις όπως οι όγκοι εγκεφάλου.

1.2: Κίνητρο της Εργασίας και Στόχοι

Η παρούσα έρευνα ξεκινά από την επιθυμία να βελτιώσει τις υπάρχουσες μεθόδους πρόβλεψης του θανάτου σε ασθενείς με εγκεφαλικούς όγκους που παρουσιάζουν μετάσταση. αξιοποιώντας την απεικόνιση χαρακτηριστικών (radiomics), μια τεχνική που εξάγει ποσοτικά χαρακτηριστικά από ιατρικές εικόνες, ο στόχος είναι να εκπαιδεύσουμε αλγόριθμους μηχανικής μάθησης να αναγνωρίζουν διακριτικά μοτίβα μέσα στον ίδιο τον όγκο που μπορεί να υποδηλώνουν μια πιο επιθετική πορεία της νόσου. Αυτά τα απεικονιστικά χαρακτηριστικά θα μπορούσαν να αποτυπώσουν πτυχές της βιολογίας του όγκου, όπως ο ρυθμός ανάπτυξής του, η κυτταρική πυκνότητα ή η κατανομή των αιμοφόρων αγγείων, τα οποία δεν είναι εύκολα αντιληπτά μόνο με την οπτική ανάλυση.

Φανταστείτε ένα σενάριο όπου σε έναν ασθενή με όγκο εγκεφάλου διαγνωστεί μετάσταση. Παραδοσιακά, οι γιατροί θα βασίζονταν σε παράγοντες όπως το μέγεθος και η θέση του όγκου, μαζί με τη γενική υγεία του ασθενούς, για να εκτιμήσουν την πιθανότητα θανάτου. Αυτή η έρευνα, ωστόσο, προτείνει μια πιο εξελιγμένη προσέγγιση. Με την τροφοδοσία απεικονιστικών χαρακτηριστικών που εξάγονται από την εξέταση MRI του ασθενούς σε ένα μοντέλο μηχανικής μάθησης, το μοντέλο θα μπορούσε δυνητικά να αναγνωρίσει κρυφά χαρακτηριστικά του μεταστατικού όγκου που συσχετίζονται με χειρότερη έκβαση για τον ασθενή. Αυτές οι πληροφορίες θα μπορούσαν στη συνέχεια να χρησιμοποιηθούν για να βελτιωθεί η διαστρωμάτωση κινδύνου των ασθενών και να καθοδηγηθούν οι αποφάσεις θεραπείας, οδηγώντας ενδεχομένως σε πιο στοχευμένες παρεμβάσεις και βελτιωμένα ποσοστά επιβίωσης.

1.3: Οργάνωση της Εργασίας

*Στο **Κεφάλαιο 2*** παρουσιάζεται το θεωρητικό υπόβαθρο σχετικά με το όργανο του εγκεφάλου και τη σημασία του καρκίνου σε αυτόν, καθώς επίσης περιγράφεται και η διαδικασία ελέγχου και απεικόνισης για τον καρκίνου του εγκεφάλου. Συγκεκριμένα στο πρώτο κομμάτι γίνεται επεξήγηση στα θέματα σχετικά με την ανατομία του εγκεφάλου, τον καρκίνο στον εγκέφαλο, τα είδη καρκίνου στον εγκέφαλο, τα συμπτώματα και τους παράγοντες κινδύνου των όγκων του εγκεφάλου, καθώς επίσης δίνεται ιδιαίτερη έμφαση στην παρουσίαση για την μετάσταση στον εγκέφαλο. Στην συνέχεια εξηγείται ο τρόπος λειτουργίας της Απεικόνισης Μαγνητικού Συντονισμού και του μαγνητικού τομογράφου και γίνεται μια συνοπτική εισαγωγή στα βήματα για την επεξεργασία των ιατρικών απεικονίσεων.

Στο ***Κεφάλαιο 3*** αναλύονται οι ενότητες της μηχανικής μάθησης και οι τεχνικές που χρησιμοποιήθηκαν για την έρευνα σε αυτή την εργασία, καθώς επίσης η χρησιμότητα και η εφαρμογή της μηχανικής μάθησης σε ιατρικά προβλήματα. Τέλος γίνεται ειδική αναφορά στην ενότητα της κατηγοριοποίησης στην μηχανική μάθηση και πως αυτή χρησιμοποιείτε στις ιατρικές εφαρμογές.

Στο ***Κεφάλαιο 4*** παρουσιάζονται οι λεπτομέρειες της υλοποίησης των αλγορίθμων στα ιατρικά δεδομένα που είχαμε στη διάθεσή μας. Εξηγείται ο τρόπος με τον οποίο λειτουργεί η εκάστοτε τεχνική, πλεονεκτήματα και μειονεκτήματα αυτών και οι παράμετροι που ορίστηκαν. Το κεφάλαιο αυτό μπορεί να λειτουργήσει ως «εισαγωγή» στο κεφάλαιο 5.

Στο ***Κεφάλαιο 5*** παρουσιάζονται τα πειραματικά αποτελέσματα όλων των διαφορετικών τεχνικών που εξετάσαμε στα Κεφάλαια 3 και 4, οι διαφορετικές παράμετροι που ορίστηκαν, οι επιδόσεις των αλγορίθμων σε διαφορετικές μετρικές, η γενίκευση των μοντέλων, η αποφυγή της υπερπροσαρμογής, η διακριτική ικανότητα των μοντέλων καθώς και η συγκριτική μελέτη μεταξύ τους. Τέλος αναφέρονται οι μελλοντικές επεκτάσεις που θα μπορούσε να λάβει η συγκεκριμένη έρευνα. Προτείνεται ένας διαφορετικός τρόπος μελέτης του αντικειμένου, όπως επίσης και πιθανές βελτιστοποιήσεις επί της παρούσας μελέτης.

Κεφάλαιο 2: Ορισμός και Περιγραφή του Προβλήματος

2.1: Καρκινικοί Όγκοι στον Εγκέφαλο

Ο όγκος είναι μια από τις πιο ευρέως αναγνωρισμένες κοινές ασθένειες. Ο Παγκόσμιος Οργανισμός Υγείας (ΠΟΥ) εκτιμά ότι η διάγνωση και η θεραπεία είναι σημαντικές για περισσότερα από 400.000 άτομα που πάσχουν από όγκους ετησίως στον κόσμο [1]. Τα κύτταρα στο σώμα μας συνεχίζουν να αναπτύσσονται και να αυξάνονται πολλαπλασιαστικά αρκετές φορές μέσα σε μία μόνο ημέρα της ζωής μας, έτσι ώστε το σώμα μας να μπορεί να λειτουργεί κανονικά. Όταν υπάρχει περιστασιακή διαταραχή κατά τη διάρκεια της διαίρεσης των κυττάρων στο σώμα μας, μπορεί να οδηγήσει στο σχηματισμό μη φυσιολογικών κυττάρων ή λανθασμένων κυττάρων και αυτό μπορεί να διαταράξει τις δραστηριότητες άλλων τμημάτων του σώματος με ανεξέλεγκτο τρόπο, γεγονός που ονομάζεται καρκίνος. Τα ανώμαλα κύτταρα που σχηματίζονται κατά τη διάρκεια της κυτταρικής διαίρεσης ονομάζονται καρκίνος και έχουν την ιδιότητα να διαπερνούν τους κοντινούς ιστούς των οργάνων, να αρχίζουν να επηρεάζουν το αίμα και το λεμφικό σύστημα, το οποίο ονομάζεται μετάσταση, μειώνοντας έτσι τη διάρκεια ζωής των ασθενών.

Ο καρκίνος μπορεί να έχει την εμφάνισή του σε οποιοδήποτε μέρος του σώματος και κατηγοριοποιούνται κυρίως από το κύτταρο από το οποίο προέρχεται. καρκίνος που προέρχεται από τον εγκέφαλο και το νευρικό σύστημα ονομάζεται καρκίνος του εγκεφάλου. Η σημαντικότερη απειλή για τη ζωή του ανθρώπου ασθένεια είναι ο όγκος του εγκεφάλου. Σύμφωνα με τις πρόσφατες στατιστικές που έχουν ληφθεί από διάφορα ερευνητικά ιδρύματα για τον καρκίνο, όπως το Εθνικό Ινστιτούτο Καρκίνου (NCI), η Αμερικανική Ένωση Εγκεφαλικών Όγκων και το Προηγμένο Κέντρο Έρευνας και Εκπαίδευσης για τη Θεραπεία του Καρκίνου (ACTREC), περίπου το 20% των θανάτων από καρκίνο οφείλεται σε όγκο στον εγκέφαλο και είναι η κύρια αιτία θανάτου που σχετίζεται με τον καρκίνο. Η συσσώρευση φυσιολογικών κυττάρων στον εγκέφαλο

σχηματίζει μια μάζα ιστού που ονομάζεται όγκος του εγκεφάλου. Κανονικά τα κύτταρα στο σώμα μας πεθαίνουν σε μια ορισμένη περίοδο και αντικαθίστανται από νέα κύτταρα. Όταν αυτός ο κύκλος διαταράσσεται, τα παλιά κύτταρα πεθαίνουν και συνεχίζουν να αναπτύσσονται. Εν τω μεταξύ, τα νέα κύτταρα συνοδεύουν αυτά τα παλιά κύτταρα και αναπτύσσονται μαζί τους και σχηματίζουν μια μάζα ιστού γνωστή ως όγκος. Ο καρκίνος του εγκεφάλου είναι ένας από τους συχνότερους καρκίνους παγκοσμίως και αυτό το είδος καρκίνου έχει το υψηλότερο ποσοστό θνησιμότητας. Η συχνότητα με την οποία εμφανίζεται ο όγκος του εγκεφάλου αυξάνεται πολύ γρήγορα και προσβάλλει περισσότερο ηλικιωμένους παρά τους νέους. Όλα τα υγιή κύτταρα του εγκεφάλου καταστρέφονται άμεσα από τους όγκους του εγκεφάλου. Η ανίχνευση του καρκίνου είναι πράγματι μια επίπονη και επαναλαμβανόμενη εργασία που χρειάζεται τη σωστή ερμηνεία των μη φυσιολογικών εικόνων του εγκεφάλου [2]. Ως αποτέλεσμα, μια τέτοια πορεία είναι πολύ δαπανηρή και επιρρεπής σε σφάλματα. Επίσης, ο ακτινολόγος πρέπει να ξοδεύει πολύ χρόνο σε αυτό. Αυτό που χρειαζόμαστε επειγόντως είναι η εισαγωγή ποιοτικής και αντικειμενικής ανάλυσης. Ως εκ τούτου, υπάρχει ανάγκη για την ανάπτυξη ενός διαγνωστικού εργαλείου υποβοηθούμενου από υπολογιστή, το οποίο θα μπορεί να διαγνώσει τον όγκο από μόνο του. Ο Παγκόσμιος Οργανισμός Υγείας (ΠΟΥ) ταξινομεί τους όγκους του εγκεφάλου σε τέσσερις τύπους. Οι όγκοι των κατηγοριών I και II δεν είναι κακοήθεις (καλοήθεις). Οι όγκοι των κατηγοριών III και IV είναι καρκινικοί και θανατηφόροι (κακοήθεις). Η έγκαιρη ανίχνευση των κακοήθων όγκων απαιτείται για τον αποτελεσματικό σχεδιασμό της θεραπείας. Ένας όγκος στον εγκέφαλο μπορεί να ανακαλυφθεί χρησιμοποιώντας διάφορες προσεγγίσεις, συμπεριλαμβανομένων των συμπτωμάτων, των τεχνικών ιατρικής απεικόνισης, των διαγνωστικών εξετάσεων κ.ο.κ. [3]. Χρησιμοποιώντας μικροσκοπική επιθεώρηση των προσβεβλημένων ιστών, μπορεί να γίνει γρήγορα η σωστή διάγνωση ενός όγκου (βιοψία). Στη σύγχρονη ιατρική, οι 4 θεραπευτικές επιλογές περιλαμβάνουν την ακτινοθεραπεία, τη χημειοθεραπεία και επέμβαση. επιτυχία της θεραπείας καθορίζεται από το μέγεθος, τη θέση, τον τύπο και το στάδιο του όγκου.

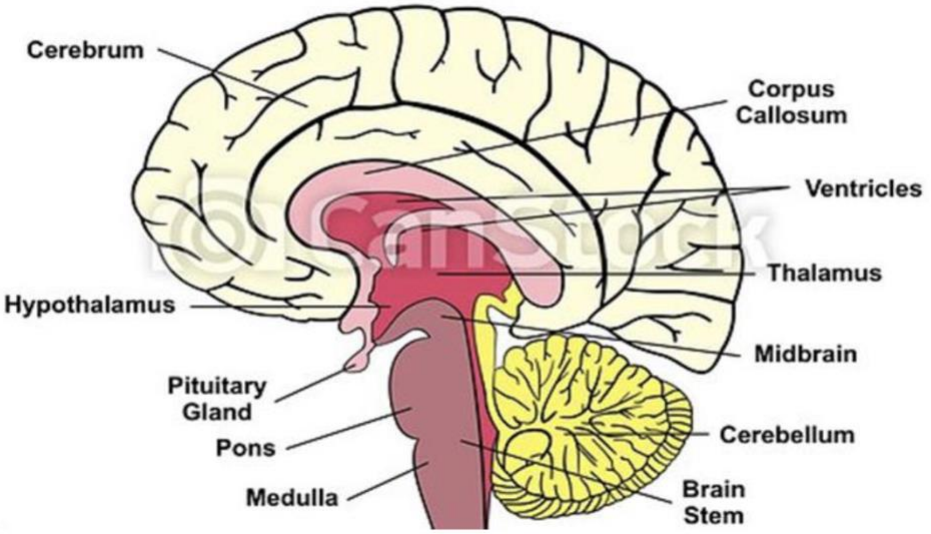
2.1.1: Ανατομία Εγκεφάλου

Ο ανθρώπινος εγκέφαλος είναι το πιο σημαντικό και πολύπλοκο όργανο του ανθρώπινου σώματος, που ελέγχει και συντονίζει όλες τις φυσικές λειτουργίες. Ο εγκέφαλος είναι ένα πολύ σύνθετη δομή που αποτελείται από δισεκατομμύρια νευρικά κύτταρα (νευρώνες), τα οποία ενώνονται μεταξύ τους για να σχηματίσουν το Κεντρικό Νευρικό Σύστημα (ΚΝΣ).

Τα κύρια μέρη του ανθρώπινου εγκεφάλου είναι η παρεγκεφαλίδα, ο εγκέφαλος και το εγκεφαλικό στέλεχος. Ο ανθρώπινος εγκέφαλος χωρίζεται σε τρία τμήματα: τον πρόσθιο

εγκέφαλο, τον μέσο εγκέφαλο και τον οπίσθιο εγκέφαλο. Ο εγκέφαλος, ο θάλαμος και ο υποθάλαμος αποτελούν μέρος του μεταιχμιακού συστήματος και απαρτίζουν τον πρόσθιο εγκέφαλο [4].

Ο μέσος εγκέφαλος αποτελείται από τον τεκτονικό και τον τεγέφυρα. Η παρεγκεφαλίδα, το pons και ο μυελός αποτελούν τον οπίσθιο εγκέφαλο. Ο μέσος εγκέφαλος, το pons και ο μυελός αναφέρονται συνήθως ως εγκεφαλικό στέλεχος. Ο εγκέφαλος ενσωματώνει διάφορες διεργασίες του ανθρώπινου σώματος, όπως η όραση, η μνήμη, η μάθηση και άλλες εργασίες. Ο εγκέφαλος είναι μια σταθερή θέση για διάφορα πρότυπα που εισέρχονται και σταθεροποιούν άλλα τμήματα. Ο εγκέφαλος είναι η μεγαλύτερη περιοχή του εγκεφάλου, ακολουθούμενη από την παρεγκεφαλίδα και το στέλεχος του εγκεφάλου.



Εικόνα 1- Ανατομία του Ανθρώπινου Εγκεφάλου από [56]

Η γενική δομή του ανθρώπινου εγκεφάλου από την πλευρά του απεικονίζεται στην Εικόνα 1, το οποίο αποτελείται από δύο ημισφαίρια, καθένα από τα οποία ελέγχει την αντίθετη πλευρά του σώματος. Δηλαδή, το δεξιό ημισφαίριο κυβερνά το αριστερό και αντίστροφα. Συνδέεται με νοητικές λειτουργίες όπως η σκέψη, η μνήμη και η ομιλία. Υπάρχουν τέσσερις λοβοί στον εγκέφαλο.

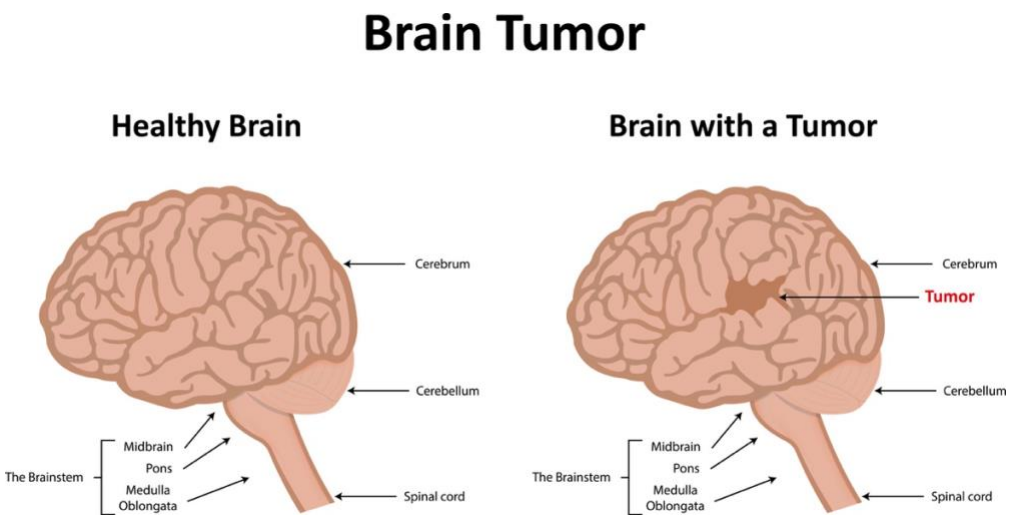
Είναι:

1. Ο πρώτος είναι ο μετωπιαίος λοβός που συνδέεται με την οργάνωση, την επίλυση προβλημάτων, τον προγραμματισμό, την επιλεκτικότητα, την προσοχή και τα συναισθήματα.

2. Ο δεύτερος είναι ο μερικός λοβός που εμπλέκεται στην κίνηση, τον προσανατολισμό, την αναγνώριση και την κατανόηση των ερεθισμάτων που ελέγχουν την αίσθηση.
3. Ο κεφαλαιώδης λοβός σχετίζεται με την προεπεξεργασία των οπτικών εικόνων.
4. Ο κροταφικός λοβός συνδέεται με την αντίληψη, και την αναγνώριση των ακουστικών ερεθισμάτων του εργοστασίου διεγείρουν και την οπτική και λεκτική μνήμη.

2.1.2: Καρκίνος Εγκεφάλου

Ένας όγκος στον εγκέφαλο αναπτύσσεται ως αποτέλεσμα ανώμαλης κυτταρικής διαίρεσης στον εγκέφαλο, προκαλώντας λανθασμένη λειτουργία του εγκεφάλου. Σήμερα είναι γνωστές πάνω από 100 διαφορετικές μορφές όγκων του εγκεφάλου. Όταν τα κύτταρα στο σώμα γερνούν, κανονικά αντικαθίστανται από νέα κύτταρα. Οι όγκοι έχουν κύτταρα που δεν αντικαθίστανται από νέα κύτταρα και πολλαπλασιάζονται γρήγορα. Αυτό θα έχει ως αποτέλεσμα τη συσσώρευση κυττάρων και τη δημιουργία εξογκωμάτων γνωστών ως όγκων. Δηλώνει μια ανώμαλη και συχνά μη ρυθμισμένη ανάπτυξη των κυττάρων. Βρίσκεται στο εσωτερικό του κρανίου και μπορεί να συμπιέσει, να μετατοπίσει ή να επιτεθεί σε υγιή εγκεφαλικό ιστό. Η Εικόνα 2: δείχνει την ανατομία του υγιούς και του ανιχνευόμενου από όγκο εγκεφάλου. Η απεικόνιση διαδραματίζει κρίσιμο ρόλο στη διάγνωση ενός όγκου στον εγκέφαλο [5]. Τα κρίσιμα δεδομένα από την άποψη αυτή περιλαμβάνουν τις εσωτερικές πληροφορίες γύρω από την ακριβή διάσταση και τη θέση του όγκου στο εσωτερικό του εγκεφάλου. Σύμφωνα με τις διαθέσιμες στατιστικές, οι όγκοι αποτελούν τη δεύτερη αιτία για την άψυχη ζωή που σχετίζεται με τον καρκίνο σε παιδιά και των δύο φύλων που είναι κάτω των 20 ετών και σε άνδρες και γυναίκες των οποίων η ηλικία είναι στην τροχιά από 20 έως 39. Αυτό το παρατηρήσιμο γεγονός αυξάνει τη σημασία της αναγνώρισης του όγκου και παρέχει αποτελεσματικά μια βοήθεια για τους γιατρούς να εντοπίσουν τη διαταραχή εγκαίρως και να πακετάρουν τις ουσιαστικές φυσικές διεργασίες [6]. Τελευταία, ο όγκος του εγκεφάλου έχει αναδειχθεί ως ένας από τους κρίσιμους λόγους για την ενίσχυση της αιτίας της αυξανόμενης θνησιμότητας μεταξύ παιδιών και ενηλίκων.

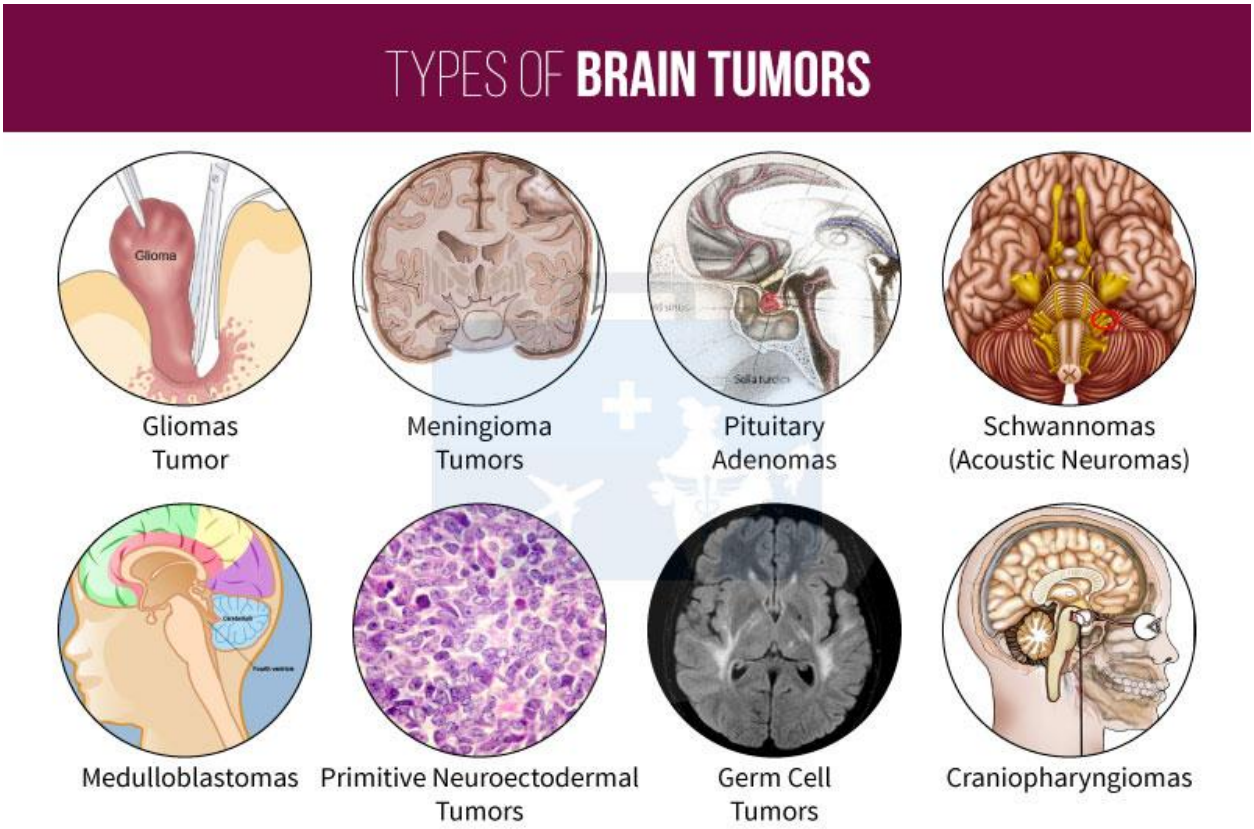


Εικόνα 2 - Υγιής Εγκέφαλος και Εγκέφαλος με Όγκο από [57]

Σύμφωνα με ορισμένες έρευνες, διαπιστώνεται ότι ο αριθμός των ατόμων που πάσχουν από όγκους του εγκεφάλου και πεθαίνουν από αυτούς αυξάνεται σε 300 ετησίως κατά τις τελευταίες δεκαετίες. Κατά τη μελέτη των ανωμαλιών που σχετίζονται με διάφορες ασθένειες που έχουν μεγάλο αντίκτυπο στον εγκέφαλο, παρατηρείται ότι ορισμένες ανωμαλίες αποτελούν πρόκληση για τη ζωή, ενώ οι άλλες δέχονται μεγαλύτερη επίδραση στις συνήθεις λειτουργίες του ανθρώπου. Από τις διαταραχές, η πιο σημαντική και επικίνδυνη διαταραχή που απειλεί την καθημερινή ζωή του ανθρώπου, γενικά, είναι ο όγκος ή ο καρκίνος.

2.1.3: Τύποι Καρκίνου στον Εγκέφαλο

Υπάρχουν 4 κύριες ταξινομήσεις όγκων του εγκεφάλου. Αυτή η ταξινόμηση των όγκων του ανθρώπινου εγκεφάλου βασίζεται στην προέλευση, τη σοβαρότητα και τις καρκινικές ιδιότητες. Στην Εικόνα 3 παρουσιάζονται οι διάφοροι τύποι όγκων του εγκεφάλου.



Εικόνα 3 - Διαφορετικοί Τύποι Όγκων Εγκεφάλου από [58]

Οι τύποι των ανθρώπινων όγκων του εγκεφάλου είναι πρωτοπαθείς, δευτεροπαθείς (μεταστατικοί), κακοήθεις και καλοήθεις

1. **Πρωτοπαθής όγκος εγκεφάλου**

Οι όγκοι του εγκεφάλου που προέρχονται από τον εγκέφαλο ονομάζονται πρωτοπαθείς όγκοι του εγκεφάλου και μπορεί να είναι είτε κακοήθεις είτε καλοήθεις. Τα γλοιώματα είναι τυπικά παραδείγματα αυτού του τύπου όγκου του εγκεφάλου και προέρχονται από τα γλοιακά κύτταρα του εγκεφάλου. Είναι πιθανό ένας πρωτοπαθής όγκος του εγκεφάλου να μετατραπεί σε δευτεροπαθή όγκο ή μεταστατικό όγκο με εξάπλωση από τον εγκέφαλο σε άλλα μέρη του σώματος.

2. **Δευτερογενής όγκος εγκεφάλου**

Οι δευτεροπαθείς (μεταστατικοί) όγκοι του εγκεφάλου προέρχονται από καρκινικά κύτταρα σε άλλο μέρος του σώματος και εξαπλώνονται στον εγκέφαλο. Τα καρκινικά κύτταρα ταξιδεύουν μέσω της λέμφου ή του αίματος για να φτάσουν στον εγκέφαλο. Όλοι οι

μεταστατικοί όγκοι είναι κακοήθεις (καρκινικοί). Οι δευτερογενείς όγκοι του εγκεφάλου εμφανίζονται συχνότερα από τους πρωτογενείς όγκους. Ο μεταστατικός καρκίνος μπορεί να εξαπλωθεί από τους πνεύμονες, τους νεφρούς, τους μαστούς ή ακόμη και από μελανώματα του δέρματος.

3. Καλοήθης όγκος εγκεφάλου

Το 50% των πρωτοπαθών όγκων του εγκεφάλου είναι καλοήθεις όγκοι του εγκεφάλου. Πρόκειται για αργά αναπτυσσόμενους όγκους, οι οποίοι έχουν συγκεκριμένο σχήμα και διακριτά όρια. Τα κύτταρα του καλοήθους όγκου φαίνονται φυσιολογικά και δεν κάνουν μεταστάσεις σε άλλα μέρη του σώματος ή σε άλλους ιστούς του εγκεφάλου. Εάν εμφανιστεί στις ζωτικές περιοχές του εγκεφάλου, μπορεί να είναι απειλητικός για τη ζωή. Η χειρουργική επέμβαση είναι μια αποτελεσματική μέθοδος για τη θεραπεία των καλοήθων όγκων του εγκεφάλου. Ο καλοήθης τύπος απεικονίζεται μέσω των ακόλουθων χαρακτηριστικών.

- Δεν εξαπλώνεται στους γειτονικούς ιστούς.
- Αναπτύσσεται πολύ αργά.
- Εντοπισμένο: Οι εντοπισμένοι όγκοι περιορίζονται σε μια συγκεκριμένη
- περιοχή
- Είναι συνήθως μη καρκινικό

4. Κακοήθης όγκος στον εγκέφαλο

Οι κακοήθεις όγκοι του εγκεφάλου είναι καρκινικοί. Είναι όγκοι που αναπτύσσονται γρήγορα και εξαπλώνονται σε άλλα τμήματα του σώματος αφού εισβάλλουν σε άλλες περιοχές του εγκέφαλο. Οι κακοήθεις όγκοι έχουν πολύ χαμηλό ποσοστό επιβίωσης. Μέσω του εγκεφαλονωτιαίου υγρού, οι όγκοι αυτοί μπορούν να ταξιδέψουν από τον εγκέφαλο στον νωτιαίο μυελό. Δεν έχουν σαφή όρια και έχουν ανομοιόμορφο σχήμα. Τα ακόλουθα χαρακτηριστικά περιγράφουν το κακοήθες είδος. Είναι ικανός να εισέλθει στο επόμενο κύτταρο με τη χρήση δύναμης και είναι πιθανό να επανεμφανιστεί ακόμη και μετά από θεραπεία.

- Είναι καρκινική στη φύση της.
- Επεμβατικό: Αυτό τείνει να εξαπλώνεται στην παρακείμενη περιοχή
- και είναι πραγματικά δύσκολο να απαλλαγούμε εντελώς από αυτά.
- Αναπτύσσεται με πολύ μεγάλη ταχύτητα.

2.1.4: Συμπτώματα Όγκων Εγκεφάλου

Τα συμπτώματα εξαρτώνται από το μέγεθος και τη θέση του όγκου. Τα συμπτώματα μπορεί να εμφανιστούν όταν ο όγκος πιέζει ένα τμήμα του εγκεφάλου και όταν εμποδίζει τη ροή του στον εγκέφαλο. Τα πιο συνηθισμένα συμπτώματα του όγκου στον εγκέφαλο είναι:

1. Πονοκέφαλος
2. Αλλαγές στην προσωπικότητα
3. Μυϊκό τράνταγμα
4. Ναυτία
5. Αλλαγή στην ομιλία, την όραση και την ακοή
6. Πρόβλημα στην ισορροπία ή στο περπάτημα
7. Μούδιασμα
8. Απώλεια μνήμης

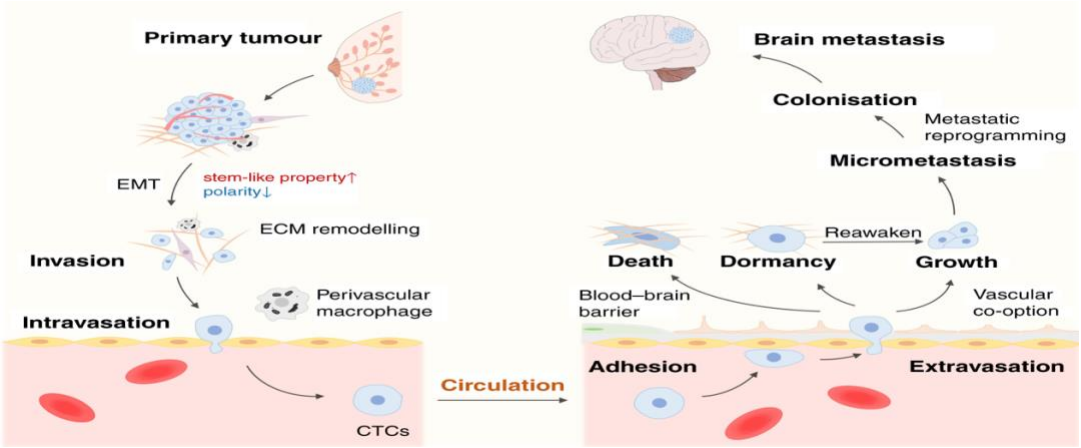
2.1.5: Παράγοντες Κινδύνου για Όγκο Εγκεφάλου

Κανείς δεν γνωρίζει τι προκαλεί τους όγκους του εγκεφάλου. Επί του παρόντος διεξάγονται πολλές έρευνες για τον προσδιορισμό των αιτιών. Αρκετές έρευνες έχουν δείξει ορισμένους από τους παράγοντες κινδύνου για όγκους του εγκεφάλου. Η ιονίζουσα ακτινοβολία από υψηλές δόσεις ακτίνων Χ και άλλες πηγές προκαλεί κυτταρική βλάβη, η οποία οδηγεί σε σχηματισμό όγκων. Τα άτομα που έχουν εκτεθεί σε ιονίζουσα ακτινοβολία διατρέχουν υψηλότερο κίνδυνο ανάπτυξης όγκων στον εγκέφαλο. Ο κίνδυνος ανάπτυξης όγκου στον εγκέφαλο αυξάνεται με την ηλικία. Το μυελοβλάστωμα είναι πιο συχνό στα παιδιά. Οι λευκοί άνθρωποι έχουν περισσότερες πιθανότητες από άλλες φυλές να εμφανίσουν όγκους στον εγκέφαλο. Οι Αφροαμερικανοί προσβάλλονται δυσανάλογα από μηνιγγίωμα. Τα κινητά τηλέφωνα, οι τραυματισμοί στο κεφάλι, η έκθεση σε χημικές ουσίες και τα μαγνητικά πεδία μελετώνται για να διαπιστωθεί αν μπορεί να αποτελούν παράγοντες κινδύνου για όγκους στον εγκέφαλο.

2.1.6: Μετάσταση Εγκεφάλου

Η μετάσταση στον εγκέφαλο αποτελεί μια σημαντική και συχνά καταστροφική διαφόρων μορφών καρκίνου, που αντιπροσωπεύει μια πολύπλοκη εκδήλωση της εξέλιξης της νόσου. Καθώς τα καρκινικά κύτταρα εξαπλώνονται από την αρχική εστία τους και εισβάλλουν στον εγκεφαλικό

παρέγχυμα, δημιουργούν δευτερεύοντες όγκους, οδηγώντας σε μια πληθώρα κλινικών προκλήσεων. Σε αυτή την έκθεση, σκοπεύουμε να διασαφηνίσουμε τη φύση της εγκεφαλικής μετάστασης, εξερευνώντας τα αίτια της, τον αντίκτυπο στους ασθενείς και τις ευρύτερες επιπτώσεις στον τομέα της ογκολογίας. [7] - [9].



Εικόνα 4 - Κύκλος Μετάστασης στον Εγκέφαλο από [59]

1. **Ορισμός και Μηχανισμοί:**

   Η εγκεφαλική μετάσταση αναφέρεται στην εξάπλωση καρκινικών κυττάρων από έναν πρωτογενή όγκο αλλού στο σώμα στον εγκέφαλο. Αυτή η διάδοση συνήθως συμβαίνει μέσω αιματογενούς ή λεμφατικής οδού, καθώς τα κακοήθη κύτταρα ταξιδεύουν μέσω της κυκλοφορίας του αίματος ή των λεμφικών αγγείων για να δημιουργήσουν δευτερεύουσες βλάβες μέσα στον εγκεφαλικό ιστό. Η διαδικασία αυτή περιλαμβάνει μια σειρά σύνθετων αλληλεπιδράσεων μεταξύ των καρκινικών κυττάρων και του μικροπεριβάλλοντος του εγκεφάλου, που καταλήγουν στο σχηματισμό μεταστατικών εστιών [4] - [6]. Διάφοροι παράγοντες επηρεάζουν την πιθανότητα και το μοτίβο της εγκεφαλικής μετάστασης, συμπεριλαμβανομένων των εγγενών μοριακών χαρακτηριστικών του πρωτογενούς όγκου, όπως ο γενετικός του αποτυπωμάτος και η τάση για μεταστατική εξάπλωση. Επιπλέον, ανατομικοί παράγοντες, όπως η παρουσία του φραγμού αίματος-εγκεφάλου, επηρεάζουν την ικανότητα των καρκινικών κυττάρων να εισβάλλουν στον εγκεφαλικό παρέγχυμα [10] - [13].

2. **Αίτια και Παράγοντες Κινδύνου:**

Αρκετοί παράγοντες συμβάλλουν στην ανάπτυξη εγκεφαλικής μετάστασης, με τον τύπο του πρωτογενούς καρκίνου να παίζει καθοριστικό ρόλο. Οι καρκίνοι που προέρχονται από τον πνεύμονα, το μαστό, το μελάνωμα, το παχύ έντερο και τον νεφρό συγκαταλέγονται στις πιο συχνές πηγές εγκεφαλικών μεταστάσεων. Η πιθανότητα εγκεφαλικής μετάστασης ποικίλλει ανάλογα με την επιθετικότητα, το στάδιο και το μοριακό προφίλ του πρωτογενούς όγκου [14] - [15]. Άλλοι παράγοντες κινδύνου περιλαμβάνουν την παρουσία συστηματικών μεταστάσεων, που υποδεικνύουν προχωρημένη νόσο, καθώς και συγκεκριμένες γενετικές μεταλλάξεις που σχετίζονται με αυξημένο μεταστατικό δυναμικό. Επιπλέον, ορισμένες φυσιολογικές καταστάσεις, όπως η ανοσοκαταστολή ή η διαταραγμένη ακεραιότητα του φραγμού αίματος-εγκεφάλου, μπορούν να διευκολύνουν την εγκατάσταση εγκεφαλικών μεταστάσεων [10] - [15].

3. **Επίδραση στους Ασθενείς:**

Η επίδραση της εγκεφαλικής μετάστασης στους ασθενείς είναι βαθιά, καθώς περιλαμβάνει ένα ευρύ φάσμα σωματικών, γνωστικών και συναισθηματικών προκλήσεων. Τα συμπτώματα ποικίλλουν ανάλογα με την εντόπιση, το μέγεθος και τον αριθμό των μεταστατικών εστιών, αλλά συχνές εκδηλώσεις περιλαμβάνουν πονοκεφάλους, επιληπτικές κρίσεις, γνωστικά ελλείμματα και κινητικές διαταραχές [12] - [16]. Επιπλέον, η εγκεφαλική μετάσταση επηρεάζει σημαντικά την ποιότητα ζωής και τη συνολική πρόγνωση των ασθενών.

Η παρουσία ενδοκρανιακής νόσου συχνά καθιστά απαραίτητες επιθετικές θεραπευτικές μεθόδους, όπως χειρουργική επέμβαση, ακτινοθεραπεία και συστηματική χημειοθεραπεία. Ωστόσο, παρά τις προόδους στη θεραπεία, η πρόγνωση για τους ασθενείς με εγκεφαλική μετάσταση παραμένει επιφυλακτική, με τα ποσοστά συνολικής επιβίωσης να μετρώνται συνήθως σε μήνες έως χρόνια [12] - [16]. Οι ψυχοκοινωνικές επιπτώσεις είναι επίσης σημαντικές, καθώς οι ασθενείς και οι οικογένειές τους έρχονται αντιμέτωποι με το συναισθηματικό φορτίο της διάγνωσης του καρκίνου, το οποίο επιδεινώνεται περαιτέρω από τις προκλήσεις που συνδέονται με την εγκεφαλική μετάσταση. Η γνωστική εξασθένηση, οι αλλαγές στην προσωπικότητα και η λειτουργική έκπτωση επιβαρύνουν περαιτέρω τους ασθενείς και τους φροντιστές τους, επισημαίνοντας τη πολυδιάστατη φύση της νόσου [16].

4. **Επιπτώσεις στην Ογκολογία:**

Η εγκεφαλική μετάσταση θέτει σημαντικές προκλήσεις για τους ογκολόγους και τους παρόχους υγειονομικής περίθαλψης, καθιστώντας απαραίτητη μια πολυεπιστημονική προσέγγιση στη διάγνωση, τη θεραπεία και την υποστηρικτική φροντίδα. Οι εξελίξεις στη νευρο απεικόνιση, τον μοριακό χαρακτηρισμό και τις στοχευμένες θεραπείες έχουν μεταμορφώσει τη διαχείριση της εγκεφαλικής μετάστασης, προσφέροντας νέους δρόμους για εξατομικευμένες θεραπευτικές στρατηγικές. Ωστόσο, παραμένουν πολλά αναπάντητα ερωτήματα σχετικά με τη βέλτιστη διαχείριση της εγκεφαλικής μετάστασης, συμπεριλαμβανομένου του ρόλου των νέων θεραπευτικών μεθόδων όπως η ανοσοθεραπεία και η εξατομικευμένη ιατρική. Επιπλέον, η βελτίωση της έγκαιρης ανίχνευσης, η ενίσχυση της θεραπευτικής αποτελεσματικότητας και η μείωση των τοξικοτήτων που σχετίζονται με τη θεραπεία αποτελούν συνεχείς τομείς ερευνητικού ενδιαφέροντος [11, 15].

5. **Αναγνώριση και Πρόβλεψη Θανάτου:**

Ο εντοπισμός και η πρόβλεψη του θανάτου σε ασθενείς με εγκεφαλική μετάσταση αποτελούν κρίσιμες πτυχές της κλινικής διαχείρισης και πρόγνωσης. Διάφοροι κλινικοί και απεικονιστικοί παράγοντες συμβάλλουν στη διαστρωμάτωση κινδύνου και στην αξιολόγηση της πρόγνωσης σε αυτόν τον πληθυσμό. Οι τεχνικές νευροαπεικόνισης, όπως η μαγνητική τομογραφία (MRI) και η αξονική τομογραφία (CT), διαδραματίζουν κεντρικό ρόλο στην ανίχνευση και τον χαρακτηρισμό των εγκεφαλικών μεταστάσεων, παρέχοντας ουσιαστικές πληροφορίες σχετικά με το μέγεθος, τον αριθμό και τη θέση των βλαβών. Επιπλέον, οι προηγμένες απεικονιστικές τεχνικές, όπως η λειτουργική μαγνητική τομογραφία και η τομογραφία εκπομπής ποζιτρονίων (PET), προσφέρουν γνώσεις σχετικά με τη βιολογία του όγκου και τη μεταβολική δραστηριότητα, βοηθώντας στον σχεδιασμό της θεραπείας και στην πρόβλεψη του αποτελέσματος.

Πέρα από τις απεικονιστικές παραμέτρους, κλινικές μεταβλητές όπως η ηλικία, η λειτουργική κατάσταση, η ιστολογική εικόνα του πρωτογενούς όγκου και η παρουσία εξωκρανιακών μεταστάσεων ενημερώνουν τα προγνωστικά μοντέλα και τους δείκτες κινδύνου που χρησιμοποιούνται για την πρόβλεψη της επιβίωσης. Επιπλέον, οι νέοι βιοδείκτες και οι τεχνικές μοριακού προφίλ παρέχουν ελπίδα για τη βελτίωση της διαστρωμάτωσης κινδύνου και τον καθορισμό εξατομικευμένων αποφάσεων θεραπείας. Ο συνδυασμός δεδομένων πολλαπλών τρόπων και η αξιοποίηση μεθόδων μηχανικής μάθησης επιτρέπουν την ανάπτυξη προβλεπτικών μοντέλων ικανών να προβλέψουν με

ακρίβεια το θάνατο σε ασθενείς με εγκεφαλική μετάσταση, διευκολύνοντας την έγκαιρη παρέμβαση και τις εξατομικευμένες θεραπευτικές στρατηγικές για τη βελτιστοποίηση της φροντίδας των ασθενών και της ποιότητας ζωής τους [11, 15].

Συμπερασματικά, η εγκεφαλική μετάσταση αποτελεί μια σοβαρή πρόκληση στον τομέα της ογκολογίας, που χαρακτηρίζεται από τη ύπουλη φύση της, την κλινική πολυπλοκότητα και τις βαθιές επιπτώσεις στη ζωή των ασθενών. Η κατανόηση των μηχανισμών που υποκρύπτονται στη διασπορά των μεταστάσεων, ο εντοπισμός τροποποιήσιμων παραγόντων κινδύνου και η ανάπτυξη καινοτόμων θεραπευτικών προσεγγίσεων είναι κεφαλαιώδους σημασίας για την αντιμετώπιση αυτής της μη καλυμμένης κλινικής ανάγκης. Διασαφηνίζοντας τα αίτια, τις επιπτώσεις και τις συνέπειες της εγκεφαλικής μετάστασης, στοχεύουμε στην ευαισθητοποίηση, την υποστήριξη και την εντατικοποίηση της έρευνας με στόχο τη βελτίωση των αποτελεσμάτων για τους ασθενείς που αντιμετωπίζουν αυτό το βαρύ φορτίο νόσου.

2.2: Έλεγχος και Απεικόνιση Καρκίνου Εγκεφάλου

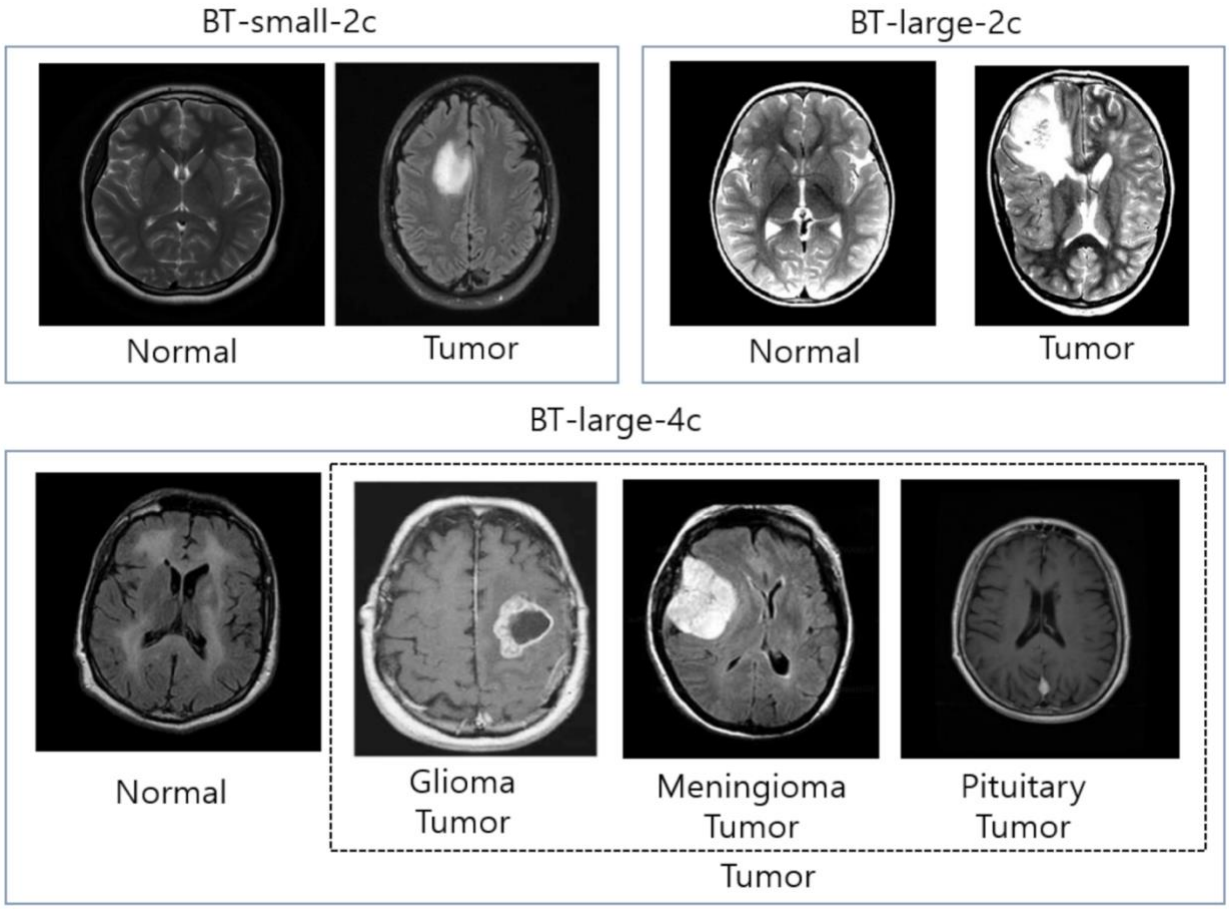
Η βιολογική ανάλυση έχει αναπτυχθεί σημαντικά τις τελευταίες δύο δεκαετίες. Η εκπληκτική ανάπτυξη νέων μεθόδων που πειραματίστηκαν στον τομέα της επεξεργασίας ιατρικών εικόνων τα τελευταία χρόνια βοήθησε στην ανάπτυξη πολλών αλγορίθμων και πακέτων λογισμικού για την επεξεργασία εικόνων. Λόγω της διαθεσιμότητας κατάλληλων ανιχνευτών, αναπτύχθηκαν πολύ γρήγορα νέες μέθοδοι για την παρακολούθηση της διάγνωσης και τη θεραπεία των ασθενών. Τον τελευταίο αιώνα, η τεχνολογία έχει εξελιχθεί από τις ακτίνες X σε ένα ευρύ φάσμα μεθόδων απεικόνισης, όπως η υπολογιστική τομογραφία (CT), η τομογραφία εκπομπής ποζιτρονίων (PET), η απεικόνιση με υπερήχους και η απεικόνιση μαγνητικού συντονισμού (MRI). Ιατρική απεικόνιση αποτελέσματα που έχουν υποστεί επαναστατικές αλλαγές, όπως η απεικόνιση μαγνητικού συντονισμού, έχουν παράσχει λεπτομερείς πληροφορίες σχετικά με διάφορες παθήσεις του εγκεφάλου και μπορούν να αξιολογήσουν την κατάσταση πολλών ασθενειών, παρέχοντας ακριβή διάγνωση.

2.2.1: Μαγνητική Τομογραφία (MRI / functional MRI)

Η μαγνητική τομογραφία (MRI) είναι μια ακτινολογική διαγνωστική διαδικασία που χρησιμοποιεί ελκυστικά πεδία και ραδιοκύματα για να παρέχει μια πλήρη εικόνα των ευαίσθητων ιστών, των εσωτερικών δομών και των οστών του σώματος. Η μαγνητική τομογραφία σαρώνει τη σπονδυλική στήλη ενεργοποιώντας σωματίδια υδρογόνου με έναν μαγνήτη που ταξιδεύει γύρω από το σώμα. Η μαγνητική τομογραφία δεν χρησιμοποιεί ακτίνες X ή ιονίζουσα ακτινοβολία, γεγονός που τη διαχωρίζει από τις αξονικές ή τις αξονικές τομογραφίες, καθώς και από τις τομογραφίες PET.

Η functional MRI θα μπορούσε να είναι ένα πολύτιμο σύστημα νευροαπεικόνισης που μετρά τη δράση του εγκεφάλου αναλύοντας τις αλλαγές εργασίας που εντοπίζονται στην κυκλοφορία του αίματος. Αυτό το πλαίσιο βασίζεται στην ίδια την αλήθεια ότι η εγκεφαλική κυκλοφορία του αίματος και η φυσική μονάδα εδάφους ενεργοποίησης των κυττάρων είναι συνδεδεμένες. Μόλις χρησιμοποιηθεί μια περιοχή του εγκεφάλου, η κυκλοφορία του αίματος σε αυτή την περιοχή θα αυξηθεί επίσης. Ο βασικός τύπος της ελκυστικής απεικόνισης αντήχησης κάνει χρήση της βελτίωσης του Seiji Ogawa που εξαρτάται από το επίπεδο αίματος-οξυγόνου (BOLD). Πρόκειται για ένα συγκεκριμένο εγκεφαλικό και ένα σώμα για τον καθορισμό της κίνησης στο εσωτερικό του μυαλού ή του νωτιαίου μυελού των ανθρώπων ή διαφορετικών πλάσματα με την απεικόνιση της αλλαγής στην κυκλοφορία του αίματος (αιμοδυναμική αντίδραση) που συνδέεται με τη χρήση ζωτικότητας από τα εγκεφαλικά κύτταρα. Η επακόλουθη αναπαράσταση του νου θα μπορούσε να παρασχεθεί διαγραμματικά με τη σκίαση του κωδικού της ποιότητας της εκκίνησης στον εγκέφαλο ή στη συγκεκριμένη περιοχή που εξετάζεται. Μπορεί ακόμη να συνδυαστεί και να συμπληρωθεί με διάφορα μέτρα της φυσιολογίας του εγκεφάλου, όπως το διάγραμμα και η φασματοσκοπία εγγύς υπέρυθρου (NIRS). η σκέψη της απεικόνισης μαγνητικού συντονισμού επεκτείνει την καινοτομία της τομογραφικής σάρωσης νωρίτερα και, επιπλέον, την αποκάλυψη των ιδιοτήτων του πλούσιου σε οξυγόνο αίματος. Η μαγνητική τομογραφία εφαρμόζεται σε διαφορετικά μεγέθη εικόνας. Ο εγκέφαλος της Τομογραφίας χρησιμοποιεί μια έντονη, αέναη, στατική ροή για να ρυθμίσει τους πυρήνες που βρίσκονται μέσα στο μυαλό που εξετάζονται. Μια άλλη ροή, το γωνιακό πεδίο, συνδέεται στη συνέχεια για να ανακαλύψει χωρικά εντελώς διαφορετικούς πυρήνες. Μόλις αποβληθεί το πεδίο RF, οι πυρήνες επανέρχονται στις μοναδικές τους καταστάσεις και επιπλέον, η ενέργεια που παράγουν μετράται με έναν βρόχο για να αναπαραχθούν οι θέσεις των πυρήνων. Η τομογραφία δίνει έτσι μια στατική βασική ανάγνωση της εγκεφαλικής ύλης. Το ουσιαστικό σημείο πίσω από την απεικόνιση

μαγνητικού συντονισμού είναι η κατασκευή της τομογραφίας για να συλλάβει χρήσιμες αλλαγές στο εσωτερικό του εγκεφάλου που προκαλούνται από τη φυσική κίνηση των κυττάρων. Οι μεταβολές στις μαγνητικές ιδιότητες μεταξύ των αγωγών (αίμα πλούσιο σε οξυγόνο) και των φλεβών (αίμα με φτωχό οξυγόνο) του αίματος δίνουν αυτή τη σύνδεση.



Εικόνα 5 Μαγνητικές Τομογραφίες για διάφορα είδη όγκων από [60]

2.2.2: Μαγνητική Τομογραφία Όγκων Εγκεφάλου

Στην εικόνα **5** απεικονίζεται μια σαρωμένη εικόνα του εγκεφάλου που έχει προσβληθεί από όγκο. Ο όγκος απεικονίζεται ως στρογγυλή λευκή κηλίδα στην εικόνα. Ένας όγκος στον εγκέφαλο μπορεί να είναι καλοήθης (μη καρκινικός) ή κακοήθης (καρκινικός). Τα καρκινικά κύτταρα δεν είναι παρόντα στους καλοήθεις όγκους του εγκεφάλου. Για την εξάλειψή του μπορούν να χρησιμοποιηθούν χειρουργικές τεχνικές. Δεν εξαπλώνεται σε άλλες περιοχές του σώματος,

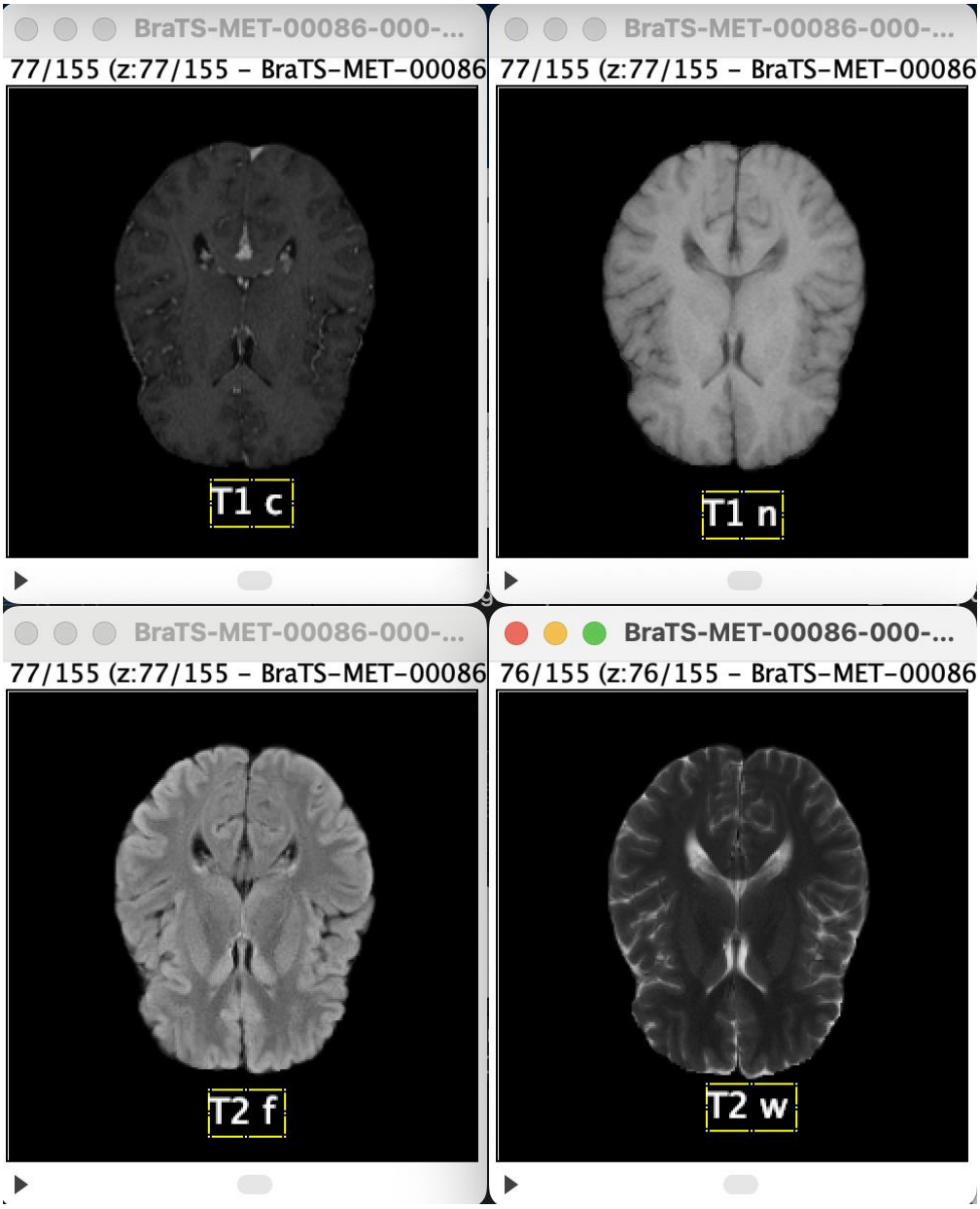
αλλά ασκεί πίεση σε μια ευαίσθητη περιοχή του εγκεφάλου, με αποτέλεσμα σοβαρές συνέπειες για την υγεία. Οι καλοήθεις όγκοι μπορεί να εξελιχθούν σε κακοήθεις όγκους. Τα καρκινικά κύτταρα βρίσκονται στους κακοήθεις όγκους του εγκεφάλου. Είναι πιο θανατηφόροι και αποτελούν σοβαρή απειλή για τη ζωή. Εξαπλώνονται γρήγορα και διηθούν τους γειτονικούς ιστούς και άλλα τμήματα του σώματος. Τα καρκινικά κύτταρα στον όγκο διαιρούνται και εξαπλώνονται σε άλλες περιοχές του σώματος. Ο βαθμός ενός όγκου στον εγκέφαλο καθορίζεται από έναν γιατρό.

Βαθμός Ι: Οι όγκοι βαθμού Ι είναι ως επί το πλείστων μη καρκινικοί και το ποσοστό επιβίωσης είναι πολύ υψηλό. Κατά την εξέταση μέσω μικροσκοπίου, τα κύτταρα παρουσιάζουν φυσιολογικές ιδιότητες. Αυτοί οι όγκοι αναπτύσσονται αργά και έχουν συγκεκριμένο σχήμα. Η χειρουργική επέμβαση είναι μια αποτελεσματική και βέλτιστη μέθοδος για τη θεραπεία των όγκων βαθμού Ι.

Βαθμός ΙΙ: Οι όγκοι εγκεφάλου βαθμού ΙΙ μπορεί να είναι είτε καρκινικοί είτε μη καρκινικοί και το ποσοστό επιβίωσης είναι μέτριο. Κατά την εξέταση μέσω μικροσκοπίου, το κύτταρο παρουσιάζει ελαφρώς ανώμαλες ιδιότητες. Αυτοί οι ιστοί μπορούν να εξαπλωθούν σε κοντινούς ιστούς και έχουν την ικανότητα να αναπαράγονται. Έχουν τη δυνατότητα να αλλάξουν ιδιότητες και να εξελιχθούν σε όγκο υψηλότερου βαθμού.

Βαθμός ΙΙΙ: Οι όγκοι βαθμού ΙΙΙ περιέχουν ανώμαλες κυτταρικές ιδιότητες και έχουν την ικανότητα να αναπαράγονται ενεργά. Τα κύτταρα αυτά αναπτύσσονται κοντά στους φυσιολογικούς ιστούς του εγκεφάλου και έχουν την ικανότητα να εξαπλώνονται και αναπαράγονται για να γίνουν όγκοι βαθμού IV. Η χειρουργική επέμβαση και η χημειοθεραπεία είναι αποτελεσματικές θεραπείες και η πιθανότητα επιβίωσης είναι πολύ χαμηλή. Μετά τη θεραπεία, η πιθανότητα επανεμφάνισης είναι υψηλή μετά από μια περίοδο παραθύρου.

Βαθμός IV: Οι όγκοι βαθμού IV είναι ο πιο καρκινικός τύπος όγκου στον εγκέφαλο. Έχουν ακανόνιστη κυτταρική εμφάνιση. Είναι σε θέση να αναπτύσσονται γρήγορα και μπορούν να επηρεάσουν τα γειτονικά φυσιολογικά κύτταρα. Σχηματίζονται νέα αιμοφόρα αγγεία για τη θρέψη και την ταχεία ανάπτυξη αυτών των κυττάρων. Ορισμένοι όγκοι έχουν περιοχές με νεκρά κύτταρα στο κέντρο και δεν έχουν συγκεκριμένο σχήμα ή μέγεθος.



Εικόνα 6 - Διαφορετικά είδη MRI

2.2.2: Επεξεργασία Ιατρικών Εικόνων

Η ιατρική απεικόνιση αποτελεί απαραίτητο εργαλείο για την ενίσχυση της διάγνωσης, της κατανόησης και της θεραπείας διαφόρων τύπων ασθενειών [17]. Η μαγνητική τομογραφία του εγκεφάλου έχει αποδειχθεί ότι είναι μια πολύ ισχυρή μέθοδος διάγνωσης και χρησιμοποιείται από τους γιατρούς για να διαπιστώσουν δομικές ανωμαλίες σε παθολογικές διαταραχές του εγκεφάλου.

Με το αυξανόμενο ενδιαφέρον στον τομέα της επεξεργασίας ιατρικών εικόνων, έχουν αναπτυχθεί διάφορα αυτόνομα και ημιαυτόματα εργαλεία που βοηθούν στην ιατρική διάγνωση. Η τρισδιάστατη ιατρική απεικόνιση υπόσχεται υψηλά αποτελέσματα στον τομέα της ιατρικής απεικόνισης. Η τρισδιάστατη απεικόνιση ήταν ο αναπτυσσόμενος τομέας για την παραγωγή εικόνων υψηλής ποιότητας. Το σήμα θα παράγει τη φωτογραφική ή αλλιώς ένα άλλο είδος σεναριακής εικόνας. Υπάρχουν τρία στάδια στην επεξεργασία εικόνας. Στο πρώτο στάδιο δίνεται η είσοδος, στη συνέχεια, στο δεύτερο στάδιο μπορεί να γίνει επεξεργασία και στο τελικό στάδιο μπορεί να ληφθεί η αποσαφηνισμένη έξοδος. Η έξοδος μπορεί να ληφθεί ανάλογα με την πηγή εισόδου που δίνεται.

- **Σύστημα λήψης εικόνας**

Ένα σύστημα λήψης εικόνας χρησιμοποιείται για τη δημιουργία φωτογραφικών εικόνων, όπως μια φυσική σκηνή ή η εσωτερική δομή ενός αντικειμένου. Η σύλληψη εικόνων, η επεξεργασία τους, η συμπίεσή τους, η αποθήκευσή τους, η εκτύπωσή τους και η παρουσίασή τους αποτελούν μέρος αυτής της διαδικασίας. Ο δικτυακός τόπος BIDC χρησιμοποιήθηκε για τη λήψη των μαγνητικών τομογραφιών εγκεφάλου που χρησιμοποιήθηκαν στην παρούσα έρευνα. Η μαγνητική τομογραφία T1-weighted διενεργήθηκε στους συμμετέχοντες σε μαγνητική τομογραφία εγκεφάλου στο προβλεπόμενο επίπεδο 1,5 T. Για να περιοριστεί το εύρος των εικονοστοιχείων του εγκεφάλου MRI, εφαρμόστηκε η τεχνική του "πρωτοτύπου κεφαλής".

- **Προεπεξεργασία**

Το αρχικό στάδιο της επεξεργασίας εικόνας χρησιμοποιείται για τη βελτίωση της ανίχνευσης ύποπτων περιοχών. Οι μικρότερες λεπτομέρειες της εικόνας αυξάνονται και ο θόρυβος στην εικόνα αφαιρείται. Η κλινική μαγνητική τομογραφία μολύνεται από θόρυβο, μειώνοντας την ακρίβεια της εικόνας. Αυτή η μείωση του θορύβου πραγματοποιείται με τη χρήση διαφόρων φίλτρων.

- **Τμηματοποίηση**

Η τεχνική της κατάτμησης χρησιμοποιείται για την απομόνωση του αντικειμένου-στόχου από την εικόνα δοκιμής. Αποσυνθέτει την εικόνα στα βασικά συναφή κομμάτια της. Το εκάστοτε πρόβλημα καθορίζει το επίπεδο στο οποίο πραγματοποιείται η υποδιαίρεση. Με βάση την ποσότητα εξαγωγής χαρακτηριστικών που απαιτείται μετά την τμηματοποίηση, μπορούμε να χωρίσουμε την τεχνική σε τεχνικές προσανατολισμένες στα εικονοστοιχεία

προσανατολισμένες στις ακμές και προσανατολισμένες στην υφή. Μπορούν επίσης να χρησιμοποιηθούν υβριδικές προσεγγίσεις, οι οποίες συνδυάζουν πολλές επεξεργασίες.

- **Εξαγωγή χαρακτηριστικών**

Το αποτέλεσμα της τμηματοποίησης, το οποίο είναι συνήθως μια ειδικά επιλεγμένη περιοχή της εικόνας που διαχωρίζεται από την υπόλοιπη εικόνα, ακολουθείται στη συνέχεια από την αναπαράσταση και την περιγραφή. Είναι ζωτικής σημασίας ο μετασχηματισμός των δεδομένων σε μορφή κατάλληλη για επεξεργασία από υπολογιστή. Ο σκοπός της εξαγωγής χαρακτηριστικών είναι να δοθεί έμφαση στις οπτικές πληροφορίες σε ένα συγκεκριμένο επίπεδο όπου μπορούν να λειτουργήσουν οι μέθοδοι που αναφέρονται παρακάτω. Ως αποτέλεσμα, οι πληροφορίες άλλων επιπέδων πρέπει να καταστέλλονται. Προσεγγίσεις για την εξαγωγή χαρακτηριστικών χρησιμοποιήθηκαν σε πολλά επίπεδα, όπως το επίπεδο δεδομένων, το επίπεδο εικονοστοιχείων, το επίπεδο ακμών, το επίπεδο υφής και το περιφερειακό επίπεδο. Η περιγραφή είναι επίσης γνωστή ως επιλογή χαρακτηριστικών. Επικεντρώνεται στην απόκτηση των χαρακτηριστικών που οδηγούν σε ποσοτικές πληροφορίες χρήσιμες για τη διαφοροποίηση ενός είδους πράγματος από ένα άλλο.

- **Ταξινόμηση**

Η ταξινόμηση, επίσης γνωστή ως ανίχνευση, είναι η διαδικασία ανάθεσης ορισμένων κατηγοριών αντικειμένων σε όλες τις συνδεδεμένες περιοχές που παράγονται από την τμηματοποίηση. Τα χαρακτηριστικά με βάση την περιοχή που αφηγούνται κατάλληλα τα χαρακτηριστικά των αντικειμένων χρησιμοποιούνται συνήθως για να βοηθήσουν τη διαδικασία ταξινόμησης.

Κεφάλαιο 3: Μηχανική Μάθηση και Κατάταξη

3.1: Μηχανική Μάθηση

3.1.1: Εισαγωγή στην Μηχανική Μάθηση

Η Μηχανική Μάθηση είναι ένα πεδίο της Επιστήμης Υπολογιστών στο οποίο μελετώνται αλγόριθμοι, οι οποίοι επιτρέπουν σε ένα υπολογιστικό σύστημα να πραγματοποιεί προβλέψεις ή να λαμβάνει αποφάσεις χωρίς να έχει προγραμματιστεί εξαρχής η ακριβής συλλογιστική πορεία που πρέπει να ακολουθήσει για να τις λάβει. Συγκεκριμένα, μια εφαρμογή ενός αλγορίθμου μηχανικής μάθησης δημιουργεί ένα μη γραμμικό μαθηματικό μοντέλο που βασίζεται σε δεδομένα - δείγματα, ή αλλιώς δεδομένα εκπαίδευσης με τα οποία τροφοδοτείται. Μοντέλα μηχανικής μάθησης, χρησιμοποιούνται σε διάφορους τομείς, όπως η πρόβλεψη των τιμών μίας μετοχής, η αυτόματη αναγνώριση ηλεκτρονικών μηνυμάτων απάτης, η επεξεργασία φυσικής γλώσσας κ.α. Ένα από τα πεδία στο οποίο βρίσκει σημαντική συνεισφορά είναι αυτό της ιατρικής και συγκεκριμένα στο πεδίο της πρόβλεψης μελλοντικών συμβάντων και επιπτώσεων καθώς και αναγνώρισης μοτίβων και προτύπων. [18]

Η Μηχανική Μάθηση επιτρέπει την ανάλυση τεράστιων ποσοτήτων δεδομένων. Αν και γενικά παρέχει ταχύτερα, πιο ακριβή αποτελέσματα για τον εντοπισμό κερδοφόρων ευκαιριών ή κινδύνων, μπορεί επίσης να απαιτήσει επιπλέον χρόνο και πόρους για να εκπαιδευτεί σωστά. Ο συνδυασμός της μηχανικής μάθησης με την τεχνητή νοημοσύνη και τις γνωστικές τεχνολογίες μπορεί να την κάνει ακόμη πιο αποτελεσματική στην επεξεργασία μεγάλου όγκου πληροφοριών. Η Μηχανική Μάθηση αποτελεί ένα δυναμικά αναπτυσσόμενο πεδίο με πλήθος εφαρμογών σε διάφορους επιστημονικούς και βιομηχανικούς τομείς. Η ικανότητα των αλγορίθμων Μηχανικής Μάθησης να "μαθαίνουν" από δεδομένα και να βελτιώνουν την απόδοσή τους με την πάροδο του χρόνου, ανοίγει τον δρόμο για νέες καινοτομίες και λύσεις σε σύνθετα προβλήματα. [18]

3.1.2: Μηχανική Μάθηση

Οι κατηγορίες μηχανικής μάθησης συνήθως ταξινομούνται σε τρεις μεγάλες κατηγορίες, ανάλογα με τη φύση των δεδομένων εισόδου που είναι διαθέσιμα σε ένα σύστημα εκμάθησης [19]. Αυτές είναι:.

Οι αλγόριθμοι μηχανικής μάθησης με επίβλεψη (supervised) μπορούν να εφαρμόσουν την γνώση που έχει αποκτηθεί στο παρελθόν σε νέα δεδομένα χρησιμοποιώντας επισημασμένα παραδείγματα για την πρόβλεψη μελλοντικών γεγονότων. Ξεκινώντας από την ανάλυση ενός γνωστού συνόλου δεδομένων, ο αλγόριθμος εκμάθησης παράγει μια συνάρτηση για να κάνει προβλέψεις σχετικά με τις τιμές εξόδου. Το σύστημα είναι σε θέση να παρέχει προβλέψεις για οποιαδήποτε νέα είσοδο μετά από επαρκή εκπαίδευση. Ο αλγόριθμος εκμάθησης μπορεί επίσης να συγκρίνει την έξοδο του με τη σωστή έξοδο και να βρει σφάλματα προκειμένου να τροποποιήσει και να βελτιστοποιήσει το εσωτερικό του μοντέλο με σκοπό την μείωση αυτού του σφάλματος.

Αντίθετα, οι μη επιτηρούμενοι (unsupervised - χωρίς επίβλεψη) αλγόριθμοι μηχανικής μάθησης βρίσκουν εφαρμογή όταν τα δεδομένα που χρησιμοποιούνται για την εκπαίδευση δεν απαιτούν επισημειωμένες ετικέτες από τον άνθρωπο. Η μη επιτηρούμενη μάθηση μελετά πώς τα συστήματα μπορούν να συμπεράνουν μια συνάρτηση για να περιγράψουν μια κρυφή δομή από δεδομένα χωρίς ετικέτα. Ο σκοπός αυτού του συστήματος δεν είναι να συγκρίνει το αποτέλεσμα που παράγει με μία βάση πραγματικών αναμενόμενων αποτελεσμάτων, αλλά να διερευνά τα δεδομένα και να μπορεί να εξαγάγει συμπεράσματα από σύνολα δεδομένων για να περιγράψει κρυφές δομές από δεδομένα χωρίς ετικέτα.

Οι ημι-επιβλεπόμενοι (semi-supervised) αλγόριθμοι μηχανικής μάθησης εμπίπτουν κάπου μεταξύ της εποπτευόμενης και της μη εποπτευόμενης μάθησης, δεδομένου ότι χρησιμοποιούν τόσο δεδομένα με ετικέτα όσο και χωρίς για την εκπαίδευση - συνήθως μια μικρή ποσότητα δεδομένων με ετικέτες και μια μεγάλη ποσότητα δεδομένων χωρίς. Τα συστήματα που χρησιμοποιούν αυτήν τη μέθοδο είναι σε θέση να βελτιώσουν σημαντικά την ακρίβεια της μάθησης. Συνήθως, η ημι-εποπτευόμενη μάθηση επιλέγεται όταν τα αποκτηθέντα επισημασμένα δεδομένα απαιτούν εξειδικευμένους και σχετικούς πόρους για να οδηγήσουν σε ένα καλό μοντέλο.

Οι αλγόριθμοι ενισχυτικής μάθησης (reinforcement learning) είναι μια μέθοδος μάθησης που αλληλεπιδρά με το περιβάλλον της παράγοντας ενέργειες και ανακαλύπτει σφάλματα ή ανταμοιβές. Η αναζήτηση δοκιμών και σφαλμάτων και η καθυστερημένη ανταμοιβή είναι τα πιο σχετικά χαρακτηριστικά της μάθησης ενίσχυσης. Αυτή η μέθοδος επιτρέπει σε μηχανές και πράκτορες λογισμικού να προσδιορίζουν αυτόματα την ιδανική συμπεριφορά σε ένα συγκεκριμένο πλαίσιο, προκειμένου να μεγιστοποιήσουν την απόδοσή της. Απαιτείται απλή ανατροφοδότηση ανταμοιβής για να μάθει ο πράκτορας ποια ενέργεια είναι καλύτερη. Αυτό είναι γνωστό ως σήμα ενίσχυσης.

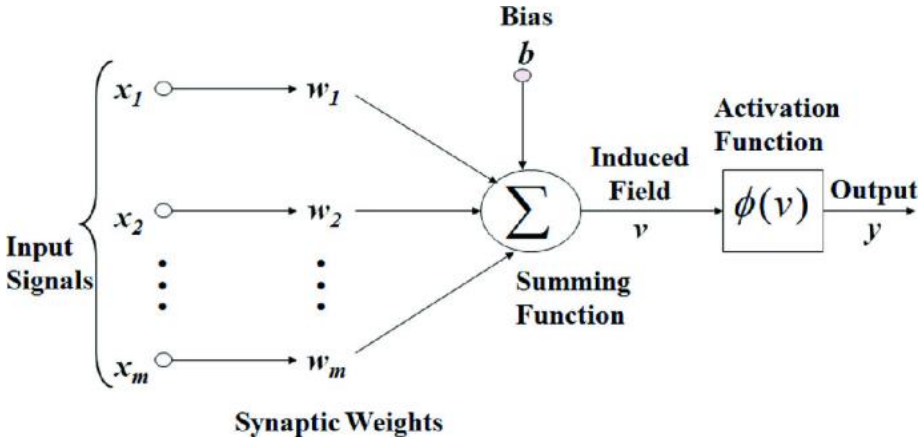
Στα πλαίσια των επιβλεπόμενων αλγορίθμων που αφορούν την παρούσα εργασία, ένα άλλο κριτήριο κατηγοριοποίησης των αλγορίθμων Μηχανικής Μάθησης είναι το είδος προβλήματος που αυτοί επιλύουν. Δύο μεγάλες ομάδες προβλημάτων είναι τα προβλήματα ταξινόμησης (classification) και τα προβλήματα παλινδρόμησης (regression). Οι αλγόριθμοι ταξινόμησης επιχειρούν να εκτιμήσουν την διακριτή ομάδα στην οποία ανήκει ένας στόχος y με βάση τις μεταβλητές εισόδου x. Για παράδειγμα, όταν έχουμε στην διάθεση μας δεδομένα από ραδιολογικά χαρακτηριστικά εικόνων μετάστασης εγκεφάλου (μεταβλητές εισόδου x) και επίσης έχοντας γνώση δεδομένων κλινικών αποτελεσμάτων σχετικά με την θνησιμότητα του ασθενή (μεταβλητές στόχου y), τότε ένας αλγόριθμος ταξινόμησης, εφόσον εκπαιδευτεί κατάλληλα στα παραπάνω δεδομένα και έχοντας γνώση των ραδιολογικών χαρακτηριστικών, μπορεί να προβλέψει με σχετική ακρίβεια την θνησιμότητα του ασθενή.

Στη μηχανική μάθηση, οι αλγόριθμοι παλινδρόμησης προσπαθούν να εκτιμήσουν τη συνάρτηση χαρτογράφησης (mapping function) f που οδηγεί από τις μεταβλητές εισόδου x σε αριθμητικές ή συνεχείς μεταβλητές εξόδου y. Σε αυτήν την περίπτωση, το y είναι ένας αριθμός, ο οποίος μπορεί να είναι ακέραια ή δεκαδική τιμή. Επομένως, τα προβλήματα πρόβλεψης παλινδρόμησης είναι συνήθως ποσότητες ή μεγέθη. Παραδείγματα των κοινών αλγορίθμων παλινδρόμησης περιλαμβάνουν την γραμμική παλινδρόμηση και τα δέντρα παλινδρόμησης. Ορισμένοι αλγόριθμοι, όπως η λογιστική παλινδρόμηση, έχουν τη λέξη παλινδρόμηση στα ονόματά τους, αλλά δεν είναι αλγόριθμοι παλινδρόμησης. [20]

3.1.3: Νευρωνικά Δίκτυα

Νευρώνας – Perceptron

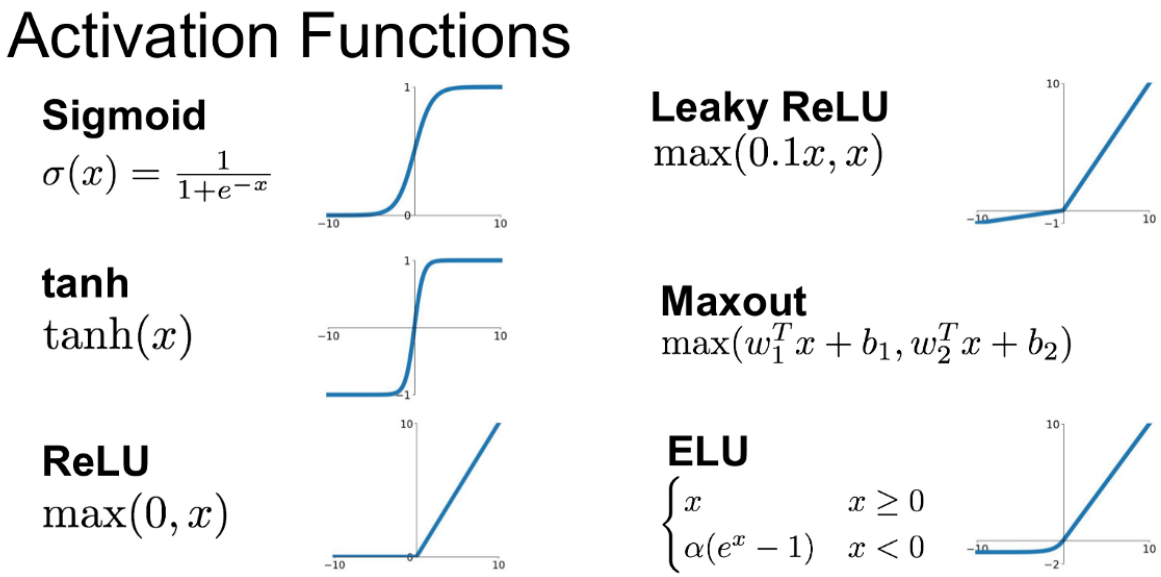
Η βασική μονάδα υπολογισμού σε ένα νευρωνικό δίκτυο είναι ο νευρώνας (Neuron), συχνά ονομαζόμενος και ως κόμβος. Το μοντέλο του Perceptron έχει περιγραφεί από τον Rosenblatt [18]. Λαμβάνει εισόδους από άλλους κόμβους ή από μια εξωτερική πηγή και υπολογίζει μια έξοδο. Κάθε είσοδος πολλαπλασιάζεται με το αντίστοιχο βάρος (Weight) και υπολογίζεται τα ολικό άθροισμα των γινομένων. Ο κόμβος εφαρμόζει μια συνάρτηση ενεργοποίησης σε αυτό το άθροισμα και έτσι προκύπτει η έξοδος του νευρώνα. Παρακάτω φαίνεται η αναπαράσταση ενός νευρώνα, καθώς και η εξίσωση της εξόδου.



Εικόνα 7 Βασική Δομή Perceptron

$$k = f(\sum_{i=0}^{n}(w_i * x_i + b_i))$$

Το παραπάνω δίκτυο παίρνει ως εισόδους τα X1, X2, ..., Xn που έχουν για βάρη τα W1, W2, ..., Wn αντίστοιχα. Επιπλέον για κάθε είσοδο υπάρχει ακόμα μια είσοδος με τιμή 1 με βάρος b η οποία ονομάζεται πόλωση (bias). Η συνάρτηση f είναι μη γραμμική και ονομάζεται συνάρτηση ενεργοποίησης (activation function). Ο σκοπός της συνάρτησης ενεργοποίησης είναι να εισάγει μη γραμμικότητα στην έξοδο ενός νευρώνα. Αυτό είναι σημαντικό καθώς σχεδόν όλα τα πραγματικά δεδομένα είναι μη γραμμικά. Παραδείγματα τέτοιων συναρτήσεων είναι η σιγμοειδής συνάρτηση, η υπερβολική εφαπτομένη και η ReLU (Rectified Linear U- nit). Η ReLu αποδίδει συχνά καλύτερα από άλλες συναρτήσεις ενεργοποίησης για πολλαπλά επίπεδα τα οποία θα παρουσιαστούν στη συνέχεια. Ο βασικός λόγος της αυξημένης απόδοσης.οφείλεται στο γεγονός ότι η ReLU είναι μια γραμμική συνάρτηση μη κορεσμού. Ο κορεσμός είναι το μεγαλύτερο πρόβλημα των δυο προηγούμενων σιγμοειδών συναρτήσεων. Σε αντίθεση λοιπόν με την υπερβολική εφαπτομένη, η ReLU δεν έρχεται σε κορεσμό στο -1, 0 ή 1. Αυτό την καθιστά την συχνότερα εμφανιζόμενη συνάρτηση ενεργοποίησης.

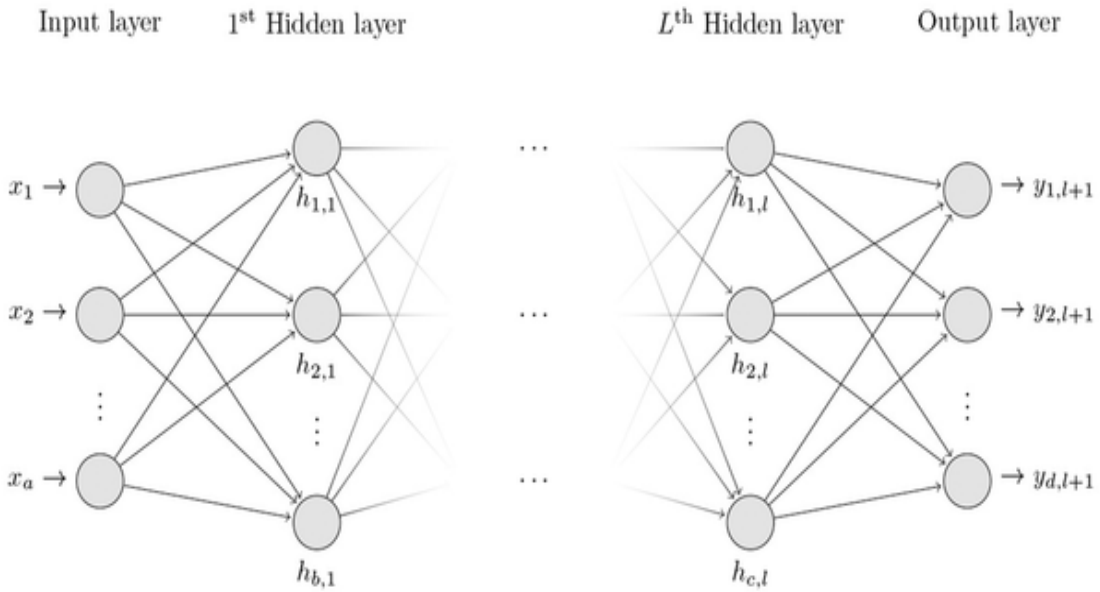


Εικόνα 8 - Συναρτήσεις Ενεργοποίησης

Δομή Νευρωνικού Δικτύου

Για την δημιουργία ενός Νευρωνικού Δικτύου, οι Τεχνητοί Νευρώνες που περιγράφηκαν παραπάνω οργανώνονται σε επίπεδα, όπου το κάθε επίπεδο επεξεργάζεται ένα σύνολο σημάτων. Το πρώτο επίπεδο ονομάζεται επίπεδο εισόδου και χρησιμοποιείται για την εισαγωγή των δεδομένων-χαρακτηριστικών στο δίκτυο. Είναι σημαντικό να αναφερθεί ότι τα στοιχεία του επιπέδου αυτού δεν είναι νευρώνες σαν αυτούς που περιγράφηκαν παραπάνω καθώς δεν εκτελούν κάποιον υπολογισμό. Στην συνέχεια, προστίθενται κρυφά επίπεδα (hidden layers), ένα ή περισσότερα, ενώ στο τέλος υπάρχει το επίπεδο εξόδου.

Οι Νευρώνες σε ένα Νευρωνικό Δίκτυο, μπορούν να χαρακτηριστούν ως μερικώς ή πλήρως συνδεδεμένοι. Έτσι, εάν όλοι οι Νευρώνες ενός επιπέδου συνδέονται με όλους τους υπόλοιπους, τότε χαρακτηρίζονται ως πλήρως συνδεδεμένοι. Σε κάθε άλλη περίπτωση χαρακτηρίζονται ως μερικώς συνδεδεμένοι. Μία τυπική περίπτωση μερικής σύνδεσης, είναι τα Δίκτυα με πρόσθια προώθηση (feedforward). Στα συγκεκριμένα, οι Νευρώνες ενός επιπέδου συνδέονται πλήρως με τους Νευρώνες στο επόμενο επίπεδο, ενώ δεν υπάρχουν συνδέσεις μεταξύ των Νευρώνων ενός επιπέδου και του προηγουμένου στο Νευρωνικό Δίκτυο. Σε αντίθετη περίπτωση το δίκτυο χαρακτηρίζεται ως δίκτυο ανατροφοδότητας (feedback network ή recurrent network), για παράδειγμα εάν υπάρχουν συνδέσεις μεταξύ νευρώνων ενός επιπέδου με το προηγούμενο επίπεδο.



Εικόνα 9 – Νευρωνικό δίκτυο πολλών επιπέδων (Multi Layer Perceptron)

Κανόνας Δέλτα

Ο κανόνας Δέλτα (Delta Rule), αναπτύχθηκε από τους Widrow και Hoff και αποτελεί γενίκευση του αλγορίθμου εκπαίδευσης Percepton. Συγκεκριμένα, είναι και αυτός καθοδηγούμενος από το σφάλμα, αφού προκύπτει από την ελαχιστοποίηση του μέσου τετραγωνικού σφάλματος (mean square error) των διανυσμάτων εκπαίδευσης. Έτσι, δείχνει τελικά πόσο αποκλίνει ένα δίκτυο από την επιθυμητή συνάρτηση. Το μέσο τετραγωνικό σφάλμα στο στοιχειώδες Percepton, για k διανύσματα εκπαίδευσης, μπορεί να υπολογιστεί από τη σχέση:

$$MSE = \frac{1}{k}\sum_{j=0}^{k}(t_j - in_j)^2$$

όπου σαν σήμα εισόδου θεωρείται το $\sum_{i=0}^{k}(w_{ki} * x_i)$ και n ο αριθμός των επιμέρους σημάτων εισόδου του νευρώνα. Θεωρώντας το διάνυσμα των βαρών (w_1, w_2, ..., w_n), μπορεί να οριστεί ο κανόνας Δέλτα που ονομάζεται και κανόνας επικλινούς μεθόδου (gradient decent), ως η αρνητική παράγωγος του μέσου τετραγωνικού σφάλματος, έτσι ώστε το διάνυσμα βαρών να προσεγγίσει σταδιακά το ιδανικό διάνυσμα. Θεωρώντας ότι

$$\Delta w = -\frac{\theta(MSE)}{\theta w_i}$$

και

$$\nabla(MSE) = (\frac{\theta(MSE)}{\theta w_i}, \ldots, \frac{\theta(MSE)}{\theta w_n})$$

προκύπτει ότι η μεταβολή στην τιμή του βάρους w_i για ένα από τα διανύσματα εκπαίδευσης x_i ως

$$\Delta w = w_{i\,(new)} - w_{i\,(old)} = \alpha * (t - input) * x_i$$

όπου *input* είναι το συνολικό σήμα του νευρώνα, t η επιθυμητή έξοδος, $w_{i\,(new)}$ και $w_{i\,(old)}$ η νέα και η παλιά τιμή του βάρους w_i , x_i η είσοδος της οποίας το βάρος αναπροσαρμόζεται και α ο ρυθμός μάθησης (learning rate), που ρυθμίζει το ρυθμό μεταβολής των βαρών. Η σταθερά α επηρεάζει την ταχύτητα σύγκλισης και καθορίζει την απόδοση του Κανόνα Δέλτα. Συγκεκριμένα, πολύ μεγάλες τιμές του α επιταχύνουν τη σύγκλιση στο ελάχιστο σφάλμα, όμως αυξάνουν τον κίνδυνο να προσπεραστεί το ελάχιστο MSE. Αντίθετα, οι μικρές τιμές του α μπορεί να αυξήσουν σημαντικά τον χρόνο εκπαίδευσης.

Επιπλέον, στην παραπάνω σχέση μπορεί αντί του σήματος εισόδου *input* να χρησιμοποιηθεί η πραγματική έξοδος του Νευρωνικού Δικτύου, λαμβάνοντας έτσι υπόψιν και τη συνάρτηση ενεργοποίησης. Η εκπαίδευση του Δικτύου σταματά όταν το μέσο τετραγωνικό σφάλμα γίνει μικρότερο από κάποια επιθυμητή τιμή. Παρόλο που ο αλγόριθμος Δέλτα, αποτελεί βελτίωση του αλγορίθμου μάθησης του στοιχειώδους percepton, δεν μπορεί να εφαρμοστεί στα δίκτυα με κρυφά επίπεδα, επειδή δεν είναι γνωστή η επιθυμητή έξοδος t σε κάθε Νευρώνα. Για την επίλυση του συγκεκριμένου προβλήματος χρησιμοποιείται ο αλγόριθμος ανάστροφης μετάδοσης λάθους

Αλγόριθμος ανάστροφης μετάδοσης λάθους (back propagation)

Για την παραπάνω τοπολογία είναι απαραίτητος ο υπολογισμός διορθώσεων στα βάρη του κάθε Νευρώνα ξεχωριστά. Για τον σκοπό αυτό γίνεται χρήση της ανάστροφης μετάδοσης λάθους (back propagation), η οποία βασίζεται στον γενικευμένο κανόνα Δέλτα (generalized Delta rule). Ο γενικευμένος κανόνας Δέλτα επιτρέπει τον καθορισμό του ποσοστού του συνολικού σφάλματος που αντιστοιχεί στα βάρη του κάθε Νευρώνα ακόμη και αν αυτά ανήκουν σε κρυφά επίπεδα των οποίων η επιθυμητή έξοδος δεν είναι γνωστή.

Για την εκπαίδευση του Multi Layer Percepton (MLP), γίνεται αρχικά ένα πρόσθιο πέρασμα (forward pass). Συγκεκριμένα, εισάγονται στην είσοδο δεδομένα από ένα διάνυσμα εκπαίδευσης και οι νευρώνες στο επίπεδο εισόδου παράγουν ένα αποτέλεσμα το οποίο στην συνέχεια αποτελεί είσοδο στο επόμενο επίπεδο. Η συγκεκριμένη διαδικασία επαναλαμβάνεται διαδοχικά για τα επόμενα κρυφά επίπεδα μέχρι το επίπεδο εξόδου. Έτσι, για n αριθμό νευρώνων του επιπέδου εισόδου, η είσοδος ενός κρυφού νευρώνα j, δίνεται από την σχέση:

$$in_j = \sum_{i=0}^{n} (w_{ij} * x_i)$$

όπου w_{ij} το βάρος της σύνδεσης μεταξύ των νευρώνων i, j και x_i το σήμα εισόδου του νευρώνα i. Αντίστοιχα, η έξοδος του συγκεκριμένου νευρώνα θα είναι

$$out_j = f(\sum_{i=0}^{n} (w_{ij} * x_i))$$

όπου f η συνάρτηση ενεργοποίησης του νευρώνα. Η έξοδος αυτή θα προωθηθεί στους νευρώνες του επόμενου επιπέδου.

Είναι σημαντικό να αναφερθεί πως η συνάρτηση ενεργοποίησης για τα δίκτυα που εκπαιδεύονται με ανάστροφη μετάδοση λάθους πρέπει να είναι μη γραμμική, μονότονα αύξουσα και παραγωγίσιμη για όλες τις τιμές εισόδου [21]

Το δίκτυο ξεκινά τους υπολογισμούς όπως και ένα απλό percepton, με τυχαίες τιμές στα βάρη των νευρώνων. Αντίστοιχα, θα υπολογιστεί και το σφάλμα εξόδου για τους Νευρώνες του επιπέδου εξόδου. Αφού υπολογιστεί το ακριβές σφάλμα στο επίπεδο εξόδου, για το οποίο είναι γνωστό το επιθυμητό αποτέλεσμα, είναι δυνατό να χρησιμοποιηθεί ο γενικευμένος κανόνας Δέλτα, για να προσαρμοστούν κατάλληλα οι τιμές των βαρών του προηγούμενου επιπέδου. Συγκεκριμένα, για k ως το επίπεδο εξόδου, και j το αμέσως προηγούμενο, μπορούμε αρχικά να ορίσουμε την ποσότητα δk, σύμφωνα με τον γενικευμένο κανόνα Δέλτα:

$$\delta_k = (t_k - out_k) * f'(in_j)$$

και να υπολογιστεί η μεταβολή στα βάρη

$$\Delta w_{jk} = \alpha * \delta_k * out_j$$

όπου α ο ρυθμός μάθησης που είναι επιθυμητός. Αντίστοιχα για το κρυφό επίπεδο j:

$$out_j = \sum_{k=1}^{m} (w_{jk} * \delta_k) * f'(in_j)$$

$$\Delta w_{jk} = \alpha * \delta_j * x_i$$

Με τον παραπάνω τρόπο, διαμορφώνονται τα βάρη όλων των κρυφών επιπέδων μέχρι το επίπεδο εισόδου. Αυτή η διαδικασία προσαρμογής των βαρών ονομάζεται πέρασμα (backward pass) ή ανάστροφη μετάδοση (back propagation). Με τον τρόπο αυτό, ο αλγόριθμος της ανάστροφης λάθους ελαχιστοποιεί το μέσο τετραγωνικό σφάλμα μεταξύ της εξόδου του δικτύου και της επιθυμητής εξόδου, για διανύσματα του συνόλου εκπαίδευσης.

Στην ουσία, με την παραπάνω διαδικασία αναζητείται το ολικό ελάχιστο της συνάρτησης σφάλματος. Η διόρθωση που γίνεται κάθε φορά προσπαθεί να κάνει εκείνες τις αλλαγές που θα μειώσουν το σφάλμα τοπικά. Αυτό είναι πιθανό να δημιουργήσει πρόβλημα, αφού υπάρχει ο κίνδυνος να εγκλωβιστεί το δίκτυο σε τοπικά ελάχιστα. Για να αντιμετωπιστεί το συγκεκριμένο πρόβλημα μπορεί να γίνει αρχικοποίηση των βαρών με διαφορετικούς τρόπους μέχρι να βρεθεί ένας ικανοποιητικός.

Ένα ακόμη σημαντικό ζήτημα που προκύπτει ορισμένες φορές είναι το δίκτυο να παραλύσει (network paralysis). Στην συγκεκριμένη περίπτωση ένα ή περισσότερα βάρη μπορεί να αποκτήσει πολύ υψηλή τιμή, με αποτέλεσμα να μην μεταβάλλεται σημαντικά στις επόμενες επαναλήψεις [22]. Το πρόβλημα αυτό μπορεί να επιλυθεί αυξάνοντας τον ρυθμό μάθησης α.

3.2: Αλγόριθμοι Μηχανικής Μάθησης και Ταξινόμηση

Στο πλαίσιο της εργασίας χρησιμοποιήθηκαν διάφορα μοντέλα επιβλεπόμενης μηχανικής μάθησης καθώς και κάποιες παραλλαγές αυτών. Η επιλογή αυτών των μοντέλων έγινε με στόχο την ύπαρξη διαφοροποίησης και ποικιλίας τεχνικών καθώς και βάσει του μεγάλου εύρους εφαρμογών που αυτοί οι αλγόριθμοι έχουν. Παρακάτω παρουσιάζονται οι βασικές υλοποιήσεις των κύριων αλγορίθμων:

3.2.1: k-Nearest Neighbors (k-NN)

Ο αλγόριθμος Κ-Κοντινότεροι Γείτονες (KNN) είναι μια θεμελιώδης και ευέλικτη τεχνική επιβλεπόμενης μάθησης που χρησιμοποιείται ευρέως στη μηχανική μάθηση τόσο για εργασίες ταξινόμησης όσο και παλινδρόμησης. Η διαισθητική του προσέγγιση, η ευκολία υλοποίησης και οι ελάχιστες υποθέσεις σχετικά με την υποκείμενη κατανομή δεδομένων τον καθιστούν δημοφιλή επιλογή για μια ποικιλία εφαρμογών. Θα εξετάσουμε βαθύτερα τις βασικές έννοιες της ταξινόμησης KNN, διερευνώντας τις λειτουργίες, τα δυνατά σημεία και τις εκτιμήσεις για τη βέλτιστη αξιοποίηση της μεθόδου.

Βασικές αρχές

Η ταξινόμηση KNN βασίζεται στην αρχή ότι τα σημεία δεδομένων με παρόμοια χαρακτηριστικά τείνουν να ανήκουν στην ίδια κλάση. Κατά τη φάση της εκπαίδευσης, ο αλγόριθμος αποθηκεύει απλώς ολόκληρο το σύνολο δεδομένων εκπαίδευσης ως αναφορά. Όταν παρουσιάζεται ένα νέο, μη επισημασμένο σημείο δεδομένων, ο αλγόριθμος KNN προσδιορίζει τους k πλησιέστερους γείτονες εντός των δεδομένων εκπαίδευσης με βάση μια επιλεγμένη μετρική απόστασης. Η απόσταση D(X,Y) μπορεί να υπολογιστεί ως η απόσταση Minkowski η οποία ορίζεται από την εξίσωση:

$$D\left(X,Y\right)=\left(\sum_{i=1}^{n}\left|x_i-y_i\right|^p\right)^{\frac{1}{p}}.$$

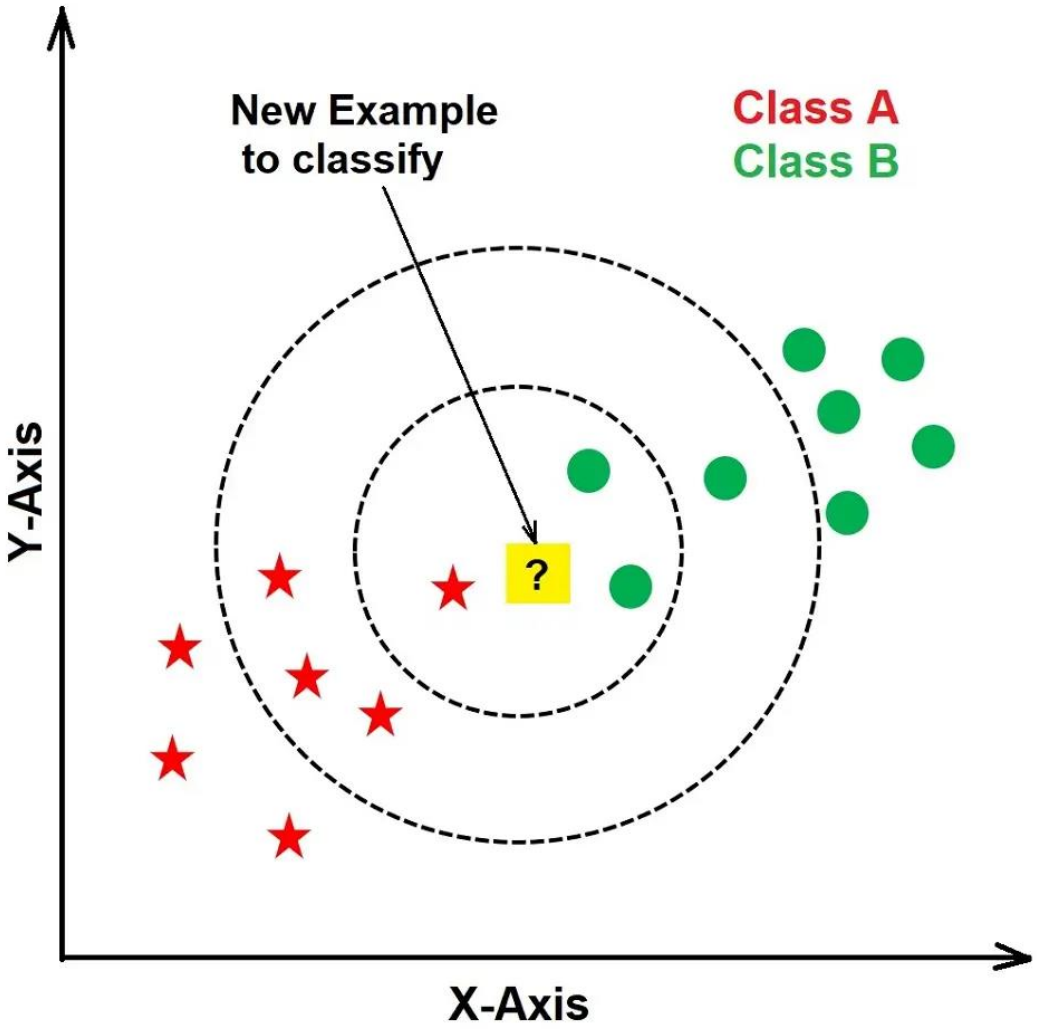
Η απόσταση Minkowski για p = 1 είναι η απόσταση Manhattan και για p = 2 είναι η Ευκλείδεια απόσταση. Στην οριακή περίπτωση που το p φτάνει στο άπειρο, λαμβάνουμε την απόσταση Chebyshev. Η ετικέτα κλάσης του νέου σημείου δεδομένων αποδίδεται στη συνέχεια με βάση

την ψήφο πλειοψηφίας (για ταξινόμηση) ή τη μέση τιμή (για παλινδρόμηση) των k γειτόνων του.

Βασικές επιλογές και χαρακτηριστικά

Αν και η μέθοδος είναι εννοιολογικά απλή, η KNN προσφέρει μια σειρά επιλογών και χαρακτηριστικών που επηρεάζουν την απόδοση και την εφαρμοσιμότητά της:

- Μετρικές απόστασης: Η επιλογή της μετρικής απόστασης επηρεάζει σημαντικά τον τρόπο με τον οποίο το KNN αντιλαμβάνεται την ομοιότητα μεταξύ σημείων δεδομένων. Οι συνήθεις μετρικές περιλαμβάνουν την ευκλείδεια απόσταση, την απόσταση Manhattan και την απόσταση Minkowski, καθεμία από τις οποίες δίνει έμφαση σε διαφορετικές πτυχές του χώρου χαρακτηριστικών. Η επιλογή της κατάλληλης μετρικής εξαρτάται από τα χαρακτηριστικά των δεδομένων.
- Τιμή του K: Η παράμετρος k καθορίζει τον αριθμό των γειτόνων που λαμβάνονται υπόψη για την ταξινόμηση. Μια υψηλή τιμή k μειώνει την ευαισθησία στο θόρυβο αλλά μπορεί να οδηγήσει σε υπερπροσαρμογή, ενώ μια χαμηλή τιμή k προσφέρει πιο λεπτομερείς ταξινομήσεις αλλά είναι ευαίσθητη στο θόρυβο και τις ακραίες τιμές. Η εύρεση του βέλτιστου k συχνά περιλαμβάνει πειραματισμό με τεχνικές διασταυρούμενης επικύρωσης.
- Βαθμισμένη ψηφοφορία: Το KNN μπορεί να ενσωματώσει βάρη στη διαδικασία ψηφοφορίας, αποδίδοντας μεγαλύτερη επιρροή στους πιο κοντινούς γείτονες. Αυτή η προσέγγιση αξιοποιεί τη διαίσθηση ότι οι κοντινότεροι γείτονες έχουν μεγαλύτερη σημασία για τον καθορισμό της ετικέτας κλάσης.
- Κατώφλι απόστασης: Αντί για μια σταθερή τιμή k, το KNN μπορεί να χρησιμοποιήσει κατώφλια απόστασης. Οι γείτονες που εμπίπτουν σε ένα συγκεκριμένο κατώφλι απόστασης λαμβάνονται υπόψη για ψηφοφορία, προσφέροντας μια πιο δυναμική προσέγγιση που προσαρμόζεται στην τοπική πυκνότητα δεδομένων.



Εικόνα 10 - Απεικόνιση K-NN Ταξινόμησης

Πλεονεκτήματα και Περιορισμοί

Παρά την εννοιολογική του απλότητα, το KNN προσφέρει μια λεπτή ισορροπία πλεονεκτημάτων και προβληματισμών. Από τη μία πλευρά, το KNN αποτελεί ένα προσιτό σημείο εκκίνησης για αρχάριους λόγω των εύκολα κατανοητών βασικών αρχών του και της απλής εφαρμογής του. Επιπλέον, ο μη παραμετρικός χαρακτήρας του επιτρέπει τον χειρισμό ποικίλων τύπων δεδομένων χωρίς να απαιτούνται συγκεκριμένες υποθέσεις σχετικά με την υποκείμενη κατανομή των δεδομένων. Αυτό καθιστά το KNN ιδιαίτερα αποτελεσματικό για δεδομένα υψηλών διαστάσεων, όπου άλλοι αλγόριθμοι μπορεί να υποφέρουν από την "κατάρα της διαστατικότητας". Επιπλέον, ο KNN επιδεικνύει κάποια ανθεκτικότητα σε ακραίες τιμές στα δεδομένα σε σύγκριση με ορισμένες μεθόδους ταξινόμησης.

Ωστόσο, ο KNN δεν είναι χωρίς περιορισμούς. Καθώς το μέγεθος των δεδομένων εκπαίδευσης αυξάνεται, ο υπολογισμός των αποστάσεων σε όλα τα σημεία δεδομένων για κάθε νέα πρόβλεψη μπορεί να γίνει υπολογιστικά δαπανηρός. Αυτό, σε συνδυασμό με τις απαιτήσεις μνήμης για την αποθήκευση ολόκληρου του συνόλου δεδομένων εκπαίδευσης, μπορεί να αποτελέσει σημαντικό εμπόδιο για μεγάλα σύνολα δεδομένων. Ενώ το KNN μπορεί να χειριστεί δεδομένα υψηλής διάστασης, η υψηλή διάσταση μπορεί να επηρεάσει αρνητικά την απόδοση λόγω των αυξημένων υπολογισμών αποστάσεων και των δυνητικά αραιών δεδομένων. Τέλος, το KNN μπορεί να είναι ευαίσθητο σε άσχετα χαρακτηριστικά στα δεδομένα. Χαρακτηριστικά που δεν έχουν καμία σχέση με την εργασία ταξινόμησης μπορούν να παραπλανήσουν τον αλγόριθμο, καθιστώντας αναγκαία την επιλογή χαρακτηριστικών ή τεχνικών κανονικοποίησης για βέλτιστη απόδοση.

Ο αλγόριθμος K-Nearest Neighbors προσφέρει ένα ισχυρό και ευέλικτο εργαλείο για εργασίες ταξινόμησης στη μηχανική μάθηση. Η ευκολία χρήσης, οι ελάχιστες παραδοχές και η αποτελεσματικότητά του με διάφορους τύπους δεδομένων τον καθιστούν πολύτιμη προσθήκη στην εργαλειοθήκη της μηχανικής μάθησης. Ωστόσο, η προσεκτική εξέταση παραγόντων όπως οι μετρικές απόστασης, η τιμή του k και το υπολογιστικό κόστος είναι ζωτικής σημασίας για την επιτυχή εφαρμογή του. Με την κατανόηση των δυνατών σημείων και των περιορισμών της, οι επιστήμονες δεδομένων μπορούν να αξιοποιήσουν το KNN για να επιτύχουν ισχυρά και αποτελεσματικά αποτελέσματα ταξινόμησης σε ένα ευρύ φάσμα εφαρμογών.

3.2.2: Support Vector Machines (SVM)

Οι Μηχανές Διανυσμάτων Υποστήριξης (SVM) είναι ένας ισχυρός και ευέλικτος αλγόριθμος επιβλεπόμενης μάθησης που χρησιμοποιείται ευρέως στη μηχανική μάθηση για εργασίες ταξινόμησης. Αυτό το κεφάλαιο παρέχει μια ολοκληρωμένη επισκόπηση των SVMs, εμβαθύνοντας στις θεωρητικές τους βάσεις, στις βασικές τους έννοιες και στις πρακτικές τους υλοποιήσεις. Εξετάζουμε τη βασική αρχή της εύρεσης του βέλτιστου υπερεπιπέδου που μεγιστοποιεί το περιθώριο μεταξύ των κλάσεων σε έναν χώρο υψηλών διαστάσεων. Το κεφάλαιο αυτό συζητά επίσης διάφορες επιλογές, παραμέτρους και χαρακτηριστικά SVM,

συμπεριλαμβανομένων συναρτήσεων πυρήνα για το χειρισμό μη γραμμικών δεδομένων και τεχνικών για την αντιμετώπιση προβλημάτων ταξινόμησης πολλαπλών κλάσεων.

Η ταξινόμηση είναι μια θεμελιώδης εργασία στη μηχανική μάθηση, με στόχο την ανάθεση νέων σημείων δεδομένων σε προκαθορισμένες κατηγορίες. Οι SVM, που εισήχθησαν από τους Vapnik et al. (1995) [23], είναι ένας εξέχων αλγόριθμος ταξινόμησης γνωστός για την αποτελεσματικότητά του σε χώρους υψηλών διαστάσεων και με περιορισμένα δεδομένα εκπαίδευσης. Αυτό το κεφάλαιο εμβαθύνει στα θεωρητικά θεμέλια και τις πρακτικές εφαρμογές των SVM για εργασίες ταξινόμησης.

Θεωρητικό Υπόβαθρο και Βασικές Έννοιες

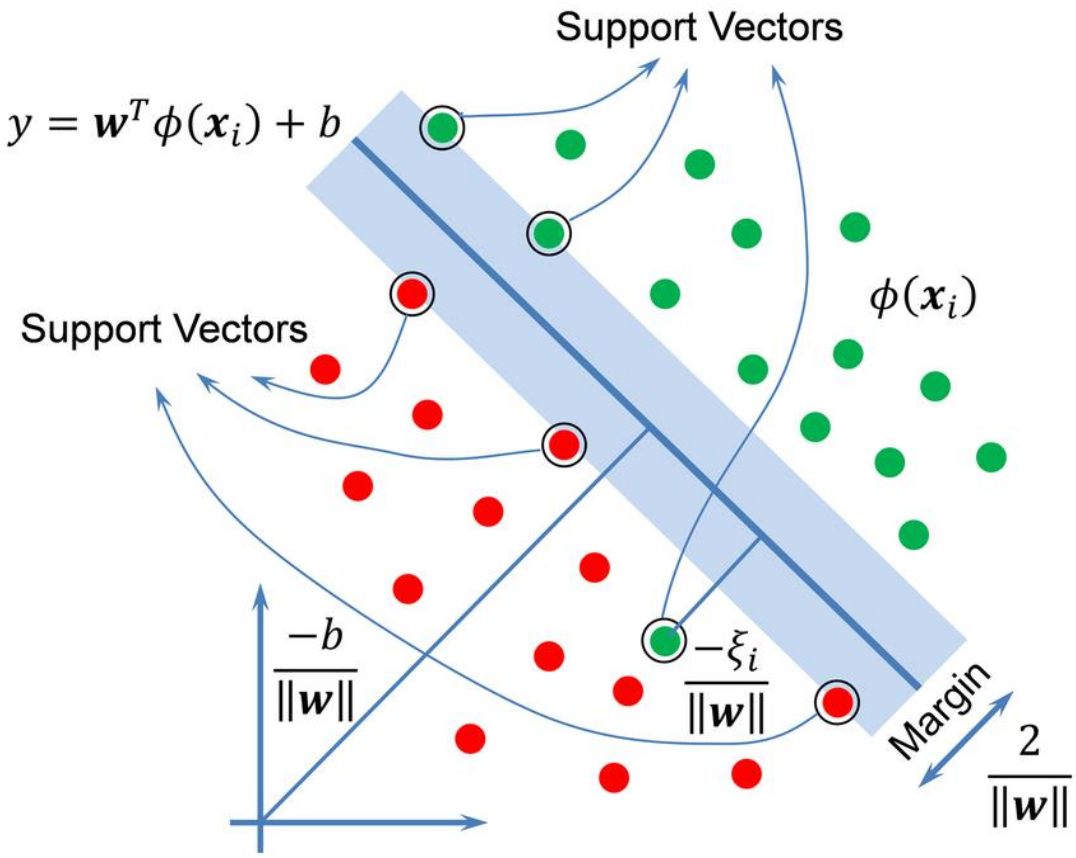
Οι SVM έχουν τις ρίζες τους στη θεωρία της στατιστικής μάθησης [23]. Λειτουργούν με την εύρεση ενός υπερεπιπέδου σε έναν χώρο χαρακτηριστικών υψηλής διάστασης που διαχωρίζει καλύτερα τα σημεία δεδομένων που ανήκουν σε διαφορετικές κλάσεις. Αυτό το υπερεπίπεδο μεγιστοποιεί το περιθώριο, το οποίο είναι η απόσταση μεταξύ του υπερεπιπέδου και των πλησιέστερων σημείων δεδομένων από κάθε κλάση, γνωστά και ως διανύσματα υποστήριξης.

Παρακάτω παρουσιάζονται κάποιες βασικές έννοιες σχετικά με τον αλγόριθμο SVM:

- **Υπερεπίπεδο** (**Hyperplane**): Ένα όριο απόφασης σε έναν χώρο υψηλών διαστάσεων που διαχωρίζει τα σημεία δεδομένων διαφορετικών κλάσεων.
- **Περιθώριο** (**Margin**): Η απόσταση μεταξύ του υπερεπιπέδου και των πλησιέστερων διανυσμάτων υποστήριξης από κάθε κλάση.
- **Διανύσματα υποστήριξης** (**Support Vectors**): Σημεία δεδομένων που καθορίζουν το περιθώριο και επηρεάζουν τη θέση του υπερεπιπέδου.
- **Συναρτήσεις πυρήνα** (**Kernel Functions**): Μαθηματικοί μετασχηματισμοί που αντιστοιχίζουν τα σημεία δεδομένων σε έναν χώρο υψηλότερων διαστάσεων, επιτρέποντας τον γραμμικό διαχωρισμό μη γραμμικά διαχωρίσιμων δεδομένων στον αρχικό χώρο.

Οι SVM έχουν σχεδιαστεί κυρίως για προβλήματα δυαδικής ταξινόμησης. Ωστόσο, διάφορες τεχνικές μπορούν να εφαρμοστούν για εργασίες ταξινόμησης πολλαπλών κλάσεων:

- **One-vs-Rest**: Εκπαίδευση πολλαπλών SVMs, καθένα από τα οποία ταξινομεί μία κλάση έναντι όλων των άλλων.
- **One-vs-One**: Εκπαίδευση πολλαπλών SVMs, καθένα από τα οποία ταξινομεί κάθε μοναδικό ζεύγος κλάσεων.



Εικόνα 11 - Απεικόνιση Δυαδικής Ταξινόμησης SVM από [61]

Επιλογές και παράμετροι SVM

Στην συνέχεια παρουσιάζονται κάποια χαρακτηριστικές παράμετροι του αλγορίθμου:

- **Τύποι πυρήνων** (**Kernel Types**): Μπορούν να χρησιμοποιηθούν διαφορετικές συναρτήσεις πυρήνα, καθεμία από τις οποίες είναι κατάλληλη για συγκεκριμένα χαρακτηριστικά δεδομένων. Οι συνήθεις επιλογές περιλαμβάνουν γραμμικούς πυρήνες για γραμμικά διαχωρίσιμα δεδομένα, μη γραμμικούς πυρήνες (π.χ. RBF, πολυώνυμα) για σύνθετα δεδομένα και πυρήνες συμβολοσειρών για δεδομένα κειμένου.
- **Παράμετρος κόστους** (**C**): Αυτή η παράμετρος ελέγχει το συμβιβασμό μεταξύ της μεγιστοποίησης του περιθωρίου και της αποδοχής των λανθασμένων ταξινομημένων σημείων. Ένα υψηλότερο C δίνει προτεραιότητα σε ένα μεγαλύτερο περιθώριο αλλά

μπορεί να οδηγήσει σε υπερπροσαρμογή, ενώ ένα χαμηλότερο C επιτρέπει περισσότερα σφάλματα αλλά μπορεί να οδηγήσει σε ένα λιγότερο ανθεκτικό μοντέλο.

- **Έψιλον** (*ε*): Αυτή η παράμετρος επιτρέπει μια ορισμένη ανοχή για εσφαλμένα ταξινομημένα σημεία κατά την εκπαίδευση. Ένα μικρότερο ***ε*** επιβάλλει αυστηρότερο διαχωρισμό αλλά μπορεί να είναι λιγότερο ευέλικτο με θορυβώδη δεδομένα.

Οι SVM προσφέρουν πολλά πλεονεκτήματα για εργασίες ταξινόμησης. Είναι ιδιαίτερα αποτελεσματικές για δεδομένα υψηλών διαστάσεων, δηλαδή μπορούν να χειριστούν μεγάλο αριθμό χαρακτηριστικών χωρίς σημαντική υποβάθμιση της απόδοσης. Αυτό τις καθιστά κατάλληλες για πολύπλοκα σύνολα δεδομένων. Επιπλέον, οι SVM τείνουν να δημιουργούν μοντέλα που γενικεύονται καλά σε αόρατα δεδομένα λόγω της εστίασής τους στη μεγιστοποίηση του περιθωρίου μεταξύ των κλάσεων. Αυτό μειώνει τον κίνδυνο υπερπροσαρμογής των δεδομένων εκπαίδευσης. Επιπλέον, οι SVM παρουσιάζουν κάποια ανθεκτικότητα στο θόρυβο των δεδομένων, λαμβάνοντας υπόψη κυρίως τα διανύσματα υποστήριξης που καθορίζουν το όριο απόφασης.

Ωστόσο, υπάρχουν επίσης περιορισμοί που πρέπει να ληφθούν υπόψη κατά τη χρήση των SVM. Η εκπαίδευση ενός μοντέλου SVM μπορεί να είναι υπολογιστικά δαπανηρή, ειδικά για μεγάλα και πολυδιάστατα σύνολα δεδομένων. Επιπλέον, η επιλογή της σωστής συνάρτησης πυρήνα μπορεί να αποτελέσει πρόκληση. Η επιλογή πυρήνα συχνά απαιτεί γνώση του τομέα ή πειραματισμό για να βρεθεί αυτός που ταιριάζει καλύτερα στα δεδομένα. Τέλος, αν και αποτελεσματικοί, οι SVM είναι λιγότερο ερμηνεύσιμοι σε σύγκριση με ορισμένους άλλους αλγορίθμους ταξινόμησης. Αυτή η φύση του "μαύρου κουτιού" μπορεί να καταστήσει δύσκολη την κατανόηση της λογικής πίσω από συγκεκριμένες προβλέψεις που γίνονται από το μοντέλο.

Οι SVM παραμένουν ένας αλγόριθμος ακρογωνιαίος λίθος στις εργασίες ταξινόμησης της μηχανικής μάθησης. Η ικανότητά τους να χειρίζονται δεδομένα υψηλών διαστάσεων, να αποδίδουν καλά με περιορισμένα δεδομένα και να παρέχουν ισχυρές θεωρητικές εγγυήσεις τους καθιστούν πολύτιμο εργαλείο. Ωστόσο, το υπολογιστικό τους κόστος και οι προκλήσεις του συντονισμού των παραμέτρων απαιτούν προσεκτική εξέταση σε πρακτικές εφαρμογές.

3.2.3: Λογιστική Παλινδρόμηση (Logistic Regression, LR)

Η λογιστική παλινδρόμηση είναι ένας θεμελιώδης εποπτευόμενος αλγόριθμος μάθησης που χρησιμοποιείται ευρέως στη μηχανική μάθηση για εργασίες ταξινόμησης. Σε αντίθεση με τη γραμμική παλινδρόμηση, η οποία προβλέπει συνεχή αποτελέσματα, η λογιστική παλινδρόμηση εκτιμά την πιθανότητα μιας παρατήρησης να ανήκει σε μια συγκεκριμένη κατηγορία. Αυτό το κεφάλαιο εμβαθύνει στις βασικές έννοιες, τις εφαρμογές και τις παραλλαγές της λογιστικής παλινδρόμησης για ταξινόμηση.

Βασικές Έννοιες

Η λογιστική παλινδρόμηση βασίζεται στις αρχές της γραμμικής παλινδρόμησης αλλά εισάγει τη λογιστική συνάρτηση για τη μετατροπή της γραμμικής εξόδου σε πιθανότητα μεταξύ 0 και 1.

Το μοντέλο μπορεί να διατυπωθεί ως εξής:

- Έστω y η δυαδική εξαρτημένη μεταβλητή (π.χ. 0 ή 1) που αντιπροσωπεύει τις ετικέτες κλάσεων.
- Έστω x ένα διάνυσμα ανεξάρτητων χαρακτηριστικών (προγνωστικών).
- Το γραμμικό μοντέλο προβλέπει μια βαθμολογία που συμβολίζεται με z:

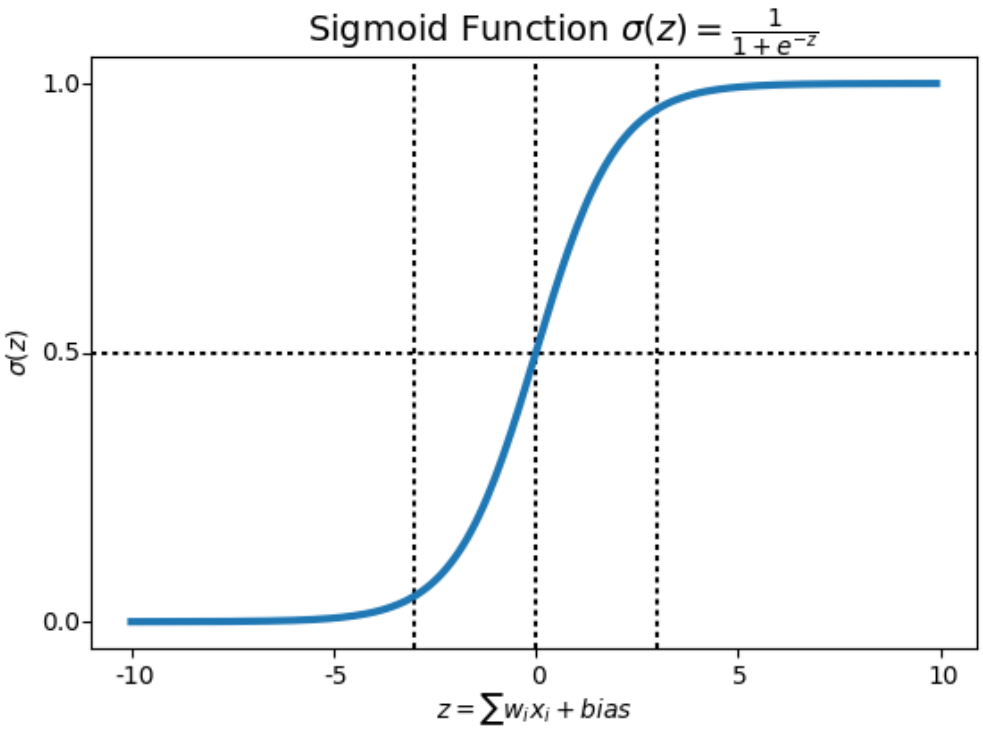
$$z = w^T * x + b$$

όπου w είναι το διάνυσμα βάρους, b είναι ο όρος πόλωσης και το ^{T} αντιπροσωπεύει τη λειτουργία μεταφοράς.

Η λογιστική συνάρτηση, που συμβολίζεται με σ(z), μετατρέπει το z σε μια εκτίμηση πιθανότητας μεταξύ 0 και 1:

$$P(y = 1 \mid x) = \sigma(z) = \frac{1}{1 + e^{-z}}$$

Αυτή η συνάρτηση, γνωστή και ως σιγμοειδής συνάρτηση, διασφαλίζει ότι η έξοδος βρίσκεται εντός του επιθυμητού εύρους πιθανότητας.



Εικόνα 12 - Ταξινόμηση με Λογιστική Παλινδρόμηση

Εκπαίδευση και Βελτιστοποίηση Μοντέλων

Η λογιστική παλινδρόμηση χρησιμοποιεί μια συνάρτηση κόστους για την αξιολόγηση της απόδοσης του μοντέλου και την καθοδήγηση της διαδικασίας βελτιστοποίησης. Μια κοινή επιλογή είναι η δυαδική συνάρτηση απώλειας διασταυρούμενης εντροπίας, η οποία μετρά τη διαφορά μεταξύ των προβλεπόμενων πιθανοτήτων και των πραγματικών ετικετών:

$$\mathrm{L}oss = -(y * log(\sigma(\mathrm{z})) + (1 - y) * log(1 - \sigma(\mathrm{z}))\,)$$

Ο στόχος είναι να ελαχιστοποιηθεί η συνάρτηση απώλειας προσαρμόζοντας το διάνυσμα βάρους w και τον όρο πόλωσης b. Αυτή η βελτιστοποίηση συνήθως επιτυγχάνεται με τη χρήση επαναληπτικών αλγορίθμων όπως η gradient descent, η οποία ενημερώνει επαναληπτικά τα βάρη προς την κατεύθυνση που ελαχιστοποιεί την απώλεια. [24]

Παραλλαγές της Λογιστικής Παλινδρόμησης

Η λογιστική παλινδρόμηση προσφέρει διάφορες παραλλαγές για την αντιμετώπιση διαφορετικών σεναρίων ταξινόμησης:

- **Πολυωνυμική Λογιστική Παλινδρόμηση**: Επεκτείνει το δυαδικό μοντέλο για να χειριστεί περιπτώσεις με περισσότερες από δύο ετικέτες κλάσεων. Αυτό επιτυγχάνεται με την προσαρμογή χωριστών μοντέλων λογιστικής παλινδρόμησης για κάθε κατηγορία έναντι όλων των άλλων κλάσεων.
- **Τακτική Λογιστική Παλινδρόμηση**: Χρησιμοποιείται για εργασίες ταξινόμησης όπου οι ετικέτες κλάσεων έχουν μια φυσική σειρά (π.χ. επίπεδα ικανοποίησης πελατών). Αυτό το μοντέλο χρησιμοποιεί μια συγκεκριμένη συνάρτηση σύνδεσης για να μοντελοποιήσει τη σχέση μεταξύ των ταξινομημένων κλάσεων και του γραμμικού προγνωστικού.
- **Συρρικνούμενη Παλινδρόμηση (Ridge, LASSO)**: Τακτοποιημένες εκδόσεις της λογιστικής παλινδρόμησης που εισάγουν ποινές στο μέγεθος του διανύσματος βάρους για τη μείωση της πολυπλοκότητας του μοντέλου και την αποφυγή υπερβολικής προσαρμογής.

Η επιλογή της παραλλαγής εξαρτάται από το συγκεκριμένο πρόβλημα και τα χαρακτηριστικά των δεδομένων.

Παράμετροι και χαρακτηριστικά

Η λογιστική παλινδρόμηση απαιτεί προσεκτική εξέταση των παραμέτρων και των χαρακτηριστικών για βέλτιστη απόδοση:

- **Ρυθμός εκμάθησης**: Ελέγχει το μέγεθος του βήματος κατά τη βελτιστοποίηση κατάβασης με κλίση. Ένας μικρός ρυθμός μάθησης διασφαλίζει τη σύγκλιση, αλλά μπορεί να είναι αργός, ενώ ένας μεγάλος ρυθμός μπορεί να οδηγήσει σε αστάθεια.
- **Παράμετρος Regularization (lambda)**: Ελέγχει την ισχύ της κανονικοποίησης στην παλινδρόμηση κορυφογραμμής και λάσο. Μια υψηλότερη τιμή λάμδα αυξάνει την ποινή στο διάνυσμα βάρους, μειώνοντας την πολυπλοκότητα του μοντέλου.
- **Επιλογή λειτουργιών**: Η επιλογή σχετικών και ενημερωτικών χαρακτηριστικών είναι ζωτικής σημασίας. Τεχνικές όπως η ανάλυση συσχέτισης, τα τεστ chi-square και οι βαθμολογίες σπουδαιότητας χαρακτηριστικών μπορούν να βοηθήσουν στον εντοπισμό

σημαντικών χαρακτηριστικών που συμβάλλουν περισσότερο στην απόδοση του μοντέλου.

- **Κλιμάκωση χαρακτηριστικών**: Η τυποποίηση χαρακτηριστικών σε κοινή κλίμακα μπορεί να βελτιώσει τη διαδικασία βελτιστοποίησης και τη σύγκλιση μοντέλων.

Η λογιστική παλινδρόμηση παραμένει ένας αλγόριθμος ακρογωνιαίος λίθος στις εργασίες ταξινόμησης λόγω της ερμηνευσιμότητας, της υπολογιστικής αποτελεσματικότητας και της στέρεας θεωρητικής βάσης. Η ευελιξία του επιτρέπει τον χειρισμό προβλημάτων δυαδικών και πολλαπλών κατηγοριών, καθιστώντας το πολύτιμο εργαλείο σε διάφορους τομείς. Κατανοώντας τις βασικές έννοιες, τις παραλλαγές, τις παραμέτρους και τα χαρακτηριστικά γνωρίσματα, οι επαγγελματίες μπορούν να αξιοποιήσουν αποτελεσματικά την λογιστική παλινδρόμηση για τη δημιουργία ισχυρών μοντέλων ταξινόμησης.

3.2.4 Random Forest (RF)

Τα Random Forests (RF) είναι μια ισχυρή τεχνική εκμάθησης συνόλου (**ensemble learning**) για εργασίες ταξινόμησης στη μηχανική μάθηση. Αξιοποιούν τα πλεονεκτήματα πολλών δέντρων αποφάσεων για να δημιουργήσουν ένα πιο ισχυρό και ακριβές μοντέλο σε σύγκριση με ένα μόνο δέντρο αποφάσεων. Αυτό το κεφάλαιο διερευνά τις θεωρητικές βάσεις, τις λεπτομέρειες εφαρμογής και τις πρακτικές εκτιμήσεις των Τυχαίων Δασών για προβλήματα ταξινόμησης.

Η μηχανική μάθηση μπορεί να κατηγοριοποιηθεί ευρέως σε εποπτευόμενη και μη εποπτευόμενη μάθηση. Η εποπτευόμενη μάθηση ασχολείται με εργασίες όπου ένα μοντέλο μαθαίνει από δεδομένα με ετικέτα για να κάνει προβλέψεις για αόρατα δεδομένα. Η ταξινόμηση, ένας υποτομέας της εποπτευόμενης μάθησης, στοχεύει στην κατηγοριοποίηση των σημείων δεδομένων σε προκαθορισμένες τάξεις.

Οι μέθοδοι συνόλου (**ensemble methods**) συνδυάζουν τις προβλέψεις πολλών αδύναμων μοντέλων για να δημιουργήσουν έναν μόνο, πιο δυνατό μαθητή. Τα Random Forests αξιοποιούν το χαρακτηριστικό των δέντρων αποφάσεων το οποίο είναι η αδυναμία μάθησης, αντιμετωπίζοντας έτσι τους περιορισμούς των μεμονωμένων δέντρων αποφάσεων, όπως η υπερπροσαρμογή (**overfitting**).

Βασικές Έννοιες

Το Random Forest είναι ένα σύνολο δέντρων αποφάσεων, το καθένα χτισμένο σε ένα τυχαίο υποσύνολο χαρακτηριστικών και σημείων δεδομένων από το σύνολο εκπαίδευσης. Όταν ένα νέο σημείο δεδομένων φτάνει για πρόβλεψη, περνά από κάθε δέντρο απόφασης στο δάσος. Η τελική πρόβλεψη της κατηγορίας καθορίζεται από την πλειοψηφία (για ταξινόμηση) ή τη μέση πρόβλεψη (για παλινδρόμηση) των μεμονωμένων δέντρων.

Υπάρχουν δύο βασικές πτυχές που εισάγουν την τυχαιότητα στον αλγόριθμο Random Forest:

- Συγκέντρωση Bootstrap (Bagging): Κατά τη δημιουργία του δάσους, κάθε δέντρο εκπαιδεύεται σε ένα τυχαίο υποσύνολο σημείων δεδομένων που σχεδιάζονται με αντικατάσταση από το αρχικό σετ εκπαίδευσης (μια τεχνική που ονομάζεται bagging). Αυτό διασφαλίζει ότι κάθε δέντρο μαθαίνει από μια ελαφρώς διαφορετική κατανομή δεδομένων, μειώνοντας τη διακύμανση του συνολικού μοντέλου.
- Τυχαία επιλογή χαρακτηριστικών: Σε κάθε κόμβο ενός δέντρου αποφάσεων, αντί να λαμβάνονται υπόψη όλα τα χαρακτηριστικά για διαχωρισμό, επιλέγεται ένα τυχαίο υποσύνολο χαρακτηριστικών. Ο καλύτερος διαχωρισμός μεταξύ αυτών των χαρακτηριστικών χρησιμοποιείται στη συνέχεια για την κατάτμηση των δεδομένων. Αυτή η διαδικασία βοηθά στην αποσυσχέτιση των δέντρων, εμποδίζοντάς τα να γίνουν υπερβολικά παρόμοια και επιρρεπή σε υπερβολική προσαρμογή.

Αλγόριθμος: Δημιουργία Τυχαίου Δάσους

Ακολουθεί μια ανάλυση του αλγόριθμου ταξινόμησης Random Forest:

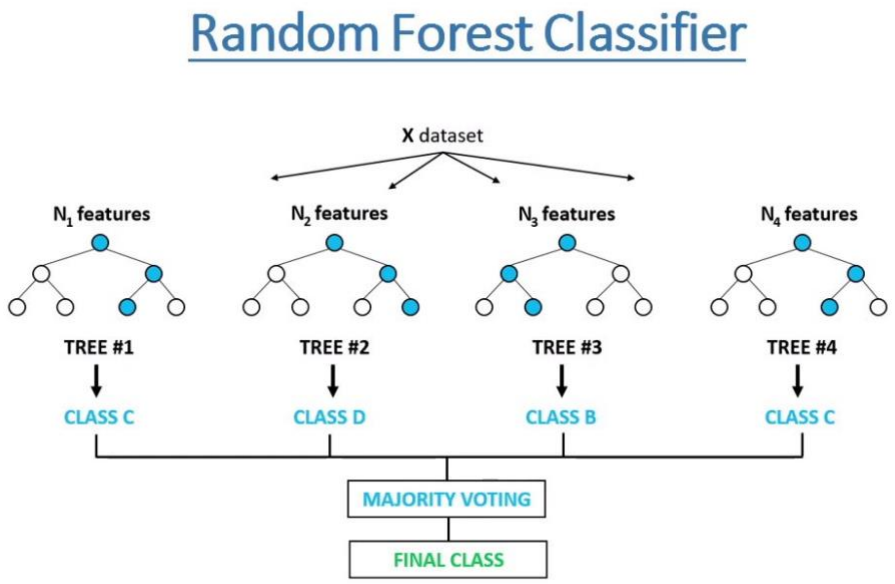
- Ορισμός παραμέτρων: Καθορίζουμε τον αριθμό των δέντρων (`**n_estimators**`) στο δάσος και τον αριθμό των χαρακτηριστικών (`**max_features**`) που λαμβάνονται υπόψη σε κάθε διαχωρισμό των δέντρων απόφασης.

- Συγκέντρωση Bootstrap: Σχεδιάζουμε τα δείγματα εκκίνησης `**n_estimators**` (μέγεθος ίσο με τα αρχικά δεδομένα) από το σετ εκπαίδευσης με αντικατάσταση.

- Δημιουργία Forest: Για κάθε δείγμα bootstrap:
    - Δημιουργία ενός δέντρου αποφάσεων χρησιμοποιώντας τα επιλεγμένα χαρακτηριστικά (**max_features**) σε κάθε σημείο διαχωρισμού.

- Επιλογή του καλύτερου διαχωρισμού με βάση ένα κριτήριο διαχωρισμού όπως η πρόσμειξη Gini (για ταξινόμηση) ή η μείωση διακύμανσης (για παλινδρόμηση).

- Πρόβλεψη: Όταν φτάσει ένα νέο σημείο δεδομένων:
  - Το εισάγουμε στο μοντέλο, δηλαδή σε κάθε δέντρο στο δάσος.
  - Για ταξινόμηση, συγκεντρώνουμε τις προβλέψεις της τάξης από κάθε δέντρο κατά πλειοψηφία.



Εικόνα 13 Random Forest Classifier [62]

Επιλογές και παράμετροι για τυχαία ταξινόμηση δασών

Αρκετές βασικές παράμετροι επηρεάζουν την απόδοση των τυχαίων δασών:

- Αριθμός δέντρων (**n_estimators**): Περισσότερα δέντρα γενικά οδηγούν σε καλύτερη απόδοση αλλά και αυξάνουν τον χρόνο προπόνησης.

- Αριθμός δυνατοτήτων σε κάθε διαχωρισμό (**max_features**): Ένα μικρότερο υποσύνολο χαρακτηριστικών σε κάθε διαχωρισμό ενθαρρύνει την ποικιλομορφία μεταξύ των δέντρων, αλλά ένας πολύ μικρός αριθμός μπορεί να περιορίσει την ικανότητα του μοντέλου να μαθαίνει πολύπλοκες σχέσεις.
- Ελάχιστα δείγματα ανά διαίρεση (**min_samples_split**): Αυτή η παράμετρος καθορίζει τον ελάχιστο αριθμό σημείων δεδομένων που απαιτούνται σε έναν κόμβο για να μπορέσει να διαχωριστεί περαιτέρω. Μια υψηλότερη τιμή αποτρέπει την υπερβολική τοποθέτηση, αλλά μπορεί να μειώσει την ευελιξία του μοντέλου.
- Ελάχιστα δείγματα ανά φύλλο (**min_samples_leaf**): Αυτή η παράμετρος ορίζει τον ελάχιστο αριθμό σημείων δεδομένων που επιτρέπεται σε έναν τερματικό κόμβο φύλλου. Παρόμοια με το «min_samples_split», επηρεάζει την πολυπλοκότητα και την υπερπροσαρμογή του μοντέλου.

Πλεονεκτήματα:

- Υψηλή Ακρίβεια: Τα τυχαία δάση συχνά επιτυγχάνουν υψηλή ακρίβεια στις εργασίες ταξινόμησης λόγω της φύσης του συνόλου και της ικανότητάς τους να χειρίζονται περίπλοκες σχέσεις.
- Στιβαρή για υπερπροσαρμογή: Η εισαγωγή της τυχαιότητας μέσω του bagging και της επιλογής χαρακτηριστικών συμβάλλει στη μείωση της υπερπροσαρμογής σε σύγκριση με τα δέντρα μεμονωμένων αποφάσεων.
- Δυνατότητα χειρισμού δεδομένων που λείπουν: Τα τυχαία δάση μπορούν εγγενώς να χειριστούν δεδομένα που λείπουν μέσα στα δέντρα λαμβάνοντας υπόψη μόνο τις διαθέσιμες λειτουργίες σε κάθε διαχωρισμό.
- Ερμηνευσιμότητα: Οι βαθμολογίες σπουδαιότητας χαρακτηριστικών μπορούν να εξαχθούν από τα Random Forests, παρέχοντας πληροφορίες για το ποια χαρακτηριστικά συμβάλλουν περισσότερο στις προβλέψεις του μοντέλου.

Περιορισμοί

- Black Box Nature: Ενώ η σημασία των χαρακτηριστικών προσφέρει κάποια ερμηνευτικότητα, η εσωτερική λειτουργία των μεμονωμένων δέντρων μέσα στο δάσος παραμένει αδιαφανής.
- Υπολογιστικά ακριβό: Η εκπαίδευση τυχαίων δασών μπορεί να είναι υπολογιστικά ακριβή, ειδικά για μεγάλα σύνολα δεδομένων με πολλά δέντρα. Συντονισμός

υπερπαραμέτρων: Η εύρεση της βέλτιστης διαμόρφωσης υπερπαραμέτρων για ένα μοντέλο Random Forest μπορεί να είναι δύσκολη. συμπέρασμα

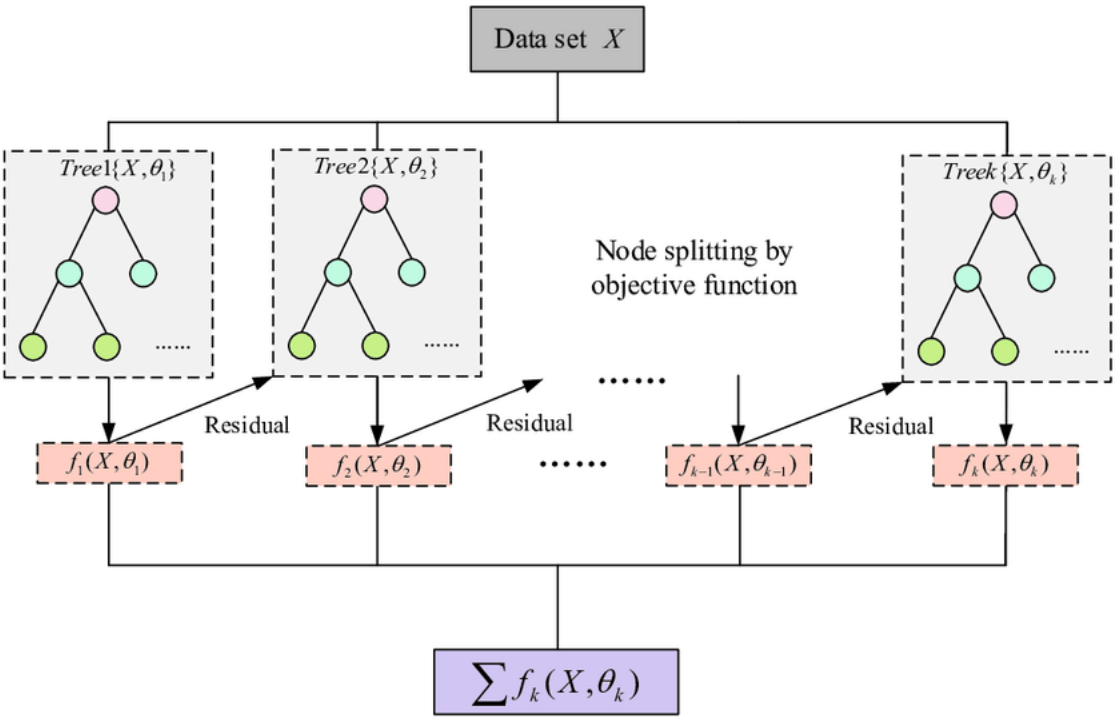
Τα Random Forests είναι ένα ισχυρό και ευέλικτο εργαλείο για εργασίες ταξινόμησης στη μηχανική μάθηση. Η συνολική φύση τους, η στιβαρότητα στην υπερβολική εφαρμογή και η ικανότητα χειρισμού

3.2.5 XGBoost (XGB)

Ο αλγόριθμος XGBoost, (συντομογραφία για το eXtreme Gradient Boosting), έχει γίνει μια κυρίαρχη δύναμη στο τοπίο της μηχανικής μάθησης, ιδιαίτερα για εργασίες ταξινόμησης. Η ικανότητά του να χειρίζεται πολύπλοκα σύνολα δεδομένων, να επιτυγχάνει υψηλή ακρίβεια και να προσφέρει αποτελεσματική εκπαίδευση το καθιστά επιλογή για τους επιστήμονες δεδομένων σε διάφορους τομείς. Αυτό το κεφάλαιο εμβαθύνει στις εσωτερικές λειτουργίες του XGBoost για ταξινόμηση, εξερευνώντας τις βασικές αρχές, τις υπερπαραμέτρους και τις λειτουργίες του.

Gradient Boosting και Ensemble Learning

Ο αλγόριθμος XGBoost εμπίπτει στην ομπρέλα της ενσωματωμένης μάθησης, ένα παράδειγμα που συνδυάζει τις προβλέψεις πολλαπλών αδύναμων μαθητών για τη δημιουργία ενός ισχυρού τελικού μοντέλου. Το Gradient boosting, μια συγκεκριμένη τεχνική ensemble, δημιουργεί επαναληπτικά μια ακολουθία μοντέλων, καθένα από τα οποία επικεντρώνεται στη διόρθωση των σφαλμάτων των προηγούμενων. Αυτή η διαδοχική βελτίωση οδηγεί σε ένα πιο ακριβές σύνολο.



Εικόνα 14 - Διάγραμμα ροής αλγορίθμου XGBoost

Οι Friedman, Hastie και Tibshirani [25] παρέχουν μια ολοκληρωμένη θεωρητική βάση για την ενίσχυση κλίσης. Παρουσιάζουν πώς ο αλγόριθμος αξιοποιεί μια συνάρτηση απώλειας για την αξιολόγηση της απόκλισης μεταξύ των προβλέψεων του μοντέλου και των πραγματικών ετικετών. Κάθε νέο μοντέλο στοχεύει στην ελαχιστοποίηση αυτής της απώλειας εστιάζοντας σε σημεία δεδομένων όπου τα προηγούμενα μοντέλα απέτυχαν.

Ο αλγόριθμος XGBoost βασίζεται στις βασικές αρχές του gradient boosting και εισάγει αρκετές βελτιώσεις που ανεβάζουν σημαντικά την απόδοσή του. Οι Chen και Guestrin (2016) [26] περιγράφουν λεπτομερώς αυτές τις βελτιώσεις, μεταξύ άλλων:

- **Ρύθμιση**: Ο XGBoost ενσωματώνει τεχνικές κανονικοποίησης L1 και L2 για την αποφυγή της υπερπροσαρμογής. Αυτές οι τεχνικές τιμωρούν τα πολύπλοκα μοντέλα, ευνοώντας τα απλούστερα που γενικεύουν καλύτερα σε αόρατα δεδομένα.

- **Αραιές αλληλεπιδράσεις**: Ο XGBoost χειρίζεται αποτελεσματικά τα αραιά δεδομένα, όπου πολλά χαρακτηριστικά έχουν ελλιπείς τιμές. Αυτό το καθιστά κατάλληλο για σύνολα δεδομένων του πραγματικού κόσμου που συχνά περιέχουν τέτοια αραιότητα.
- **Συρρίκνωση (ρυθμός μάθησης)**: Ο XGBoost χρησιμοποιεί μια παράμετρο συρρίκνωσης, γνωστή και ως ρυθμός μάθησης, η οποία ελέγχει την επίδραση κάθε νέου μοντέλου στο σύνολο. Ένας χαμηλότερος ρυθμός μάθησης μειώνει το μέγεθος του βήματος που λαμβάνεται σε κάθε επανάληψη, οδηγώντας σε ένα πιο συντηρητικό και ανθεκτικό μοντέλο.
- **Παραλληλοποίηση**: Ο XGBoost αξιοποιεί την παράλληλη επεξεργασία για να επιταχύνει την εκπαίδευση σε συστήματα πολλαπλών πυρήνων, καθιστώντας το αποτελεσματικό για το χειρισμό μεγάλων συνόλων δεδομένων.

XGBoost για ταξινόμηση: Υπερπαράμετροι και επιλογές

Ο αλγόριθμος XGBoost προσφέρει ένα πλούσιο σύνολο υπερπαραμέτρων και λειτουργιών που δίνουν τη δυνατότητα στους χρήστες να ρυθμίσουν το μοντέλο για βέλτιστη απόδοση στη συγκεκριμένη εργασία ταξινόμησης. Ακολουθεί μια επισκόπηση ορισμένων βασικών πτυχών:

- **Παράμετροι επιπέδου δέντρου**:
    - **max_depth**: Ελέγχει το μέγιστο βάθος κάθε δέντρου απόφασης στο σύνολο. Τα βαθύτερα δέντρα μπορούν να συλλάβουν πολύπλοκες σχέσεις, αλλά μπορεί επίσης να οδηγήσουν σε υπερπροσαρμογή.
    - **eta (ρυθμός μάθησης)**: Όπως αναφέρθηκε προηγουμένως, αυτή η παράμετρος ρυθμίζει το μέγεθος του βήματος που λαμβάνει κάθε δέντρο στη διαδικασία ενίσχυσης.
    - **gamma**: Ελάχιστη μείωση των απωλειών που απαιτείται για μια διάσπαση στο δέντρο απόφασης. Υψηλότερες τιμές αποτρέπουν την υπερπροσαρμογή αποφεύγοντας τις περιττές διασπάσεις.

- **Παράμετροι εξομάλυνσης:**
    - **lambda**: Παράμετρος κανονικοποίησης L2 που τιμωρεί την πολυπλοκότητα του μοντέλου.
    - **alpha**: Παράμετρος κανονικοποίησης L1 που προωθεί τη σπανιότητα συρρικνώνοντας τα βάρη των χαρακτηριστικών προς το μηδέν.

- **Γενικές παράμετροι:**
    - **n_estimators**: Αριθμός δέντρων που θα συμπεριληφθούν στο τελικό σύνολο. Περισσότερα δέντρα οδηγούν γενικά σε καλύτερη ακρίβεια, αλλά μπορεί επίσης να αυξήσουν το χρόνο εκπαίδευσης.
    - **subsample**: Αναλογία των δειγμάτων δεδομένων που χρησιμοποιούνται για την εκπαίδευση κάθε δέντρου. Αυτό εισάγει τυχαιότητα και συμβάλλει στην αποφυγή υπερβολικής προσαρμογής.

- **Προηγμένες επιλογές**: Ο XGBoost παρέχει μια πληθώρα προηγμένων επιλογών για έμπειρους χρήστες. Αυτές περιλαμβάνουν:

    - **Πρόωρη διακοπή**: Σταματά την εκπαίδευση μόλις η απόδοση επικύρωσης δεν βελτιωθεί για έναν προκαθορισμένο αριθμό επαναλήψεων.
    - **Σημαντικότητα χαρακτηριστικών**: Αναλύει τη σχετική συμβολή κάθε χαρακτηριστικού στις προβλέψεις του μοντέλου.
    - **Προσαρμοσμένες αντικειμενικές συναρτήσεις**: Επιτρέπει στους χρήστες να ορίζουν προσαρμοσμένες συναρτήσεις απωλειών προσαρμοσμένες σε συγκεκριμένα προβλήματα ταξινόμησης.

Για λεπτομερείς περιγραφές και έναν πλήρη κατάλογο υπερπαραμέτρων, ανατρέξτε στην τεκμηρίωση του XGBoost στην σχετική αναφορά [27].

Η αποτελεσματικότητα του XGBoost πηγάζει από την ικανότητά του να συνδυάζει τα δυνατά σημεία του gradient boosting με καινοτόμες βελτιώσεις. Η ευελιξία, η αποδοτικότητα και η ακρίβειά του το καθιστούν μια συναρπαστική επιλογή για ένα ευρύ φάσμα εργασιών ταξινόμησης. Με την κατανόηση των βασικών αρχών, των υπερπαραμέτρων και των προηγμένων χαρακτηριστικών του XGBoost, οι επιστήμονες δεδομένων μπορούν να αξιοποιήσουν τη δύναμή του για τη δημιουργία ισχυρών και επιτυχημένων μοντέλων μηχανικής μάθησης.

3.3: Μηχανική Μάθηση σε Ιατρικές Εφαρμογές

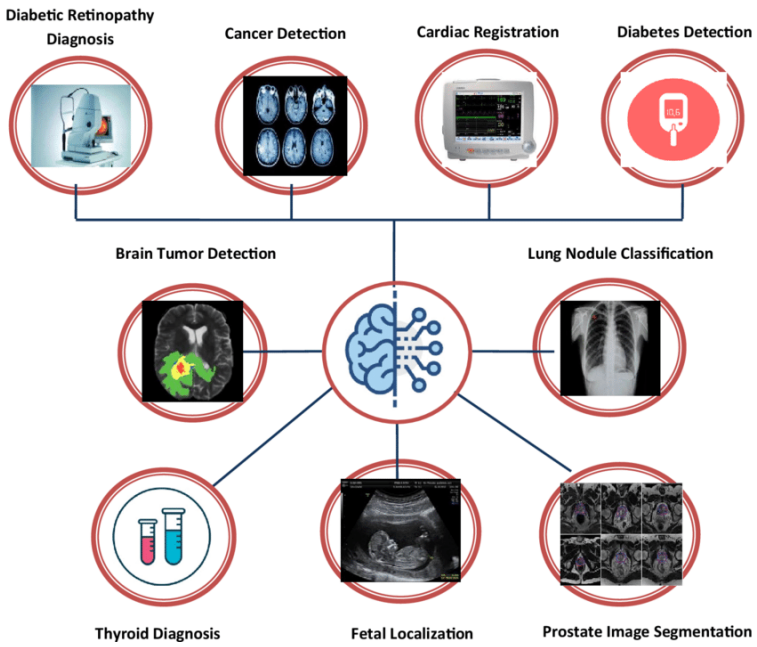
Τα τελευταία χρόνια, η ενσωμάτωση των τεχνικών μηχανικής μάθησης σε διάφορους τομείς έχει φέρει επανάσταση στον τρόπο με τον οποίο προσεγγίζουμε τις διαδικασίες επίλυσης προβλημάτων και λήψης αποφάσεων. Ένας τομέας όπου η μηχανική μάθηση έχει δείξει σημαντικές υποσχέσεις και αντίκτυπο είναι η ιατρική επιστήμη και η υγειονομική περίθαλψη. Από τη διάγνωση και την πρόγνωση έως τη βελτιστοποίηση της θεραπείας και την εξατομικευμένη ιατρική, οι αλγόριθμοι μηχανικής μάθησης γίνονται ολοένα και περισσότερο απαραίτητα εργαλεία για τους επαγγελματίες του τομέα της υγείας.

Οι αλγόριθμοι μηχανικής μάθησης διακρίνονται για την εξαγωγή πολύτιμων πληροφοριών από πολύπλοκα σύνολα δεδομένων, γεγονός που είναι ιδιαίτερα σημαντικό στην υγειονομική περίθαλψη, όπου παράγονται καθημερινά τεράστιες ποσότητες δεδομένων μέσω αρχείων ασθενών, ιατρικών απεικονίσεων, γενετικής αλληλουχίας και φορητών συσκευών. Αξιοποιώντας αυτές τις πηγές δεδομένων, τα μοντέλα μηχανικής μάθησης μπορούν να βοηθήσουν τους κλινικούς γιατρούς να κάνουν πιο ακριβείς διαγνώσεις, να προβλέπουν τα αποτελέσματα των ασθενών, να εντοπίζουν πληθυσμούς υψηλού κινδύνου και να προσαρμόζουν τις στρατηγικές θεραπείας σε μεμονωμένους ασθενείς [28] .

Η μηχανική μάθηση έχει σημαντική πρόοδο στην ταξινόμηση των όγκων του εγκεφάλου, επιτρέποντας ακριβέστερες, μη επεμβατικές διαγνώσεις που υποστηρίζουν κλινικές αποφάσεις. Τα συνελικτικά νευρωνικά δίκτυα (CNN) βρίσκονται στην πρώτη γραμμή αυτής της προόδου, επιτυγχάνοντας υψηλές ακρίβειες, όπως 97-99%, σε εργασίες όπως η διάκριση γλοιωμάτων, μηνιγγιωμάτων και όγκων της υπόφυσης, καθώς και η διαφοροποίηση μεταξύ γλοιωμάτων χαμηλού και υψηλού βαθμού. Τα αποτελέσματα αυτά προκύπτουν από τη συστηματική ρύθμιση των υπερπαραμέτρων, η οποία ενισχύει τις δυνατότητες γενίκευσης των μοντέλων, ελαχιστοποιώντας παράλληλα την υπολογιστική πολυπλοκότητα [29],[31]. Επιπλέον, οι συστηματικές αξιολογήσεις των αποτελεσμάτων αυτών των μοντέλων, υπογραμμίζουν τις κλινικές δυνατότητες των τεχνικών βαθιάς μάθησης, με πολυάριθμες μελέτες να αναφέρουν σχεδόν τέλειες μετρήσεις επιδόσεων. Ωστόσο, πρέπει να αντιμετωπιστούν προκλήσεις όπως η μεταβλητότητα στα σύνολα δεδομένων, οι ασυνέπειες στις μεθόδους επικύρωσης και η περιορισμένη κλινική χρηστικότητα. Για παράδειγμα, οι ερευνητές υποστηρίζουν την ισχυρή διασταυρούμενη επικύρωση και τα ποικίλα σύνολα δεδομένων για την καλύτερη απεικόνιση των συνθηκών του πραγματικού κόσμου [30],[31]. Αυτές οι εξελίξεις υπογραμμίζουν τις μετασχηματιστικές δυνατότητες της μηχανικής μάθησης στη διάγνωση

όγκων στον εγκέφαλο, τοποθετώντας την ως ακρογωνιαίο λίθο για τις μελλοντικές εξελίξεις στην ιατρική απεικόνιση.

Ένας από τους πρωταρχικούς τομείς στους οποίους η μηχανική μάθηση έχει σημειώσει σημαντικά βήματα προόδου είναι το πρόβλημα της ταξινόμησης εγκεφαλικών όγκων ιατρική διάγνωση. Οι παραδοσιακές διαγνωστικές μέθοδοι βασίζονται συχνά σε υποκειμενικές ερμηνείες των συμπτωμάτων και των ιατρικών εξετάσεων, οδηγώντας σε μεταβλητότητα και πιθανά σφάλματα στη διάγνωση. Οι αλγόριθμοι μηχανικής μάθησης, από την άλλη πλευρά, μπορούν να αναλύσουν μεγάλους όγκους δεδομένων ασθενών για τον εντοπισμό μοτίβων και συσχετίσεων που μπορεί να μην είναι άμεσα εμφανείς στους ανθρώπινους παρατηρητές. Αυτό μπορεί να οδηγήσει σε έγκαιρη ανίχνευση ασθενειών, ακριβέστερη ταξινόμηση των καταστάσεων και βελτιωμένα αποτελέσματα για τους ασθενείς.



Εικόνα 15 Εφαρμογές της Μηχανική Μάθησης στην Ιατρική από [63]

Πέρα από τη διάγνωση, η μηχανική μάθηση διαδραματίζει κρίσιμο ρόλο στην προγνωστική μοντελοποίηση και πρόγνωση. Αναλύοντας κλινικά δεδομένα και βιοδείκτες, τα μοντέλα μηχανικής μάθησης μπορούν να προβλέψουν την πιθανότητα εξέλιξης της νόσου, υποτροπής ή ανεπιθύμητων συμβάντων σε μεμονωμένους ασθενείς. Αυτές οι πληροφορίες

μπορούν να βοηθήσουν τους κλινικούς γιατρούς να προσαρμόσουν τα σχέδια θεραπείας, να κατανείμουν αποτελεσματικότερα τους πόρους και να παρέμβουν προληπτικά για τη βελτίωση των αποτελεσμάτων των ασθενών [29] .

Μία από τις πιο υποσχόμενες εφαρμογές της μηχανικής μάθησης στην υγειονομική περίθαλψη είναι η εξατομικευμένη ιατρική. Αναλύοντας δεδομένα ειδικά για τον ασθενή, ιλαμβανομένων των γενετικών πληροφοριών, του ιατρικού ιστορικού και των παραγόντων του τρόπου ζωής, τα μοντέλα μηχανικής μάθησης μπορούν να προτείνουν θεραπείες που βελτιστοποιούνται για μεμονωμένους ασθενείς [33]-[35] . Αυτό μπορεί να οδηγήσει σε πιο αποτελεσματικές θεραπείες, μειωμένες παρενέργειες και βελτιωμένη ικανοποίηση των ασθενών. Ενώ τα δυνητικά οφέλη από την ενσωμάτωση της μηχανικής μάθησης στην υγειονομική περίθαλψη είναι τεράστια, υπάρχουν αρκετές προκλήσεις και προβληματισμοί που πρέπει αντιμετωπιστούν. Σε αυτές περιλαμβάνονται ζητήματα που σχετίζονται με την ποιότητα και τη διαλειτουργικότητα των δεδομένων, την προστασία της ιδιωτικής ζωής και τη συγκατάθεση των ασθενών, τη διαφάνεια και την ερμηνευσιμότητα των αλγορίθμων και τη συμμόρφωση με τις κανονιστικές διατάξεις. Επιπλέον, υπάρχει ανάγκη για ισχυρή επικύρωση και αξιολόγηση των μοντέλων μηχανικής μάθησης ώστε να διασφαλιστεί η αξιοπιστία και η γενίκευσή τους σε διαφορετικούς πληθυσμούς ασθενών και ρυθμίσεις υγειονομικής περίθαλψης [37]-[38].

Η μηχανική μάθηση έχει τη δυνατότητα να μεταμορφώσει την υγειονομική περίθαλψη, επιτρέποντας ακριβέστερη διάγνωση, εξατομικευμένη θεραπεία και βελτιωμένα αποτελέσματα για τους ασθενείς. Ωστόσο, η αξιοποίηση αυτού του δυναμικού απαιτεί την αντιμετώπιση διαφόρων τεχνικών, ηθικών και κανονιστικών προκλήσεων. Με τη συνεργατική αντιμετώπιση αυτών των προκλήσεων και την αξιοποίηση της δύναμης της μηχανικής μάθησης, μπορούμε να φέρουμε επανάσταση στην παροχή υγειονομικής περίθαλψης και να εγκαινιάσουμε μια νέα εποχή της ιατρικής ακριβείας.

3.3.1: Επεξεργασία Ιατρικών Εικόνων

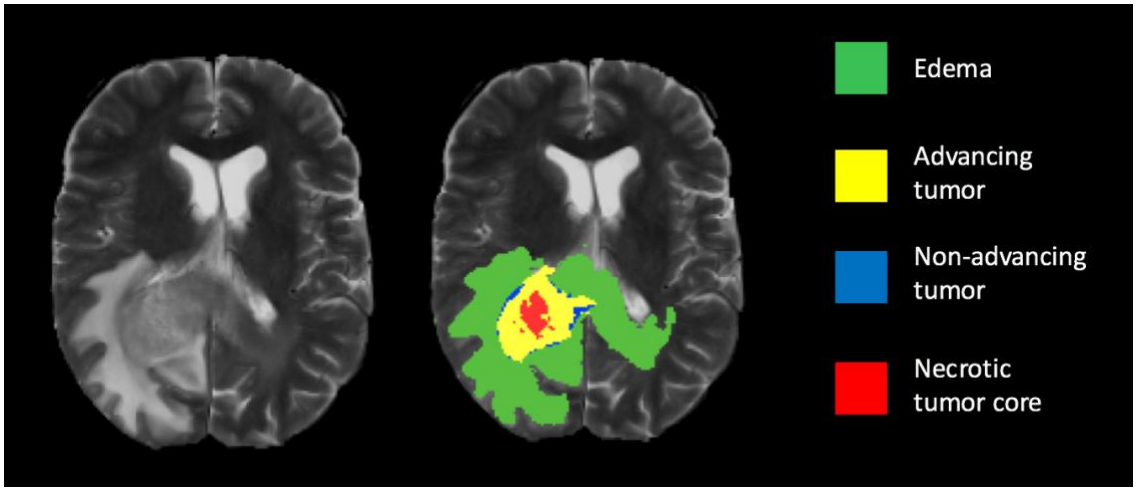
Η ιατρική απεικόνιση διαδραματίζει καθοριστικό ρόλο στη διάγνωση, τη θεραπεία και την παρακολούθηση διαφόρων ιατρικών καταστάσεων. Με την πρόοδο της τεχνολογίας, οι μέθοδοι ιατρικής απεικόνισης, όπως οι ακτίνες X, η μαγνητική τομογραφία, η αξονική τομογραφία και οι υπέρηχοι, έχουν γίνει απαραίτητα εργαλεία στην υγειονομική περίθαλψη.

Ωστόσο, η ερμηνεία των ιατρικών εικόνων μπορεί να αποτελέσει πρόκληση λόγω της πολυπλοκότητας και της μεταβλητότητας των ανατομικών δομών και των παθολογικών καταστάσεων. Η επεξεργασία ιατρικών εικόνων, η οποία περιλαμβάνει την ανάλυση και τον χειρισμό ιατρικών εικόνων με τη χρήση υπολογιστικών τεχνικών, έχει αναδειχθεί ως μια πολύτιμη προσέγγιση για τη βελτίωση της χρησιμότητας και της ακρίβειας των δεδομένων ιατρικής απεικόνισης.

Η επεξεργασία ιατρικών εικόνων περιλαμβάνει ένα ευρύ φάσμα εργασιών, όπως η βελτίωση της εικόνας, η τμηματοποίηση, η καταχώριση και η ταξινόμηση, μεταξύ άλλων. Αυτές οι εργασίες είναι απαραίτητες για την εξαγωγή σημαντικών πληροφοριών από τις ιατρικές εικόνες, βοηθώντας τους κλινικούς ιατρούς στη διάγνωση, τον προγραμματισμό της θεραπείας και την πρόβλεψη των αποτελεσμάτων. Με την εφαρμογή προηγμένων αλγορίθμων επεξεργασίας εικόνας, οι ερευνητές και οι κλινικοί γιατροί μπορούν να βελτιώσουν την ποιότητα των ιατρικών εικόνων, να εξάγουν ποσοτικές μετρήσεις και να εντοπίσουν λεπτά μοτίβα και ανωμαλίες που μπορεί να μην είναι εύκολα διακριτά μόνο από τους ανθρώπινους παρατηρητές.

Οι τεχνικές βελτίωσης εικόνας αποσκοπούν στη βελτίωση της οπτικής ποιότητας των ιατρικών εικόνων με τη μείωση του θορύβου, την ενίσχυση της αντίθεσης και την ανάδειξη των σχετικών χαρακτηριστικών. Διάφορες μέθοδοι, όπως η εξίσωση ιστογράμματος, το φιλτράρισμα και η προσαρμοστική επεξεργασία, έχουν χρησιμοποιηθεί για τη βελτίωση της διαγνωστικής αξίας των ιατρικών εικόνων. Για παράδειγμα, η προσαρμοστική εξίσωση ιστογράμματος έχει αποδειχθεί ότι ενισχύει την ορατότητα λεπτών λεπτομερειών σε εικόνες μαγνητικής τομογραφίας και αξονικής τομογραφίας, οδηγώντας σε βελτιωμένη διαγνωστική ακρίβεια [39].

Η τμηματοποίηση εικόνας (Image Segmentation) περιλαμβάνει την κατάτμηση μιας εικόνας σε πολλαπλές περιοχές ή τμήματα με βάση ορισμένα κριτήρια, όπως η ένταση, η υφή ή τα ανατομικά όρια. Η τμηματοποίηση διαδραματίζει κρίσιμο ρόλο στην οριοθέτηση δομών ενδιαφέροντος, όπως όγκοι, όργανα ή ανατομικά ορόσημα, για περαιτέρω ανάλυση και ποσοτικοποίηση. Διάφοροι αλγόριθμοι τμηματοποίησης, όπως η κατωφλίωση, η αύξηση περιοχών και τα μοντέλα ενεργού περιγράμματος, έχουν αναπτυχθεί και εφαρμοστεί σε ιατρικές εικόνες με πολλά υποσχόμενα αποτελέσματα. Για παράδειγμα, τα μοντέλα ενεργού περιγράμματος έχουν χρησιμοποιηθεί για την ακριβή τμηματοποίηση όγκων του εγκεφάλου σε εικόνες MRI, διευκολύνοντας τη μέτρηση του όγκου του όγκου και το σχεδιασμό της θεραπείας [40].



Εικόνα 16 Τμηματοποίηση Εικόνας (Image Segmentation) από [64]

Η καταχώριση εικόνας (Image Registration) είναι η διαδικασία ευθυγράμμισης πολλαπλών εικόνων του ίδιου θέματος που αποκτήθηκαν σε διαφορετικές χρονικές στιγμές ή με διαφορετικές μεθόδους. Η καταχώριση επιτρέπει την ενσωμάτωση πληροφοριών από πολλαπλές εικόνες, διευκολύνοντας τη σύγκριση, τη συγχώνευση και τη διαχρονική ανάλυση. Οι τεχνικές καταχώρισης κυμαίνονται από άκαμπτους και συγγενείς μετασχηματισμούς έως παραμορφώσιμες μεθόδους καταχώρισης που βασίζονται σε μαθηματικά μοντέλα ή χαρακτηριστικά της εικόνας. Στο πλαίσιο της ιατρικής απεικόνισης, η καταγραφή εικόνων έχει εφαρμοστεί σε διάφορες εργασίες, όπως η συγχώνευση εικόνων για πολυτροπική απεικόνιση, η διόρθωση κίνησης σε δυναμική απεικόνιση και η παρακολούθηση της εξέλιξης της νόσου με την πάροδο του χρόνου [41].

Τα τελευταία χρόνια, οι τεχνικές μηχανικής μάθησης εφαρμόζονται όλο και περισσότερο σε εργασίες επεξεργασίας ιατρικών εικόνων, αξιοποιώντας τη δύναμη των μοντέλων που βασίζονται σε δεδομένα για την αυτοματοποίηση και τη βελτίωση της ανάλυσης εικόνων. Η βαθιά μάθηση, ειδικότερα, έχει επιδείξει αξιοσημείωτη επιτυχία σε εργασίες όπως η ταξινόμηση, η κατάτμηση και η ανακατασκευή εικόνων. Τα συνελικτικά νευρωνικά δίκτυα (CNNs) έχουν αναδειχθεί ως η πλέον σύγχρονη προσέγγιση για πολλές εργασίες επεξεργασίας ιατρικών εικόνων, επιτυγχάνοντας επιδόσεις συγκρίσιμες με τις επιδόσεις των ανθρώπινων εμπειρογνωμόνων ή και καλύτερες από αυτές [42].

Η επεξεργασία ιατρικών εικόνων διαδραματίζει κρίσιμο ρόλο στην εξαγωγή πολύτιμων πληροφοριών από ιατρικές εικόνες, στην ενίσχυση της διαγνωστικής ακρίβειας και στην

υποβοήθηση της λήψης κλινικών αποφάσεων. Με την αξιοποίηση προηγμένων υπολογιστικών τεχνικών, οι ερευνητές και οι κλινικοί γιατροί μπορούν να ξεκλειδώσουν το πλήρες δυναμικό των δεδομένων ιατρικής απεικόνισης, οδηγώντας σε βελτιωμένα αποτελέσματα για τους ασθενείς και εξελίξεις στην υγειονομική περίθαλψη.

3.3.2: Τμηματοποίηση και Καταχώρηση

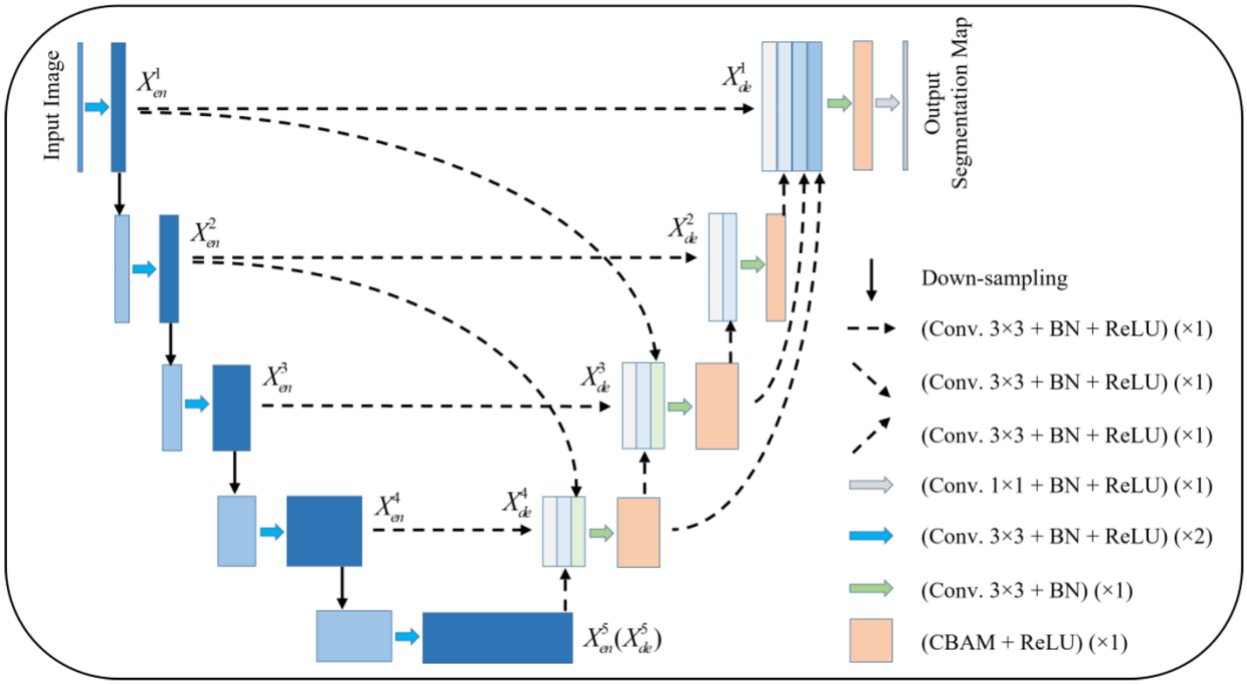
Η τμηματοποίηση και η καταχώριση ιατρικών εικόνων διαδραματίζουν καθοριστικό ρόλο στον τομέα της ανάλυσης ιατρικών εικόνων. Η τμηματοποίηση περιλαμβάνει την κατάτμηση μιας εικόνας σε περιοχές ή δομές με νόημα, ενώ η καταχώριση περιλαμβάνει την ευθυγράμμιση και τη σύγκριση εικόνων από διαφορετικές μεθόδους ή χρονικές στιγμές. Οι διαδικασίες αυτές είναι απαραίτητες για εργασίες όπως η διάγνωση, ο σχεδιασμός θεραπείας και η παρακολούθηση της εξέλιξης της νόσου. Σε αυτό το κεφάλαιο, θα εμβαθύνουμε στις τεχνικές και τις μεθοδολογίες που χρησιμοποιούνται στην κατάτμηση και την καταχώριση ιατρικών εικόνων, δίνοντας έμφαση στη σημασία τους στο πλαίσιο εφαρμογών μηχανικής μάθησης για την ανίχνευση μεταστάσεων όγκων στον εγκέφαλο.

Η τμηματοποίηση ιατρικών εικόνων αποσκοπεί στην οριοθέτηση ανατομικών δομών ή παθολογικών περιοχών ενδιαφέροντος εντός ιατρικών εικόνων. Έχουν αναπτυχθεί διάφορες τεχνικές τμηματοποίησης, που κυμαίνονται από παραδοσιακές μεθόδους έως πιο προηγμένες προσεγγίσεις βαθιάς μάθησης.

Οι παραδοσιακές μέθοδοι τμηματοποίησης περιλαμβάνουν κατωφλίωση, ανάπτυξη περιοχών, ανίχνευση ακμών και αλγορίθμους ομαδοποίησης, όπως k-means και fuzzy C-means. Αυτές οι μέθοδοι βασίζονται σε προκαθορισμένους κανόνες, κατώτατα όρια ή χαρακτηριστικά για την κατάτμηση εικόνων. Ωστόσο, μπορεί να δυσκολεύονται με πολύπλοκες δομές ή θορυβώδεις εικόνες.

Οι τεχνικές τμηματοποίησης που βασίζονται στη βαθιά μάθηση, ιδίως τα νευρωνικά δίκτυα συνελίξεων (CNN), έχουν επιδείξει αξιοσημείωτη επιτυχία τα τελευταία χρόνια. Τα CNN μπορούν να μάθουν αυτόματα ιεραρχικά χαρακτηριστικά από ιατρικές εικόνες και να τμηματοποιήσουν με ακρίβεια δομές ενδιαφέροντος. Αξιοσημείωτες αρχιτεκτονικές περιλαμβάνουν τα U-Net, SegNet και DeepLab. Αυτά τα μοντέλα έχουν εφαρμοστεί σε διάφορες μεθόδους ιατρικής απεικόνισης, συμπεριλαμβανομένης της απεικόνισης μαγνητικού συντονισμού (MRI), της υπολογιστικής τομογραφίας (CT) και της τομογραφίας εκπομπής ποζιτρονίων (PET).

Για παράδειγμα, το U-Net, το οποίο εισήχθη από τους Ronneberger et al. (2015) [43], έχει υιοθετηθεί ευρέως για την κατάτμηση βιοϊατρικών εικόνων λόγω της ικανότητάς του να συλλαμβάνει τόσο τις τοπικές λεπτομέρειες όσο και το παγκόσμιο πλαίσιο. Η αρχιτεκτονική του δικτύου αποτελείται από ένα συρρικνούμενο μονοπάτι για τη σύλληψη του πλαισίου και ένα συμμετρικό επεκτεινόμενο μονοπάτι για τον ακριβή εντοπισμό.



Εικόνα 17 Δομή συνελικτικού δικτύου U-net από [65]

Η καταγραφή ιατρικών εικόνων περιλαμβάνει την ευθυγράμμιση εικόνων που έχουν αποκτηθεί από διαφορετικές μεθόδους ή χρονικά σημεία για τη διευκόλυνση της σύγκρισης, της συγχώνευσης ή της ανάλυσης. Η καταγραφή είναι ζωτικής σημασίας για την ενσωμάτωση πληροφοριών από πολλαπλές μεθόδους απεικόνισης, όπως η μαγνητική τομογραφία και το PET, ή για την παρακολούθηση της εξέλιξης της νόσου με την πάροδο του χρόνου.

Οι μέθοδοι καταχώρισης μπορούν να κατηγοριοποιηθούν σε άκαμπτους, συγγενείς και μη άκαμπτους μετασχηματισμούς με βάση το βαθμό παραμόρφωσης που επιτρέπεται. Η άκαμπτη καταγραφή διατηρεί το σχήμα των δομών, ενώ η καταγραφή κατά συγγένεια λαμβάνει υπόψη την κλιμάκωση, την περιστροφή και τη διάτμηση. Η μη άκαμπτη καταχώριση επιτρέπει πιο σύνθετες παραμορφώσεις, κατάλληλες για την ευθυγράμμιση εικόνων με σημαντικές ανατομικές διαφορές.

Για την καταχώριση χρησιμοποιούνται διάφορες τεχνικές βελτιστοποίησης, συμπεριλαμβανομένων μεθόδων που βασίζονται στην ένταση, μεθόδων που βασίζονται σε χαρακτηριστικά και μεθόδων που βασίζονται στην αμοιβαία πληροφορία. Αλγόριθμοι βελτιστοποίησης όπως η κάθοδος κλίσης, οι γενετικοί αλγόριθμοι και η βελτιστοποίηση σμήνους σωματιδίων χρησιμοποιούνται συνήθως για την εύρεση των βέλτιστων παραμέτρων μετασχηματισμού.

Η βαθιά μάθηση έχει επίσης δείξει υποσχέσεις σε εργασίες καταχώρισης ιατρικών εικόνων. Τα συνεπτυγμένα νευρωνικά δίκτυα μπορούν να μάθουν πολύπλοκους μετασχηματισμούς εικόνας απευθείας από τα δεδομένα, εξαλείφοντας την ανάγκη για ρητή εξαγωγή χαρακτηριστικών ή μοντέλα καταχώρισης. Για παράδειγμα, η εργασία των Balakrishnan κ.ά. (2018) [44] πρότεινε το VoxelMorph, ένα πλαίσιο καταχώρισης βασισμένο σε CNN που μαθαίνει έναν χωρικό μετασχηματισμό μεταξύ δύο εικόνων απευθείας από τις τιμές της έντασής τους.

Στο πλαίσιο της ανίχνευσης μεταστάσεων όγκων στον εγκέφαλο, η ακριβής τμηματοποίηση και η εγγραφή είναι ζωτικής σημασίας για την ακριβή οριοθέτηση των περιοχών του όγκου και την παρακολούθηση της εξέλιξης της νόσου. Με την τμηματοποίηση όγκων από ιατρικές εικόνες, όπως μαγνητικές τομογραφίες, οι κλινικοί γιατροί μπορούν να ποσοτικοποιήσουν το μέγεθος, το σχήμα και τη χωρική κατανομή του όγκου, βοηθώντας στο σχεδιασμό της θεραπείας και στην αξιολόγηση της ανταπόκρισης στη θεραπεία.

Επιπλέον, οι τεχνικές καταγραφής επιτρέπουν τη σύγκριση εικόνων που αποκτήθηκαν σε διαφορετικές χρονικές στιγμές, επιτρέποντας στους κλινικούς γιατρούς να παρακολουθούν τις αλλαγές στο μέγεθος, το σχήμα και τη θέση του όγκου με την πάροδο του χρόνου. Αυτή η διαχρονική ανάλυση παρέχει πολύτιμες πληροφορίες σχετικά με την εξέλιξη της νόσου και την αποτελεσματικότητα των θεραπευτικών παρεμβάσεων.

Η ενσωμάτωση αλγορίθμων μηχανικής μάθησης, όπως οι αλγόριθμοι Naive Bayes, Random Forest, Support Vector Machines (SVMs), XGBoost και Neural Networks, σε αγωγούς τμηματοποίησης και καταχώρησης μπορεί να βελτιώσει την ακρίβεια και την αποτελεσματικότητα της ανίχνευσης και παρακολούθησης όγκων. Αυτοί οι αλγόριθμοι μπορούν να μάθουν πολύπλοκα μοτίβα και σχέσεις από δεδομένα ιατρικών εικόνων, επιτρέποντας αυτοματοποιημένες και ισχυρές διαδικασίες τμηματοποίησης και καταχώρησης.

Η τμηματοποίηση και η καταχώριση ιατρικών εικόνων είναι θεμελιώδεις εργασίες στην ανάλυση ιατρικών εικόνων, με σημαντικές επιπτώσεις στη διάγνωση, το σχεδιασμό θεραπείας και την παρακολούθηση ασθενειών. Οι παραδοσιακές μέθοδοι και οι προηγμένες τεχνικές βαθιάς μάθησης διαδραματίζουν συμπληρωματικό ρόλο στην αντιμετώπιση των προκλήσεων

που σχετίζονται με την τμηματοποίηση πολύπλοκων ανατομικών δομών και την ευθυγράμμιση εικόνων από διαφορετικές μεθόδους ή χρονικές στιγμές. Στο πλαίσιο της ανίχνευσης μετάστασης όγκου εγκεφάλου, η ακριβής τμηματοποίηση και καταχώριση είναι κρίσιμης σημασίας για τη βελτίωση των αποτελεσμάτων των ασθενών και την καθοδήγηση της λήψης κλινικών αποφάσεων.

Κεφάλαιο 4: Ανάλυση Δεδομένων και Επιλογή Μοντέλου

Το παρόν κεφάλαιο παρουσιάζει τα βήματα και τις τεχνικές για το πειραματικό μέρος της διπλωματικής εργασίας. Γίνεται εκ βαθέως ανάλυση των δεδομένων που χρησιμοποιήθηκαν στο ερευνητικό κομμάτι της εργασίας. Από την συλλογή και την επεξεργασία των δεδομένων μέχρι την εξαγωγή των απαραίτητων χαρακτηριστικών (radiomics). Στην συνέχεια διεξάγεται διερευνητική ανάλυση των χαρακτηριστικών (Exploratory Data Analysis) με σκοπό να γίνει η σωστή αντίληψη των δεδομένων και η κατάλληλη βελτιστοποίηση αυτών. Τέλος εξετάζεται η διαδικασία για την επιλογή, εκπαίδευση και βελτιστοποίηση των μοντέλων πρόβλεψης χρησιμοποιώντας της κατάλληλες μετρικές. Ο σκοπός αυτού του κεφαλαίου είναι να αναδείξει την έρευνα γύρω από το πειραματικό μέρος της διπλωματικής εργασίας εξετάζοντας πολλαπλές τεχνικές και μεθόδους που αναφέρθηκαν στα προηγούμενα κεφάλαια.

4.1: Δεδομένα Μοντέλου

Όπως αναλύθηκε εκτενώς στα προηγούμενα κεφάλαια το επίκεντρο της εργασίας είναι οι καρκινικοί όγκοι στον εγκέφαλο και οι επιδράσεις τους. Έτσι για το πειραματικός μέρος της εργασίας επιλέχθηκε μία συλλογή δεδομένων που συλλέχθηκε από ασθενείς με μετάσταση στον εγκέφαλο. Από εδώ και στο εξής και για όλο το πειραματικό μέρος θα κάνουμε χρήση αυτών των δεδομένων για όλη την διερευνητική ανάλυση και την εκπαίδευση των μοντέλων μηχανικής μάθησης.

4.1.1: Συλλογή Δεδομένων

Η συλλογή δεδομένων που επιλέχθηκε για το πειραματικό μέρος, προέρχεται από το Εθνικό Ινστιτούτο για τον Καρκίνο (**NCI**), των Ηνωμένων Πολιτειών. Το NCI συντονίζει το Εθνικό Πρόγραμμα

για τον Καρκίνο των Ηνωμένων Πολιτειών και αποτελεί μέρος του Εθνικού Ινστιτούτου Υγείας (NIH), το οποίο είναι ένας από τους έντεκα φορείς που αποτελούν μέρος του Υπουργείου Υγείας και Ανθρωπίνων Υπηρεσιών των ΗΠΑ. Το NCI διεξάγει και υποστηρίζει έρευνα, εκπαίδευση, διάδοση πληροφοριών για την υγεία και άλλες δραστηριότητες που σχετίζονται με τα αίτια, την πρόληψη, τη διάγνωση και τη θεραπεία του καρκίνου, την υποστηρικτική φροντίδα ασθενών με καρκίνο και των οικογενειών του και την επιβίωση από τον καρκίνο.

Τα δεδομένα είναι διαθέσιμα μέσω μίας υπηρεσίας του NCI, το Αρχείο Απεικόνισης του Καρκίνου (**TCIA**) [45]. Το TCIA είναι μια υπηρεσία που φιλοξενεί ένα μεγάλο αρχείο ιατρικών εικόνων καρκίνου, προσβάσιμες για δημόσια χρήση. Τα δεδομένα οργανώνονται ως «συλλογές». τυπικά η απεικόνιση ασθενών που σχετίζεται με μια κοινή ασθένεια (π.χ. καρκίνος του πνεύμονα), τον τρόπο ή τον τύπο της εικόνας (MRI, CT, ψηφιακή ιστοπαθολογία κ.α.).Η μορφή αποθήκευσης DICOM είναι η κύρια μορφή αρχείου που χρησιμοποιείται από το TCIA για την ακτινολογική απεικόνιση. Υποστηρικτικά δεδομένα που σχετίζονται με τις εικόνες, όπως αποτελέσματα ασθενών, λεπτομέρειες θεραπείας, γονιδιωματική και αναλύσεις ειδικών παρέχονται επίσης όταν είναι διαθέσιμα.

Στην παρούσα εργασία έγινε επιλογή μίας συλλογής δεδομένων από το TCIA, και εμπεριέχει δημογραφικά δεδομένα και δεδομένα απεικόνισης από 3D εικόνες που έχουν τμηματοποιηθεί (Segmentation). Το σημείο ενδιαφέροντος που εξετάζει είναι η περιοχή του εγκεφάλου και συνολικά στην συλλογή υπάρχουν δεδομένα από 200 συμμετέχοντες. Για όλα τα δεδομένα έχουν αφαιρεθεί κάθε είδος προσωπικό στοιχείο ή πληροφορία έτσι ώστε να διασφαλίζεται η ανωνυμία των συμμετεχόντων. Τα δεδομένα των ασθενών συλλέχθηκαν από 3 διαφορετικές πηγές: από το Yale New Haven Health database (2013-2021), από το Yale tumor board registry (2021), και από το Yale gamma knife registry (2017-2021).

Τα κριτήρια ένταξης περιλάμβαναν παθολογικά αποδεδειγμένη διάγνωση εγκεφαλικής μετάστασης και διαθεσιμότητα σάρωσης πριν από τη θεραπεία με τις συνήθεις ακολουθίες μαγνητικής τομογραφίας (T1 w, T1 post-gadolinium, T2w και FLAIR). Τα κριτήρια αποκλεισμού περιλάμβαναν την έλλειψη σάρωσης προθεραπείας ή μίας από τις τυπικές ακολουθίες μαγνητικής τομογραφίας και σημαντικά τεχνήματα κίνησης σε οποιαδήποτε από τις τυπικές ακολουθίες. Οι εικόνες αποκτήθηκαν με τη χρήση σαρωτή 1,5T ή 3T. Για τους περισσότερους ασθενείς οι εικόνες αποκτήθηκαν με T1 post-gadolinium ακολουθίες με την χρήση του πρωτοκόλλου MPRAGE, ενώ ένα μικρό υποσύνολο έγινε με την χρήση spin echoes. Οι τμηματοποιήσεις του πυρήνα και των νεκρωτικών μερών του όγκου πραγματοποιήθηκαν σε ακολουθίες T1 post-gadolinium, ενώ ολόκληρος ο όγκος τμηματοποιήθηκε σε ακολουθίες FLAIR. Όλες οι τμηματοποιήσεις εφαρμόστηκαν χειροκίνητα σε εικόνες DICOM (Digital Imaging and Communications in Medicine) Όλες οι τμηματοποιήσεις εγκρίθηκαν από δύο νευροακτινολόγους.

Οι τέσσερις πρότυπες ακολουθίες για κάθε ασθενή εξήχθησαν από το ερευνητικό PACS σε μορφή NifTI, συν-καταχωρίστηκαν στο ανατομικό πρότυπο SRI24, έγινε επαναληπτική δειγματοληψία σε

ομοιόμορφη ισοτροπική ανάλυση (1 mm3) και απογύμνωση κρανίου. Οι μάσκες τμηματοποίησης του πυρήνα, της νέκρωσης και ολόκληρης της εικόνας εξήχθησαν μεμονωμένα από το ερευνητικό PACS σε μορφή NifTI, συνδυάστηκαν σε μία μάσκα διατηρώντας τις θέσεις τους στον εγγενή χώρο και καταχωρήθηκαν στο ανατομικό πρότυπο SRI24 πριν από την επαναληπτική δειγματοληψία σε ομοιόμορφη ισοτροπική ανάλυση. Τα αρχεία NifTI ακολουθίας και κατάτμησης ελέγχθηκαν και διορθώθηκαν χειροκίνητα από νευροακτινολόγο στο λογισμικό ITK-SNAP.

Τα δημογραφικά δεδομένα, συμπεριλαμβανομένου του φύλου, της εθνικότητας, της ηλικίας κατά τη διάγνωση, του ιστορικού καπνίσματος σε πακέτα ετών και της παρουσίας εξωσωματικών μεταστάσεων ελήφθησαν με τη χρήση του ηλεκτρονικού ιατρικού φακέλου (EMR - electronic medical record). Η επιβίωση υπολογίστηκε αφαιρώντας την ημερομηνία διάγνωσης από την ημερομηνία θανάτου ή από την ημερομηνία της τελευταίας σημείωσης του EMR για τους λογοκριμένους ασθενείς. Τα ποιοτικά απεικονιστικά χαρακτηριστικά, συμπεριλαμβανομένης της ενδοκοιλιακής συμμετοχής και της ενδοογκικής ευαισθησίας σε τουλάχιστον μία βλάβη στην ακολουθία SWI, προέκυψαν από την οπτική αξιολόγηση των εικόνων. Τα ποσοτικά απεικονιστικά χαρακτηριστικά εξήχθησαν από τη μάσκα τμηματοποίησης NifTI για κάθε ασθενή και περιλαμβάνουν τον συνολικό μέρος του όγκου που έχει ενισχυθεί, τον συνολικό κομμάτι του νεκρωτικού όγκου, τον συνολικό μέρος του περιτοναϊκού οιδήματος, τον λόγο του νεκρωτικού προς τον όγκο που ενισχύεται, τον λόγο του περιτοναϊκού οιδήματος προς τον όγκο που ενισχύεται, τον αριθμό των ενισχυτικών βλαβών, τον αριθμό των νεκρωτικών βλαβών και τον αριθμό των βλαβών με περιτοναϊκό οίδημα. Τέλος, η προέλευση της μετάστασης λήφθηκε επίσης από το EMR από προηγούμενες ογκολογικές και/ή παθολογικές εκθέσεις. Οι ημερομηνίες γέννησης, διάγνωσης, θανάτου και τελευταίας σημείωσης ήταν όλες ανώνυμες διατηρώντας τη διάρκεια μεταξύ των συμβάντων. Οι πληροφορίες για τα κλινικά και απεικονιστικά χαρακτηριστικά αποθηκεύονται σε αρχείο Excel.

Το συγκεκριμένο σύνολο δεδομένων είναι μοναδικό καθώς περιλαμβάνει πολλές εγκεφαλικές μεταστάσεις κάτω του εκατοστού, για τις οποίες δεν υπάρχουν ακόμη ισχυροί αλγόριθμοι. Επιπλέον, το σύνολο δεδομένων περιλαμβάνει πολλές βλάβες με νεκρωτικές κατατμήσεις, οι οποίες ιστορικά δεν παρέχονται με τις κατατμήσεις του συνόλου δεδομένων. Επίσης περιλαμβάνονται αποτελέσματα επιβίωσης που η λήψη τους έγινε χειροκίνητα από το EMR, επιτρέποντας τη μελλοντική συσχέτιση των απεικονιστικών δεδομένων με κλινικά αποτελέσματα. Επιπλέον, η ετερογένεια των πηγών από τις οποίες έχουν ληφθεί τα δεδομένα των ασθενών, θα επιτρέψει την εκπαίδευση των αλγορίθμων σε ένα πραγματικό σύνολο νοσοκομειακών δεδομένων που δεν είναι ιδιαίτερα επιμελημένο, ενισχύοντας έτσι τη δυνατότητα εφαρμογής του μοντέλου στην κλινική πρακτική. [45]

4.1.2: Προετοιμασία Δεδομένων

Έχοντας επιλέξει το σύνολο δεδομένων στο οποίο θα εργαστούμε, ξεκινάμε να εξετάζουμε τα αρχικά δεδομένα. Όπως αναφέραμε προηγουμένως, το σύνολο δεδομένων αποτελείται από 1. **Κλινικά/δημογραφικά δεδομένα** και **3D εικόνες και τμηματοποιήσεις** αυτών από συνολικά **200** ασθενείς/δείγματα.

Τα κλινικά δεδομένα είναι διαθέσιμα μέσω ενός αρχείου Excel. Στην πρώτη καρτέλα με όνομα "Data" έχουμε τα κλινικά δεδομένα που συλλέχθηκαν για κάθε έναν ασθενή και στην δεύτερη καρτέλα με όνομα "Description" έχουμε μία σύντομη περιγραφή των στηλών της καρτέλας "Data". Ας ρίξουμε μία γρήγορη ματιά στα κλινικά δεδομένα:

Εικόνα 18 - Κλινικά δεδομένα

	Column	Description	Values
0	BraTS_MET_ID	anonymized patient identifier with the format "BraTS-MET-00XXX-000" where XXX is the number displayed in the column	NaN
1	Age	age (years) at diagnosis of brain metastasis; blank if not available in electronic medical record	NaN
2	Sex	sex	0 = Male; 1 = Female
3	Ethnicity	ethnicity; blank if patient declined to state or information was not available in electronic medical record	0 = White; 1 = White Hispanic; 2 = White/Black; 3 = Black; 4 = Asian/Pacific Islander
4	Smoke	smoking history at diagnosis (pack-years); blank if not available in electronic medical record	NaN
5	Primary	primary tumor origin of metastasis	1 = Breast cancer; 2 = Gastrointestinal cancers; 3 = Small cell lung cancer; 4 = Melanoma; 5 = Non small cell lung cancer; 6 = Renal cell carcinoma; 7 = Other
6	Extranodal	presence of extranodal metastasis; blank if not available in electronic medical record	0 = No; 1 = Yes
7	Survival	time (months) to death or last note in electronic medical record as of July 2022; blank if not available in electronic medical record	NaN
8	Death	whether death occurred as of July 2022 or patient was censored (i.e. only date of last note in electronic medical record was available); blank if not available in electronic medical record	0 = No; 1 = Yes
9	Infratentorial	presence of metastasis in infratentorial region on imaging	0 = No; 1 = Yes
10	ET_num	number of enhancing lesions on T1 post-contrast imaging	NaN
11	Nec_num	number of lesions with necrosis on T1 post-contrast imaging	NaN
12	Edema_num	number of lesions with peritumoral edema on FLAIR imaging	NaN
13	ET_vol	volume (cm^3) of enhancing tumor on T1 post-contrast imaging (excluding necrotic portions and intrinsic T1 hyperintensity)	NaN
14	Nec_vol	volume (cm^3) of necrotic tumor on T1 post-contrast imaging	NaN
15	Edema_vol	volume (cm^3) of peritumoral edema on FLAIR imaging	NaN
16	Nec_ET_ratio	ratio of necrotic volume to enhancing volume (including necrotic portions but excluding intrinsic T1 hyperintensity) on T1 post-contrast imaging	NaN
17	Edema_ET_ratio	ratio of peritumoral edema volume on FLAIR imaging to enhancing volume (including necrotic portions but excluding intrinsic T1 hyperintensity) on T1 post-contrast imaging	NaN

Εικόνα 19 - Περιγραφές των δημογραφικών δεδομένων

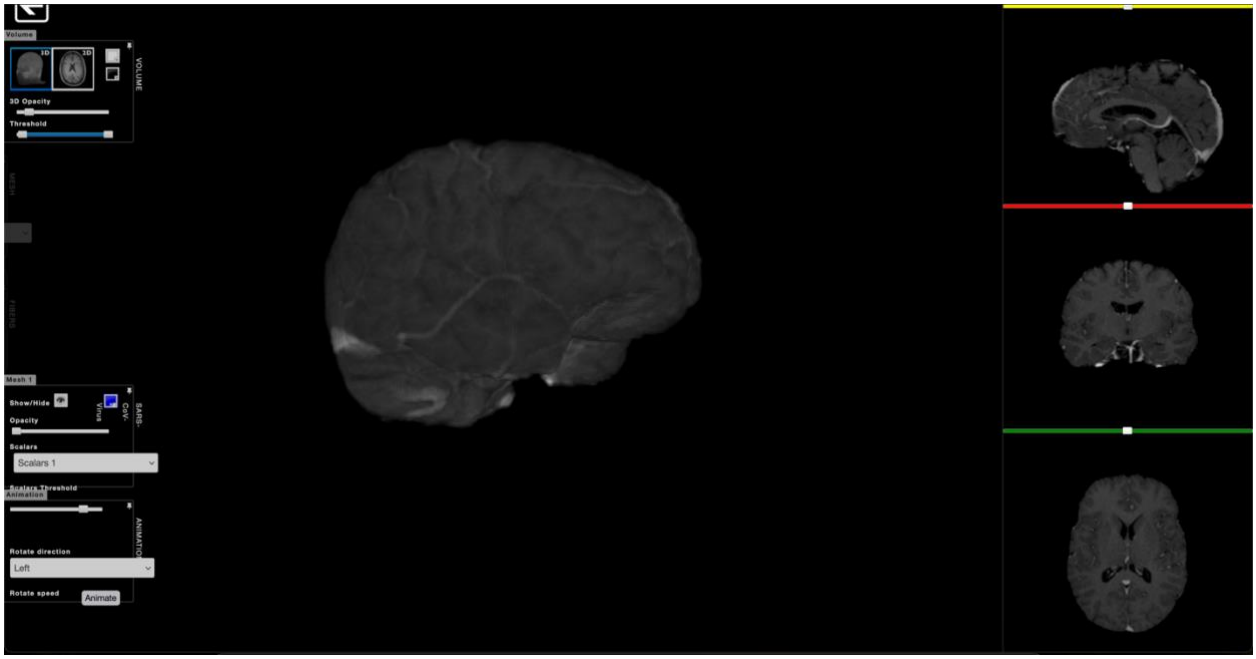
Τα συγκεκριμένα δεδομένα, θα συμβάλουν στο σύνολο των χαρακτηριστικών που θα διαμορφώσουμε, όπως θα δούμε και θα αναλύσουμε περισσότερο στις επόμενες ενότητες. Η στήλη που μας ενδιαφέρει περισσότερο και για εμάς θα αποτελεί την μεταβλητή στόχο, είναι τα δεδομένα σχετικά με το "Death". Πρόκειται για μία δυαδική μεταβλητή που εκφράζει εάν εάν ο θάνατος συνέβη τον Ιούλιο του 2022 ή ο ασθενής λογοκρίθηκε (δηλαδή ήταν διαθέσιμη μόνο η ημερομηνία τελευταίας σημείωσης στον ηλεκτρονικό ιατρικό φάκελο). Εάν η μεταβλητή είναι κενή τότε η πληροφορία δεν είναι διαθέσιμη στον ηλεκτρονικό ιατρικό φάκελο. Έτσι η τιμή 0 σημαίνει πως δεν επήλθε ο θάνατος και η τιμή 1 σημαίνει πως ο θάνατος συνέβη για τον ασθενή.

Το δεύτερο κομμάτι των δεδομένων μας είναι οι 3D εικόνες και οι τμηματοποιήσεις αυτών. Για να έχουμε πρόσβαση και να μπορώ να έχουμε τα δεδομένα στο δικό μας σύστημα, κάνουμε χρήση του

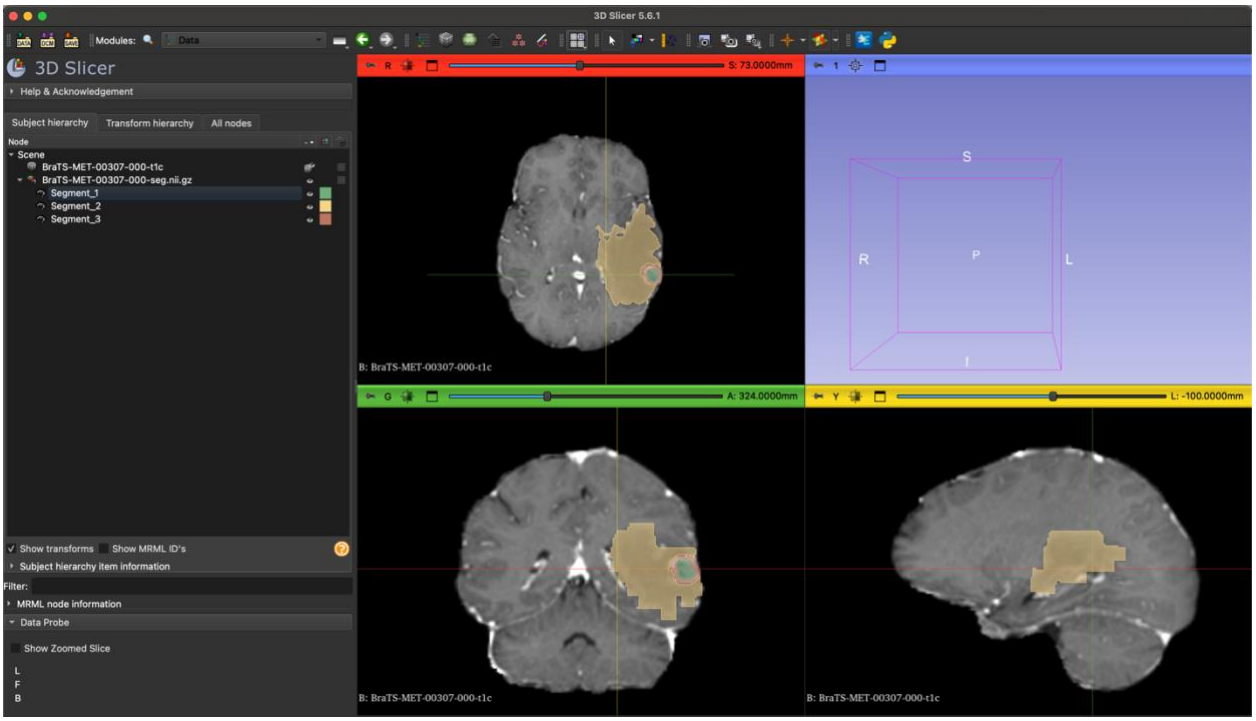
βοηθητικού εργαλείου IBM Aspera. Αντίστοιχα με τα κλινικά δεδομένα έτσι και εδώ, έχουμε δεδομένα για 200 συμμετέχοντες σε 200 φακέλους, ένας για κάθε ασθενή, με συνολικά 1000 αρχεία σε μορφή αρχείου nii.gz. Κάθε φάκελος περιέχει 5 αρχεία, 4 για κάθε μία από τις διαφορετικές MRI ακολουθίες (T1 w, T1 post-gadolinium, T2w, and FLAIR), και ένα αρχείο για την τμηματοποίηση (segmentation). Τοποθετούμε όλα τα δεδομένα σε έναν φάκελο της επιλογής μας (για εμάς ***Pretreat-MetsToBrain-Masks***) με σκοπό την αρχική εξερεύνηση και επεξεργασία των δεδομένων. Για εμάς τα δεδομένα στον συγκεκριμένο φάκελο θα αποτελούν την αρχική πηγή της αλήθειας.

4.1.3: Απεικόνιση και Επεξεργασία Δεδομένων

Για να απεικονίσουμε τα δεδομένα κάνουμε χρήση ενός 3D viewer λογισμικού. Αυτό μας βοηθάει να δούμε την συλλογή δεδομένων στην αρχική της μορφή και να εξετάσουμε γρήγορα ότι έχουμε ακριβώς τα δεδομένα που περιμέναμε, να καταλάβουμε καλύτερα τα διαφορετικά είδη δεδομένων και πιθανόν αν χρειάζεται να απορρίψουμε κάποια δείγματα που δεν θα μας βοηθήσουν στην ανάλυση μας. Υπάρχουν πολλές επιλογές λογισμικών απεικόνισης διαθέσιμες προς χρήση. Συγκεκριμένα στην δικιά μας περίπτωση έγινε χρήση των λογισμικών 3D Slicer και Fiji, λογισμικά που είναι ελεύθερα προς χρήση από την κοινότητα και επαρκούν για τον δικό μας σκοπό. Στην συνέχεια βλέπουμε 2 παραδείγματα για διαφορετικές απεικονίσεις στα λογισμικά που χρησιμοποιήθηκαν.



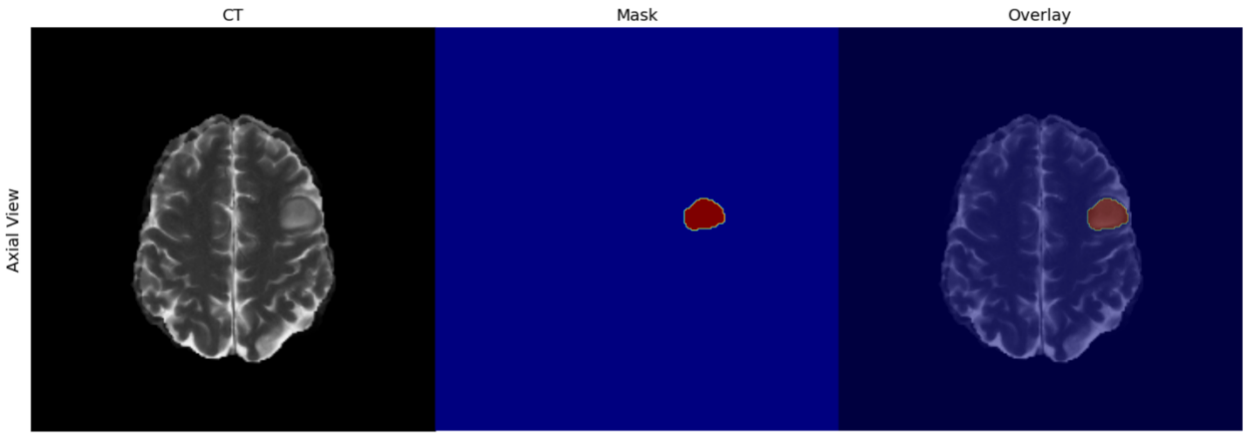
Εικόνα 20 – 3D Απεικόνιση NifTI αρχείου



Εικόνα 21 - Απεικόνιση στο λογισμικό Slicer

Πριν προχωρήσουμε στην επεξεργασία των δεδομένων και την εξαγωγή χαρακτηριστικών πρέπει να φέρουμε τα δεδομένα στην κατάλληλη μορφή και να κρατήσουμε αυτά που μας ενδιαφέρουν για την ανάλυση μας. Απαραίτητα στοιχεία για την εξαγωγή των χαρακτηριστικών είναι τουλάχιστον μία ακολουθία και η μάσκα της τμηματοποίησης. Όπως προαναφέραμε, έχουμε στην διάθεση μας 4 διαφορετικές ακολουθίες του μαγνητικού τομογράφου, T1 w, T1 post-gadolinium, T2w και FLAIR . Στην δική μας ανάλυση επιλέγουμε να εργαστούμε πάνω στην ακολουθία T2w. Επιπλέον, όπως είναι ορατό στην [Εικόνα 21] υπάρχουν δείγματα στην συλλογή μας με περισσότερες από μία κατηγορίες για την περιοχή ενδιαφέροντος (ROI). Για να επιτύχουμε την εξαγωγή των χαρακτηριστικών, ήταν απαραίτητο να συγχωνεύσουμε τις διαφορετικές κατηγορίες των ROI σε μία.

Σε αυτό το σημείο είναι σημαντικό να κάνουμε αναφορά στα εργαλεία που χρησιμοποιήθηκαν για την επίτευξη των βημάτων που περιεγράφηκαν παραπάνω. Ένα πολύ σημαντικό και ισχυρό εργαλείο είναι το simpleITK. Το SimpleITK είναι μια δωρεάν βιβλιοθήκη λογισμικού ανοικτού κώδικα που έχει σχεδιαστεί για να απλοποιεί την πρόσβαση στις εκτεταμένες δυνατότητες επεξεργασίας και ανάλυσης εικόνας του Insight Segmentation and Registration Toolkit (ITK). Προσφέρει μια φιλική προς την Python διεπαφή, γεγονός που την καθιστά δημοφιλή επιλογή για ερευνητές και προγραμματιστές που εργάζονται στην ανάλυση ιατρικών εικόνων και σε συναφείς τομείς. Το SimpleITK έχει πολλά χαρακτηριστικά που απλοποιούν τις λειτουργίες που απαιτούνται για την επεξεργασία ιατρικών εικόνων. Το SimpleITK διαβάζει και γράφει απρόσκοπτα μια μεγάλη ποικιλία μορφών ιατρικών εικόνων, συμπεριλαμβανομένων των NifTI, DICOM, NRRD, TIFF, κ.α. . Αυτή η ευελιξία μας επιτρέπει να εργαζόμαστε με δεδομένα από διάφορες πηγές. Επίσης παρέχει ένα πλούσιο σύνολο φίλτρων επεξεργασίας εικόνας για εργασίες όπως εξομάλυνση, μείωση θορύβου, καταχώρηση (registration), τμηματοποίηση (segmentation) και φιλτράρισμα. Επιπλέον το SimpleITK περιλαμβάνει βασικές δυνατότητες οπτικοποίησης. Μπορεί να εμφανίζει εικόνες μέσα στο περιβάλλον Python, αξιοποιώντας συχνά το Fiji (Fiji Is Just ImageJ) για διαδραστική προβολή. [46]



Εικόνα 22 - Οπτικοποίηση με το εργαλείο SimpleITK

4.1.4: Εξαγωγή Χαρακτηριστικών

Σε αυτό το κεφάλαιο, θα εξερευνήσουμε τη διαδικασία εξαγωγής χαρακτηριστικών ραδιομικής από τις τρισδιάστατες εικόνες όγκων εγκεφάλου NifTI και τις αντίστοιχες τμηματοποιήσεις τους χρησιμοποιώντας την εργαλειοθήκη PyRadiomics για την γλώσσα προγραμματισμού Python. Αυτά θα είναι τα χαρακτηριστικά που θα μας χρησιμεύσουν ως βάση για τη δημιουργία των μοντέλων μηχανικής μάθησης για την πρόβλεψη των αποτελεσμάτων ως προς την μεταβλητή στόχο, που στην περίπτωση μας είναι η επιβίωση του ασθενή.

Το PyRadiomics είναι ένα πακέτο ανοικτού κώδικα για την γλώσσα Python που έχει σχεδιαστεί ειδικά για την εξαγωγή ευρέος φάσματος ποσοτικών χαρακτηριστικών από ιατρικές εικόνες. Προσφέρει μια τυποποιημένη προσέγγιση για την ανάλυση των radiomics, εξασφαλίζοντας την αναπαραγωγιμότητα και διευκολύνοντας τις συγκρίσεις μεταξύ μελετών. Το PyRadiomics υποστηρίζει την εξαγωγή χαρακτηριστικών τόσο από 2D όσο και από 3D εικόνες. Επιτρέπει τον υπολογισμό χαρακτηριστικών για μια καθορισμένη περιοχή ενδιαφέροντος (ROI) εντός της εικόνας, που συνήθως αντιστοιχεί στην κατάτμηση του όγκου, ή για ολόκληρο τον όγκο της εικόνας.

Το PyRadiomics προσφέρει ένα ολοκληρωμένο σύνολο κλάσεων χαρακτηριστικών που περιλαμβάνουν διάφορες πτυχές της υφής της εικόνας, της κατανομής της έντασης και των χαρακτηριστικών σχήματος. Αναλυτικά ορισμένες βασικές κατηγορίες χαρακτηριστικών είναι: Τα χαρακτηριστικά πρώτης τάξης που αποτυπώνουν τη συνολική κατανομή της έντασης εντός της ROI, όπως ο μέσος όρος, το ελάχιστο, το μέγιστο και η τυπική απόκλιση. Τα χαρακτηριστικά με βάση το σχήμα, τα οποία ποσοτικοποιούν τις γεωμετρικές ιδιότητες της ROI, συμπεριλαμβανομένου του όγκου, της σφαιρικότητας και της επιφάνειας. Τα χαρακτηριστικά υφής, όπου περιγράφουν τη χωρική κατανομή των

εντάσεων εντός της ROI. Δημοφιλή παραδείγματα περιλαμβάνουν τα χαρακτηριστικά Haralick, τα χαρακτηριστικά GLCM (gray level co-occurrence matrix) και τα χαρακτηριστικά Gabor. Τέλος τα χαρακτηριστικά ανώτερης τάξης, τα οποία αποτυπώνουν πιο σύνθετες σχέσεις μεταξύ των εντάσεων των voxel, όπως η εντροπία, η λοξότητα και η κύρτωση. Μία ακόμα δυνατότητα από το συγκεκριμένο εργαλείο είναι να παρέχει λειτουργίες για τον υπολογισμό παράγωγων εικόνων από τα αρχικά δεδομένα εικόνας. Αυτές οι παράγωγες εικόνες δημιουργούνται με την εφαρμογή διαφόρων φίλτρων (π.χ. Laplacian of Gaussian, Wavelet decomposition) που μπορούν να αναδείξουν συγκεκριμένα χαρακτηριστικά της εικόνας και να οδηγήσουν στην εξαγωγή πρόσθετων χαρακτηριστικών.

Μερικά από τα πολλά πλεονεκτήματα από τη χρήση του PyRadiomics για την εξαγωγή χαρακτηριστικών είναι η τυποποιημένη προσέγγιση για τον υπολογισμό των χαρακτηριστικών, διασφαλίζοντας τη συνέπεια και διευκολύνοντας τις συγκρίσεις με άλλες μελέτες, η πλήρης κοινή χρήση του κώδικα και την αναπαραγωγή της διαδικασίας εξαγωγής χαρακτηριστικών, το ευρύ φάσμα καθιερωμένων χαρακτηριστικών ραδιομικής, μειώνοντας την ανάγκη για χειροκίνητη μηχανική των χαρακτηριστικών και τέλος η ευκολία χρήσης που απλοποιεί τη ροή εργασίας, καθιστώντας το εργαλείο προσιτό σε ερευνητές με διαφορετικά επίπεδα εμπειρίας στον προγραμματισμό.

Η διαδικασία εξαγωγής χαρακτηριστικών με το PyRadiomics μπορεί να συνοψιστεί στα ακόλουθα βήματα:

Φόρτωση εικόνων και τμηματοποίηση: Το PyRadiomics υποστηρίζει τη φόρτωση διαφόρων μορφών ιατρικών εικόνων, συμπεριλαμβανομένου του NifTI. Θα χρειαστεί να φαρτώσουμε τόσο την τρισδιάστατη εικόνα του όγκου του εγκεφάλου όσο και την αντίστοιχη μάσκα τμηματοποίησης.

Καθορισμός της περιοχής ενδιαφέροντος (ROI): Η μάσκα τμηματοποίησης ορίζει την ROI, που συνήθως αντιπροσωπεύει την περιοχή του όγκου.

Ορισμός παραμέτρων εξαγωγής χαρακτηριστικών: Το PyRadiomics επιτρέπει την προσαρμογή της εξαγωγής χαρακτηριστικών μέσω διαφόρων παραμέτρων. Μπορούμε να καθορίσουμε τις επιθυμητές κλάσεις χαρακτηριστικών, τις επιλογές φιλτραρίσματος και τις μεθόδους επαναδειγματοληψίας.

Υπολογισμός χαρακτηριστικών: Χρησιμοποιώντας τη φορτωμένη εικόνα, τη μάσκα και τις επιλεγμένες παραμέτρους, το PyRadiomics υπολογίζει τα καθορισμένα ακτινολογικά χαρακτηριστικά.

Έξοδος χαρακτηριστικών: Το PyRadiomics παρέχει τα εξαγόμενα χαρακτηριστικά ως λεξικό Python ή πλαίσιο δεδομένων pandas, επιτρέποντας την εύκολη ενσωμάτωση με αγωγούς μηχανικής μάθησης.

Συγκεκριμένα για την δικιά μας εφαρμογή, για να εξάγουμε τα επιθυμητά χαρακτηριστικά ραδιομικής από τις τρισδιάστατες εικόνες NifTI με την βοήθεια του εργαλείου pyradiomics, δημιουργούμε

ένα αρχείο τύπου yaml όπου εκεί ορίζουμε όλες τις ρυθμίσεις για την εξαγωγή των χαρακτηριστικών. Υπάρχουν 4 βασικές κατηγορίες ρυθμίσεων για την εξαγωγή:

1-Image Type: Καθορισμός των τύπων εικόνων (πρωτότυπων/προερχόμενων) που θα χρησιμοποιηθούν για την εξαγωγή χαρακτηριστικών από

2-Feature Class: Καθορίζοντας ποιο χαρακτηριστικό (κλάση) θα εξαχθεί

3-Setting: Καθορισμός ρυθμίσεων, οι οποίες ελέγχουν την προεπεξεργασία και προσαρμόζουν τη συμπεριφορά των ενεργοποιημένων φίλτρων και κατηγοριών χαρακτηριστικών.

4-Voxel Based: Καθορισμός των συγκεκριμένων ρυθμίσεων που βασίζονται σε voxel, οι οποίες χρειάζονται μόνο όταν χρησιμοποιείτε το PyRadiomics για τη δημιουργία χαρτών χαρακτηριστικών. Στην δικιά μας περίπτωση αυτή η ρύθμιση δεν θα χρησιμοποιηθεί.

Παρακάτω παρουσιάζονται τα περιεχόμενα των ρυθμίσεων που χρησιμοποιήθηκαν για την εφαρμογή μας.

Image Type

```
v2 > params > ! params_1.yaml
1   # ############################  Extracted using PyRadiomics version: 1  ####################################
2
3   imageType:
4     Original: {}
5     LoG:
6       # If you include sigma values >5, remember to also increase the padDistance. Because of resampling to (2, 2, 2), the
7       # use of sigmas < 2 mm is not recommended.
8       sigma: [2.0, 3.0, 4.0, 5.0]
9     Wavelet: {}
10
```

Εικόνα 23 - Ρύθμιση Εξαγωγής Χαρακτηριστικών - Image Type

Αυτοί είναι οι τύποι εικόνων (είτε η αρχική εικόνα είτε οι παράγωγες εικόνες που χρησιμοποιούν φίλτρα) που μπορούν να χρησιμοποιηθούν για την εξαγωγή χαρακτηριστικών από. Στη ρύθμιση που ορίζουμε εδώ για τον τύπο της εικόνας, είναι ενεργοποιημένοι 3 τύποι εικόνας ("Original", "LoG" (Laplacian of Gaussian) και "Wavelet"), με προσαρμοσμένες ρυθμίσεις που καθορίζονται για "LoG" ("sigma") .

Feature Class

Οι κατηγορίες χαρακτηριστικών με τις επιμέρους τάξεις είναι:

- `First Order Statistics` (19 features)
- `Shape-based (3D)` (16 features)
- `Shape-based (2D)` (10 features)
- `Gray Level Co-occurrence Matrix` (24 features)
- `Gray Level Run Length Matrix` (16 features)
- `Gray Level Size Zone Matrix` (16 features)
- `Neighbouring Gray Tone Difference Matrix` (5 features)
- `Gray Level Dependence Matrix` (14 features)

Εικόνα 24 – Κατηγορίες χαρακτηριστικών και πλήθος των επιμέρους τάξεων

v2 > params > ! params_1.yaml

```yaml
  featureClass:
    # redundant Compactness 1, Compactness 2 an Spherical Disproportion features are disabled by default, they can be
    # enabled by specifying individual feature names (as is done for glcm) and including them in the list.
    firstorder:
    shape:
    glcm:  # Disable SumAverage and Maximal Correlation Coefficient by specifying all other GLCM features available
      - 'Autocorrelation'
      - 'JointAverage'
      - 'ClusterProminence'
      - 'ClusterShade'
      - 'ClusterTendency'
      - 'Contrast'
      - 'Correlation'
      - 'DifferenceAverage'
      - 'DifferenceEntropy'
      - 'DifferenceVariance'
      - 'JointEnergy'
      - 'JointEntropy'
      - 'Imc1'
      - 'Imc2'
      - 'Idm'
      - 'Idmn'
      - 'Id'
      - 'Idn'
      - 'InverseVariance'
      - 'MaximumProbability'
      - 'SumEntropy'
      - 'SumSquares'
    glrlm:
    glszm:
    gldm:
    ngtdm:
```

Εικόνα 25 - Ρύθμιση Εξαγωγής Χαρακτηριστικών - Feature Class

Αυτά είναι τα χαρακτηριστικά που εξάγονται από κάθε (πρωτότυπο και/ή προερχόμενο) τύπο εικόνας. Τα διαθέσιμα χαρακτηριστικά καθορίζονται δυναμικά και ταξινομούνται σε κατηγορίες χαρακτηριστικών. Στη συνέχεια, ορίζονται 7 κατηγορίες χαρακτηριστικών. Τα "firstorder", "shape", "glcm", "glrlm", "glszm" kai "ngtdm" είναι ενεργοποιημένα με όλες τις πιθανές τάξεις στην αντίστοιχη κλάση, ενώ μόνο το "Maximum Correlation Coefficient" και το "SumAverage" δεν είναι ενεργοποιημένα για το "glcm". Ορίζοντας την κατηγορία χαρακτηριστικών χωρίς να αναφέρουμε τις επιμέρους τάξεις, επιλέγουμε όλες τις επιλογές.

Setting

v2 > params > ! params_1.yaml

```yaml
setting:
  # Normalization:
  # MR signal is usually relative, with large differences between scanners and vendors. By normalizing the image before
  # feature calculation, this confounding effect may be reduced. However, if only one specific scanner is used, or the
  # images reflect some absolute world value (e.g. ADC maps, T2maps (NOT T2 weighted)), consider disabling the
  # normalization.
  normalize: true
  normalizeScale: 100  # This allows you to use more or less the same bin width.

  # Resampling:
  # If slices are very thin (~1mm), such as in 3D scanned (isotropic) volumes, resampledPixelSpacing may be reduced to
  # (1, 1, 1). Furthermore, in case of isotropic volumes, consider disabling resampling.
  # On a side note: increasing the resampled spacing forces PyRadiomics to look at more coarse textures, which may or
  # may not increase accuracy and stability of your extracted features.
  interpolator: 'sitkBSpline'
  resampledPixelSpacing: [2, 2, 2]

  # Mask validation:
  # correctMask and geometryTolerance are not needed, as both image and mask are resampled, if you expect very small
  # masks, consider to enable a size constraint by uncommenting settings below:
  #minimumROIDimensions: 2
  #minimumROISize: 50

  # Image discretization:
  # The ideal number of bins is somewhere in the order of 16~128 bins. A possible way to define a good binwidt is to
  # extract firstorder:Range from the dataset to analyze, and choose a binwidth so, that range/binwidth remains approximately
  # in this range of bins.
  binWidth: 5

  # Misc:
  # default label value. Labels can also be defined in the call to featureextractor.execute, as a commandline argument,
  # or in a column "Label" in the input csv (batchprocessing)
  label: 1
```

Εικόνα 26 - Ρύθμιση Εξαγωγής Χαρακτηριστικών - Setting

Εκτός από τις ρυθμίσεις για το τι πρέπει να εξαχθεί (τύποι εικόνων, χαρακτηριστικά), το PyRadiomics διαθέτει διάφορες ρυθμίσεις προσαρμόζοντας τον τρόπο εξαγωγής των χαρακτηριστικών. Αυτές οι ρυθμίσεις λειτουργούν σε διαφορετικά επίπεδα. Π.χ. Η επαναδειγματοληψία πραγματοποιείται αμέσως μετά τη φόρτωση των εικόνων (στον εξαγωγέα χαρακτηριστικών), επομένως οι ρυθμίσεις που ελέγχουν την επαναδειγματοληψία λειτουργούν μόνο στο επίπεδο εξαγωγής χαρακτηριστικών. Τέλος για

την δικιά μας εφαρμογή, καθορίζονται 5 ρυθμίσεις: "normalize", "normalizeScale", "interpolator", "resampledPixelSpacing", "binWidth" και "label". Για περισσότερες πληροφορίες σχετικά με την επιλογή των ρυθμίσεων και των παραμέτρων, μπορούμε να βρούμε στην αναφορά [50] στο σχετικό κομμάτι: Customizing the Extraction.

Έχοντας ορίσει όλες τις ρυθμίσεις και τις παραμέτρους για την εξαγωγή δεδομένων, αρκεί τώρα να κάνουμε χρήση του εξαγωγέα χαρακτηριστικών του εργαλείου PyRadiomics. Έτσι για κάθε ζευγάρι τρισδιάστατων εικόνων και της αντίστοιχης τμηματοποίησης, παράγουμε τα χαρακτηριστικά. Το αποτέλεσμα για όλα τα δείγματα αποθηκεύεται σε ένα πίνακα, όπως φαίνεται παρακάτω:

	original_shape_Elongation	original_shape_Flatness	original_shape_LeastAxisLength	original_shape_MajorAxisLength	original_shape_Maximum2DDiameterColumn	original_shape_Maximum2DDiameterRow	original_shape_Maximum2DDiameterSlice	original_shape_Maximum3DDiameter	original_sh
0	0.743201	0.342911	14.042317	40.950293	34.928498	40.249224	34.928498	42.379240	
1	0.786585	0.470726	39.115621	83.096308	76.026311	86.556340	94.762862	95.540567	
2	0.804832	0.676209	50.273607	74.346296	75.286121	92.541882	95.036835	96.726418	
3	0.873655	0.642660	8.244739	12.829073	12.806248	14.422205	12.649111	15.620499	
4	0.755626	0.352583	39.270515	111.379385	104.690019	112.445542	105.848949	119.431989	
...	...	...	...	...	...	...	...	...	
195	0.666965	0.401032	58.450399	145.749926	128.062485	147.959454	154.983870	164.401946	
196	0.775088	0.509471	21.146181	41.506154	38.418745	45.650849	42.190046	48.785244	
197	0.273180	0.166535	27.351589	164.239088	116.017240	75.073298	72.801099	118.473626	
198	0.752908	0.517420	38.664430	74.725478	76.941536	88.814413	74.966659	88.904443	
199	0.260107	0.156114	5.780493	37.027325	18.439089	10.770330	8.944272	33.823069	

200 rows × 1215 columns

Εικόνα 27 - Χαρακτηριστικά Ραδιομικής για Όλα τα Δείγματα

4.2: Διερευνητική Ανάλυση και Προεπεξεργασία Χαρακτηριστικών

4.2.1: Χωρισμός σε Training και Test Set

Στη μηχανική μάθηση, τα σετ εκπαίδευσης (training set) και δοκιμών (testing set) είναι θεμελιώδη στοιχεία για τη δημιουργία ισχυρών και γενικεύσιμων μοντέλων. Αυτά τα σύνολα δεδομένων εξυπηρετούν διαφορετικούς σκοπούς στη διαδικασία ανάπτυξης του μοντέλου, οδηγώντας τελικά σε ένα μοντέλο που μπορεί να αποδώσει αποτελεσματικά σε άγνωστα δεδομένα. Ο πρωταρχικός λόγος για τη χρήση ξεχωριστών συνόλων εκπαίδευσης και δοκιμών είναι η αξιολόγηση της ικανότητας ενός μοντέλου να γενικεύει (generalization) σε νέα δεδομένα. Επίσης το σετ δοκιμών βοηθά στον εντοπισμό της υπερπροσαρμογής (overfitting) εκθέτοντας το μοντέλο σε δεδομένα που δεν έχει δει στο παρελθόν.

Το αρχικό σύνολο δεδομένων χωρίζεται συνήθως σε δύο υποσύνολα: το σύνολο εκπαίδευσης (συνήθως μεγαλύτερο, συχνά 70% έως 90% των δεδομένων) και το σύνολο δοκιμής (μικρότερο, συχνά 10% έως 30% των δεδομένων τα δεδομένα). Για τη δημιουργία αντιπροσωπευτικών υποσυνόλων χρησιμοποιούνται τεχνικές όπως η τυχαία δειγματοληψία (random sampling) ή η στρωματοποιημένη δειγματοληψία (stratified sampling) (εξασφαλίζοντας ότι το σύνολο δοκιμής αντικατοπτρίζει την κατανομή κλάσης των αρχικών δεδομένων).

Το σετ εκπαίδευσης χρησιμοποιείται για την εκπαίδευση του μοντέλου μηχανικής εκμάθησης. Το μοντέλο μαθαίνει μοτίβα και σχέσεις μέσα στα δεδομένα, προσαρμόζοντας τις εσωτερικές του

παραμέτρους για να ελαχιστοποιήσει μια συνάρτηση απώλειας. Αυτή η διαδικασία επαναλαμβάνεται μέχρι να συγκλίνει το μοντέλο.

Αφού εκπαιδευτεί, η απόδοση του μοντέλου αξιολογείται στο σετ δοκιμής. Το μοντέλο κάνει προβλέψεις στα σημεία δεδομένων δοκιμής και αυτές οι προβλέψεις συγκρίνονται με τις πραγματικές ετικέτες για τον υπολογισμό των μετρήσεων αξιολόγησης. Η απόδοση του συνόλου δοκιμής αντικατοπτρίζει τη γενίκευση και την ικανότητα του μοντέλου να χειρίζεται αόρατα δεδομένα.

Συγκεκριμένα στην δικιά μας εφαρμογή, έγινε διαχωρισμός με αναλογία 70% για το σετ εκπαίδευσης και 30% για το σετ δοκιμής. Κατά τον διαχωρισμό διατηρήθηκε η αναλογία των 2 κλάσεων (Death Yes / No) και στα δύο σετ.

4.2.2: Αφαίρεση Ακραίων Τιμών

Η παρουσία ακραίων τιμών στα χαρακτηριστικά radiomics μπορεί να επηρεάσει σημαντικά την απόδοση των μοντέλων μηχανικής μάθησης που χρησιμοποιούνται για εργασίες όπως η πρόβλεψη των αποτελεσμάτων των ασθενών. Σε αυτή την ενότητα αναλύεται η λογική πίσω από την αφαίρεση των ακραίων τιμών και διερευνώνται τρεις κοινές μέθοδοι για την ανίχνευση ακραίων τιμών στο πλαίσιο της ανάλυσης χαρακτηριστικών.

Οι ακραίες τιμές είναι σημεία δεδομένων που αποκλίνουν σημαντικά από την πλειονότητα των δεδομένων. Στο πλαίσιο της ραδιομικής, οι ακραίες τιμές στα εξαγόμενα χαρακτηριστικά μπορεί να προκύψουν πιθανόν λόγω κάποιου προβλήματος κατά την απόκτηση της εικόνας, όπως η δυσλειτουργία του σαρωτή ή κίνηση του ασθενούς, μπορούν να εισάγουν μη φυσιολογικές τιμές έντασης. Ακόμα μία περίπτωση είναι οι ατελείς μάσκες τμηματοποίησης που μπορεί να περιλαμβάνουν θόρυβο ή να αποκλείουν σχετικές περιοχές όγκου, οδηγώντας σε χαρακτηριστικά που δεν αντιπροσωπεύουν την πραγματική βιολογία του όγκου.

Η παρουσία ακραίων τιμών μπορεί να επηρεάσει αρνητικά τα μοντέλα μηχανικής μάθησης με διάφορους τρόπους όπως οι ανακριβείς προβλέψεις λόγω της στρέβλωσης κατά την διαδικασία μάθησης του μοντέλου. Επίσης οι ακραίες τιμές μπορεί να καταστήσουν το μοντέλο υπερβολικά ευαίσθητο σε συγκεκριμένα σημεία δεδομένων, με αποτέλεσμα ασταθείς προβλέψεις. Τέλος οι ακραίες τιμές μπορεί να παρεμποδίσουν την ερμηνευσιμότητα του μοντέλου αποκρύπτοντας τις υποκείμενες σχέσεις μεταξύ χαρακτηριστικών και αποτελεσμάτων.Ως εκ τούτου, η αφαίρεση των ακραίων τιμών είναι ένα κρίσιμο βήμα στην προεπεξεργασία των χαρακτηριστικών της ραδιομικής για να διασφαλιστεί η ευρωστία και η δυνατότητα γενίκευσης της επακόλουθης ανάλυσης μηχανικής μάθησης.

Για την δικιά μας εφαρμογή, διερευνούμε τρεις ευρέως χρησιμοποιούμενες μεθόδους ανίχνευσης ακραίων τιμών κατάλληλες για την ανάλυση των χαρακτηριστικών μας. Οι μέθοδοι είναι:

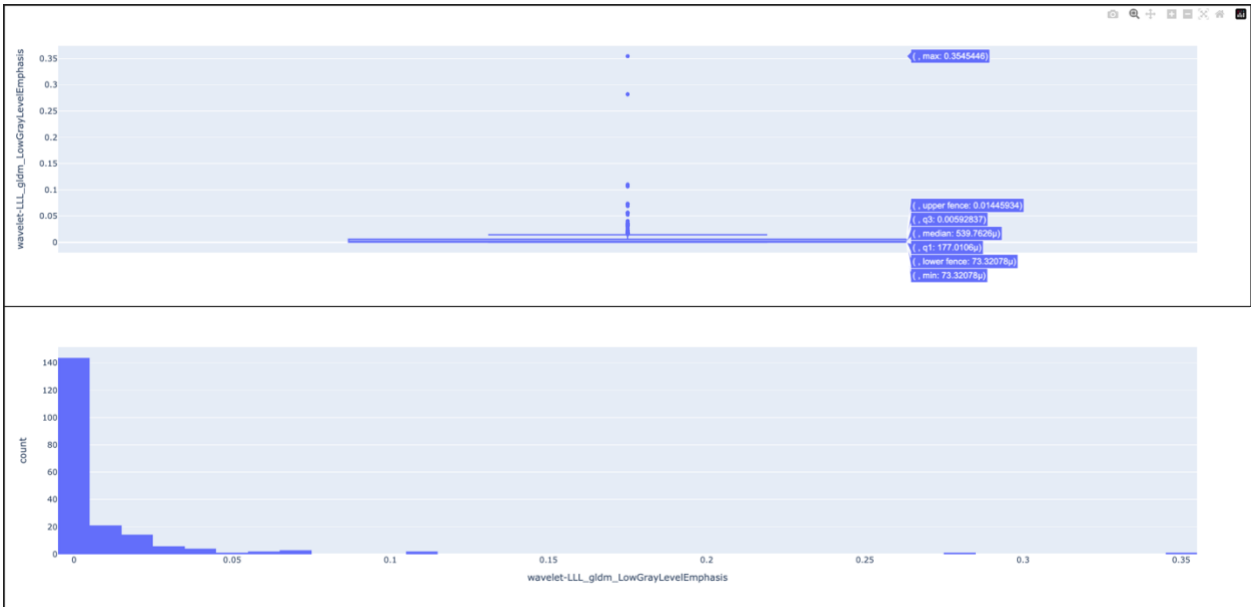
Isolation Forest(IF): Αυτή η μέθοδος χωρίς επίβλεψη απομονώνει τις ακραίες τιμές με τυχαίο διαμερισμό των δεδομένων σε υποδιαστήματα. Τα σημεία που απαιτούν λιγότερες κατατμήσεις για να απομονωθούν θεωρούνται ακραία. Η IF είναι αποτελεσματική για μεγάλα σύνολα δεδομένων και ανθεκτική σε άσχετα χαρακτηριστικά.

Elliptic Envelope (EE): Η μέθοδος αυτή υποθέτει ότι τα δεδομένα των παρεκκλίσεων ακολουθούν κατανομή Gauss και ορίζει ένα ελλειψοειδές όριο γύρω από τα δεδομένα. Τα σημεία που βρίσκονται εκτός του ορίου θεωρούνται ακραία. Η EE είναι κατάλληλη για δεδομένα υψηλών διαστάσεων, αλλά ευαίσθητη σε ασύμμετρες κατανομές.

Local Outlier Factor (LOF): Αυτή η μέθοδος συγκρίνει την τοπική πυκνότητα ενός σημείου δεδομένων με τα γειτονικά του σημεία. Σημεία με σημαντικά χαμηλότερη τοπική πυκνότητα επισημαίνονται ως ακραία σημεία. Ο LOF είναι αποτελεσματικός στον εντοπισμό ακραίων τιμών σε συστάδες, αλλά είναι υπολογιστικά δαπανηρός για μεγάλα σύνολα δεδομένων.

Εφαρμόζοντας τεχνικές εντοπισμού ακραίων τιμών όπως αυτές που περιγράφονται παραπάνω, μπορούμε να διασφαλίσουμε την ποιότητα και την αξιοπιστία των χαρακτηριστικών μας, οδηγώντας τελικά σε καλύτερη απόδοση και ερμηνευσιμότητα των μοντέλων μηχανικής μάθησης στην πρόβλεψη των αποτελεσμάτων των ασθενών.

Ένα παράδειγμα ακραίων τιμών στην δικιά μας συλλογή χαρακτηριστικών παρουσιάζεται παρακάτω και συγκεκριμένα για το χαρακτηριστικό ***wavelet-LLL-gldm_LowGrayLevelEmphasis*** βλέπουμε το boxplot και το ιστόγραμμα:



Εικόνα 28 - Οπτικοποίηση Ακραίων Τιμών με Boxplot & Histogram

4.2.3: Κανονικοποίηση Δεδομένων

Με αυτή την τεχνική ελέγχουμε τον τρόπο με τον οποίο κλιμακώνονται τα αριθμητικά χαρακτηριστικά στο σύνολο δεδομένων σας κατά την προεπεξεργασία δεδομένων. Η κανονικοποίηση στοχεύει να φέρει τα χαρακτηριστικά σε παρόμοιο εύρος, αποτρέποντας χαρακτηριστικά με μεγαλύτερη κλίμακα από το να κυριαρχούν στη διαδικασία εκμάθησης του μοντέλου. Αυτό μπορεί να βελτιώσει την απόδοση ορισμένων αλγορίθμων μηχανικής εκμάθησης.

Επιλέξαμε να εξετάσουμε 4 τεχνικές, που δυνητικά μπορούν να προσφέρουν καλύτερα αποτελέσματα:

- Z-score: Αυτή είναι η τυπική κανονικοποίηση της βαθμολογίας z. Αφαιρεί τον μέσο όρο κάθε χαρακτηριστικού από τις τιμές του και στη συνέχεια διαιρεί με την τυπική απόκλιση. Τα προκύπτοντα κανονικοποιημένα χαρακτηριστικά έχουν μέσο όρο 0 και τυπική απόκλιση 1. Αυτή είναι μια κοινή και αποτελεσματική επιλογή για πολλές εφαρμογές μηχανικής εκμάθησης.

- Min-Max: Η κανονικοποίηση Min-max κλιμακώνει κάθε χαρακτηριστικό ξεχωριστά σε ένα εύρος από 0 έως 1. Αυτό μπορεί να είναι χρήσιμο όταν θέλουμε όλα τα χαρακτηριστικά να συνεισφέρουν εξίσου, ανεξάρτητα από τις αρχικές τους μονάδες.

- Max-absolute: Η κανονικοποίηση της μέγιστης απόλυτης τιμής κλιμακώνει και μεταφράζει κάθε χαρακτηριστικό έτσι ώστε η μέγιστη απόλυτη τιμή κάθε χαρακτηριστικού να γίνεται 1,0. Αυτό είναι παρόμοιο με την κανονικοποίηση min-max, αλλά χρησιμοποιεί την απόλυτη μέγιστη τιμή αντί για το εύρος.

- Robust normalization: Αυτή η μέθοδος πραγματοποιεί μία εύρωστη (robust) κανονικοποίηση βασισμένη στο Interquartile Range (IQR). Για κάθε χαρακτηριστικό υπολογίζεται το IQR, δηλαδη η διαφορά μεταξύ του 75^{ου} και 25^{ου} εκατοστημόριου (percentile), στην συνέχεια εντοπίζει τις ακραίες τιμές, οι οποίες είναι σημεία που βρίσκονται εκτός ενός εύρους βασισμένο στο IQR (τυπικά αυτό είναι 1,5 φορά τις τιμής του IQR κάτω από το 25^{ο} εκατοστημόριο ή πάνω το 75^{ο} εκατοστημόριο). Τέλος κλιμακώνει τα χαρακτηριστικά χρησιμοποιώντας μία παραλλαγή του Min-Max, με στόχο την μείωση της επίδρασης των ακραίων τιμών.

Η κανονικοποίηση δεν βελτιώνει απαραίτητα την απόδοση του μοντέλου για όλους τους αλγόριθμους. Ορισμένοι αλγόριθμοι (π.χ. δέντρα αποφάσεων) είναι λιγότερο ευαίσθητοι στην κλιμάκωση χαρακτηριστικών. Έτσι για κάθε ξεχωριστό πρόβλημα η επιλογή της τεχνικής πρέπει να προσαρμόζεται στην απαιτήσεις και την φύση της εργασίας. Συγκεκριμένα στο δικό μας πρόβλημα, μετά από συγκριτική μελέτη, έγινε χρήση της μεθόδου Z-score καθώς σε σχέση με τις υπόλοιπες δεν υπήρχε κάποια σημαντική διαφορά στο αποτέλεσμα, η μέθοδος αυτή είναι πιο εύκολα ερμηνεύσιμη και φέρει λιγότερη πολυπλοκότητα στο συνολικό σύστημα μας.

4.2.4: Εκκαθάριση Συγγραμικότητας

Ένα άλλο κρίσιμο βήμα περιλαμβάνει την αντιμετώπιση της συγγραμικότητας και της συσχέτισης μεταξύ των χαρακτηριστικών. Αυτή η ενότητα διερευνά την έννοια της συγγραμικότητας και της συσχέτισης, τον αντίκτυπό τους στα μοντέλα μηχανικής μάθησης και παρουσιάζει τη μέθοδο ελάχιστης απόλυτης συσχέτισης Pearson (MAPC) για την επιλογή χαρακτηριστικών.

Κατά την εξαγωγή χαρακτηριστικών από ιατρικές εικόνες, καταλήγουμε να έχουμε έναν πολύ μεγάλο αριθμό χαρακτηριστικών. Ενώ αυτή η ολοκληρωμένη ανάλυση είναι πολύτιμη, τα χαρακτηριστικά μπορεί να παρουσιάζουν υψηλή συγγραμικότητα ή συσχέτιση. Η συγγραμικότητα ή συσχέτιση συμβαίνει όταν δύο ή περισσότερα χαρακτηριστικά περιέχουν σχεδόν πανομοιότυπες πληροφορίες. Με απλούστερους όρους, ένα χαρακτηριστικό μπορεί να προβλεφθεί τέλεια από ένα άλλο. Συγκεκριμένα αυτό αναφέρεται σε μια στατιστική σχέση μεταξύ δύο χαρακτηριστικών. Δείχνει το βαθμό στον οποίο τα χαρακτηριστικά κινούνται μαζί, είτε θετικά (αύξηση μαζί) είτε αρνητικά (αντίθετες τάσεις).

Η παρουσία συγγραμικότητας και συσχέτισης στα ραδιομετρικά χαρακτηριστικά μπορεί να επηρεάσει αρνητικά τα μοντέλα μηχανικής μάθησης με διάφορους τρόπους, όπως η αυξημένη διακύμανση και η δυσκολία στην ερμηνεία του μοντέλου και η υπολογιστική αναποτελεσματικότητα. Επομένως, η αντιμετώπιση της συγγραμικότητας και της συσχέτισης με την επιλογή ενός μη πλεονάζοντος συνόλου χαρακτηριστικών είναι απαραίτητη για την οικοδόμηση αξιόπιστων και ερμηνεύσιμων μοντέλων μηχανικής μάθησης.

Μια κοινή προσέγγιση για τον χειρισμό της συγγραμικότητας και της συσχέτισης είναι η χρήση χαρακτηριστικών ενός υποσυνόλου πληροφοριακών και μη πλεονασματικών χαρακτηριστικών. Εδώ, θα συζητήσουμε τη μέθοδο ελάχιστης απόλυτης συσχέτισης Pearson (MAPC).

Η μέθοδος MAPC λειτουργεί ως εξής:

- Υπολογίζει τον συντελεστή συσχέτισης Pearson μεταξύ όλων των πιθανών ζευγών χαρακτηριστικών. Ο συντελεστής συσχέτισης Pearson κυμαίνεται από -1 (τέλεια αρνητική συσχέτιση) έως +1 (τέλεια θετική συσχέτιση), ενώ το 0 υποδηλώνει μηδενική συσχέτιση.
- Λαμβάνει την απόλυτη τιμή των συντελεστών συσχέτισης για να αφαιρέσει την διεύθυνση της σχέσης (θετική ή αρνητική) από την ανάλυση. Μας ενδιαφέρει μόνο η ισχύς της σχέσης.
- Εντοπίζει τα χαρακτηριστικά με τους υψηλότερους απόλυτους συντελεστές συσχέτισης (δηλ. με τη μεγαλύτερη συσχέτιση).
- Από αυτά τα χαρακτηριστικά με υψηλή συσχέτιση, διατηρεί μόνο ένα χαρακτηριστικό, συνήθως αυτό με την υψηλότερη διακύμανση (αποτυπώνει τις περισσότερες πληροφορίες) ή αυτό που είναι πιο σχετικό με το βιολογικό πλαίσιο της μελέτης σας.

- Αυτή η διαδικασία επαναλαμβάνεται επαναληπτικά, αφαιρώντας τα χαρακτηριστικά με υψηλή συσχέτιση μέχρι να επιτευχθεί ένα προκαθορισμένο κατώφλι μέγιστης συσχέτισης.

Ο υπολογισμός του συντελεστή συσχέτισης Pearson υπολογίζεται σύμφωνα με την παρακάτω φόρμουλα.

$$r = \frac{cov_{xy}}{s_x s_y} = \frac{\sum(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2\,(y_i - \bar{y})^2}}$$

Όπου:

- r ο συντελεστής συσχέτισης Pearson
- x και y είναι τα 2 χαρακτηριστικά
- cov(x,y) είναι η συνδιακύμανση του x και y
- s_x είναι η τυπική απόκλιση του x
- s_y είναι η τυπική απόκλιση του y
- x_i είναι η κάθε παρατήρηση του x
- y_i είναι η κάθε παρατήρηση του y
- $\bar{x}$ είναι η μέση τιμή του x
- $\bar{y}$ είναι η μέση τιμή του y

4.2.5: Μείωση Διαστατικότητας

Στις προηγούμενες ενότητες συζητήθηκαν στρατηγικές για τον χειρισμό των ακραίων τιμών και των συσχετισμένων χαρακτηριστικών στην προεπεξεργασία χαρακτηριστικών ραδιομικής. Το παρόν κεφάλαιο διερευνά τεχνικές μείωσης της διαστατικότητας, εστιάζοντας ειδικότερα στην ανάλυση κύριων συνιστωσών (PCA) με πυρήνες γραμμικής και ακτινικής συνάρτησης βάσης (RBF).

Όταν εξάγουμε ακτινογνωστικά χαρακτηριστικά προερχόμενα από ιατρικές εικόνες, μπορεί να χρειάζεται να διαχειριστούμε έναν μεγάλο αριθμό χαρακτηριστικών. Ενώ αυτή η ολοκληρωμένη ανάλυση προσφέρει ένα πλούσιο σύνολο πληροφοριών και χαρακτηριστικών, μπορεί επίσης να παρουσιάσει μια πρόκληση γνωστή ως "κατάρα της διαστατικότητας". Αυτή αναφέρεται στο φαινόμενο όπου τα μοντέλα μηχανικής μάθησης δυσκολεύονται με δεδομένα υψηλής διάστασης, οδηγώντας σε αυξημένο κόστος όπου απαιτούνται σημαντικοί υπολογιστικοί πόρους και χρόνος, σε υπερβολική ευαισθησία στα δεδομένα εκπαίδευσης με αποτέλεσμα την υπερπροσαρμογή και επίσης την δυσκολία κατανόησης και ερμηνείας του μοντέλου. Οι τεχνικές μείωσης διαστάσεων αντιμετωπίζουν αυτά τα ζητήματα μετασχηματίζοντας τον χώρο των χαρακτηριστικών υψηλής διάστασης σε έναν υποχώρο χαμηλότερης διάστασης, διατηρώντας παράλληλα τις πιο σχετικές πληροφορίες. Αυτό μας δίνει την δυνατότητα να έχουμε βελτιωμένη απόδοση

μοντέλου με ταχύτερους χρόνους εκπαίδευσης και ενδεχομένως καλύτερη γενίκευση με τον μετριασμό της υπερπροσαρμογής.

Η PCA είναι μια δημοφιλής και αποτελεσματική τεχνική για τη μείωση της διαστατικότητας. Μετατρέπει τον αρχικό χώρο χαρακτηριστικών υψηλών διαστάσεων σε ένα νέο σύστημα συντεταγμένων, όπου οι άξονες (κύρια στοιχεία) ταξινομούνται με βάση την ποσότητα διακύμανσης που καταγράφουν από τα δεδομένα. Σε αυτή την ενότητα και για την δικιά μας εφαρμογή θα εξετάσουμε 2 μορφές της μεθόδου PCA, γραμμική και μη γραμμική. Για την γραμμική εκδοχή τα βήματα εκτέλεσης έχουν ως εξής:

1.**Κεντράρισμα των δεδομένων**: Τα χαρακτηριστικά γνωρίσματα κεντράρονται πρώτα αφαιρώντας τη μέση τιμή από κάθε χαρακτηριστικό γνώρισμα.

2.**Υπολογισμός του πίνακα συνδιακύμανσης**: Ο πίνακας συνδιακύμανσης αποτυπώνει τις γραμμικές σχέσεις μεταξύ όλων των ζευγών χαρακτηριστικών.

3.**Αποσύνθεση ιδιοτιμών**: Ο πίνακας συνδιακύμανσης αναλύεται σε ιδιοτιμές και ιδιοδιανύσματα. Οι ιδιοτιμές αντιπροσωπεύουν τη διακύμανση που εξηγείται από κάθε ιδιοδιάνυσμα.

4.**Επιλογή κύριων συνιστωσών**: Τα ιδιοδιανύσματα που αντιστοιχούν στις μεγαλύτερες ιδιοτιμές αντιπροσωπεύουν τις κατευθύνσεις της μεγαλύτερης διακύμανσης. Επιλέγουμε ένα υποσύνολο αυτών των ιδιοδιανυσμάτων (Principal Components - PCs) για να ορίσουμε τον υποχώρο χαμηλότερης διάστασης.

5.**Προβολή δεδομένων σε PCs**: Τα αρχικά σημεία δεδομένων προβάλλονται στη συνέχεια στα επιλεγμένα PCs, με αποτέλεσμα μια χαμηλότερης διάστασης αναπαράσταση που διατηρεί τις σημαντικότερες πληροφορίες.

Παρόλο που η γραμμική εκδοχή της PCA υποθέτει γραμμική σχέση μεταξύ των χαρακτηριστικών, σε ορισμένες περιπτώσεις, οι σχέσεις μεταξύ των χαρακτηριστικών που προέρχονται από ιατρικές εικόνες μπορεί να είναι μη γραμμικές.

$$K(x\,,y) \;=\; exp\left(-\,\frac{\|x-y\|^2}{2\sigma^2}\right)$$

Η μη γραμμική εκδοχή της PCA χρησιμοποιεί το "τέχνασμα πυρήνα" – RBF (Radial Basis Function) Kernel, όπως περιγράφεται από την εξίσωση παραπάνω, για να αποτυπώσει την υποκείμενη δομή των δεδομένων. Τα επιμέρους βήματα εκτέλεσης αυτής της εκδοχής είναι:

1.**Υπολογισμός πυρήνα**: Υπολογισμός του πίνακα πυρήνα RBF για τα δεδομένα εισόδου, ο οποίος αποτυπώνει τις ομοιότητες μεταξύ των σημείων δεδομένων στον πολυδιάστατο χώρο χαρακτηριστικών.

2.**Κεντράρισμα του πίνακα πυρήνα**: Ο πίνακας πυρήνα κεντράρεται για να εξασφαλιστεί ότι τα δεδομένα έχουν μηδενική μέση τιμή στο χώρο χαρακτηριστικών.

3.**Αποσύνθεση ιδιοτιμών:** Πραγματοποιείται ιδιοαποσύνθεση στον κεντραρισμένο πίνακα πυρήνα για να ληφθούν οι ιδιοτιμές και τα ιδιοδιανύσματα.

4.**Προβολή**: Προβολή των αρχικών δεδομένων στις κύριες συνιστώσες που ορίζονται από τα ιδιοδιανύσματα στο χώρο του πυρήνα.

Σε αυτό το σημείο είναι σημαντικό να αναφέρουμε τον συμβιβασμό μεταξύ του αριθμού των κύριων συνιστωσών που επιλέγονται και της ποσότητας της πληροφορίας που διατηρείται. Η επιλογή πολύ λίγων συνιστωσών μπορεί να οδηγήσει σε απώλεια πληροφορίας, ενώ οι πάρα πολλές μπορεί να μην προσφέρουν σημαντικά οφέλη και να αυξήσουν την υπολογιστική πολυπλοκότητα. Η επιλογή του βέλτιστου αριθμού συνιστωσών συχνά περιλαμβάνει την αξιολόγηση της απόδοσης του μοντέλου σε ένα σύνολο επικύρωσης.

4.2.6: Διόρθωση της Ανισορροπίας με Συνθετικά Σεδομένα

Στον τομέα της ιατρικής όπου είναι συχνή η χρήση της ραδιομικής και των ακτινολογικών χαρακτηριστικών, αντιμετωπίζουμε συχνά μια σημαντική πρόκληση, αυτή της ανισορροπίας των τάξεων. Αυτό συμβαίνει όταν μια κατηγορία, όπως οι ασθενείς με έναν συγκεκριμένο τύπο όγκου ή εκείνοι που ανταποκρίνονται σε μια συγκεκριμένη θεραπεία, υποεκπροσωπούνται σημαντικά σε σύγκριση με άλλες κατηγορίες. Αυτή η ανισορροπία μπορεί να παρεμποδίσει σοβαρά την απόδοση των μοντέλων μηχανικής μάθησης, οδηγώντας σε μεροληπτικές προβλέψεις και ανακριβείς κλινικές αποφάσεις [49].

Μια ισχυρή τεχνική για την αντιμετώπιση αυτού του προβλήματος είναι η *Τεχνική συνθετικής υπερδειγματοληψίας μειοψηφιών* (SMOTE). Εισήχθη από τους [49], η SMOTE παράγει συνθετικά νέα δείγματα εντός του χώρου χαρακτηριστικών της μειονοτικής κλάσης αντί να αντιγράφει απλώς τα υπάρχοντα δείγματα της μειονοτικής κλάσης. Αυτή η προσέγγιση εξισορροπεί αποτελεσματικά την κατανομή της κλάσης, ενώ εισάγει νέα σημεία δεδομένων, βελτιώνοντας έτσι την ικανότητα του μοντέλου να γενικεύει και να μαθαίνει τα υποκείμενα πρότυπα της κλάσης μειονότητας.

Για να χρησιμοποιήσουμε τον αλγόριθμο SMOTE θα χρειαστεί αρχικά να προσδιορίσουμε την μειοψηφική κλάση. Στο σύνολο δεδομένων μας με τα χαρακτηριστικά που εξάγονται από τις τρισδιάστατες σαρώσεις εγκεφαλικών μεταστάσεων, η μειοψηφική κλάση αντιπροσωπεύει τους ασθενείς που μετά το πέρας της έρευνας απεβίωσαν. Στην συνέχεια, για κάθε δείγμα στην κατηγορία μειονότητας, ο SMOTE εντοπίζει τους k-κοντινότερους γείτονές του στο χώρο των χαρακτηριστικών. Αυτοί οι γείτονες είναι άλλα σημεία δεδομένων με παρόμοιες τιμές χαρακτηριστικών. Αυτό το βήμα χρησιμοποιεί μια μετρική απόστασης όπως η ευκλείδεια απόσταση για την εύρεση των πλησιέστερων γειτόνων [49]. Ο αλγόριθμος δημιουργεί νέα συνθετικά δείγματα κατά μήκος των τμημάτων γραμμής που συνδέουν το δείγμα της μειονοτικής κατηγορίας με τους πλησιέστερους γείτονές του. Αυτό περιλαμβάνει την τυχαία επιλογή ενός

σημείου στο ευθύγραμμο τμήμα και τον υπολογισμό των τιμών των χαρακτηριστικών του με βάση την παρεμβολή μεταξύ των δύο αρχικών δειγμάτων. Τέλος, επαναλαμβάνοντας αυτή τη διαδικασία για πολλαπλά δείγματα της τάξης μειονότητας, ο SMOTE γεμίζει αποτελεσματικά το χώρο χαρακτηριστικών με συνθετικά σημεία δεδομένων, εξισορροπώντας την κατανομή των τάξεων και παρέχοντας στο μοντέλο μηχανικής μάθησης πιο ποικίλα παραδείγματα για μάθηση.

Μερικά από τα πλεονεκτήματα του αλγορίθμου SMOTE είναι να μειώνει τον κίνδυνο υπερπροσαρμογής, όπου το μοντέλο απομνημονεύει τα δεδομένα εκπαίδευσης αντί να μαθαίνει γενικεύσιμα πρότυπα [49]. Αυτό είναι ιδιαίτερα κρίσιμο στις εφαρμογές βασισμένες στα radiomics, όπου τα σύνολα δεδομένων μπορεί να είναι σχετικά μικρά. Επίσης μπορεί να βελτιώσει σημαντικά την ευαισθησία και την ειδικότητα των μοντέλων μηχανικής μάθησης στην ανίχνευση της τάξης μειονότητας, οδηγώντας σε ακριβέστερες διαγνώσεις και σχεδιασμό θεραπείας [48]. Αυτό που είναι επίσης σημαντικό είναι η διατήρηση της δομής των δεδομένων και των σχέσεων εντός του χώρου χαρακτηριστικών, διασφαλίζοντας την ακεραιότητα της ανάλυσης ραδιομικής.

Σε εφαρμογές όπου το σύνολο δεδομένων έχει περιορισμένο αριθμό ασθενών, ο αλγόριθμος SMOTE μπορεί να χρησιμοποιηθεί για τη δημιουργία συνθετικών δειγμάτων αυτής της μειονοτικής κατηγορίας. Αυτό θα επιτρέψει στο μοντέλο μηχανικής μάθησης να μάθει καλύτερα τα ξεχωριστά χαρακτηριστικά που σχετίζονται με την πλήρη ανταπόκριση, βελτιώνοντας τελικά την ακρίβεια της πρόβλεψης της θεραπευτικής ανταπόκρισης. Η προσέγγιση αυτή έχει εφαρμοστεί με επιτυχία σε διάφορες μελέτες ιατρικής απεικόνισης, συμπεριλαμβανομένης της διάγνωσης και πρόγνωσης του καρκίνου [50].

Ο SMOTE είναι ένα πολύτιμο εργαλείο για την αντιμετώπιση της ανισορροπίας των κλάσεων στην ανάλυση και στην εκπαίδευση των μοντέλων μηχανικής μάθησης. Με τη δημιουργία συνθετικών δειγμάτων εντός του χώρου χαρακτηριστικών, ο SMOTE ενισχύει την ικανότητα του μοντέλου να μαθαίνει από την κλάση μειονότητας, οδηγώντας σε πιο ακριβείς και αξιόπιστες προβλέψεις. Καθώς τα εξαγώμενα χαρακτηριστικά συνεχίζουν να διαδραματίζουν ολοένα και σημαντικότερο ρόλο στην εξατομικευμένη ιατρική, τεχνικές όπως ο SMOTE θα είναι ζωτικής σημασίας για να διασφαλιστεί ότι τα μοντέλα μηχανικής μάθησης μπορούν να αξιοποιήσουν αποτελεσματικά τις πληροφορίες που εμπεριέχονται σε πολύπλοκα δεδομένα ιατρικής απεικόνισης.

4.2.7: Επιλογή Χαρακτηριστικών με την Μέθοδο LGBM

Το LightGBM, συντομογραφία του Light Gradient Boosting Machine [51], είναι ένα ισχυρό εργαλείο, που χρησιμοποιεί την τεχνική gradient boosting, γνωστό για την ταχύτητα και την αποτελεσματικότητά του. Ενώ είναι εγγενώς ικανό να χειρίζεται μεγάλα σύνολα χαρακτηριστικών, η

επιλογή χαρακτηριστικών μπορεί να βελτιώσει περαιτέρω την απόδοση του μοντέλου, την ερμηνευσιμότητα και τον χρόνο εκπαίδευσης [52]. Παρακάτω θα εξηγήσουμε και θα εμβαθύνουμε στις μεθόδους αξιοποίησης των δυνατοτήτων του LightGBM για αποτελεσματική επιλογή χαρακτηριστικών.

Το LightGBM προσφέρει δύο πρωταρχικούς τρόπους για την αξιολόγηση της σημασίας των χαρακτηριστικών [51]: Σημασία με βάση τον διαχωρισμό και σημασία με βάση το κέρδος. Η σημασία βάσει διαχωρισμού μετράει πόσο συχνά χρησιμοποιείται ένα χαρακτηριστικό για διαχωρισμό σε όλα τα δέντρα του μοντέλου. Μια υψηλότερη μέτρηση υποδηλώνει μεγαλύτερη συμμετοχή στη διαδικασία λήψης αποφάσεων. Η σημασία με βάση το κέρδος μετρά το συνολικό κέρδος (βελτίωση της ακρίβειας) που επιτυγχάνεται από τον διαχωρισμό με τη χρήση ενός συγκεκριμένου χαρακτηριστικού. Προσφέρει μια πιο διαφοροποιημένη άποψη εξετάζοντας την ποιότητα των διαχωρισμών.

Ακολουθούν οι κύριες τεχνικές για την επιλογή χαρακτηριστικών με τη χρήση του LightGBM: Built-in Feature Importance, Recursive Feature Elimination (RFE) και Permutation Importance.

Για να χρησιμοποιήσουμε τη μέθοδο Built-in Feature Importance, εκπαιδεύουμε πρώτα το μοντέλο LightGBM στο σύνολο δεδομένων με όλα τα χαρακτηριστικά. Στη συνέχεια, εξάγουμε είτε βαθμολογίες σημαντικότητας με βάση τον «διαχωρισμό» είτε με βάση το «κέρδος». Τέλος, ταξινομούμε τα χαρακτηριστικά με βάση τη σημασία και επιλέγουμε τα K καλύτερα χαρακτηριστικά ή εκείνα που βρίσκονται πάνω από ένα ορισμένο κατώφλι.[51][52]

Η τεχνική RFE [53] είναι μια επαναληπτική προσέγγιση που αφαιρεί αναδρομικά τα λιγότερο σημαντικά χαρακτηριστικά σύμφωνα με την αξιολόγηση του μοντέλου. Πρώτον, εκπαιδεύουμε το μοντέλο LightGBM και λαμβάνουμε τις τιμές σημαντικότητας των χαρακτηριστικών. Στη συνέχεια, αφαιρούμε τα λιγότερο σημαντικά χαρακτηριστικά. Εκπαιδεύουμε εκ νέου το μοντέλο στο μειωμένο σύνολο χαρακτηριστικών και επαναλαμβάνουμε τα βήματα 1 και 2 έως ότου απομείνει ο επιθυμητός αριθμός χαρακτηριστικών.

Η τεχνική Permutation Importance μετρά τη σημασία ενός χαρακτηριστικού υπολογίζοντας τη μείωση της απόδοσης του μοντέλου όταν οι τιμές του χαρακτηριστικού ανακατεύονται τυχαία. Για να χρησιμοποιήσουμε αυτή την τεχνική, ξεκινάμε με την εκπαίδευση του μοντέλου LightGBM. Για κάθε χαρακτηριστικό, αναδιατάσσουμε τυχαία τις τιμές του και αξιολογούμε την απόδοση του μοντέλου στα ανακατεμένα δεδομένα. Η πτώση της απόδοσης αντιπροσωπεύει τη σημασία του χαρακτηριστικού. [54]

Κατά την εκτέλεση της επιλογής χαρακτηριστικών με το LightGBM, λάβμάνουμε υπόψην την επιλογή της μετρικής σπουδαιότητας, την κανονικοποίηση, τη διασταυρούμενη επικύρωση, την εμπειρογνωμοσύνη στον τομέα και να θυμάστε ότι πρόκειται για μια επαναληπτική διαδικασία. Η σημασία «κέρδους» παρέχει συχνά ένα πιο αξιόπιστο μέτρο της επιρροής του χαρακτηριστικού. Επίσης μπορούμε να εξετάζουμε το ενδεχόμενο κανονικοποίησης των βαθμολογιών σπουδαιότητας χαρακτηριστικών για ευκολότερη σύγκριση. Χρησιμοποιούμε τη διασταυρούμενη επικύρωση για να λάβουμε πιο αξιόπιστες

εκτιμήσεις της σημασίας χαρακτηριστικών. Συνδυάζουμε τη σημασία χαρακτηριστικών του LightGBM με τη γνώση του τομέα για τεκμηριωμένη επιλογή χαρακτηριστικών. Η επιλογή χαρακτηριστικών είναι συχνά μια επαναληπτική διαδικασία οπότε μπορούμε να πειραματιστούμε με διαφορετικές τεχνικές και κατώτατα όρια για να βρούμε το βέλτιστο υποσύνολο χαρακτηριστικών.[51]

Η ενσωματωμένη σημασία χαρακτηριστικών του LightGBM και η συμβατότητα με τεχνικές όπως η RFE προσφέρουν μια ισχυρή εργαλειοθήκη για την επιλογή χαρακτηριστικών. Εφαρμόζοντας με σύνεση αυτές τις μεθόδους, μπορούμε να βελτιώσουμε την απόδοση του μοντέλου, να μειώσουμε την υπερπροσαρμογή και να αποκτήσουμε πολύτιμες πληροφορίες για τα δεδομένα.

4.3: Ανάλυση Μοντέλων Ταξινόμησης

4.3.1: Εκπαίδευση Μοντέλων

Έχοντας δημιουργήσει ένα ισχυρό και κατατοπιστικό σύνολο χαρακτηριστικών μέσω των τεχνικών προεπεξεργασίας, όπως η αφαίρεση ακραίων τιμών, η ανάλυση συσχέτισης και η μείωση της διαστατικότητας, μπαίνουμε τώρα στην ουσία της προσπάθειάς μας για μηχανική μάθηση: την εκπαίδευση του μοντέλου. Σε αυτή την ενότητα διερευνάμε την έννοια της εκπαίδευσης του μοντέλου και τη σημασία της για τη δημιουργία ενός μοντέλου ικανού να προβλέψει τη θνησιμότητα των ασθενών με βάση τα ακτινολογικά χαρακτηριστικά των όγκων του εγκεφάλου.

Η εκπαίδευση μοντέλου αναφέρεται ουσιαστικά στη διαδικασία καθοδήγησης ενός αλγορίθμου μηχανικής μάθησης ώστε να μάθει τις υποκείμενες σχέσεις μεταξύ των χαρακτηριστικών και της μεταβλητής-στόχου (στην προκειμένη περίπτωση, την πιθανότητα θανάτου). Αυτή η διαδικασία εκμάθησης περιλαμβάνει την έκθεση του αλγορίθμου σε ένα σύνολο δεδομένων με ετικέτες, όπου κάθε δείγμα έχει μια γνωστή έκβαση. Το μοντέλο αναλύει αυτά τα δείγματα εκπαίδευσης, εντοπίζοντας μοτίβα και σχέσεις μεταξύ των χαρακτηριστικών που μπορούν να χρησιμοποιηθούν για την πρόβλεψη της έκβασης για άγνωστα δεδομένα.

Όπως προαναφέραμε έχουμε χωρίσει το σύνολο των δεδομένων μας σε 2 σετ. Σε αυτό το στάδιο της εκπαίδευσης του μοντέλου θα χρησιμοποιήσουμε το σετ εκπαίδευσης. Αποτελεί το 70% από την ολόκληρη συλλογή δεδομένων των χαρακτηριστικών αφού έχουν εφαρμοστεί όλα τα βήματα της προεπεξεργασίας.

Τα μοντέλα που χρησιμοποιήθηκαν για την πειραματική διαδικασίας παρουσιάζονται στον παρακάτω πίνακα:

	Name	Reference
lr	Logistic Regression	sklearn.linear_model._logistic.LogisticRegression
knn	K Neighbors Classifier	sklearn.neighbors._classification.KNeighborsClassifier
nb	Naive Bayes	sklearn.naive_bayes.GaussianNB
dt	Decision Tree Classifier	sklearn.tree._classes.DecisionTreeClassifier
svm	SVM - Linear Kernel	sklearn.linear_model._stochastic_gradient.SGDClassifier
rbfsvm	SVM - Radial Kernel	sklearn.svm._classes.SVC
gpc	Gaussian Process Classifier	sklearn.gaussian_process._gpc.GaussianProcessClassifier
mlp	MLP Classifier	sklearn.neural_network._multilayer_perceptron.MLPClassifier
ridge	Ridge Classifier	sklearn.linear_model._ridge.RidgeClassifier
rf	Random Forest Classifier	sklearn.ensemble._forest.RandomForestClassifier
qda	Quadratic Discriminant Analysis	sklearn.discriminant_analysis.QuadraticDiscriminantAnalysis
ada	Ada Boost Classifier	sklearn.ensemble._weight_boosting.AdaBoostClassifier
gbc	Gradient Boosting Classifier	sklearn.ensemble._gb.GradientBoostingClassifier
lda	Linear Discriminant Analysis	sklearn.discriminant_analysis.LinearDiscriminantAnalysis
et	Extra Trees Classifier	sklearn.ensemble._forest.ExtraTreesClassifier
xgboost	Extreme Gradient Boosting	xgboost.sklearn.XGBClassifier
lightgbm	Light Gradient Boosting Machine	lightgbm.sklearn.LGBMClassifier
catboost	CatBoost Classifier	catboost.core.CatBoostClassifier
dummy	Dummy Classifier	sklearn.dummy.DummyClassifier

Εικόνα 29 - Μοντέλα Εκπαίδευσης για την Πειραματική Διαδικασία

Στην πρώτη στήλη έχουμε την λέξη-κλειδί που χρησιμοποιούμε για συντομία όταν κάνουμε αναφορά στο κάθε μοντέλο. Στην δεύτερη στήλη έχουμε ολόκληρο το όνομα του μοντέλου και στην τρίτη στήλη την αναφορά στην βάση κώδικα που υλοποιεί την δομή του μοντέλου και μας επιτρέπει να το χρησιμοποιήσουμε. Παρόλο που αυτή την στιγμή έχουμε να εξερευνήσουμε και να πειραματιστούμε ανάμεσα σε 19 επιλογές, θα δούμε στις επόμενες ενότητες πως πολύ γρήγορα και άμεσα θα μειώσουμε τον χώρο αναζήτησης.

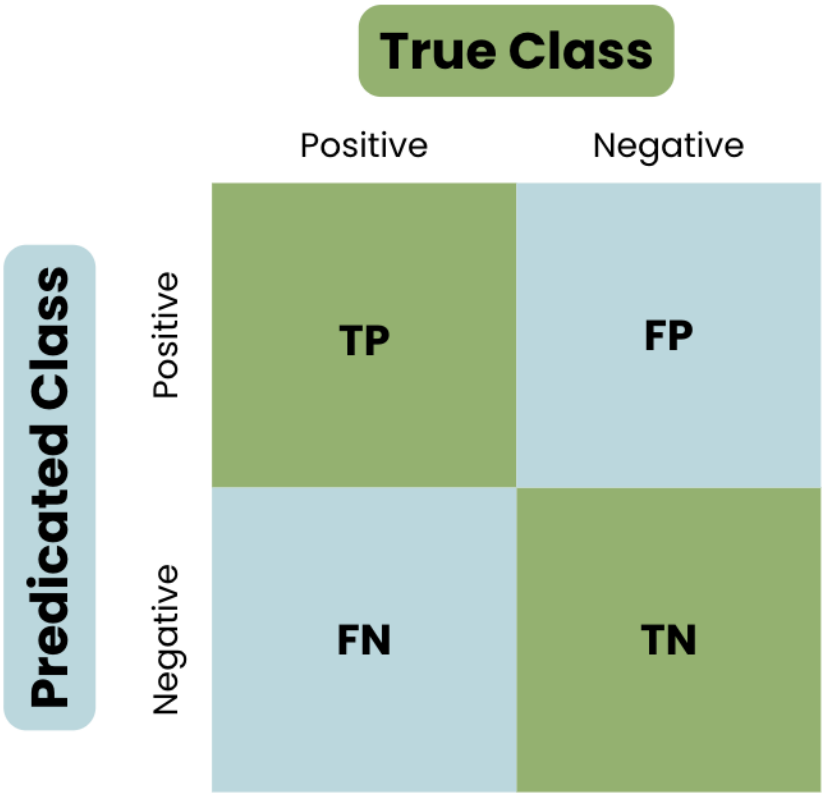
Αφού έχουμε ορίσει τα μοντέλα εκπαίδευσης και το σετ εκπαίδευσης, εκτελούμε την εκπαίδευση για κάθε μοντέλο. Η εκπαίδευση είναι μια επαναληπτική διαδικασία. Για να μπορούμε να συγκρίνουμε τα μοντέλα μεταξύ τους χρησιμοποιούμε κάποιες μετρικές αξιολόγησης. Με βάση τις μετρικές αξιολόγησης, το μοντέλο μπορεί να επανεκπαιδευτεί με προσαρμογές στις υπερπαραμέτρους ή ακόμη και στον ίδιο τον επιλεγμένο αλγόριθμο. Ο στόχος είναι να επιτευχθεί ένα μοντέλο που όχι μόνο αποδίδει καλά στα δεδομένα εκπαίδευσης αλλά και γενικεύει αποτελεσματικά για την δικιά μας εφαρμογή που είναι η πρόβλεψη της θνησιμότητας νέων ασθενών με όγκο στον εγκέφαλο.

Στη επόμενη ενότητα θα εμβαθύνουμε στις τεχνικές αξιολόγησης του μοντέλου, που είναι ζωτικής σημασίας για την αξιολόγηση της γενικευσιμότητας και της κλινικής συνάφειας του εκπαιδευμένου μοντέλου.

4.3.2: Μετρικές Αξιολόγησής Μοντέλων

Σε αυτή την κοινότητα, θα δούμε διάφορες μετρικές για την αξιολόγηση της απόδοσης των μοντέλων μηχανικής μάθησης που χρησιμοποιούνται για την πρόβλεψη των αποτελεσμάτων επιβίωσης των ασθενών με βάση τις εικόνες NifTI όγκων εγκεφάλου και τις τμηματοποιήσεις τους. Η επιλογή των κατάλληλων μετρικών είναι ζωτικής σημασίας για την αξιολόγηση της αποτελεσματικότητας των διαφόρων μοντέλων και της καταλληλότητάς τους για το συγκεκριμένο έργο μας.

Δεδομένου ότι τα μοντέλα μας προβλέπουν την πιθανότητα θανάτου (αποτέλεσμα επιβίωσης), το οποίο είναι συνήθως ένα δυαδικό πρόβλημα ταξινόμησης (επιβίωση ή θάνατος), θα επικεντρωθούμε σε μετρικές ταξινόμησης. Συνήθως για προβλήματα δυαδικής ταξινόμησης, αρχικά χρησιμοποιούμε έναν πίνακα που ονομάζεται πίνακας σύγχυσης (confusion matrix), ο οποίος βασίζεται στην σύγκριση του συνόλου των πραγματικών τιμών της συλλογής δεδομένων και του συνόλου των προβλέψεων μέσω του εκπαιδευμένου μοντέλου. Μία γενικευμένη εικόνα για τον πίνακα σύγχυσης παρουσιάζεται στον παρακάτω πίνακα.



Εικόνα 30 - Πίνακας Σύγχυσης (Confusion Matrix)

Οι 4 τιμές που εμπεριέχονται στον παραπάνω πίνακα είναι:

- TP (True Positives): Ο αριθμός των σωστά ταξινομημένων θετικών δειγμάτων
- FP (False Positives): Ο αριθμός των λανθασμένα ταξινομημένων θετικών δειγμάτων

- FN (False Negatives): Ο αριθμός των λανθασμένα ταξινομημένων αρνητικών δειγμάτων
- TN (True Negatives): Ο αριθμός των σωστά ταξινομημένων αρνητικών δειγμάτων

Το μέγεθος του πίνακα αντιστοιχούν στον αριθμό των διαφορετικών κατηγοριών της μεταβλητής στόχου. Στο πρόβλημα που εξετάζουμε, ένα δυαδικό πρόβλημα (0 ή 1 / Αρνητικό ή Θετικό), ο πίνακας έχει 2 γραμμές και στήλες όπου αντιπροσωπεύουν τις διαφορετικές κατηγορίες μεταβλητών βάση της πρόβλεψης ή των πραγματικών τιμών, αντίστοιχα.

Παρόλο που ο πίνακας συνοψίζει τις πληροφορίες και τα αποτελέσματα για τον προσδιορισμό της απόδοσης του μοντέλου ταξινόμησης, τις περισσότερες φορές μία μετρική που είναι στοχευμένη μπορεί να αποδώσει την πληροφορία που αναζητούμε καλύτερα. Παρακάτω παρουσιάζονται οι πιο συχνές μετρικές για ένα πρόβλημα ταξινόμησης:

Ορθότητα/Accuracy: Η ακρίβεια είναι η πιο βασική μετρική, η οποία αντιπροσωπεύει το ποσοστό των σωστά ταξινομημένων περιπτώσεων. Υπολογίζει το ποσοστό των ασθενών όπου η πρόβλεψη του μοντέλου (επιβίωση/θάνατος) ταιριάζει με το πραγματικό αποτέλεσμα. Αν και φαινομενικά απλή, η ακρίβεια μπορεί να είναι παραπλανητική σε ανισόρροπα σύνολα δεδομένων, όπου μια κλάση (π.χ. θάνατος) μπορεί να είναι λιγότερο συχνή.

$$Accuracy = \frac{Αριθμός \text{ σωστά ταξινομημένων δειγμάτων}}{Αριθμός \ \ όλων \ των \ δειγμάτων} = \frac{TP + TN}{TP + FP + TN + FN}$$

Ακρίβεια/Precision: Η ακρίβεια μετρά το ποσοστό των αληθώς θετικών μεταξύ όλων των θετικών προβλέψεων. Στην περίπτωσή μας, αντικατοπτρίζει το ποσοστό των ασθενών που προβλέπεται να πεθάνουν (θετικό) και οι οποίοι πράγματι απεβίωσαν.

$$Precision = \frac{Αριθμός \text{ σωστά ταξινομημένων } θετικών \text{ δειγμάτων}}{Αριθμός \ \ θετικά \ ταξινομημένων \ δειγμάτων} = \frac{TP}{TP + FP}$$

Ευαισθησία/Sensitivity: Αυτή η μετρική παρέχει μια πιο διαφοροποιημένη άποψη της απόδοσης του μοντέλου, ιδιαίτερα σημαντική για ανισόρροπα σύνολα δεδομένων. Η ευαισθησία, επίσης γνωστή ως ανάκληση, μετρά το ποσοστό των αληθώς θετικών μεταξύ όλων των πραγματικών θετικών περιπτώσεων. Αντιπροσωπεύει την ικανότητα του μοντέλου να εντοπίζει σωστά τους ασθενείς που απεβίωσαν. Αυτή η μετρική θεωρείται ως μία από τις πιο σημαντικές, όταν το περιεχόμενο του προβλήματος είναι ιατρικής φύσεως, καθώς ο στόχος είναι να μην παραλείψουμε κανένα θετικό περιστατικό (θάνατος), που μετραφράζεται σε υψηλή ευαισθησία.

$$Sensitivity/Recall = \frac{Αριθμός\ σωστά\ ταξινομημένων\ θετικών\ δειγμάτων}{Αριθμός\ \ όλων\ των\ θετικών\ δειγμάτων} = \frac{TP}{TP + FN}$$

Ειδικότητα/Specificity: Η ειδικότητα, ή το ποσοστό των αληθώς αρνητικών (True Negative Rate), είναι η αντίστοιχη μετρική της για την αρνητική κλάση και δηλώνει το ποσοστό των αρνητικών δειγμάτων που έχουν ταξινομηθεί σωστά.

$$\text{Specificity} = \frac{Αριθμός\ σωστά\ ταξινομημένων\ θετικών\ δειγμάτων}{Αριθμός\ \ όλων\ των\ αρνητικών\ δειγμάτων} = \frac{TP}{TN + FP}$$

F1-Score: Το F1-score είναι ένας αρμονικός μέσος όρος της ακρίβειας και της ανάκλησης, παρέχοντας μια ισορροπημένη εικόνα της απόδοσης του μοντέλου, ειδικά όταν η ακρίβεια και η ανάκληση διαφέρουν σημαντικά.

$$\text{F1} - \text{Score} = 2\,\frac{Precision\ *\ Recall}{Precision\ +\ Recall}$$

Εμβαδόν κάτω από την καμπύλη ROC (AUC): Η καμπύλη ROC (Receiver Operating Characteristic) απεικονίζει το συμβιβασμό μεταξύ του αληθώς θετικού ποσοστού (ανάκληση) και του ψευδώς θετικού ποσοστού (1 εξειδίκευση) για διαφορετικά κατώτατα όρια ταξινόμησης. Η AUC αντιπροσωπεύει την εμβαδόν κάτω από την καμπύλη ROC, αποτυπώνοντας την ικανότητα του μοντέλου να διακρίνει μεταξύ ζωντανών και νεκρών ασθενών.

Η επιλογή της μετρικής αξιολόγησης εξαρτάται από τους συγκεκριμένους στόχους της μελέτης μας και τα χαρακτηριστικά του συνόλου δεδομένων σας. Κάποια χαρακτηριστικά που πρέπει να ληφθούν υπόψη για την επιλογή της κατάλληλης μετρικής είναι η ύπαρξη ανισορροπίας δεδομένων, δηλαδή εάν το σύνολο δεδομένων μας έχει σημαντική ανισορροπία κλάσεων (π.χ. λιγότεροι ασθενείς που έχουν αποβιώσει), πιθανόν η ακρίβεια ενδέχεται να είναι παραπλανητική. Σε τέτοιες περιπτώσεις, συνιστάται η εστίαση στην ακρίβεια, την ανάκληση και το F1-score. Ένας ακόμα παράγοντας είναι η κλινική συνάφεια. Όταν πρόκειται για πρόβλεψη επιβίωσης ασθενών, μετρικές όπως η AUC παρέχουν πολύτιμες πληροφορίες σχετικά με την ικανότητα του μοντέλου να διακρίνει μεταξύ διαφορετικών προτύπων επιβίωσης.

Χρησιμοποιώντας έναν συνδυασμό αυτών των μετρικών, μπορούμε να αξιολογήσουμε ολοκληρωμένα την απόδοση των μοντέλων μηχανικής μάθησης για την πρόβλεψη της επιβίωσης των ασθενών με όγκους εγκεφάλου. Αυτή η αξιολόγηση θα μας οδηγήσει στην επιλογή του πιο

αποτελεσματικού μοντέλου για περαιτέρω κλινική διερεύνηση και πιθανή ενσωμάτωση στις διαδικασίες λήψης αποφάσεων.

4.3.3: Βελτιστοποίηση Μοντέλων

Στις προηγούμενες ενότητες, συζητήσαμε μεθόδους για την εκπαίδευση και την αξιολόγηση της απόδοσης των μοντέλων μηχανικής μάθησης που χρησιμοποιούνται για την πρόβλεψη της επιβίωσης ασθενών με όγκους εγκεφάλου. Σε αυτή την ενότητα εξετάζονται διάφορες τεχνικές για τη βελτιστοποίηση των υπερπαραμέτρων αυτών των μοντέλων, με στόχο την επίτευξη της καλύτερης δυνατής απόδοσης στην πρόβλεψη.

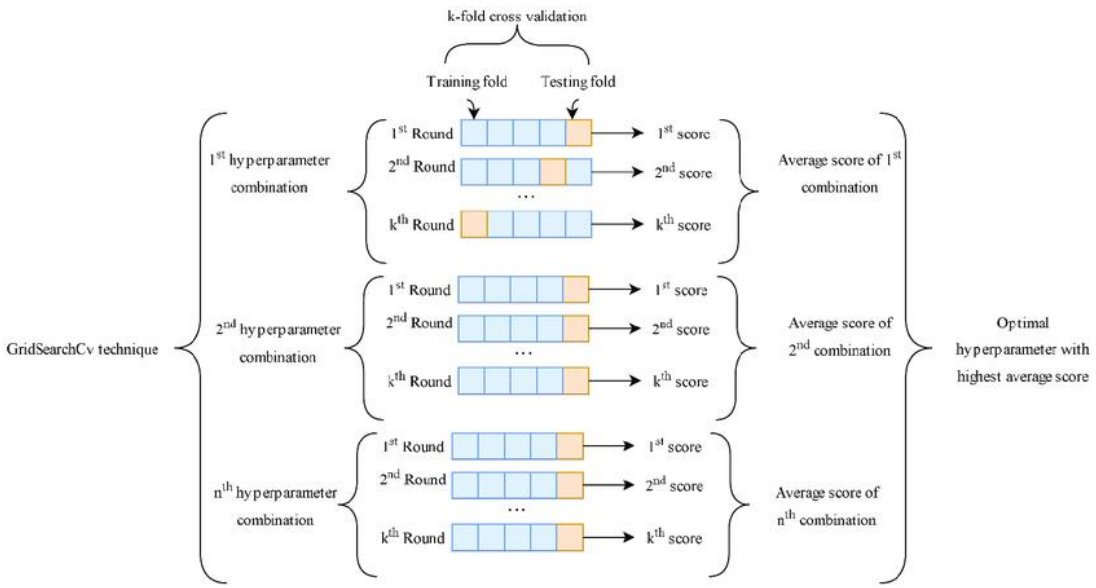
Τα μοντέλα μηχανικής μάθησης βασίζονται σε παραμέτρους που μαθαίνονται από τα δεδομένα εκπαίδευσης. Ωστόσο, υπάρχουν επίσης υπερπαράμετροι, οι οποίες είναι ρυθμίσεις που ελέγχουν την ίδια τη διαδικασία μάθησης. Αυτές οι υπερπαράμετροι δεν μαθαίνονται από τα δεδομένα αλλά προκαθορίζονται από τον ερευνητή πριν από την έναρξη της εκπαίδευσης. Η επιλογή των βέλτιστων τιμών υπερπαραμέτρων είναι καθοριστικής σημασίας για την απόδοση ενός μοντέλου. Επηρεάζουν σημαντικά παράγοντες όπως η πολυπλοκότητα του μοντέλου, η ικανότητα γενίκευσης και ο χρόνος εκπαίδευσης.

Στην δικιά μας σειρά ενεργειών της πειραματικής διαδικασίας θα χρησιμοποιήσουμε τεχνικές βελτιστοποίησης σε 2 από τα στάδια. Το πρώτο στάδιο που συναντάμε μία τεχνική βελτιστοποίησης είναι κατά την διάρκεια της εκπαίδευσης του μοντέλου κάνοντας χρήση της διασταυρούμενης επικύρωσης **K**-πλαισίων (K-fold Cross Validation). Το δεύτερο στάδιο που εφαρμόζουμε βελτιστοποίηση είναι στη ρύθμιση των παραμέτρων των μοντέλων με αναζήτηση (Parameter Grid Search).

Η τεχνική της διασταυρούμενης επικύρωσης **K**-πλαισίων χωρίζει το σετ εκπαίδευσης σε K πλαίσια και εκτελεί K φορές την εκπαίδευση με δεδομένα εκπαίδευσης τα K-1 πλαίσια και δεδομένα δοκιμής το 1 υπολειπόμενο πλαίσιο. Για κάθε μία από τις K επαναλήψεις υπολογίζουμε τις μετρικές που μας ενδιαφέρουν και στο τέλος παίρνουμε των μέσο όρο των K αποτελεσμάτων των μετρικών για να υπολογίζουμε την συγκεντρωτική μετρική. Τα θετικά αυτής της τεχνικής είναι πως δεν συμπεριλαμβάνουμε το αρχικό σετ δοκιμής στην εκπαίδευση επιτυγχάνοντας αμεροληψία και επίσης εξασφαλίζουμε ένα πιο γενικευμένο και εύρωστο μοντέλο καθώς όλα τα μέρη του σετ δεδομένων μας έχουν συμβάλει στην εκπαίδευση αλλά και την επικύρωση.

Εικόνα 31 - K-Fold Cross Validation (10 folds)

Η μέθοδος της αναζήτησης σε πλέγμα είναι μία εξαντλητική αναζήτηση που αξιολογεί όλους τους δυνατούς συνδυασμούς τιμών υπερπαραμέτρων εντός ενός προκαθορισμένου πλέγματος. Κατά την διάρκεια αυτής της τεχνικής χρησιμοποιούμε παράλληλα την τεχνική της διασταυρούμενης επικύρωσης για να αξιολογήσουμε της επιλογή της υπερπαραμέτρου. Αν και περιεκτική, η αναζήτηση πλέγματος μπορεί να είναι υπολογιστικά δαπανηρή, ειδικά για μοντέλα με πολλές υπερπαραμέτρους. Εφαρμόσαμε αυτή την τεχνική στο βήμα μετά την διαδικασία εκπαίδευσης των μοντέλων και σε συγκεκριμένα μοντέλα που διακρίθηκαν ώστε να ταιριάζουν περισσότερο στην δικιά μας εφαρμογή



Εικόνα 32 - Grid Search Cross Validation

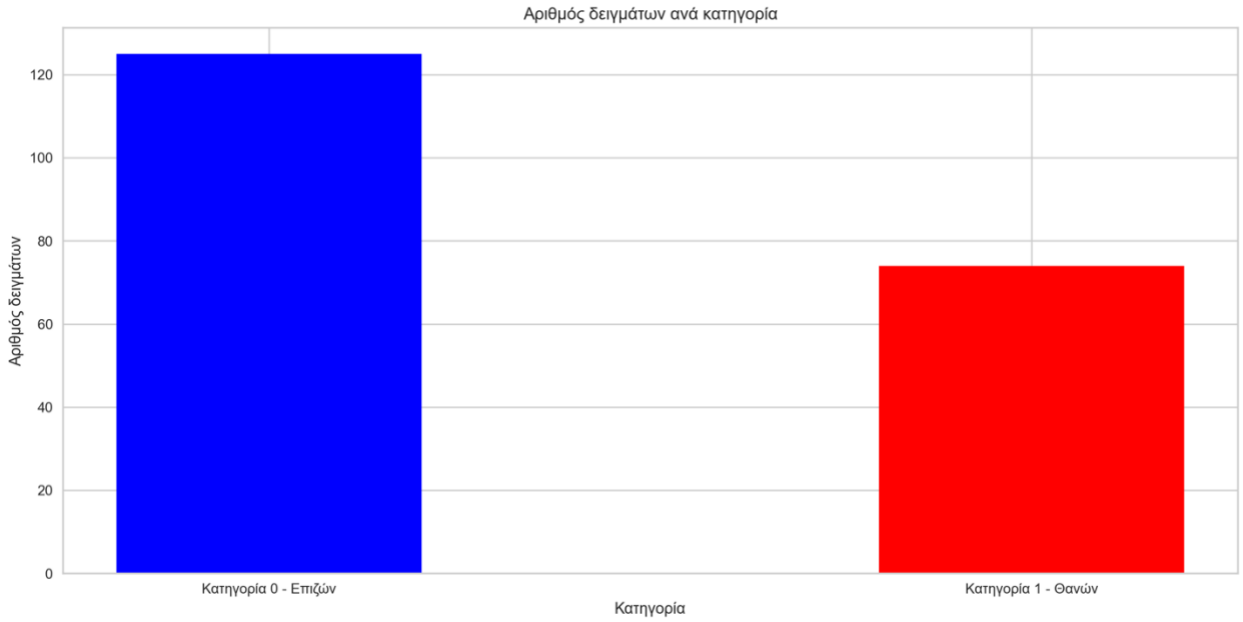
Κεφάλαιο 5: Αποτελέσματα, Παρατηρήσεις και Συμπεράσματα

Το τελικό κεφάλαιο της διπλωματικής εργασίας συνθέτει όλα τα πειραματικά ευρήματα από μια ολοκληρωμένη σύγκριση της απόδοσης των μοντέλων μηχανικής μάθησης μέσω μιας σειράς πειραμάτων με στόχο τη βελτίωση της πρόβλεψης της θνησιμότητας (θανάτου) σε ασθενείς με εγκεφαλικές μεταστάσεις με βάση ακτινολογικά και κλινικά δεδομένα. Αρχικά παρουσιάζουμε κάποια στατιστικά για το σύνολο δεδομένων μας μαζί με κάποιες πιθανές προκλήσεις που πιθανόν να συναντήσουμε. Στην συνέχεια αξιολογούμε τις επιδράσεις διαφόρων τεχνικών προεπεξεργασίας και βελτιστοποίησης στην απόδοση του μοντέλου σε διάφορες μετρικές, όπως η ευαισθησία (Sensitivity), το F1-score, η ακρίβεια(Accuracy), η AUC και η Precision. Τα πειράματα εξελίσσονται σταδιακά, ενσωματώνοντας πιο προηγμένες τεχνικές σε κάθε βήμα, επιτρέποντάς μας να παρατηρήσουμε τον αντίκτυπο κάθε παρέμβασης. Θα επισημάνουμε επίσης τυχόν ενδιαφέρουσες παρατηρήσεις σχετικά με τα πλεονεκτήματα και τις αδυναμίες συγκεκριμένων μοντέλων με βάση τις επιλεγμένες μεθόδους αξιολόγησης.

Στη συνέχεια, το κεφάλαιο θα εξετάσει μερικές τεχνικές βελτιστοποίησης που χρησιμοποιήσαμε για να βελτιώσουμε την απόδοση του μοντέλου. Θα παρουσιάσουμε τον αντίκτυπο της ρύθμισης των υπερπαραμέτρων, της μάθησης με ευαισθησία στο κόστος, την επιλογή των κατάλληλων τεχνικών προεπεξεργασίας δεδομένων και της επιλογής των πιο σημαντικών χαρακτηριστικών και μοντέλων ως προς την πρόβλεψη της θνησιμότητας. Εδώ, θα συζητήσουμε πώς αυτές οι βελτιστοποιήσεις επηρέασαν την ικανότητα των μοντέλων να εντοπίζουν σωστά τους ασθενείς με υψηλότερο κίνδυνο θανάτου. Τέλος, το κεφάλαιο θα ολοκληρωθεί με την παρουσίαση του καταλληλότερου μοντέλου για τη συγκεκριμένη εφαρμογή μας, λαμβάνοντας υπόψη τόσο την απόδοση όσο και την ερμηνευσιμότητα. Επίσης θα συζητήσουμε τους περιορισμούς της μελέτης μας και τους πιθανούς τομείς για μελλοντική έρευνα.

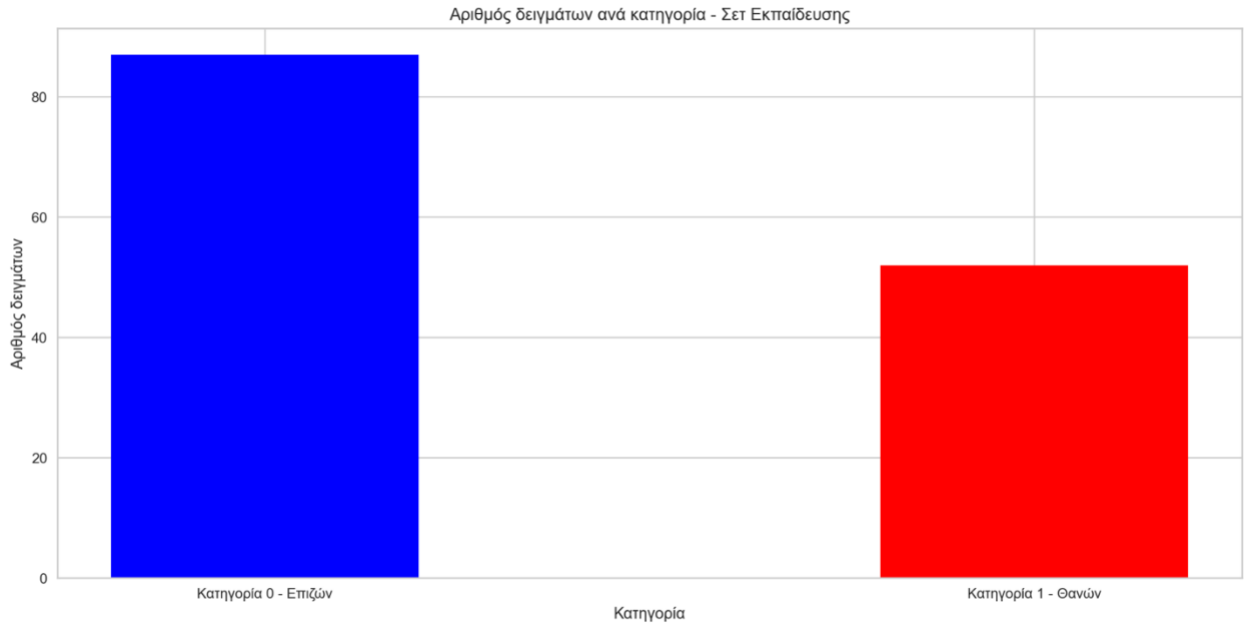
5.1: Αποτελέσματα

Το σημαντικότερο πρόβλημα στην ανάπτυξη του μοντέλου πρόβλεψης για την θνησιμότητα ενός ασθενούς με όγκο στον εγκέφαλο, με το παρόν σύνολο δεδομένων, ήταν η σχετική ανισορροπία των δύο κατηγοριών, με την κατηγορία επιζών να έχει σχεδόν διπλάσια ποσότητα δειγμάτων από την κατηγορία θανών. Εξαιτίας αυτού του γεγονότος τα μοντέλα που εκπαιδεύονται σε μη ισορροπημένα δεδομένα εμφανίζουν ιδιαίτερα χαμηλές τιμές στην ευαισθησία, δηλαδή στον εντοπισμό της κατηγορίας θανών σε νέα, άγνωστα δεδομένα. Ένα ακόμα επίσης σημαντικό πρόβλημα ήταν ο περιορισμένος αριθμός συνολικών δειγμάτων σε σχέση με τα παραγόμενα χαρακτηριστικά radiomics. Για να αντιμετωπίσουμε αυτά τα πρόβλημα εφαρμόσαμε τεχνικές, όπως αναλύθηκαν στο προηγούμενο κεφάλαιο. Η σχετική κατανομή των κατηγοριών φαίνεται στο παρακάτω σχήμα.

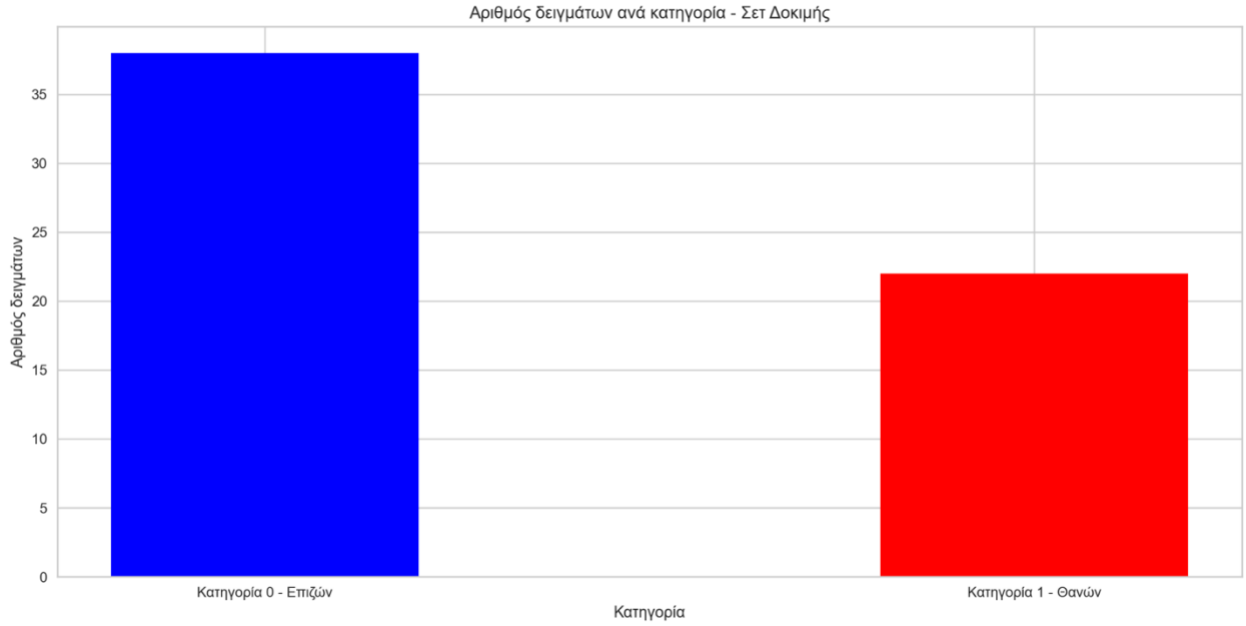


Εικόνα 33 - Κατανομή των Κατηγοριών των Δειγμάτων

Τα ποσοστά των δύο κατηγοριών μέσα στο συνολικό σύνολο δεδομένων είναι 63% για την κατηγορία 0 και το υπολειπόμενο 37% για την κατηγορία 1. Αυτά τα ποσοστά διατηρούνται ακόμα και μετά τον χωρισμό στα σετ εκπαίδευσης και δοκιμής. Συγκεκριμένα οι κατανομές των δειγμάτων για κάθε σετ φαίνεται παρακάτω:



Εικόνα 35 - Σετ Εκπαίδευσης - Κατανομή των Κατηγοριών των Δειγμάτων



Εικόνα 34 - Σετ Δοκιμής - Κατανομή των Κατηγοριών των Δειγμάτων

Στο επόμενο στάδιο θα παρουσιάσουμε τα διαφορετικά αποτελέσματα κατά την διαδικασία της προεπεξεργασίας με τα διαφορετικά χαρακτηριστικά και πως το κάθε ένα μας ωφελεί. Σε όλα τα πειράματα

χρησιμοποιήθηκε μια στρατηγική διαστρωματικής διασταυρούμενης επικύρωσης **K**-πλαισίων (Stratified K-fold Cross Validation) για να εξασφαλιστούν αξιόπιστες και αμερόληπτες εκτιμήσεις επιδόσεων καθώς επίσης και την εφαρμογή ισχυρής κανονικοποίησης (Robust Normalization) στα χαρακτηριστικά για τον περιορισμό της επίδρασης των ακραίων τιμών. Τα σχετικά βήματα για τα πειράματα, που ενσωματώνονται σταδιακά περιγράφονται από τις ακόλουθες τεχνικές:

Βασικές (Basic): Robust Normalization και Stratified K-Fold cross-validation.

Αφαίρεση ακραίων τιμών (Remove Outliers): Αφαίρεση των πιο ακραίων σημείων σε ποσοστό 1% των δεδομένων.

Εκκαθάριση της πολυσυγγραμμικότητας (Remove Multicollinearity): Αφαίρεση των πλεοναζουσών εξαιρετικά συσχετισμένων χαρακτηριστικών με κατώφλι συσχέτισης 0,75.

Διόρθωση της ανισορροπίας με συνθετικά δεδομένα SMOTE (Fix Imbalance with Synthetic Data): Χρήση της τεχνικής συνθετικής δειγματοληψίας μειοψηφίας (SMOTE) για τη δημιουργία συνθετικών σημείων δεδομένων και την αντιμετώπιση της ανισορροπίας κλάσεων.

Γραμμική PCA (Linear PCA): Εφαρμογή της ανάλυσης κύριων συνιστωσών (PCA) για μείωση της διαστατικότητας στα χαρακτηριστικά.

Επιλογή χαρακτηριστικών (Feature Selection): Χρήση ενός ταξινομητή Light Gradient Boosting Machine (LGBM) για την επιλογή των σημαντικότερων χαρακτηριστικών.

Στην συνέχεια για κάθε ένα από τα πειράματα που περιγράφονται παραπάνω θα εκπαιδεύσουμε τους αλγόριθμους που είδαμε στον πίνακα [- *Μοντέλα Εκπαίδευσης για την Πειραματική* Διαδικασία] και θα δούμε τα αποτελέσματα τους και θα τα αξιολογήσουμε με βάση τις μετρικές που μας ενδιαφέρουν. Ως μετρική βελτιστοποίησης επιλέξαμε την μετρική της ευαισθησίας (Sensitivity) που εστιάζει στο ποσοστό των True Positives, καθώς θέλουμε να εξασφαλίσουμε πως θα προβλέπουμε σωστά τις περιπτώσεις που έχουν όγκο στον εγκέφαλο και αντιμετωπίζουν κίνδυνο θανάτου. Με την αύξηση του sensitivity χάνουμε στην ακρίβεια του μοντέλου καθώς εξ ορισμού οι δύο μετρικές είναι αντιστρόφως ανάλογες μεταξύ τους. Στην δικιά μας έρευνα δώσαμε μεγάλη σημασία στην πρόβλεψη της κατηγορίας 1, που εκφράζει την πιθανότητα θανάτου του ασθενή, ακόμα και αν αυτή η πρόβλεψη είναι λανθασμένη, με στόχο να προλαμβάνουμε μία κατάσταση για έναν ασθενή ακόμα και αν αυτή δεν τελικά δεν καταλήγει στον θάνατο.

5.1.1: Πείραμα 1: Βασική Προεπεξεργασία (Robust Normalization, Stratified K-Fold Cross-Validation)

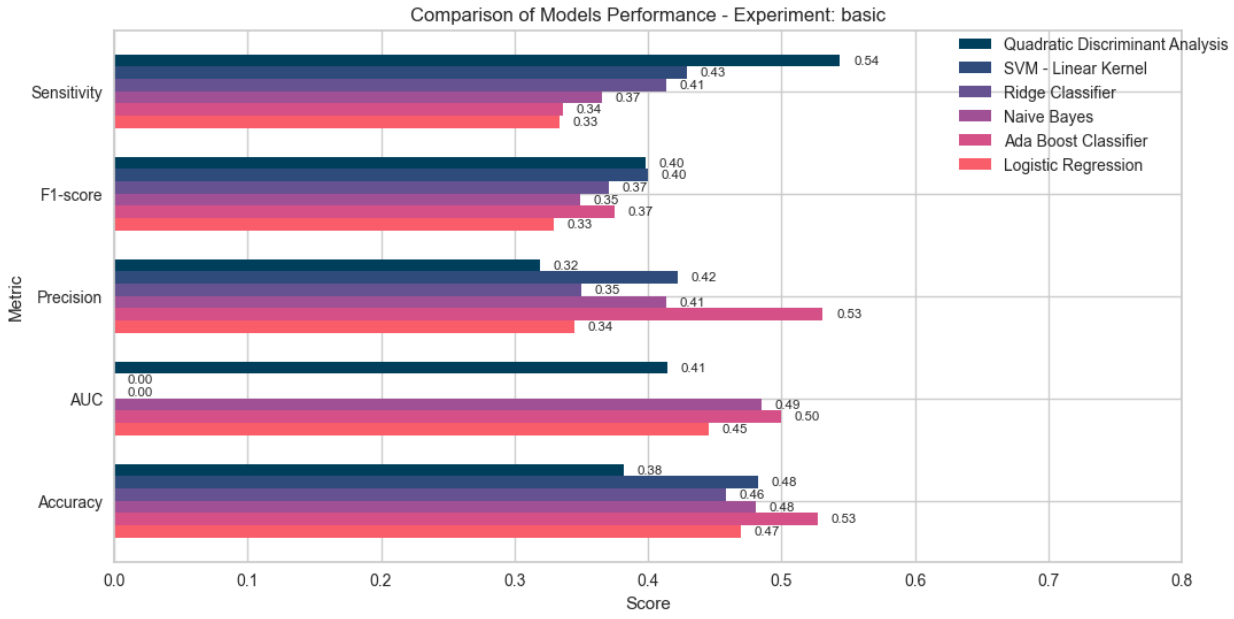
Στο βασικό πείραμα, εφαρμόσαμε ελάχιστη προεπεξεργασία με ισχυρή κανονικοποίηση και χρησιμοποιήσαμε στρωματοποιημένη K-πτυχή διασταυρούμενης επικύρωσης για να εξασφαλίσουμε ισορροπημένη δειγματοληψία των κλάσεων-στόχων.

	Description	Value
0	Session id	42
1	Target	Death
2	Target type	Binary
3	Original data shape	(198, 1198)
4	Transformed data shape	(198, 1198)
5	Numeric features	1197
6	Normalize	True
7	Normalize method	robust
8	Fold Generator	StratifiedKFold
9	Fold Number	10
10	Use GPU	True
11	Experiment Name	basic

Εικόνα 36 - Ρυμθίσεις Πειράματος 1 - Basic

Στην Εικόνα 36 παρουσιάζονται οι τεχνικές που χρησιμοποιήθηκαν όπως προαναφέρθηκαν παραπάνω μαζί με τις πληροφορίες σχετικά με τις διαστάσεις και τον τύπο των δεδομένων. Βλέπουμε πως έχουμε στην διάθεση μας 198 δείγματα με διάσταση χαρακτηριστικών 1197 κάθως και την μεταβλητή στόχο μας.

Μετά την αξιολόγηση διάφορων μοντέλων, με τα καλύτερα 6 που ξεχωρίζουν να είναι: Quadratic Discriminant Analysis, SVM (Linear Kernel), Ridge Classifier, Naive Bayes, Ada Boost Classifier και Logistic Regression.



Εικόνα 37 - Επιδόσεις Καλύτερων Μοντέλων Πειράματος 1 - Basic

Παρατηρούμε πως τα περισσότερα μοντέλα έχουν χαμηλές επιδόσεις και συγκεκριμένα η ακρίβεια να κυμαίνεται γύρω στο 40-48 % με το μοντέλο Ada Boost να ξεχωρίζει με 53%. Το μοντέλο Quadratic Discriminant Analysis πέτυχε την υψηλότερη βαθμολογία ευαισθησίας 54%, υποδηλώνοντας ότι είναι το καλύτερο στην καταγραφή των αληθώς θετικών αποτελεσμάτων σε αυτό το πείραμα με την αρχική βασική ρύθμιση. Όσων αφορά τις μετρικές F1-score και AUC, όλα τα μοντέλα δυσκολεύτηκαν και ήταν αξιοσημείωτα αδύναμα, υπογραμμίζοντας την γενικά χαμηλή δυνατότητα διάκρισης μεταξύ των κλάσεων στο βασικό μας πείραμα.

Παρά την απλότητα αυτού του πειράματος, λειτούργησε ως θεμελιώδες σημείο αναφοράς για την αξιολόγηση της επίδρασης μελλοντικών τεχνικών.

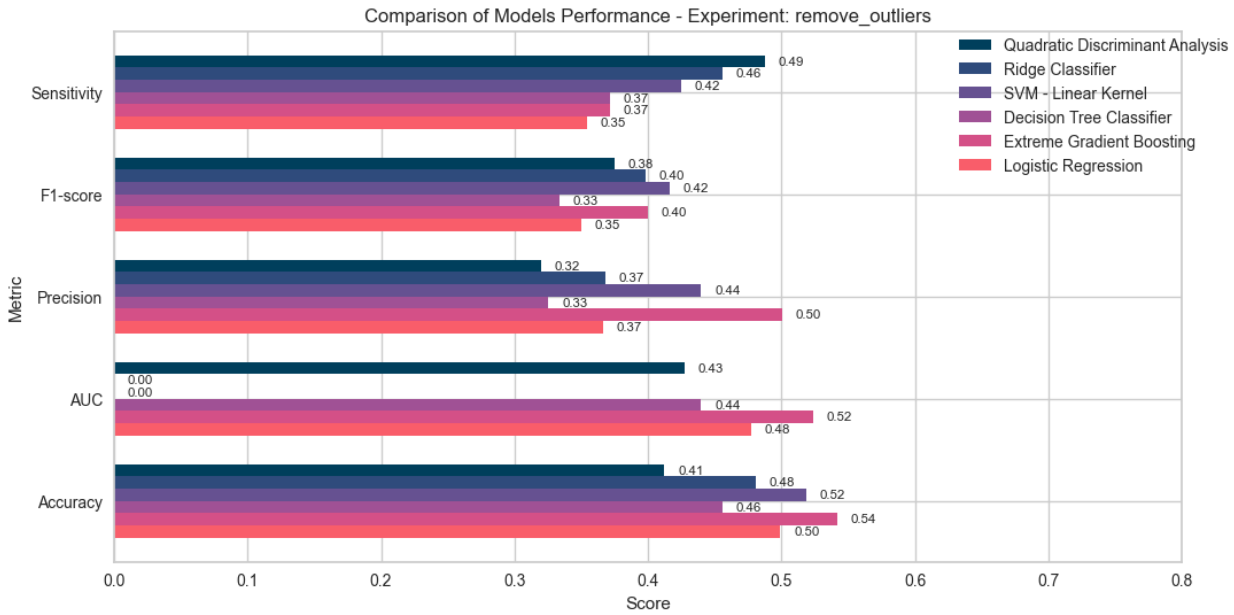
5.1.2: Πείραμα 2: Αφαίρεση Ακραίων Τιμών

Οι ακραίες τιμές μπορούν να αλλοιώσουν την απόδοση του μοντέλου, ιδίως σε σύνολα ιατρικών δεδομένων όπου τα ασυνήθιστα σημεία δεδομένων μπορεί να μην αντιπροσωπεύουν πραγματικές κλινικές τάσεις αλλά μάλλον σφάλματα στη συλλογή δεδομένων. Σε αυτό το πείραμα, το 1% των πιο ακραίων ακραίων τιμών αφαιρέθηκε από το σύνολο δεδομένων.

	Description	Value
0	Session id	42
1	Target	Death
2	Target type	Binary
3	Original data shape	(198, 1198)
4	Transformed data shape	(196, 1198)
5	Numeric features	1197
6	Remove outliers	True
7	Outliers threshold	0.010000
8	Normalize	True
9	Normalize method	robust
10	Fold Generator	StratifiedKFold
11	Fold Number	10
12	Use GPU	True
13	Experiment Name	remove_outliers

Εικόνα 38 - Ρυμθίσεις Πειράματος 2 - Remove Outliers

Η ουσιαστική διαφορά για το δεύτερο πείραμα είναι η αφαίρεση 2 δειγμάτων από το σύνολο των δεδομένων. Παρακάτω παρουσιάζονται οι επιδόσεις των καλύτερων μοντέλων μεταξύ των μετρικών.



Εικόνα 39 - Επιδόσεις Καλύτερων Μοντέλων Πειράματος 2 - Remove Outliers

Η ευαισθησία καθώς και η ακρίβεια παρέμειναν στα ίδια επίπεδα για τα περισσότερα μοντέλα, με τα μοντέλο QDA και XGBoost να πετυχαίνουν τις καλύτερες τιμές 49% και 54% αντίστοιχα. Το αποτέλεσμα είναι απολύτως αναμενόμενο, καθώς το σύνολο δεδομένων μας είναι κατά 99% ίδιο με αυτό του προηγούμενου πειράματος, και έτσι δεν παρατηρείται κάποια σημαντική βελτίωση προς την σωστή κατεύθυνση.

Συγκεκριμένα, συγκριτικά με το Πείραμα 1, η αφαίρεση των ακραίων τιμών οδήγησε σε μία μέτρια βελτίωση σε όλες τις μετρήσεις. Ενώ η συνολική ακρίβεια παρέμεινε παρόμοια με το βασικό πείραμα, οι βελτιώσεις στην ευαισθησία, την AUC και την ακρίβεια ήταν περισσότερο ορατές αλλά και πάλι όχι αξιοσημείωτες. Αυτό καθιστά την προσέγγιση αφαίρεσης των ακραίων τιμών πιο αποτελεσματική στη σταθεροποίηση των επιδόσεων χωρίς ριζική μεταβολή της συνολικής δομής του συνόλου δεδομένων.

5.1.3: Πείραμα 3: Αφαίρεση της Πολυσυγγραμμικότητας

Η πολυσυγγραμμικότητα μεταξύ των χαρακτηριστικών μπορεί να επισκιάσει την ερμηνευσιμότητα του μοντέλου και να διογκώσει τη σημασία ορισμένων μεταβλητών σε σύνολα ιατρικών δεδομένων, όπου πολλές κλινικές μεταβλητές συχνά συσχετίζονται. Σε αυτό το πείραμα αφαιρέθηκαν τα χαρακτηριστικά με κατώτατο όριο συσχέτισης 0,75 ή υψηλότερο.

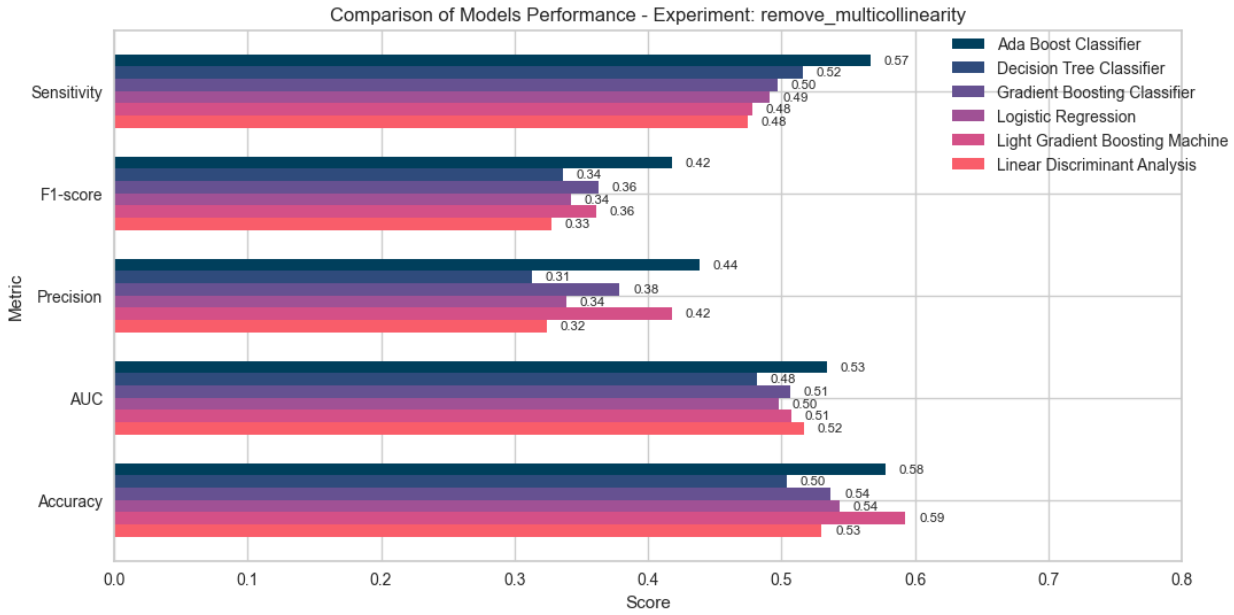
	Description	Value
0	Session id	42
1	Target	Death
2	Target type	Binary
3	Original data shape	(198, 1198)
4	Transformed data shape	(196, 74)
5	Numeric features	1197
6	Remove multicollinearity	True
7	Multicollinearity threshold	0.750000
8	Remove outliers	True
9	Outliers threshold	0.010000
10	Normalize	True
11	Normalize method	robust
12	Fold Generator	StratifiedKFold
13	Fold Number	10
14	Use GPU	True
15	Experiment Name	remove_multicollinearity

Εικόνα 40 - Ρυμθίσεις Πειράματος 3 - Remove Multicollinearity

Στην Εικόνα 40 παρουσιάζονται οι ρυθμίσεις που έχουμε εφαρμόσει μέχρι στιγμής, οι ίδιες με τα προηγούμενα πειράματα με επιπλέον βήμα την αφαίρεση της πολυσυγγραμμικότητας. Βλέπουμε πως μετά την εφαρμογή αυτών των τεχνικών, η διάσταση των χαρακτηριστικών μας μειώθηκε δραστικά, με συνολικά 73 χαρακτηριστικά από 1197 που ήταν αρχικά, μία μείωση περίπου 93%.

Τα συγκεντρωτικά αποτελέσματα παρουσιάζονται στην Εικόνα 41. Η ευαισθησία και το F1-Score παρουσίασαν περαιτέρω βελτίωση, ιδίως στα μοντέλα που βασίζονται σε δέντρα αποφάσεων, με την ευαισθησία να αυξάνεται σε 52%. Αυτό υποδηλώνει ότι η αφαίρεση περιττών χαρακτηριστικών διευκόλυνε τα μοντέλα να αποτυπώσουν τις σχέσεις μεταξύ ανεξάρτητων μεταβλητών και αποτελεσμάτων. Η μετρική AUC βελτιώθηκε οριακά αλλά σημαντικά για αρκετά μοντέλα, υποδεικνύοντας ότι η αφαίρεση συσχετιζόμενων μεταβλητών βοήθησε τα μοντέλα να διαφοροποιούνται καλύτερα μεταξύ των κλάσεων. Η ακρίβεια παρουσίασε σημαντική βελτίωση για όλα τα μοντέλα, καθώς αυτά πλέον δεν επηρεάζονται αρνητικά από τις έντονα συσχετιζόμενες μεταβλητές και έτσι έγιναν πιο ακριβείς στις προβλέψεις τους.

Όταν συγκρίνουμε το παρόν πείραμα με το προηγούμενο, το βασικό πλεονέκτημα αυτού του πειράματος ήταν η καλύτερη ερμηνευσιμότητα των μοντέλων λόγω της εξάλειψης των συσχετισμένων μεταβλητών. Ενώ τα κέρδη ευαισθησίας και ακρίβειας ήταν σημαντικά, οι συνολικές βελτιώσεις ήταν σταδιακές. Αυτή η μέθοδος λειτουργεί καλά για σύνολα δεδομένων με πολλές συσχετιζόμενες κλινικές μεταβλητές, αλλά δεν άλλαξε δραστικά το Precision ή το F1-score.



Εικόνα 41 - Επιδόσεις Καλύτερων Μοντέλων Πειράματος 3 - Remove Multicollinearity

5.1.4: Πείραμα 4: Διόρθωση της Ανισορροπίας με Συνθετικά Δεδομένα (SMOTE)

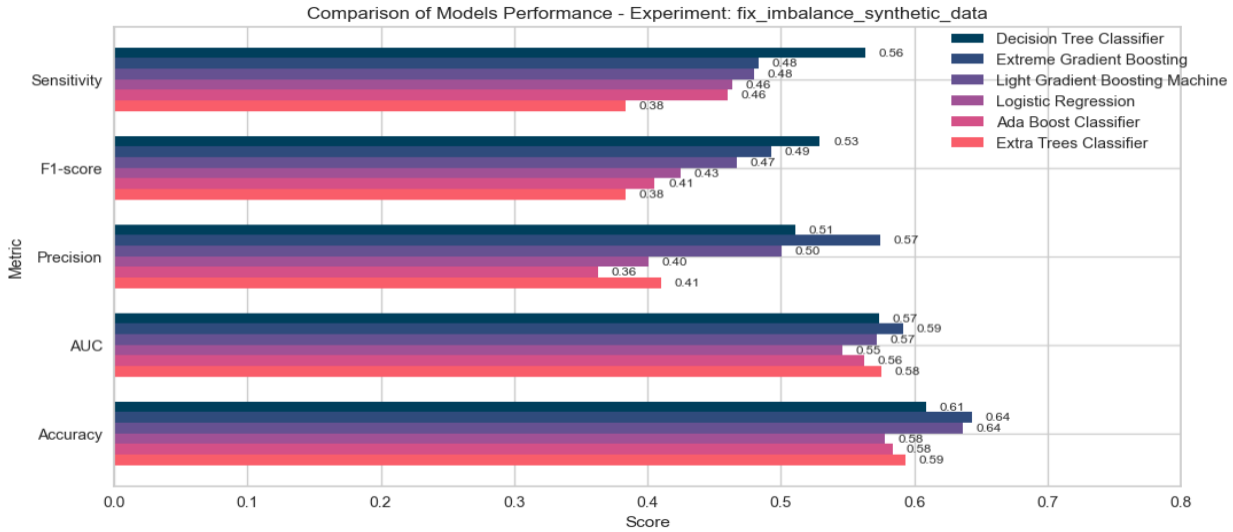
Τα σύνολα ιατρικών δεδομένων είναι συχνά ιδιαίτερα ανισόρροπα, όπου ορισμένες κατηγορίες υπό-εκπροσωπούνται σημαντικά. Το ίδιο παρατηρούμε και στο δικό μας σύνολο δεδομένων όπου η κατηγορία 1 έχει σημαντικά λιγότερα δείγματα από την κατηγορία 0. Έτσι και σε αυτό το πείραμα, εφαρμόστηκε η τεχνική SMOTE (Synthetic Minority Over-sampling Technique) για την εξισορρόπηση του συνόλου δεδομένων με τη δημιουργία συνθετικών δειγμάτων για την τάξη της μειονότητας. Για το παρόν πείραμα παρατηρούμε τα εξής αποτελέσματα όσων αφορά τις μετρικές:

Η ευαισθησία σημείωσε μεγάλη διακύμανση σε αυτό το πείραμα, με τα μοντέλα να επιτυγχάνουν σκορ ευαισθησίας που κυμαίνεται από 38% έως και 56% με κορυφαίο το μοντέλο Decision Tree Classifier. Αυτό ήταν κρίσιμο για τη βελτίωση της ανίχνευσης των υποεκπροσωπούμενων κλάσεων. Το F1-Score βελτιώθηκε επίσης σημαντικά, αντανακλώντας μια καλύτερη ισορροπία μεταξύ ακρίβειας και ανάκλησης. Οι μετρικές Precision και AUC συνέχισαν να παρουσιάζουν βελτίωση, ιδίως στα μοντέλα που βασίζονται σε δέντρα. Το μοντέλο Extreme Gradient Boosting είχε εξαιρετικά καλές επιδόσεις με AUC 59%. Η ακρίβεια παρέμεινε σταθερή ανάμεσα στα μοντέλα του πειράματος αλλά βελτιώθηκε ελαφρώς σε σύγκριση με προηγούμενα πειράματα, υπογραμμίζοντας τη σημασία της εστίασης σε μετρικές όπως το F1-score και η ευαισθησία σε μη ισορροπημένα σύνολα δεδομένων.

	Description	Value
0	Session id	42
1	Target	Death
2	Target type	Binary
3	Original data shape	(198, 1198)
4	Transformed data shape	(232, 74)
5	Numeric features	1197
6	Remove multicollinearity	True
7	Multicollinearity threshold	0.750000
8	Remove outliers	True
9	Outliers threshold	0.010000
10	Fix imbalance	True
11	Fix imbalance method	SMOTE
12	Normalize	True
13	Normalize method	robust
14	Fold Generator	StratifiedKFold
15	Fold Number	10
16	Use GPU	True
17	Experiment Name	fix_imbalance_synthetic_data

Εικόνα 42 - Ρυμθίσεις Πειράματος 4 - Fix Imbalance with Synthetic Data

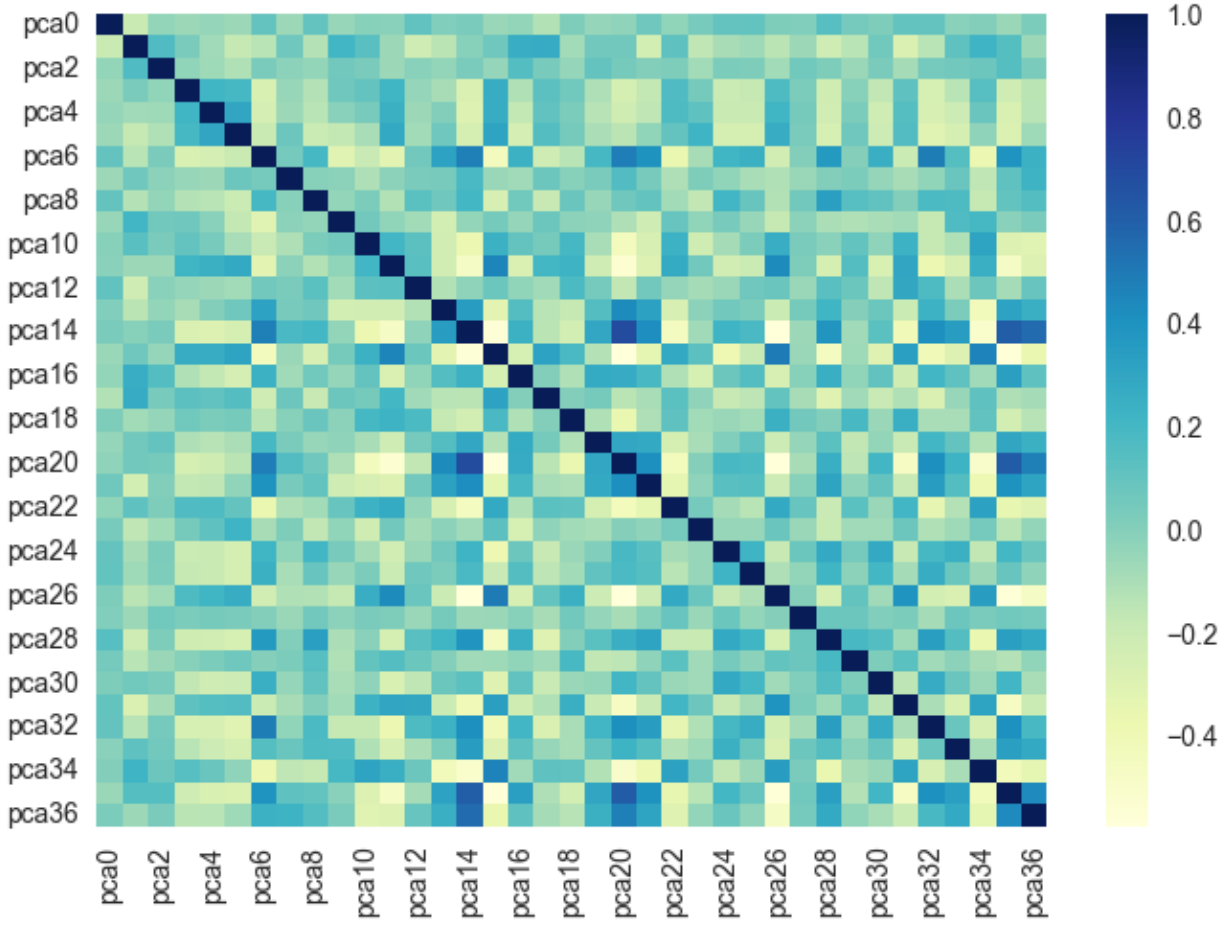
Το μεγαλύτερο πλεονέκτημα σε αυτό το πείραμα ήταν η σημαντική βελτίωση στην ανίχνευση της μειοψηφούσας κατηγορίας 1. Η ευαισθησία σημείωσε σημαντικό άλμα, πράγμα που σημαίνει ότι τα μοντέλα ήταν σε θέση να ανιχνεύσουν περισσότερες από τις θετικές περιπτώσεις. Παρόλο που υπάρχει ένας μικρός κίνδυνος υπερπροσαρμογής σε συνθετικά δεδομένα, τα οφέλη της αυξημένης ευαισθησίας και του F1-score υπερκαλύπτουν κατά πολύ αυτόν τον κίνδυνο, ιδίως για ιατρικές προβλέψεις όπου η κλάση μειονότητας αντιπροσωπεύει κρίσιμες περιπτώσεις.



Εικόνα 43 - Επιδόσεις Καλύτερων Μοντέλων Πειράματος 4 - Fix Imbalance with Synthetic Data

5.1.5: Πείραμα 5: Μείωση Διαστάσεων με Χρήση PCA

Η Ανάλυση Κύριων Συνιστωσών (PCA) εισήχθη ως τεχνική μείωσης της διαστατικότητας μετά την επιλογή χαρακτηριστικών. Η τεχνική αυτή αποσκοπούσε στη μείωση του χώρου των χαρακτηριστικών διατηρώντας όσο το δυνατόν μεγαλύτερη διακύμανση. Για το δικό μας σύνολο δεδομένων επιλέχθηκε η μέθοδο PCA με πυρήνες γραμμικής συνάρτησης βάσης (RBF) με την υποθεση της γραμμική σχέση μεταξύ των χαρακτηριστικών. Στην παρακάτω εικόνα παρουσιάζονται οι κύριες συνιστώσες και ο βαθμός συσχέτισης μεταξύ τους σε μορφή πίνακα.

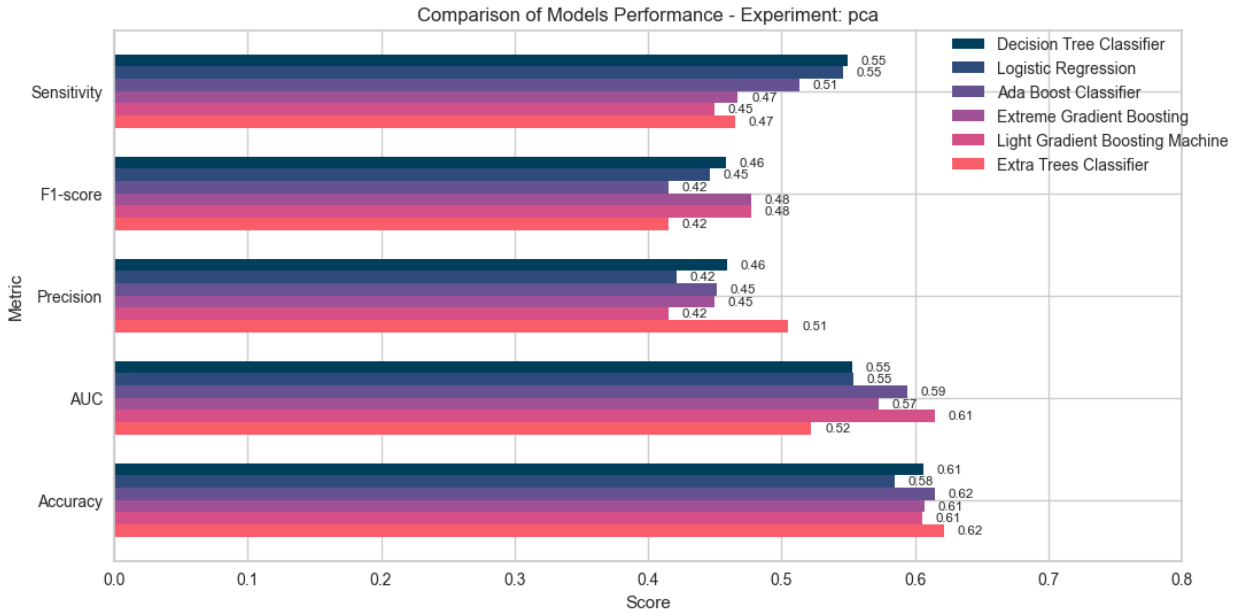


Εικόνα 44 - Πίνακας Συσχέτισης Πειράματος 4 - PCA

	Description	Value
0	Session id	42
1	Target	Death
2	Target type	Binary
3	Original data shape	(198, 1198)
4	Transformed data shape	(232, 38)
5	Numeric features	1197
6	Remove multicollinearity	True
7	Multicollinearity threshold	0.750000
8	Remove outliers	True
9	Outliers threshold	0.010000
10	Fix imbalance	True
11	Fix imbalance method	SMOTE
12	Normalize	True
13	Normalize method	robust
14	PCA	True
15	PCA method	linear
16	PCA components	37
17	Fold Generator	StratifiedKFold
18	Fold Number	10
19	Use GPU	True
20	Experiment Name	pca

Εικόνα 45 - Ρυμθίσεις Πειράματος 5 - PCA

Από τα αποτελέσματα που συγκεντρώνουμε βλέπουμε πως το μοντέλο Decision Tree Classifier διατήρησε την ισχυρή του απόδοση με ευαισθησία 55%, δείχνοντας ευρωστία ακόμη και μετά τη μείωση της διαστατικότητας. Υπήρξε μια μικρή αύξηση αλλά γενικά σε χαμηλά επίπεδα στα F1-scores με ένα εύρος τιμών από 42%-48%, υποδεικνύοντας κάποια απώλεια στην ισορροπία μεταξύ ακρίβειας και ανάκλησης. Οι βαθμολογίες AUC παρέμειναν σχετικά σταθερές σε όλα τα μοντέλα, με τα Ada Boost Classifier και Light Gradient Boosting Machine να διατηρούν υψηλές επιδόσεις με 59% και 61%, αντίστοιχα. Τέλος το μοντέλο με τις καλύτερες γενικά επιδόσεις, Decision Tree Classifier, πέτυχε από τις υψηλότερες τιμές για την ακρίβεια 61%, υποδεικνύοντας ότι η PCA βοήθησε στη διατήρηση της προβλεπτικής ισχύος παρά τη μείωση της διαστατικότητας.



Εικόνα 46 - Επιδόσεις Καλύτερων Μοντέλων Πειράματος 5 - PCA

Ενώ η PCA βοήθησε στον εξορθολογισμό του χώρου χαρακτηριστικών, ο αντίκτυπός της μέθοδο στην απόδοση του μοντέλου είχε συγκριτικά μικρότερη επίδραση σε σχέση με τον προηγούμενο πείραμα της συμπλήρωσης της ανισορροπίας. Ωστόσο, συνέβαλε στη σταθερότητα του μοντέλου και στη βελτίωση της υπολογιστικής απόδοσης.

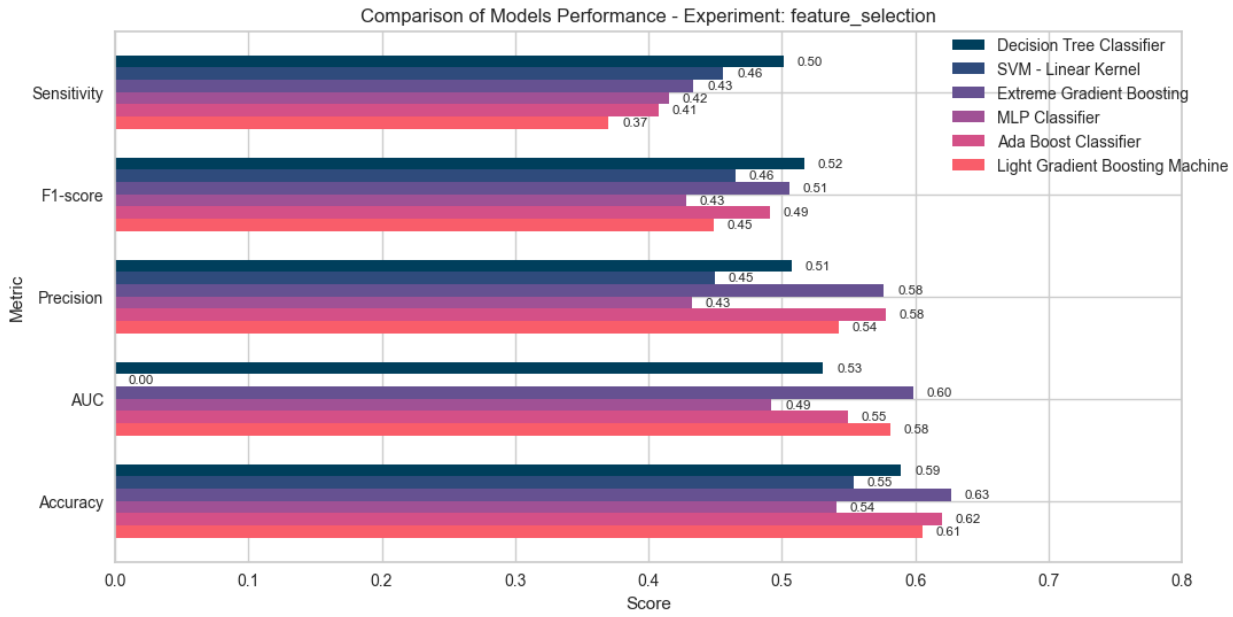
5.1.6: Πείραμα 6: Επιλογή Χαρακτηριστικών με τον Ταξινομητή LGBM

Σε αυτό το πείραμα, εισαγάγαμε την επιλογή χαρακτηριστικών με τη χρήση της επιλογής με βάση τη σημασία του ταξινομητή LightGBM για την εξάλειψη των άσχετων χαρακτηριστικών.

	Description	Value
0	Session id	42
1	Target	Death
2	Target type	Binary
3	Original data shape	(198, 1198)
4	Transformed data shape	(232, 21)
5	Numeric features	1197
6	Remove multicollinearity	True
7	Multicollinearity threshold	0.750000
8	Remove outliers	True
9	Outliers threshold	0.010000
10	Fix imbalance	True
11	Fix imbalance method	SMOTE
12	Normalize	True
13	Normalize method	robust
14	PCA	True
15	PCA method	linear
16	PCA components	37
17	Feature selection	True
18	Feature selection method	classic
19	Feature selection estimator	lightgbm
20	Number of features selected	20
21	Fold Generator	StratifiedKFold
22	Fold Number	10
23	Use GPU	True
24	Experiment Name	feature_selection

Εικόνα 47 - Ρυμθίσεις Πειράματος 6 - Feature Selection

Η επιλογή χαρακτηριστικών με LGBM φαίνεται πως έχει συνολικά καλύτερα αποτελέσματα όσον αφορά την ακρίβεια, την AUC και το Precision. Το LGBM υπερτερεί σταθερά έναντι άλλων μοντέλων μετά την επιλογή χαρακτηριστικών, γεγονός που υποδηλώνει ότι η εξάλειψη των άσχετων χαρακτηριστικών έχει ισχυρότερη επίδραση από την απλή μείωση των διαστάσεων μέσω PCA. Ωστόσο στην συγκεκριμένη περίπτωση υπάρχει και ο κίνδυνος να βρισκόμαστε σε μία κατάσταση υπερπροσαρμογής (overfitting). Αυτό μπορούμε να το υποθέσουμε από την σχετική διαφορά μεταξύ των μετρικών AUC και της ακρίβειας, κάτι που υποδηλώνει αδυναμία γενίκευσης του μοντέλου. Υπάρχουν διάφοροι λόγοι γιατί αυτό συμβαίνει, μερικοί από αυτούς είναι η πολυπλοκότητα που έχουμε εισάγει, το ρίσκο χαμηλής γενίκευσης λόγω μεγάλης απώλειας χαρακτηριστικών, κάποιες φορές υψηλή διακύμανση στην απόδοση των μοντέλων, η χρήση συνθετικών δεδομένων, κτλ.



Εικόνα 48 - Επιδόσεις Καλύτερων Μοντέλων Πειράματος 6 - Feature Selection

Σε αυτό το σημείο έχουμε μια γενικότερη εικόνα ως προς ποιες τεχνικές προεπεξεργασίας είχαν σημαντικό αντίκτυπο στην επίδοση αλλά επίσης και ποια μοντέλα παρατηρούμε να έχουν επαναλαμβανόμενη καλή επίδοση και καλό συνδυασμό μετρικών. Στο επόμενο κεφάλαιο θα εξετάσουμε κάποιες πιθανές βελτιστοποιήσεις ώστε να πετύχουμε καλύτερα αποτελέσματα.

5.2: Βελτιστοποίηση Αποτελεσμάτων

Σε αυτή την ενότητα, διερευνούμε και εφαρμόζουμε μια σειρά από τεχνικές βελτιστοποίησης που αποσκοπούν στη βελτίωση της απόδοσης και της γενίκευσης των μοντέλων μηχανικής μάθησης που εφαρμόσαμε στα προηγούμενα βήματα. Αφού πραγματοποιήσουμε προεπεξεργασία και επιλογή χαρακτηριστικών, όπως αναλύθηκε σε προηγούμενες ενότητες, θα επικεντρωθούμε τώρα στις ακόλουθες 3 μεθόδους βελτιστοποίησης:

1. Ενσωμάτωση δημογραφικών δεδομένων στο σύνολο χαρακτηριστικώ

2.Συνδυασμοί μοντέλων με την χρήση των μεθόδων blend και stack

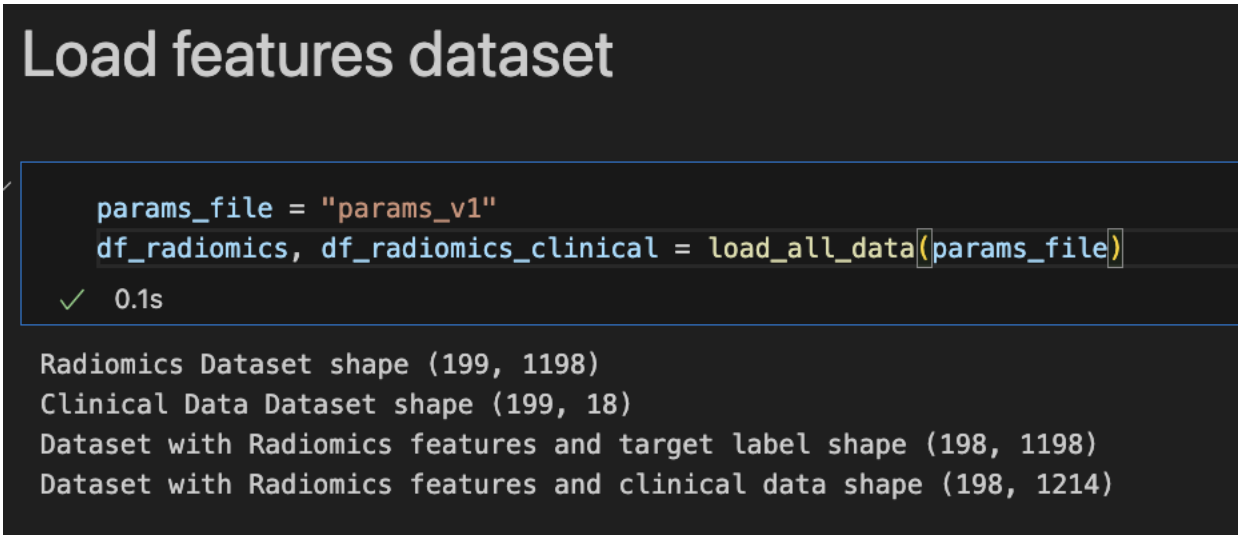
3.Αναζήτηση βέλτιστων υπερπαραμέτρων με την χρήση της αναζήτησης πλέγματος και την βελτιστοποίηση Bayes

Αυτές οι τεχνικές έχουν σχεδιαστεί για να βελτιώσουν τόσο την ακρίβεια πρόβλεψης όσο και τη σταθερότητα των μοντέλων, διασφαλίζοντας ότι παρέχουν αξιόπιστα και ερμηνεύσιμα αποτελέσματα, ιδίως για κρίσιμες προβλέψεις που σχετίζονται με τα αποτελέσματα των ασθενών.

5.2.1: Ενσωμάτωση Δημογραφικών δεδομένων

Ενώ τα χαρακτηριστικά που εξάγονται από τρισδιάστατες σαρώσεις παρέχουν εξαιρετικά λεπτομερείς ποσοτικές πληροφορίες, συχνά επωφελούνται από τη συμπερίληψη των δημογραφικών δεδομένων, όπως η ηλικία, το φύλο, η πρωτογενής προέλευση της μετάστασης και άλλα. Η ενσωμάτωση αυτών των πρόσθετων δεδομένων επιτρέπει στο μοντέλο να αξιοποιήσει ευρύτερες, συναφείς πληροφορίες, καταγράφοντας ενδεχομένως πιο ολιστικά μοτίβα στην πρόβλεψη της μεταβλητής-στόχου.

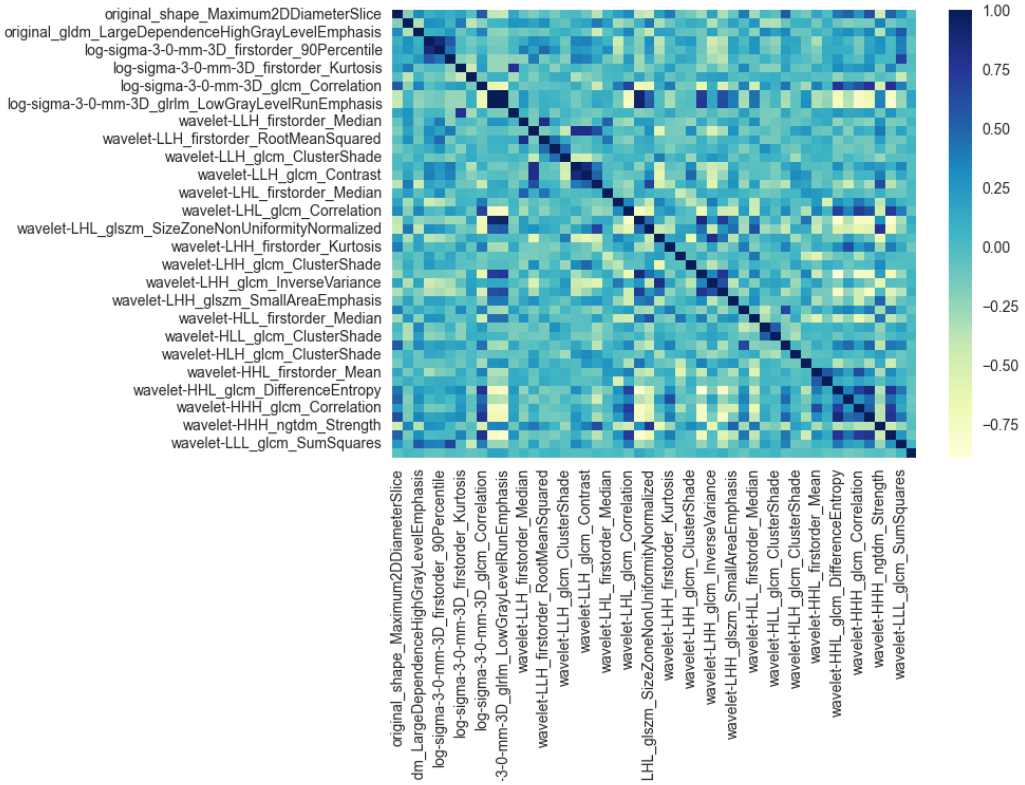
Για το βήμα αυτό, επεκτείναμε το σύνολο των χαρακτηριστικών συνδυάζοντας τα ακτινολογικά χαρακτηριστικά με κλινικά δεδομένα. Αυτά τα χαρακτηριστικά υποβλήθηκαν σε προεπεξεργασία ανάλογα με τη φύση τους: Συγκεκριμένα οι κατηγοριακές μεταβλητές, όπως π.χ. το φύλο, κωδικοποιήθηκαν με ένα σημείο για να διατηρηθεί η συμβατότητα του μοντέλου, ενώ οι συνεχείς μεταβλητές όπως π.χ. η ηλικία, κανονικοποιήθηκαν με την ίδια τεχνική (Robust Normalization) που εφαρμόστηκε στα ακτινολογικά χαρακτηριστικά για την αποφυγή προβλημάτων κλιμάκωσης.



Εικόνα 49 - Ενσωμάτωση Δημογραφικών Δεδομένων στο Σύνολο Χαρακτηριστικών

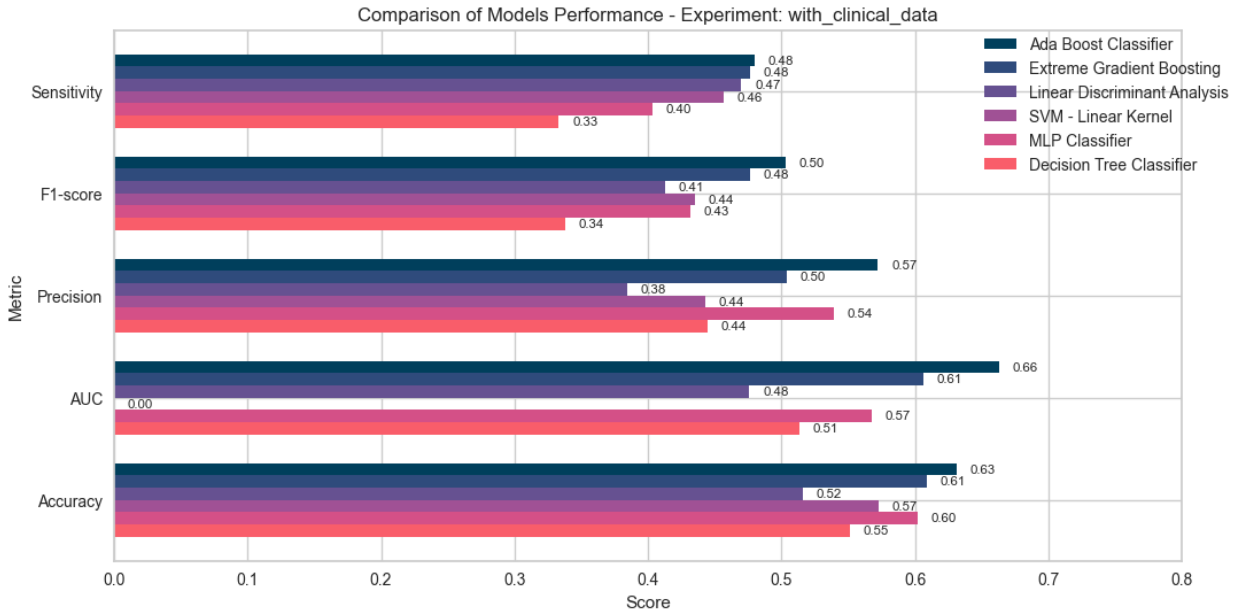
	Description	Value
0	Session id	42
1	Target	Death
2	Target type	Binary
3	Original data shape	(198, 1214)
4	Transformed data shape	(232, 51)
5	Fix imbalance	True
6	Fix imbalance method	SMOTE
7	Normalize method	robust
8	Feature selection method	classic
9	Feature selection estimator	lightgbm
10	Number of features selected	50
11	Fold Generator	StratifiedKFold
12	Fold Number	10
13	CPU Jobs	-1
14	Use GPU	True
15	Log Experiment	False
16	Experiment Name	with_clinical_data
17	USI	abdb

Εικόνα 50 - Ρυθμίσεις Εκπαίδευσης Μοντέλων για Σύνολο με Κλινικά Δεδομένα



Εικόνα 51 - Πίνακας Συσχέτισης – Σύνολο με Κλινικά Δεδομένα

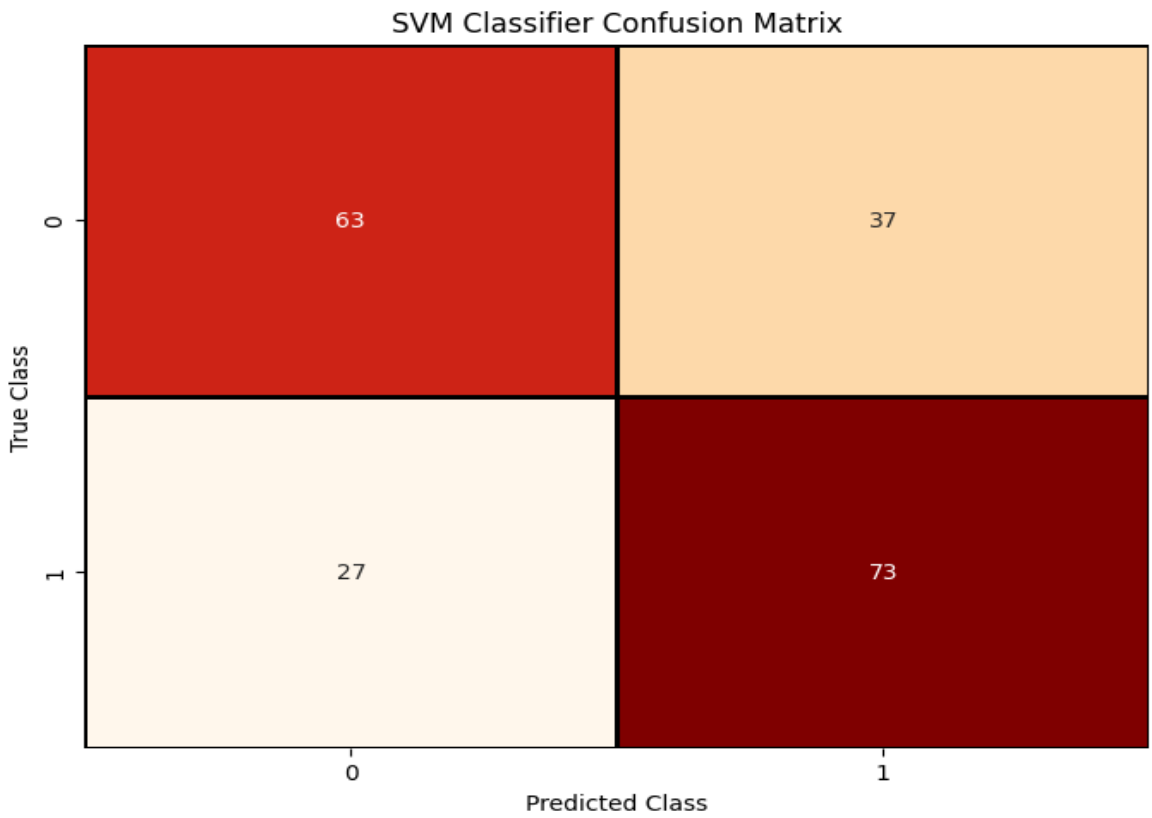
Αφού αυξήσαμε το σύνολο δεδομένων με κλινικά δεδομένα σε συνολικά 1213 χαρακτηριστικά, προσπαθούμε να μειώσουμε την πολυπλοκότητα των σταδίων προεπεξεργασίας κρατώντας μόνο τα βήματα κανονικοποίησης, διόρθωσης ανισορροπίας και επιλογής χαρακτηριστικών σε συνολικά 50, εκπαιδεύσαμε εκ νέου τα μοντέλα με τα καλύτερους 6 να φαίνονται παρακάτω:



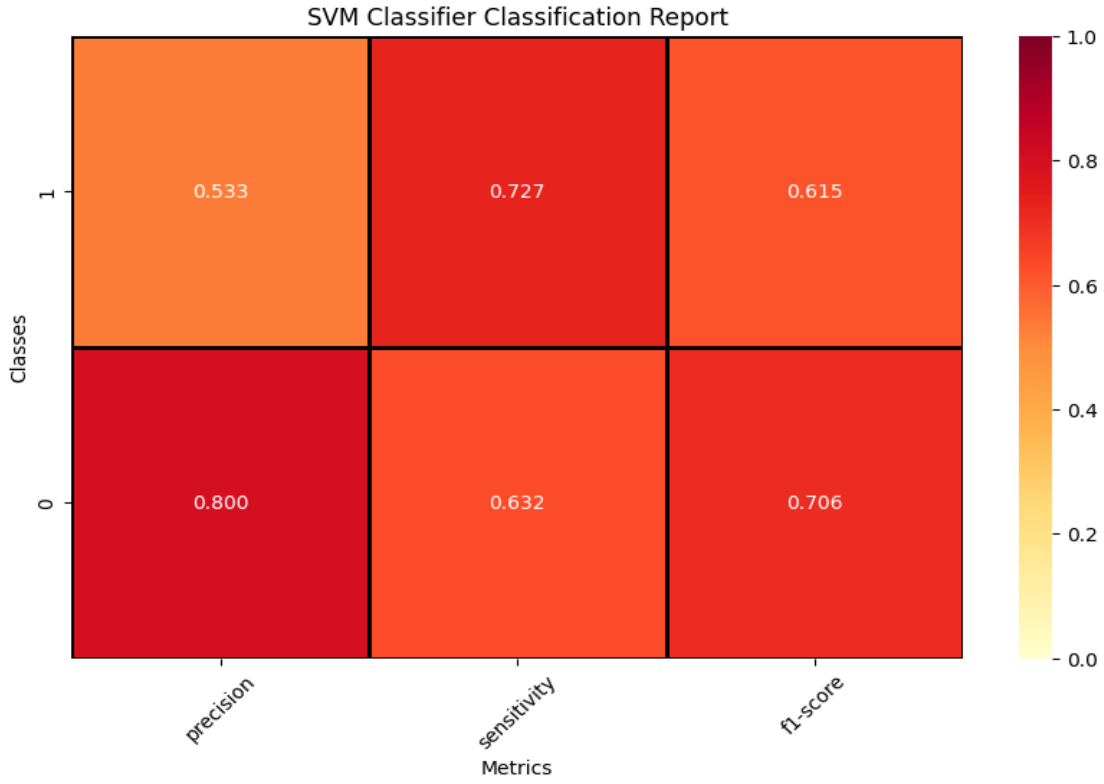
Εικόνα 52 - Επιδόσεις Καλύτερων Μοντέλων - Σύνολο με Κλινικά Δεδομένα

Η ενσωμάτωση δημογραφικών δεδομένων οδήγησε σε παρόμοια βελτίωση της γενικότερης απόδοσης των μοντέλων με αυτή που είδαμε στο Πείραμα 6. Οι μετρικές που αντικατοπτρίζουν τη δυνατότητα γενίκευσης του μοντέλου, όπως η AUC και η Precision παρουσιάζουν αυξημένες τιμές. Το προστιθέμενο σετ δεδομένων από τα κλινικά χαρακτηριστικά φάνηκε να βοηθά τα μοντέλα να κάνουν πιο τεκμηριωμένες προβλέψεις, ιδίως σε οριακές περιπτώσεις όπου τα ακτινολογικά χαρακτηριστικά από μόνα τους μπορεί να μην επαρκούσαν για να αποτυπώσουν πλήρως την πολυπλοκότητα των αποτελεσμάτων των ασθενών.

Ψάχνοντας εις βάθος τα μοντέλα παρατηρούμε πως το μοντέλο SVM ξεχωρίζει αρκετά ως προς τον πίνακα σύγχυσης και συγκεκριμένα διακρίνουμε την σωστή πρόβλεψη των 2 κλάσεων σημαντικά καλύτερα και ειδικότερα αυτή της κατηγορίας 1. Παρακάτω παρουσιάζονται τα σχετικά αποτελέσματα:

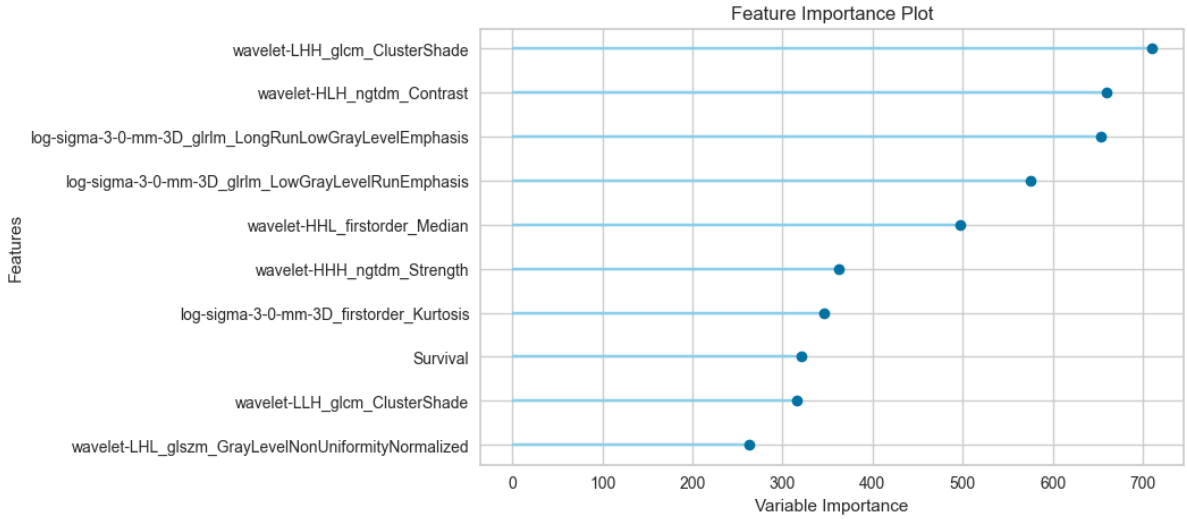


Εικόνα 54 - SVM Confusion Matrix - Σύνολο με Κλινικά Δεδομένα



Εικόνα 53 - SVM Μετρικές ανά κλάση - Σύνολο με Κλινικά Δεδομένα

Παρατηρούμε πως το συγκεκριμένο μοντέλο και για την κατηγορία 1, καταφέρνουμε να πετύχουμε υψηλές τιμές στις μετρικές και ειδικότερα στην μετρική της ευαισθησίας, κάτι πολύ σημαντικό για την φύση της εφαρμογής μας.



Εικόνα 55 - SVM Σημαντικότητα Χαρακτηριστικών Σύνολο με Κλινικά Δεδομένα

Τα κλινικά δεδομένα συνέβαλαν σημαντικά στην ελαχιστοποίηση των λανθασμένων ταξινομήσεων, ιδίως σε περιπτώσεις όπου τα ακτινολογικά χαρακτηριστικά ήταν διφορούμενα.

5.2.2: Συνδυασμοί Μοντέλων: Ensemble, Blend και Stack

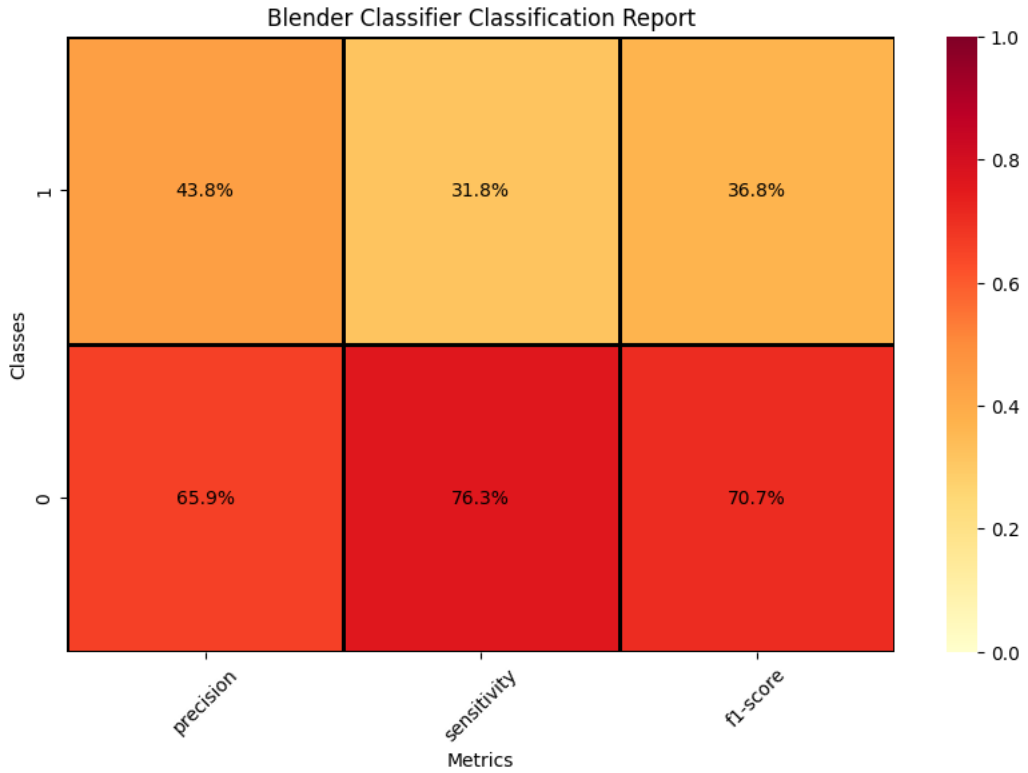
Το Ensembling είναι μια ισχυρή τεχνική για την ενίσχυση της απόδοσης των μοντέλων συνδυάζοντας προβλέψεις από πολλαπλά μοντέλα. Η μέθοδος αυτή εκμεταλλεύεται την ποικιλομορφία των διαφόρων αλγορίθμων για τη μείωση της διακύμανσης και της μεροληψίας, οδηγώντας σε ένα πιο ισχυρό τελικό μοντέλο. Σε αυτή την ενότητα, διερευνούμε διάφορες τεχνικές ensemble, συμπεριλαμβανομένων των ***blending*** και ***stacking***, οι οποίες αποσκοπούν στη συγκέντρωση των πλεονεκτημάτων πολλαπλών μοντέλων. Οι τεχνικές ensembling εφαρμόστηκαν σε συνδυασμούς των μοντέλων SVM-Liner Kernel, XGBoost και MLP, DecisionTree και LGBM, AdaBoost. Επιλέχθηκαν αυτά τα μοντέλα καθώς σημείωσαν κάποιες από τις καλύτερες επιδόσεις που εντοπίστηκαν σε προηγούμενα πειράματα.

Αρχικά θα εξετάσουμε την τεχνική του blender classifier. Σε αυτό το βήμα εξετάστηκαν αρκετοί συνδυασμοί των παραπάνω μοντέλων με διαφορετικές ρυθμίσεις, όπως την μέθοδο ψήφου για τα αποτελέσματα των προβλέψεων που είτε είναι μέσω της πλειοψηφίας (hard voting) είτε μέσω του μέσου όρου (argmax των αθροισμάτων) των πιθανοτήτων (soft voting) ή όπως το βάρος που αποτελεί μία ακολουθία βαρών, είτε για τη στάθμιση των εμφανίσεων των ετικετών (labels) πρόβλεψης της κάθε

κατηγορίας ετικέτες (hard voting) ή για την στάθμιση των πιθανοτήτων της κάθε κατηγορίας πριν από τον μέσο όρο (soft voting). Στα αποτελέσματα μας παρουσιάζουμε δύο εκδοχές με διαφορετικά αποτελέσματα για διαφορετική εφαρμογή το κάθε ένα. Στην πρώτη εκδοχή επικεντρωθήκαμε στην γενικότερη βελτίωση των επιδόσεων των μετρικών. Επιλέχθηκαν τα μοντέλα LGBM, AdaBoost και MLP. Αρχικά η συμμετοχή του κάθε μοντέλου ήταν ισοβαρής αλλά μετά από βελτιστοποίηση των βαρών καταλήξαμε στα [0.84, 0.47, 0.24]

```
VotingClassifier
VotingClassifier(estimators=[('Light Gradient Boosting Machine',
                              LGBMClassifier(boosting_type='gbdt',
                                             class_weight=None,
                                             colsample_bytree=1.0,
                                             importance_type='split',
                                             learning_rate=0.1, max_depth=-1,
                                             min_child_samples=20,
                                             min_child_weight=0.001,
                                             min_split_gain=0.0,
                                             n_estimators=100, n_jobs=-1,
                                             num_leaves=31, objective=None,
                                             random_state=42, reg_alpha=0.0,
                                             reg_lambda...
                                             hidden_layer_sizes=(100,),
Light Gradient Boosting Machine | Ada Boost Classifier | MLP Classifier
LGBMClassifier | AdaBoostClassifier | MLPClassifier
LGBMClassifier(n_jobs=-1, random_state=42) | AdaBoostClassifier(random_state=42) | MLPClassifier(max_iter=500, random_state=42)
```

Εικόνα 56 – Ρυθμίσεις μοντέλου blender_v1



Εικόνα 57 - Μετρικές ανά κατηγορία για το μοντέλο blender_v1

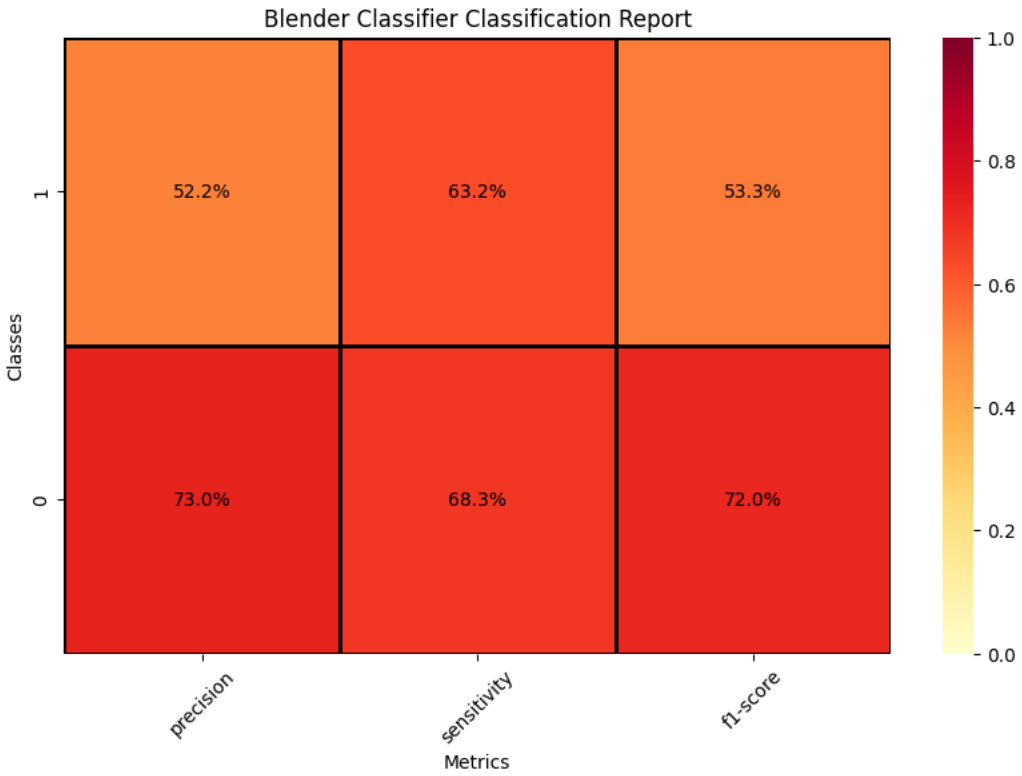
	Accuracy	AUC	F1	Sensitivity
Fold				
0	0.5714	0.4667	0.5000	0.6000
1	0.5000	0.3556	0.2222	0.2000
2	0.5000	0.7111	0.4615	0.6000
3	0.5714	0.7111	0.2500	0.2000
4	0.7857	0.6667	0.6667	0.6000
5	0.2857	0.3556	0.1667	0.2000
6	0.7143	0.6875	0.6000	0.5000
7	1.0000	1.0000	1.0000	1.0000
8	0.8462	0.6750	0.7500	0.6000
9	0.6154	0.5750	0.6154	0.8000
Mean	0.6390	0.6204	0.5232	0.5300
Std	0.1933	0.1833	0.2471	0.2532

Εικόνα 58 - Επιδόσεις Μετρικών κατα την εκπαίδευση του μοντέλου blender_v1

Σημαντική βελτίωση αυτή της ακρίβειας και του F1-score, το οποίο χαρακτηρίζει το παρόν μοντέλο με περισσότερη ευρωστία, αλλά το μειονέκτημα είναι πως το μοντέλο επικεντρώνεται περισσότερο στην κατηγορία 0 και “παραμελεί” την κατηγορία 1, με κίνδυνο της αύξηση των ψευδώς αρνητικών αποτελεσμάτων, κάτι που είναι πρακτικώς πολύ κακό.

Στην δεύτερη εκδοχή επικεντρωθήκαμε στην βελτίωση της ικανότητας του μοντέλου να διαχωρίζει τις δύο κλάσεις και να ελαχιστοποιεί τα ψευδώς αρνητικά αποτελέσματα. Επιλέχθηκαν τα μοντέλα SVM-Liner Kernel, XGBoost και MLP. Αρχικά η συμμετοχή του κάθε μοντέλου ήταν ισοβαρής αλλά μετά από βελτιστοποίηση των βαρών καταλήξαμε στα [0.68, 0.47, 0.12]

Εικόνα 59 - Ρυθμίσεις μοντέλου blender_v2



Εικόνα 60 - Μετρικές ανά κατηγορία για το μοντέλο blender_v2

Στα αποτελέσματα από αυτή την εκδοχή, είναι προφανές πως το μοντέλο SVM συνεισφέρει με την ικανότητα του να διαχωρίζει τις 2 κατηγορίες. Παρατηρούμε βελτιωμένες τιμές για την μετρική της ευαισθησίας και μειωμένα ποσοστά για τα ψευδώς αρνητικά αποτελέσματα. Από την άλλη μεριά, η επιλογή αυτών των μοντέλων για τη σύνθεση του **blender,** χάνει σε ευρωστία και δυνατότητα γενίκευσης, με μικρότερη τιμή ακρίβειάς (60.3%) σε σχέση με το μοντέλο της πρώτης εκδοχής, καθώς είναι μεροληπτικό προς την κατηγορία 1.

Οι μέθοδοι ensemble υπερείχαν ιδιαίτερα στη βελτίωση της ευαισθησίας των μοντέλων, η οποία είναι ζωτικής σημασίας για ιατρικές εφαρμογές όπου πρέπει να ελαχιστοποιηθούν τα ψευδώς αρνητικά αποτελέσματα (πρόβλεψη επιβίωσης για έναν ασθενή που αργότερα πεθαίνει). Ειδικότερα η τεχνική blend, επέδειξε ισορροπία μεταξύ ακρίβειας και ανάκλησης, καθιστώντας την κατάλληλη επιλογή για κλινική χρήση.

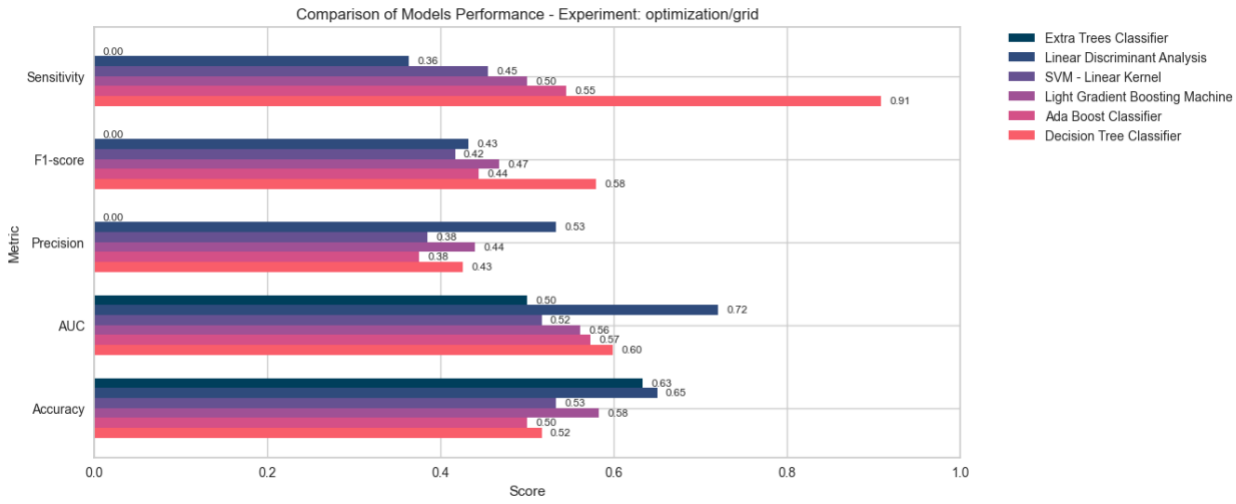
5.2.3: Αναζήτηση Βέλτιστων Υπερπαραμέτρων

Για την περαιτέρω βελτίωση των μοντέλων και τη βελτιστοποίηση των προβλεπτικών τους δυνατοτήτων, πραγματοποιήθηκε αναζήτηση για τις βέλτιστες υπερπαραμέτρους. Αυτό το βήμα είναι

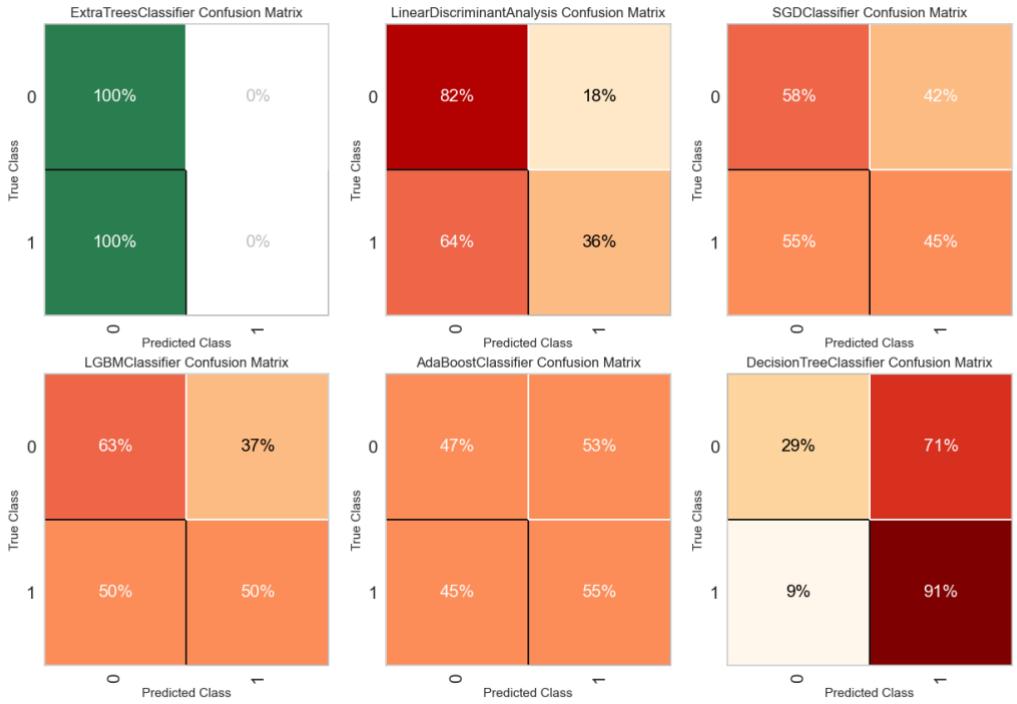
κρίσιμο καθώς τα μοντέλα μηχανικής μάθησης, ειδικά τα πολύπλοκα όπως το LGBM και το Extreme Gradient Boosting, έχουν πολλαπλές υπερπαραμέτρους που επηρεάζουν σημαντικά την απόδοσή τους.

Χρησιμοποιήθηκαν δύο προσεγγίσεις: πρώτα αυτή της Αναζήτησης Πλέγματος (Grid Search), όπου πραγματοποιήθηκε εξαντλητική αναζήτηση σε έναν προκαθορισμένο χώρο υπερπαραμέτρων για κάθε μοντέλο. Παράμετροι όπως ο ρυθμός μάθησης, το μέγιστο βάθος, και ο αριθμός των εκτιμητών συντονίστηκαν για μοντέλα που βασίζονται σε δέντρα, όπως το Decision Tree και το AdaBoost. Η δεύτερη προσέγγιση είναι αυτή της Βελτιστοποίησης κατά Bayes. Σε αντίθεση με την αναζήτηση πλέγματος, η βελτιστοποίηση Bayes περιορίζει επαναληπτικά τον χώρο αναζήτησης εστιάζοντας στις περιοχές τιμών υπερπαραμέτρων που είναι πιο πιθανό να αποφέρουν βελτιώσεις. Για την εφαρμογή αυτής της στρατηγικής χρησιμοποιήθηκε η Optuna, μια σύγχρονη βιβλιοθήκη βελτιστοποίησης. Και οι 2 τεχνικές βελτιστοποίησης εφαρμόστηκαν σε επιλεγμένα 6 από τα συνολικά 19 μοντέλα όπου αυτά ξεχώρισαν γενικοτέρα σε επιδόσεις σε σχέση με τα υπόλοιπα. Επίσης για αυτά τα μοντέλα όπου ο χώρος υπερπαραμέτρων είναι μεγάλος και η εξαντλητική αναζήτηση είναι υπολογιστικά δαπανηρή, αυτές οι μέθοδοι βελτιστοποίησης έχουν χώρο εφαρμογής.

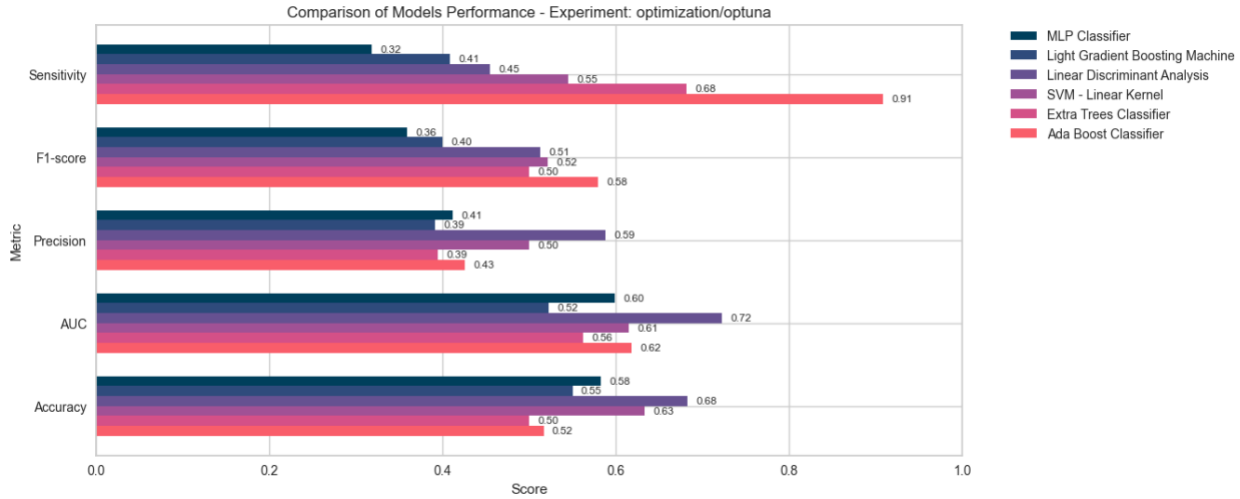
Παρακάτω παρουσιάζονται συγκεντρωτικά τα αποτελέσματα για τις δύο προσεγγίσεις των μεθόδων αναζήτησης υπερπαραμέτρων:



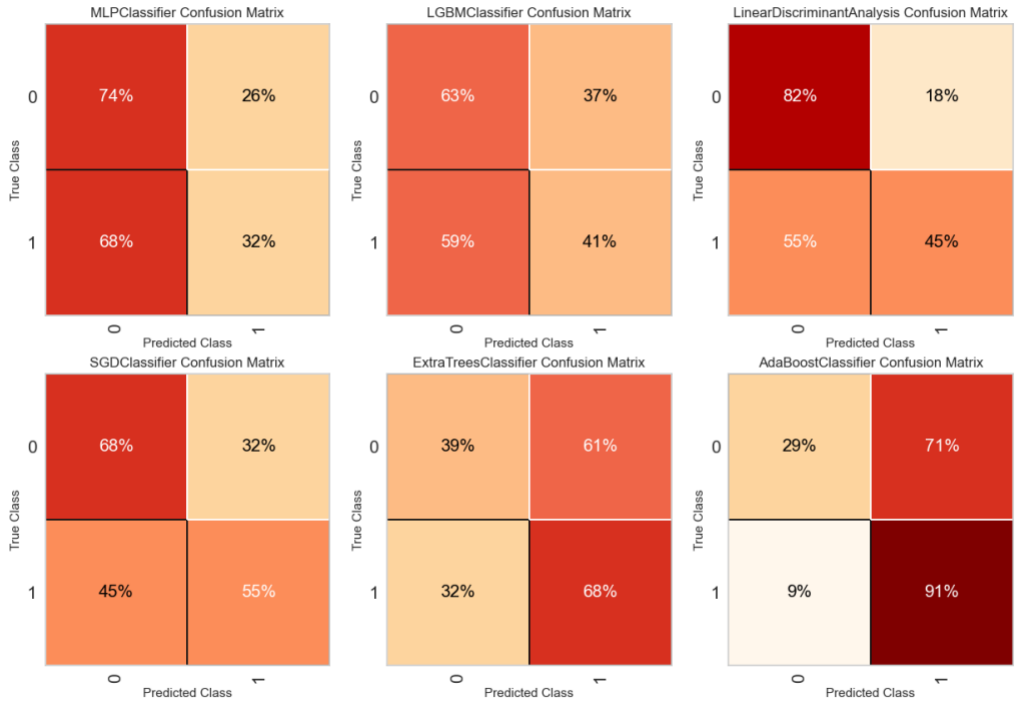
Εικόνα 61 - Επιδόσεις Μοντέλων - Grid Search



Εικόνα 62 - Συγκριτικοί Πίνακες Σύγχυσης - Grid Search



Εικόνα 63 - Επιδόσεις Μοντέλων - Bayesian Search



Εικόνα 64 - Συγκριτικοί Πίνακες Σύγχυσης - Bayesian Search

Η βελτιστοποίηση Bayesian ξεπέρασε την αναζήτηση πλέγματος (Grid) τόσο από άποψη υπολογιστικού χρόνου όσο και από άποψη απόδοσης του μοντέλου. Οι υπερπαράμετροι που επιλέχθηκαν μέσω της βελτιστοποίησης Bayes επέτρεψαν στα μοντέλα να συγκλίνουν ταχύτερα και να παράγουν πιο αξιόπιστες προβλέψεις. Για παράδειγμα, στην περίπτωση του μοντέλου ExtraTreesClassifier, η ρύθμιση των παραμέτρων class weight, max_features, min_impurity_decrease και και min_sample_leafs, καθώς και για το μοντέλο SVM η ρύθμιση των παραμέτρων alpha, eta0, l1_ratio και του ρυθμού μάθησης βελτίωσε σημαντικά την ικανότητα του μοντέλου να γενικεύει, αυξάνοντας την AUC και την ακρίβειά του χωρίς να θυσιάσει την ευαισθησία, κάτι πολύ σημαντικό για την εφαρμογή μας.

Συνολικά, ο συντονισμός των υπερπαραμέτρων οδήγησε σε σημαντικές βελτιώσεις των επιδόσεων σε όλα τα μοντέλα, επιβεβαιώνοντας τη σημασία της βελτιστοποίησης ειδικά για το μοντέλο για την πλήρη αξιοποίηση των δυνατοτήτων των αλγορίθμων μηχανικής μάθησης.

AdaBoost

	bayesian	optuna
algorithm	SAMME	SAMME
base_estimator	deprecated	deprecated
estimator	None	None
learning_rate	0.010984	0.001753
n_estimators	294	167
random_state	42	42

ExtraTreesClassifier

	bayesian	optuna
bootstrap	True	False
ccp_alpha	0.000000	0.000000
class_weight	None	balanced_subsample
criterion	gini	gini
max_depth	2	5
max_features	0.400000	0.463893
max_leaf_nodes	None	None
max_samples	None	None
min_impurity_decrease	0.500000	0.043950
min_samples_leaf	5	2
min_samples_split	2	2
min_weight_fraction_leaf	0.000000	0.000000
n_estimators	300	279
n_jobs	-1	-1
oob_score	False	False
random_state	42	42
verbose	0	0
warm_start	False	False

SVM Liner

	bayesian	optuna
alpha	0.048145	0.019744
average	False	False
class_weight	None	None
early_stopping	False	False
epsilon	0.100000	0.100000
eta0	0.005188	0.003949
fit_intercept	True	True
l1_ratio	1.000000	0.270928
learning_rate	optimal	constant
loss	hinge	hinge
max_iter	1000	1000
n_iter_no_change	5	5
n_jobs	-1	-1
penalty	l1	l1
power_t	0.500000	0.500000
random_state	42	42
shuffle	True	True
tol	0.001000	0.001000
validation_fraction	0.100000	0.100000
verbose	0	0
warm_start	False	False

Εικόνα 65 - Σύγκριση Υπερπαραμέτρων Επιλεγμένων Μοντέλων - Βελτιστοποίηση

Για να αναδείξουμε την διαφορετικότητα των δύο τεχνικών βελτιστοποίησης, συγκρίνουμε με περισσότερη λεπτομέρεια τις διαφορές μεταξύ τριών μοντέλων. Συγκεκριμένα για το μοντέλο AdaBoost παρατηρούμε πολύ μεγάλες τιμές στην ευαισθησία, κάτι που αρχικά φαίνεται καλό για την εφαρμογή μας, αλλά στην πραγματικότητα το μοντέλο μας είναι σε μεγάλο βαθμό προκατειλημμένο ως προς την κατηγορία 1 και ειδικότερα στηρίζεται μόνο σε ένα χαρακτηριστικό του συνόλου δεδομένων, την εκτίμηση επιβίωσης σε μήνες (Survival), χαρακτηριστικό των δημογραφικών δεδομένων, που έγινε από τον επιβλέπονται ιατρό. Ένα τέτοιο αποτέλεσμα πρακτικά δεν μας ωφελεί σε κάτι παράπανω, πρόκειτα για μία περίπτωση υπερπροσαρμογής και δεν έχει καμία διαφορά από την εκτίμηση ενός έμπειρου γιατρού. Για το μοντέλο ExtraTreesClassifier παρατηρούμε πως στην βελτιστοποίηση πλέγματος, δεν καταφέρνουμε κάποιο ιδιαίτερα καλό αποτέλεσμα. Αντιθέτως με την πιθανοτική αναζήτηση κατά Bayes, καταφέρνουμε βρούμε μία βέλτιστη δομή των υπερπαραμέτρων κάτι που μεταφράζεται και στις επιδόσεις του μοντέλου

και στην ικανότητα του να διακρίνει τις 2 κατηγορίες και να γενικεύει καλύτερα. Τέλος για το μοντέλο SVM παρατηρούμε πιθανό την πιο ισορροπημένη εικόνα μεταξύ διάκρισης των 2 κατηγοριών, της γενίκευσης, των γενικότερων επιδόσεων, της αποφυγής της υπερπροσαρμογής και της βέλτιστης απόδοσης στην πρόβλεψη ιατρικών αποτελεσμάτων.

5.3: Συμπεράσματα και Μελλοντικές Επεκτάσεις

Το αντικείμενο της παρούσης εργασίας ήταν η δημιουργία ενός μοντέλου μηχανικής μάθησης με στόχο την πρόβλεψη του θανάτου για ασθενείς με όγκο στον εγκέφαλο. Ο πρωταρχικός στόχος μας είναι να βοηθήσουμε στην πρόληψη του γεγονότος του πιθανού θανάτου και όχι στην διάγνωση της τελικής κατάστασης του ασθενούς. Έτσι και για την δικιά μας εφαρμογή, χρησιμοποιήσαμε τα radiomics χαρακτηριστικά των τρισδιάστατων εικόνων, ώστε να βρούμε τις πιο σημαντικές πληροφορίες που συμβάλουν στην πρόβλεψη του τελικού αποτελέσματος σε συνδυασμό με ένα εύρωστο και αποτελεσματικό μοντέλο που γενικεύει.

Τα δεδομένα προέρχονται από μία συγκεκριμένη πηγή [4.1: Δεδομένα Μοντέλου] όπου οι συμμετέχοντες ασθενείς προέρχονται από τον γενικό πληθυσμό. Ως αποτέλεσμα μπορούμε να πούμε πως αντιμετωπίσαμε κάποιους περιορισμούς ως προς το πλήθος των δεδομένων, την ποικιλότητα των δεδομένων και ως προς την ισορροπία των κατηγοριών μέσα στο σύνολο των δεδομένων, και συγκεκριμένα με μειωμένη συμμετοχή για την κατηγορία 1 (Θανών). Αυτοί οι περιορισμοί κάνουν δύσκολη την εκπαίδευση ενός γενικευμένου μοντέλου με την χρήση μηχανικής μάθησης για την αναγνώριση νέων άγνωστων περιπτώσεων της κατηγορίας 1, όπως παρατηρήθηκε και στην δικιά μας περίπτωση κατά την πειραματική διαδικασία και αντιμετωπίστηκε με τις τεχνικές προεπεξεργασίας και βελτιστοποίησης [4.2: Διερευνητική Ανάλυση και Προεπεξεργασία Χαρακτηριστικών - 4.3.3: Βελτιστοποίηση Μοντέλων].

Ως μετρική αξιολόγησης χρησιμοποιήθηκε η ευαισθησία, ώστε η επιλογή του μοντέλου να στοχεύει και να εστιάζει σε όσο το δυνατόν μεγαλύτερο ποσοστό σωστών προβλέψεων για την κατηγορία 1. Κατά την διάρκεια της επιλογής του μοντέλου, βάσει των μετρικών, δόθηκε επίσης ιδιαίτερη μέριμνα ώστε να επιλέξουμε αυτά τα μοντέλα που έχουν σαφώς μειωμένα ποσοστά για την λάθος ταξινόμηση της κατηγορίας 1 σε κατηγορία 0 και αντίστοιχα σαφώς αυξημένα ποσοστά για την λάθος ταξινόμηση της κατηγορίας 0 σε κατηγορία 1. Αυτό συμβαίνει γιατί, όπως προαναφέραμε, από την μία μεριά θέλουμε ιδανικά να ελαχιστοποιήσουμε τις περιπτώσεις εκείνες όπου ο ασθενής συντρέχει σοβαρό κίνδυνο της ζωής του και το συμβουλευόμενο μοντέλο τον ταξινομεί στην κατηγορία 0 και από την άλλη μεριά παρόλο που ο ασθενείς εν τέλει δεν διατρέχει κάποιον θανατηφόρο κίνδυνο ο σκοπός μας είναι να κάνουμε πρόληψη του προβλήματος και να παροτρύνουμε σε περαιτέρω διερεύνηση και εξέταση του ασθενούς.

Εικόνα 66 - Σύγκριση Μετρικών ανα Μοντέλο και Πείραμα

Κατά την πειραματική διαδικασία αναπτύχθηκαν και εκπαιδεύτηκαν περισσότερα από 100 διαφορετικά μοντέλα με διάφορους συνδυασμούς παραμέτρων, τεχνικών προεπεξεργασίας δεδομένων και βελτιστοποίησης. Στην Εικόνα 66 παρουσιάζεται συγκεντρωτικά η πορεία της εκπαίδευσης των μοντέλων για όλα τα πειράματα και πως αυτά συγκρίνονται με την μέση επίδοση στην κάθε μετρική ανάμεσα σε όλα τα μοντέλα. Παρατηρούμε γενικά μία ανοδική πορεία σε όλα τις μετρικές κάτι που δείχνει την θετική επίδραση των τεχνικών που προστίθενται σε κάθε βήμα. Παρόλο που σε όλα τα βήματα και τα διαφορετικά πειράματα που δοκιμάσαμε, τα μοντέλα που ξεχώριζαν συνεχώς εναλλάσσονταν, στο τέλος καταλήγουμε σε 2 πιο πιθανά υποψήφια μοντέλα, για τα οποία τα αποτελέσματα ήταν ενθαρρυντικά αλλά για να χρησιμοποιηθούν με αρκετή εμπιστοσύνη, σε πραγματικές κλινικές εφαρμογές και ως υποβοήθηση για την διάγνωση και την πρόληψη, θα χρειαστεί να βελτιωθούν και γίνουν πιο αξιόπιστα. Αυτά είναι τα μοντέλα ExtraTreesClassifier και SVM-Liner Kernel.

Για περαιτέρω βελτίωση της επίδοσης των μοντέλων θα μπορούσαν να χρησιμοποιηθούν επιπλέον σύνολα δεδομένων με μεγαλύτερη ποσότητα και με επιπρόσθετη ποικιλότητα (ηλικία, φύλο, καταγωγή, κτλ.) δεδομένων. Σχετικά με τα παραγόμενα χαρακτηριστικά και τις διαθέσιμες κλινικές πληροφορίες, θα συνέβαλε πολύ στα αποτελέσματα η ποσότητα των ποιοτικών χαρακτηριστικών καθώς επίσης και

καινούργια χαρακτηριστικά που δυνητικά θα ενισχύσουν την διακριτική ικανότητα των μοντέλων (όπως χαρακτηριστικά αιματολογικών εξετάσεων, οικογενειακό ιστορικό, συνθήκες ζωής και καθημερινότητας, στοιχεία προσωπικότητας).

Σχετικά με την βελτιστοποίηση των αποτελεσμάτων όσων αφορά τις τεχνικές και τους αλγορίθμους που χρησιμοποιήθηκαν, θα μπορούσαν δυνητικά να χρησιμοποιηθούν καλύτερες μέθοδοι για την εύρεση των κατάλληλων υπερπαραμέτρων, η εύρεση και εξαγωγή καλύτερων χαρακτηριστικών ή ακόμα τεχνικές βαθιάς μάθησης και συγκεκριμένων συνελικτικών μοντέλων μηχανικής μάθησης. Σύμφωνα με το [55] όπου εφαρμόζουν το μη-γραμμικό νευρωνικό δίκτυο DeepSurv για την πρόβλεψη της επιβίωσης ενός ασθενή με μετάσταση στον εγκάφαλο, τα αποτελέσματά υποδηλώνουν ότι η χρήση τεχνικών βαθιάς μάθησης μπορούν να ξεπεράσουν τα παραδοσιακά μοντέλα, ειδικά στις περίπτωσης όπου μια ακριβής πρόγνωση είναι πολύ σημαντική κλινικά.

Συμπερασματικά, η χρήση αλγορίθμων μηχανικής μάθησης μπορεί να αποτελέσει χρήσιμο εργαλείο στα χέρια των ειδικών τόσο για την πρόβλεψη όσο και για την πρόληψη των θανάτων που οφείλονται στην μετάσταση εγκεφάλου. Η λήψη προληπτικών μέτρων θα έχει μεγάλο όφελος για τη ζωή κάθε ατόμου, την ποιότητα ζωής του, το οικογενειακό και κοινωνικό του περιβάλλον, αλλά βεβαίως και για τα ασφαλιστικά συστήματα των κρατών. Η μελέτη αυτή συνηγορεί μαζί με πολλές άλλες ότι είναι εφικτό με την επιλογή κατάλληλων χαρακτηριστικών και αλγορίθμων μηχανικής μάθησης.

Βιβλιογραφία

[1] Sung, H, Ferlay, J, Siegel, RL, Laversanne, M, Soerjomataram, I, Jemal, A & Bray, F 2021, ***Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers 185 countries****, CA Cancer J Clin., vol. 71, pp. 209-249.*

[2] Yamini Sharma & Yogesh K Meghrajani 2014, ***Extraction of gray scale brain tumor in magnetic resonance image****, International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, no. 11, pp. 8542-8544.*

[3] Bhagat, JV & Dhaigude, NB 2017, ***A survey on brain tumor detection techniques****, International Research Journal of Engineering and Technology, vol. 4, no. 3, pp. 1795-1796.*

[4] Cha, S 2006, ***Review article: Update on brain tumor imaging: From anatomy to physiology****, of Neuro Radiology, vol. 27, pp. 475-487.*

[5] Rakesh, M & Ravi, T 2012, ***Image segmentation and detection of tumor objects in MR brain images using Fuzzy C-Means (FCM) algorithm****, International Journal of Engineering Research and Applications (IJERA), vol. 2, no. 3, pp. 2088-2094.*

[6] Azhari, EE, Hatta M, Htike MM & Win, SL 2014, ***Brain tumor detection and localization in magnetic resonance imaging****, International Journal of Information Technology Convergence and Services (IJITCS), vol. 4, no. 1, pp. 2231-1939.*

[7] Barnholtz-Sloan, J. S., Sloan, A. E., Davis, F. G., Vigneau, F. D., Lai, P., & Sawaya, R. E. (2004). ***Incidence proportions of brain metastases in patients diagnosed (1973 to 2001) in the Metropolitan Detroit Cancer Surveillance System.*** *Journal of Clinical Oncology, 22(14), 2865-2872.*

[8] Preusser, M., Capper, D., Ilhan-Mutlu, A., Berghoff, A. S., & Birner, P. (2015). ***Brain metastases: pathobiology and emerging targeted therapies.*** *Acta neuropathologica, 131(6), 803-823.*

[9] Lin, N. U., Lee, E. Q., Aoyama, H., Barani, I. J., Barboriak, D. P., Baumert, B. G., ... & Brown, P. D. (2015). ***Response assessment criteria for brain metastases: proposal from the RANO group.*** *The Lancet Oncology, 16(6), e270-e278.*

[10] Nayak, L., & Lee, E. Q. (2010). ***Epidemiology of brain metastases.*** *Current oncology reports, 12(4), 1-9.*

[11] Soffietti, R., Abacioglu, U., Baumert, B., Combs, S. E., Kinhult, S., Kros, J. M., ... & Weber, D. C. (2017). ***Diagnosis and treatment of brain metastases from solid tumors: guidelines from the European Association of Neuro-Oncology (EANO).*** *Neuro-Oncology, 19(2), 162-174.*

[12] Wefel, J. S., Kesler, S. R., Noll, K. R., & Schagen, S. B. (2015). ***Clinical characteristics, pathophysiology, and management of noncentral nervous system cancer-related cognitive impairment in adults****. CA: a cancer journal for clinicians, 65(2), 123-138.*

[13] Gaspar, L., Scott, C., Rotman, M., Asbell, S., Phillips, T., Wasserman, T., ... & Curran Jr, W. (1997). ***Recursive partitioning analysis (RPA) of prognostic factors in three Radiation Therapy Oncology Group (RTOG) brain metastases trials****. International journal of radiation oncology, biology, physics, 37(4), 745-751.*

[14] Sperduto, P. W., Kased, N., Roberge, D., Xu, Z., Shanley, R., Luo, X., ... & Mehta, M. (2012). ***Summary report on the graded prognostic assessment: an accurate and facile diagnosis-specific***

tool to estimate survival for patients with brain metastases. Journal of Clinical Oncology*, 30(4), 419-425.*

[15] Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., van Stiphout, R. G., Granton, P., ... & Deist, T. M. (2012). ***Radiomics: extracting more information from medical images using advanced feature analysis. European Journal of Cancer****, 48(4), 441-446.*

[16] Kickingereder, P., Neuberger, U., Bonekamp, D., Piechotta, P. L., Götz, M., Wick, A., ... & Radbruch, A. (2016). ***Radiomic subtyping improves disease stratification beyond key molecular, clinical, and standard imaging characteristics in patients with glioblastoma****. Neuro-Oncology, 20(6), 848-857.*

[17] Chenxi Zhang, Manning Wang & Zhijian Song 2011, ***A brain-deformation framework based on a linear elastic model and evaluation using clinical data****, IEEE Transaction on Biomedical Engneering, vol. 58, pp. 191-199.*

[18] Machine Learning | Artificial Inteligence . Encyclopedia Britannica

[19] Russell, Stuart J. (2002). ***Artificial intelligence : a modern approach*** *(2nd edition). New Delhi: Prentice-Hall of India*

[20] Bishop, Christopher M. ***Pattern Recognition and Machine Learning****. New York :Springer, 2006.*

[21] Ν. Βασιλειάδης Φ. Κόκκορας Η. Σακελλαρίου Ι. Βλαχάβας, Π. Κεφαλάς, ***Τεχνητή Νοημοσύνη.*** *Εκδόσεις Πανεπιστημίου Μακεδονίας, 2006.*

[22] Simon O. Haykin, ***Neural Networks and Learning Machines****, Third Edition, Pearson Education, McMaster University, Canada, 2009.*

[23] Vapnik, V., Cortes, C., & Vapnik, S. (1995). ***Support-vector networks****. Machine learning, 20(3), 273-297.*

[24] Hosmer, D. W., & Lemeshow (2000) ***Applied Logistic Regression***

[25] Friedman, J. H., Hastie, T., & Tibshirani, R. (2001). ***The elements of statistical learning*** *(Vol. 1). Springer series in statistics New York, NY, USA:.*

[26] Chen, T., & Guestrin, C. (2016, August). ***Xgboost: A scalable tree boosting system.***

[27] Moshe Sipper (2022, July) ***A Large-Scale Study of Hyperparameter Tuning for Machine Learning Algorithms***

[28] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). ***A guide to deep learning in healthcare.*** *Nature medicine, 25(1), 24-29.*

[29] Xie, Y.; Zaccagna, F.; Rundo, L.; Testa, C.; Agati, R.; Lodi, R.; Manners, D.N.; Tonon, C. ***Convolutional Neural Network Techniques for Brain Tumor Classification (from 2015 to 2022): Review, Challenges, and Future Perspectives****. Diagnostics* ***2022****, 12, 1850. https://doi.org/10.3390/diagnostics12081850*

[30] Aamir, M.; Namoun, A.; Munir, S.; Aljohani, N.; Alanazi, M.H.; Alsahafi, Y.; Alotibi, F. ***Brain Tumor Detection and Classification Using an Optimized Convolutional Neural Network.*** *Diagnostics* ***2024****, 14, 1714. https://doi.org/10.3390/diagnostics14161714*

[31] K. Neamah et al., ***Brain Tumor Classification and Detection Based DL Models: A Systematic Review,*** *in IEEE Access, vol. 12, pp. 2517-2542, 2024, doi: 10.1109/ACCESS.2023.3347545*

[32] Rajkomar, A., Dean, J., & Kohane, I. (2019). ***Machine learning in medicine.*** *New England Journal of Medicine, 380(14), 1347-1358.*

[33] Beam, A. L., & Kohane, I. S. (2018). ***Big data and machine learning in health care.*** *JAMA, 319(13), 1317-1318.*

[34] *Obermeyer, Z., & Emanuel, E. J. (2016).* ***Predicting the future—big data, machine learning, and clinical medicine.*** *New England Journal of Medicine, 375(13), 1216-1219.*

[35] *Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015).* ***Machine learning applications in cancer prognosis and prediction****. Computational and structural biotechnology journal, 13, 8-17.*

[36] *Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017).* ***A survey on deep learning in medical image analysis****. Medical image analysis, 42, 60-88.*

[37] *Cabitza, F., Rasoini, R., & Gensini, G. F. (2017).* ***Unintended consequences of machine learning in medicine****. JAMA, 318(6), 517-518.*

[38] *Chen, J. H., Asch, S. M.,* ***Machine Learning and Prediction in Medicine*** *- Beyond the Peak of Inflated Expectations. N Engl J Med, 381(23), 2281-2284.*

[39] *Srinivasan, P., & Acharya, U. R. (2018).* ***A review on image enhancement techniques for diagnostic radiography.*** *Journal of Medical Imaging and Health Informatics, 8(4), 894-902.*

[40] *Zhang, Z., Liang, Y., Chen, J., Zhang, L., & Zhang, Y. (2015).* ***Automated brain tumor segmentation using kernel dictionary learning and superpixel-level features****. Computerized Medical Imaging and Graphics, 46, 129-138.*

[41] *Maintz, J. B., & Viergever, M. A. (1998).* ***A survey of medical image registration****. Medical image analysis, 2(1), 1-36.*

[42] *Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017).* ***A survey on deep learning in medical image analysis****. Medical image analysis, 42, 60-88.*

[43] *Ronneberger, O., Fischer, P., & Brox, T. (2015).* ***U-Net: Convolutional Networks for Biomedical Image Segmentation.*** *In International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI).*

[44] *Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., & Dalca, A. V. (2018).* ***An Unsupervised Learning Model for Deformable Medical Image Registration.*** *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[45] *Ramakrishnan, D., Jekel, L., Chadha, S., Janas, A., Moy, H., Maleki, N., Sala, M., Kaur, M., Petersen, G. C., Merkaj, S., von Reppert, M., Baid, U., Bakas, S., Kirsch, C., Davis, M., Bousabarah, K., Holler, W., Lin, M., Westerhoff, M., Aneja, S., Memon, F., Aboian, M. (2023).* ***A Large Open Access Dataset of Brain Metastasis 3D Segmentations on MRI with Clinical and Imaging Feature Information (Version 1) [dataset]****. The Cancer Imaging Archive. DOI: https://doi.org/10.7937/6be1-r748*

[46] ***SimpleITK Python toolkit Documentation*** *- https://simpleitk.readthedocs.io/en/master/*

[47] ***Pyradiomics Python toolkit Documentation*** *- https://readthedocs.org/projects/pyradiomics/*

[48] *Blagus, R., & Lusa, L. (2013).* ***SMOTE for high-dimensional class-imbalanced data****. BMC bioinformatics, 14(1), 106.*

[49] *Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002).* ***SMOTE: synthetic minority over-sampling technique.*** *Journal of artificial intelligence research, 16, 321-357.*

[50] *Feng, L., Liu, Z., Wang, J., Zhao, L., & Shen, C. (2019).* ***Radiomics analysis of magnetic resonance imaging and its clinical application in cancer****. Quantitative imaging in medicine and surgery, 9(8), 1327.*

[51] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). **Lightgbm: A highly efficient gradient boosting decision tree**.

[52] *Kuhn, M., & Johnson, K. (2019).* ***Feature engineering and selection: A practical approach for predictive models****. CRC press.*

[53] *Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002).* ***Gene selection for cancer classification using support vector machines****. Machine learning, 46(1), 389-422.*

[54] *Louppe, G. (2014).* ***Understanding random forests: From theory to practice****. arXiv preprint arXiv:1407.7502.*

[55] ***Deep learning-based survival analysis for brain metastasis patients with the national cancer database*** *- Noah Bice, Neil Kirby,Tyler Bahr, Karl Rasmussen, Daniel Saenz, Timothy Wagner, Niko Papanikolaou and Mohamad Fakhreddine 2020 Aug 13. DOI: 10.1002/acm2.12995*

[56] *Anatomy of human brain https://www.canstockphoto.com/human-brain-anatomy-diagram-48996036.html*

[57] *Image of healthy and tumor detect brain https://www.medanta.org/patient-education-blog/8-silent-signs-of-brain-tumour-you-should-know/*

[58] *Different types of brain tumor* ***https://blkhospitals.wordpress.com/2019/03/25/different-types-of-brain-tumor/***

[59] *Li, Ting & Sun, Shichen & Li, Yubing & Zhang, Yanyu & Wei, Linlin. (2024).* ***Immunotherapy revolutionizing brain metastatic cancer treatment: personalized strategies for transformative outcomes****. Frontiers in Immunology. 15. 10.3389/fimmu.2024.1418580.*

[60] *Kang, J.; Ullah, Z.; Gwak, J.* ***MRI-Based Brain Tumor Classification Using Ensemble of Deep Features and Machine Learning Classifiers. Sensors*** *2021, 21, 2222 https://doi.org/10.3390/s21062222*

[61] *Liu, Rong & Lin, Sijie & Rallo, Robert & Zhao, Yan & Damoiseaux, Robert & Xia, Tian & Lin, Shuo & Nel, Andre & Cohen, Yoram. (2012).* ***Automated Phenotype Recognition for Zebrafish Embryo Based In Vivo High Throughput Toxicity Screening of Engineered Nano-Materials****. PloS one. 7. e35014. 10.1371/journal.pone.0035014.*

[62] *Antunes, Jose. (2021).* ***To supervise or to self-supervise: a machine learning based comparison on credit supervision****. Financial Innovation. 7. 10.1186/s40854-021-00242-4.*

[63] *El Naqa, Issam & Hu, Qiyuan & Chen, Weijie & Li, Hui & Fuhrman, Jordan & Gorre, Naveena & Giger, Maryellen. (2021).* ***Lessons learned in transitioning to AI in the medical imaging of COVID-19****. Journal of Medical Imaging. 8. 10.1117/1.JMI.8.S1.010902.*

[64] *Daniel, Roshan & Tavares, Mithisha & Desphande, Vinayak. (2023).* ***UW -MADISON GI TRACT IMAGE SEGMENTATION TO TRACK HEALTHY ORGANS TO IMPROVE CANCER TREATMENT.***

[65] Xu, Yang and Hou, Shi-ke and Wang, Xiang-yu and Li, Duo and Lu, Lu, **C+Ref-UNet: A Novel Approach for Medical Image Segmentation Based on Multi-Scale Connected UNet and CBAM.** http://dx.doi.org/10.2139/ssrn.4192614