ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΒΙΟΪΑΤΡΙΚΗΣ

Αυτόματη Κατηγοριοποίηση της Σοβαρότητας της Πνευμονίας COVID-19 χρησιμοποιώντας Ψηφιακή Ακτινολογική Απεικόνιση και Σύγχρονες Τεχνικές Μηχανικής Μάθησης

Νέστορας Παπαδόπουλος
Αριθμός Μητρώου: 19388116

Επιβλέπων Καθηγητής
Εμμανουήλ Αθανασιάδης, Επίκουρος Καθηγητής

Αθήνα 16/10/2024

Η Τριμελής Εξεταστική Επιτροπή

Ο Επιβλέπων Καθηγητής

Εμμανουήλ Αθανασιάδης	Σπυρίδων Κωστόπουλος	Δημήτριος Γκλώτσος
Επίκουρος Καθηγητής	Αναπληρωτής Καθηγητής	Καθηγητής
[ΨΗΦΙΑΚΗ ΥΠΟΓΡΑΦΗ]	[ΨΗΦΙΑΚΗ ΥΠΟΓΡΑΦΗ]	[ΨΗΦΙΑΚΗ ΥΠΟΓΡΑΦΗ]
Digitally signed by EMMANOUIL ATHANASIADIS Date: 2024.11.18 14:33:46 +02'00'	Digitally signed by Spyridon Kostopoulos Date: 2024.11.18 13:33:17 +02'00'	Digitally signed by DIMITRIOS GKLOTSOS Date: 2024.11.18 13:18:28 +02'00'

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Ο υπογράφων Νέστορας Παπαδόπουλος του Γεωργίου, με αριθμό μητρώου 19388116, φοιτητής του Τμήματος Μήχανικών Βιοιατρικής της Σχολής Μηχανικών του Πανεπιστημίου Δυτικής Αττικής, δηλώνω υπεύθυνα ότι:

«Είμαι συγγραφέας αυτής της διπλωματικής εργασίας και κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος.

Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του διπλώματός μου».

Ημερομηνία 11/10/2024 Ο Δηλών

Νέστορας Παπαδόπουλος

ΠΕΡΙΛΗΨΗ

Η πανδημία COVID-19 έχει δημιουργήσει σημαντικές προκλήσεις στον τομέα της υγειονομικής περίθαλψης, ιδίως στην ακριβή διάγνωση και κατηγοριοποίηση της σοβαρότητας της πνευμονίας που προκαλείται από τον ιό. Αυτή η διπλωματική εργασία παρουσιάζει μια αυτοματοποιημένη προσέγγιση για την κατηγοριοποίηση της σοβαρότητας της πνευμονίας COVID-19 χρησιμοποιώντας ακτινογραφίες θώρακος και τεχνικές μηχανικής μάθησης. Χρησιμοποιήθηκαν δύο σύνολα δεδομένων: το πρώτο σύνολο βασίστηκε στη μέθοδο κατωφλίωσης του για τη δημιουργία μασκών και το δεύτερο σύνολο περιλάμβανε χειροκίνητη τμηματοποίηση των πνευμόνων. Χρησιμοποιώντας τις δυαδικές μάσκες εξάχθηκαν τα χαρακτηριστικά radiomics. Εφαρμόστηκαν τεχνικές επιλογής και μείωσης χαρακτηριστικών όπως η Ανάλυση Κυρίων Συνιστωσών (PCA), η Επαναλαμβανόμενη Αποβολή Χαρακτηριστικών (RFE) και η ανάλυση συσχέτισης. Υλοποιήθηκαν αρκετοί αλγόριθμοι μηχανικής μάθησης, όπως Υποστηρικτικές Μηχανές Διανυσμάτων (SVM), Τυχαία Δάση (Random Forests), Λογιστική Παλινδρόμηση, Δέντρα Απόφασης (CART),Ταξινομητής Ελάχιστης Απόστασης (MDC), Ταξινομητής Bayes,Ανάλυση Γραμμικού Διαχωρισμού (LDA) και Perceptron.

Τα μοντέλα αξιολογήθηκαν χρησιμοποιώντας cross-validation και στατιστική ανάλυση για τον εντοπισμό των πιο σχετικών χαρακτηριστικών και της απόδοσης κάθε ταξινομητή. Τα αποτελέσματα ανέδειξαν συγκεκριμένα χαρακτηριστικά, όπως RunLengthNonUniformity, Ενέργεια, CusterProminence, Entropy τα οποία είναι κρίσιμα για τη διάκριση μεταξύ υγιών ατόμων και ασθενών με COVID-19. Τα ευρήματα υποδεικνύουν ότι τα μοντέλα μπορούν να προσφέρουν πολύτιμες πληροφορίες για τη λήψη κλινικών αποφάσεων σχετικά με την εκτίμηση της σοβαρότητας της COVID-19.

Αυτή η εργασία επιδεικνύει το δυναμικό της ενσωμάτωσης της ραδιομικής ανάλυσης και της μηχανικής μάθησης για τη βελτίωση της διαγνωστικής ακρίβειας στην ιατρική απεικόνιση και παρέχει ένα πλαίσιο για μελλοντικές μελέτες στην αυτόματη κατηγοριοποίηση ασθενειών μέσω ψηφιακής ακτινογραφίας.

Λέξεις Κλειδιά: *Πνευμονία COVID-19, Μηχανική Μάθηση, Radiomics, Ακτινογραφίες Θώρακος, Τμηματοποίηση Εικόνας*

Abstract

The COVID-19 pandemic has posed significant challenges in healthcare, especially in the accurate diagnosis and severity classification of pneumonia caused by the virus. This thesis presents an automated approach for classifying the severity of COVID-19 pneumonia using chest X-rays and machine learning techniques. Two datasets were utilized: the first set used thresholding to create masks, and the second set involved manual segmentation of lung regions. Radiomic features were extracted from the masked regions to quantify texture and intensity variations.

To reduce dimensionality and improve model performance, feature selection techniques such as Principal Component Analysis (PCA), Recursive Feature Elimination (RFE), and correlation analysis were applied. Several machine learning algorithms were implemented, including Support Vector Machines (SVM), Random Forests, Logistic Regression, Decision Trees (CART), Minimum Distance Classifier (MDC), Bayesian Classifier, Linear Discriminant Analysis (LDA), and Perceptron.

The models were evaluated using cross-validation and statistical analysis to identify the most relevant features and the performance of each classifier. The results highlight specific radiomic features, such as run-length non-uniformity, energy,ClusterProminence and Entropy which are critical in distinguishing between healthy controls and COVID-19 patients. The findings suggest the models can provide valuable insights for clinical decision-making in COVID-19 severity assessment.

This work demonstrates the potential of integrating radiomics and machine learning for improving diagnostic accuracy in medical imaging and provides a framework for future studies in automatic disease classification using digital radiography.

Keywords: *COVID-19 Pneumonia, Machine Learning, Radiomics, Chest X-rays Image Segmentation*

Ευχαριστίες:
Θα ήθελα να εκφράσω τις ιδιαίτερες μου ευχαριστίες στον επιβλέποντα καθηγητή για την καθοδήγηση και την υποστήριξη κατά την διάρκεια της εκπόνησης της παρούσας εργασίας.

Περιεχόμενα

Κατάλογος Συντομογραφιών

COVID-19 - Coronavirus Disease 2019

WHO - World Health Organization

ECMO - ExtraCorporeal Membrane Oxygenation

PCA - Principal Component Analysis

RFE - Recursive Feature Elimination

SVM - Support Vector Machines

MDC - Minimum Distance Classifier

LDA - Linear Discriminant Analysis

GGO - Ground Glass Opacity

CT - Computed Tomography

MRI - Magnetic Resonance Imaging

PET - Positron Emission Tomography

ROI - Region of Interest

ECMO - Extracorporeal Membrane Oxygenation

RT-PCR - Reverse Transcription Polymerase Chain Reaction

STD - Standard Deviation

PA - Posterior-Anterior (X-ray view)

AP - Anterior-Posterior (X-ray view)

GLCM - Gray Level Co-occurrence Matrix

GLRLM - Gray Level Run Length Matrix

GLSZM - Gray Level Size Zone Matrix

NGTDM - Neighboring Gray Tone Dependence Matrix

GLDM - Gray Level Dependence Matrix

ΕΙΣΑΓΩΓΗ

Η πανδημία της ασθένιας COVID-19 έφερε την ιατρική κοινότητα αντιμέτωπη με νέες προκλήσεις στην αντιμετώπιση και διάγνωση της νόσου. Ένα από τα κύρια διαγνωστικά εργαλεία είναι η ψηφιακή ακτινολογική απεικόνιση, η οποία προσφέρει άμεση και ακριβή εικόνα της κατάστασης των πνευμόνων των ασθενών. Η παρούσα διπλωματική εργασία επιδιώκει να αναπτύξει μια αυτόματη μέθοδο κατηγοριοποίησης της σοβαρότητας της πνευμονίας COVID-19 χρησιμοποιώντας ακτινολογικές εικόνες και σύγχρονες τεχνικές μηχανικής μάθησης. Συγκεκριμένα, στόχος είναι η δημιουργία ενός εργαλείου που να μπορεί να διακρίνει τα διάφορα στάδια της πνευμονίας COVID-19 με ακρίβεια και συνέπεια, βοηθώντας τους ιατρούς στη λήψη αποφάσεων.

Για την επίτευξη του στόχου, χρησιμοποιήθηκαν ακτινολογικές εικόνες πνευμόνων από ασθενείς με COVID-19. Οι εικόνες αυτές αναλύθηκαν χρησιμοποιώντας την βιβλιοθήκη "radiomics" της Python, η οποία επιτρέπει την εξαγωγή χαρακτηριστικών από τις ιατρικές εικόνες. Τα εξαγόμενα χαρακτηριστικά στη συνέχεια αποτέλεσαν τα δεδομένα εισόδου για τους αλγόριθμους μηχανικής μάθησης που αναπτύχθηκαν.

Στην εργασία χρησιμοποιήθηκαν μηχανές υποστήριξης διανυσμάτων (Support Vector Machines - SVM), τυχαία δάση (Random Forests) και άλλες τεχνικές μηχανικής μάθησης που δεν περιλαμβάνουν βαθιά μάθηση. Η επιλογή αυτή έγινε ώστε να αποφευχθούν τα προβλήματα υπερπροσαρμογής (overfitting) και να εξασφαλιστεί η δυνατότητα εκπαίδευσης των μοντέλων ακόμη και με περιορισμένα δεδομένα.

Κατά την ανάπτυξη της μεθόδου, αντιμετωπίστηκαν αρκετές δυσκολίες. Πρώτον, η εξασφάλιση της ποιότητας των δεδομένων ήταν κρίσιμη, καθώς οποιαδήποτε λάθος επισήμανση μπορούσε να επηρεάσει αρνητικά την εκπαίδευση των μοντέλων. Δεύτερον, η διαδικασία εξαγωγής χαρακτηριστικών μέσω της βιβλιοθήκης "radiomics" απαιτούσε ενδελεχή ρύθμιση των παραμέτρων για την εξαγωγή των πιο σχετικών χαρακτηριστικών.

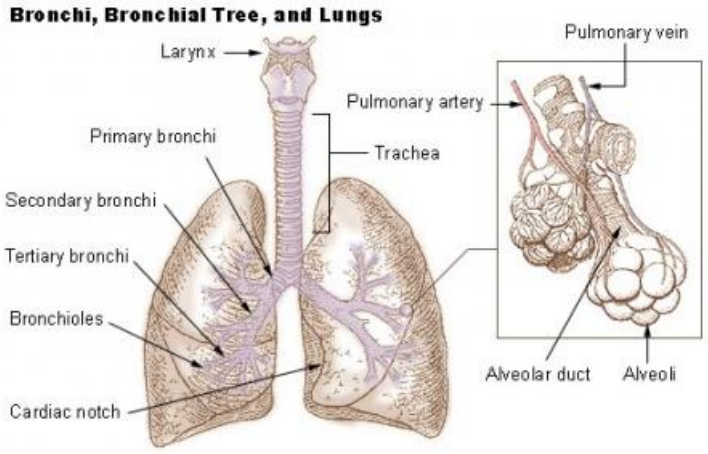
Παρουσιάζεται μια αναλυτική επισκόπηση της πνευμονίας που προκαλείται από τον ιό COVID-19, με έμφαση στη διάγνωση μέσω ακτινολογικών εικόνων. Εξετάζεται το θεωρητικό υπόβαθρο των μεθόδων μηχανικής μάθησης που χρησιμοποιήθηκαν στην εργασία, με αναφορές στη χρήση της Python και των σχετικών βιβλιοθηκών. Περιγράφεται η διαδικασία εξαγωγής χαρακτηριστικών από τις ακτινολογικές εικόνες χρησιμοποιώντας την βιβλιοθήκη "radiomics" της Python. Αναλύεται η διαδικασία ανάπτυξης των αλγορίθμων μηχανικής μάθησης, οι παράμετροι που χρησιμοποιήθηκαν, και η εκπαίδευση των μοντέλων. Παρουσιάζονται τα αποτελέσματα των μοντέλων, με έμφαση στην ακρίβεια και τη συνέπεια της κατηγοριοποίησης των σταδίων της πνευμονίας COVID-19. Συνοψίζονται τα συμπεράσματα της εργασίας και προτείνονται κατευθύνσεις για μελλοντική έρευνα, ιδίως σχετικά με την ανάγκη για μεγαλύτερα και πιο ποιοτικά σύνολα δεδομένων.

Σε αυτήν την εργασία, το ταξίδι ξεκινά με μια επισκόπηση της COVID-19 πνευμονίας, εστιάζοντας στη διάγνωσή της μέσω ακτινολογικής απεικόνισης, και στη συνέχεια εμβαθύνει στο θεωρητικό πλαίσιο της μηχανικής μάθησης, χρησιμοποιώντας “Python”. Η καρδιά της εργασίας περιλαμβάνει την ανάπτυξη και τη βελτίωση ενός κώδικα για την ταξινόμηση ακτινολογικών εικόνων ασθενών, αντιμετώπιση προκλήσεων όπως η υπερπροσαρμογή (overfitting) και ο περιορισμός δεδομένων, ιδιαίτερα σε ορισμένα στάδια της νόσου. Παρά τα εμπόδια αυτά, γίνονται σταδιακές βελτιώσεις. Η διατριβή ολοκληρώνεται υπογραμμίζοντας την ανάγκη για ένα πιο εκτεταμένο σύνολο δεδομένων για την ενίσχυση της αποτελεσματικότητας του κώδικα στην ακριβή ταξινόμηση των σταδίων της νόσου.

1. Νόσος του κορωνοϊού 2019 (COVID-19)

1.1. Ανατομία των πνευμόνων

Κάθε πνεύμονας διαιρείται σε λοβούς μέσω σχισμών. Και οι δύο πνεύμονες περιέχουν λοξές σχισμές, ενώ ο δεξιός πνεύμονας έχει επιπλέον μια οριζόντια σχισμή. Η λοξή σχισμή στον αριστερό πνεύμονα χωρίζει τον άνω από τον κάτω λοβό. Στον δεξιό πνεύμονα, η λοξή και η εγκάρσια σχισμή τον διαχωρίζουν σε άνω, μεσαίο και κάτω λοβό. Επομένως, ο δεξιός πνεύμονας έχει τρεις λοβούς, ενώ ο αριστερός μόνο δύο. Κάθε λοβός λαμβάνει αερισμό μέσω ενός βρόγχου. Οι λοβοί, με τη σειρά τους, χωρίζονται σε βρογχοπνευμονικά τμήματα που εξυπηρετούνται από τμηματικούς βρόγχους.



Εικόνα 1.1 Ανατομία των πνευμόνων *[1]*

Όλες οι αναπνευστικές οδοί από την τραχεία έως τους αναπνευστικούς βρόγχους ονομάζονται τραχειοβρογχικό δέντρο. Η τραχεία διαιρείται στο επίπεδο του στέρνου σε δεξιό και αριστερό κύριο βρόγχο που πηγαίνουν στον δεξιό και αριστερό πνεύμονα. Κάθε βρόγχος εισέρχεται στον πνεύμονα μέσω μιας εγκοπής που ονομάζεται πύλη. Σε αυτό το σημείο, τα αιμοφόρα αγγεία και τα νεύρα συνδέονται επίσης με τους πνεύμονες, και μαζί με τον βρόγχο σχηματίζουν τη ρίζα των πνευμόνων. Ο δεξιός κύριος βρόγχος έχει μεγαλύτερη διάμετρο και είναι πιο κάθετος σε σχέση με τον αριστερό, καθιστώντας τον πιο ευθυγραμμισμένο με την τραχεία. Ως εκ τούτου, κατά λάθος εισπνεόμενα αντικείμενα είναι πιο πιθανό να παγιδευτούν στον δεξιό κύριο βρόγχο. Οι κύριοι βρόγχοι διακλαδίζονται σε λοβιαίους (δευτερογενείς) βρόγχους εντός κάθε πνεύμονα. Ο αριστερός πνεύμονας έχει δύο λοβαίους βρόγχους, ενώ ο δεξιός έχει τρεις. Αυτοί οι λοβαίοι βρόγχοι με τη σειρά τους διαχωρίζονται σε τμηματικούς (τριτογενείς) βρόγχους, οι οποίοι τροφοδοτούν τα βρογχοπνευμονικά τμήματα[1].

1.2.Βασικές πληροφορίες για τη νόσο

1.2.1.Ορισμός και γενική επισκόπηση

Η πνευμονία που προκαλείται από τη λοίμωξη από τον ιό σοβαρού αναπνευστικού συνδρόμου κατά του οποίου ορισμένοι αναφέρονται ως SARS-CoV-2 εμφανίστηκε στην πόλη Γουχάν της Κίνας τον Δεκέμβριο του 2019.

Στις 11 Φεβρουαρίου 2020, ο Παγκόσμιος Οργανισμός Υγείας (WHO) ονόμασε επίσημα τη νόσο που προκαλείται από τον ιό SARS-CoV-2 ως νόσο κορωνοϊού 2019 (COVID-19). Η COVID-19 εμφανίζει μια ποικιλία κλινικών συμπτωμάτων, όπως πυρετό, ξηρό βήχα και κόπωση, με συχνή εμπλοκή των πνευμόνων. Ο SARS-CoV-2 είναι εξαιρετικά μεταδοτικός, και το μεγαλύτερο μέρος του πληθυσμού είναι ευάλωτο στη μόλυνση.

Οι περισσότεροι άνθρωποι που μολύνονται από τον ιό θα εμφανίσουν ήπια έως μέτρια αναπνευστική νόσο και θα αναρρώσουν χωρίς την ανάγκη ειδικής θεραπείας. Ωστόσο, κάποιοι θα νοσήσουν σοβαρά και θα χρειαστούν ιατρική φροντίδα. Οι ηλικιωμένοι και όσοι πάσχουν από υποκείμενες ασθένειες, όπως καρδιοπάθειες, διαβήτη, χρόνια αναπνευστική νόσο ή καρκίνο, έχουν μεγαλύτερη πιθανότητα να αναπτύξουν σοβαρές επιπλοκές.

Η ιογενής λοίμωξη προκαλεί βλάβη στα επιθηλιακά κύτταρα των αεραγωγών και στις κυψελίδες των πνευμόνων. Ωστόσο, όπως και στον SARS-CoV, η αντίδραση του ανοσοποιητικού συστήματος μπορεί να παίξει κρίσιμο ρόλο στην παθογένεση της COVID-19, ειδικά σε εκείνους με σοβαρή νόσο [2,3].

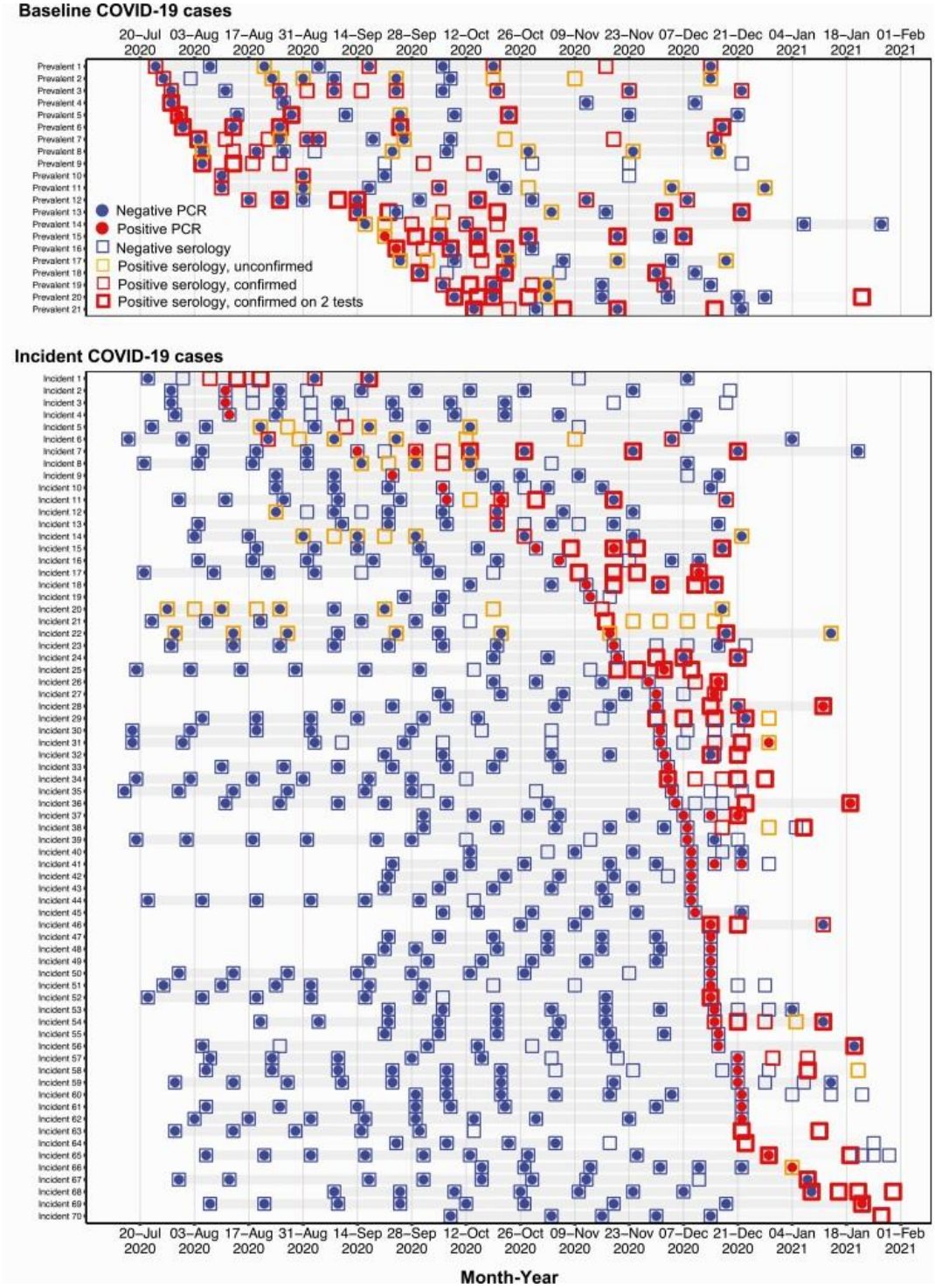
1.2.2. Επιπολασμός και συχνότητα εμφάνισης της νόσου

Σε μια έρευνα από τον [3], εντοπίστηκαν 21 από τα 2435 άτομα με ενδείξεις COVID-19 κατά την έναρξη και εκτιμήθηκε επιπολασμός 0,86% (95% CI, 0,53%–1,32%). Κατά τη διάρκεια της παρακολούθησης, εντοπίστηκαν 70 από τα 2.414 άτομα (2,9%) με επεισόδιο COVID-19, και εκτιμήθηκε ένα συνολικό ποσοστό επίπτωσης 9,11 περιστατικών ανά 100 ανθρώπινα έτη (95% CI, 7,11–11,52). Ο αριθμός των νέων κρουσμάτων αυξήθηκε ανάλογα με τον επιπολασμό της COVID-19 στην περιοχή των 8 κομητειών όπου διεξήχθη η μελέτη (Σχήμα 2). Οι εκτιμήσεις του ποσοστού επίπτωσης δεν διέφεραν ανάλογα με το φύλο, την εθνικότητα/φυλή ή το επάγγελμα (Συμπληρωματικό Σχήμα 1).
Από τα 21 κρούσματα COVID-19 που ανιχνεύθηκαν κατά την αρχική εξέταση, όλα πληρούσαν τα κριτήρια κρούσματος με θετικά ανοσολογικά αποτελέσματα· μόνο 3 είχαν επίσης θετικό αποτέλεσμα RT-PCR. Η πλειοψηφία των 70 νέων κρουσμάτων επιβεβαιώθηκε μέσω RT-PCR (53 από 70 [76%]), είτε συνοδευόμενα είτε όχι από θετικά ανοσολογικά αποτελέσματα. Από τα 17 άτομα (24%) με θετικά ανοσολογικά αποτελέσματα, μόνο 2 (12%) εμφάνισαν θετικό αποτέλεσμα RT-PCR σε μεταγενέστερη επίσκεψη (2 ή 5 εβδομάδες μετά το θετικό ανοσολογικό αποτέλεσμα).
Πραγματοποιήθηκε μια ανάλυση ευαισθησίας χρησιμοποιώντας έναν εναλλακτικό ορισμό των περιστατικών COVID-19 που περιλάμβανε όλα τα μη επιβεβαιωμένα θετικά ανοσολογικά αποτελέσματα ως κρούσματα, με αποτέλεσμα την αύξηση του επιπολασμού κατά την αρχική εξέταση σε 1,07% (95% CI, 0,79%–1,56%) και την αύξηση της αθροιστικής επίπτωσης σε 9,26 περιστατικά ανά 100 ανθρώπινα έτη (95% CI, 7,24–11,69).Για να εξεταστεί η επίδραση πιθανών ψευδώς θετικών αποτελεσμάτων RT-PCR, πραγματοποιήθηκε μια δεύτερη ανάλυση ευαισθησίας χρησιμοποιώντας

έναν δεύτερο εναλλακτικό ορισμό των κρουσμάτων, που απέκλεισε 7 κρούσματα που πληρούσαν αυτόν τον ορισμό. Αυτό μείωσε το συγκεντρωτικό ποσοστό επίπτωσης σε 8,18 περιστατικά ανά 100 ανθρώπινα έτη (95% CI, 6,29–10,4).

Συνολικά, η απόδοση των ελέγχων της κοόρτης περιστατικών ήταν σχετικά χαμηλή: μόνο 30 από τις 12.007 δοκιμές RT-PCR (0,25%) που πραγματοποιήθηκαν σε ασυμπτωματικούς συμμετέχοντες είχαν θετικά αποτελέσματα, και 7 από τα 30 (23%) πληρούσαν τον ορισμό του ψευδός-θετικού περιστατικού.

Το Σχήμα 1 δείχνει τη χρονική αλληλουχία των αποτελεσμάτων των ελέγχων στο επίπεδο των συμμετεχόντων για όλα τα αρχικά προϋπάρχοντα περιστατικά και όλα τα επεισόδια περιστατικά. Βρήκαμε σημαντική εξέλιξη των αντισωματικών αντιδράσεων με την πάροδο του χρόνου: από τα 56 περιστατικά που αρχικά διαγνώστηκαν με RT-PCR, 11 είχαν τουλάχιστον ένα θετικό αντίσωμα κατά τη διάγνωση. Μέχρι το τέλος της παρακολούθησης, αυτός ο αριθμός αυξήθηκε σε 27 άτομα με τουλάχιστον ένα θετικό αποτέλεσμα αντισώματος. Επιπλέον, 11 άτομα που είχαν ανιχνευθεί με αντισώματα με μία μέθοδο, αργότερα βρέθηκαν θετικά με άλλη μέθοδο[4].



Εικόνα 1.2 Επίπτωση και επιπολασμός της COVID-19

[4]

Ο χρονισμός και η ακολουθία των θετικών τεστ μεταξύ των εργαζομένων στον τομέα της υγείας που είχαν COVID-19 παρουσιάζεται με γραμμές, καθεμία από τις οποίες αντιπροσωπεύει τα αποτελέσματα των τεστ για κάθε συμμετέχοντα. Το διάγραμμα απεικονίζει τον χρόνο παρακολούθησης με γκρι σκίαση για κάθε άτομο, ενώ οι κουκκίδες δείχνουν τα αποτελέσματα της RT-PCR, και τα κουτιά δείχνουν τα ανοσολογικά αποτελέσματα.

- Μπλε χρώμα: αρνητικά αποτελέσματα RT-PCR ή ανοσολογικών τεστ.

- Κόκκινο χρώμα: θετικά αποτελέσματα RT-PCR ή επιβεβαιωμένα θετικά ανοσολογικά αποτελέσματα.
- Πορτοκαλί κουτιά: μη επιβεβαιωμένα θετικά ανοσολογικά αποτελέσματα.

Το πάχος των κόκκινων κουτιών σχετίζεται με τον αριθμό των επιβεβαιωμένων θετικών τεστ αντισωμάτων, για παράδειγμα, 2 ή 3 θετικά τεστ αντισωμάτων. Αυτός ο τρόπος απεικόνισης επιτρέπει την παρακολούθηση της εξέλιξης των τεστ και τον χρονισμό των θετικών αποτελεσμάτων κατά τη διάρκεια της μελέτης.

1.2.3. Σημασία μελέτης της COVID-19

Η γνώση για την COVID-19 παίζει κρίσιμο ρόλο στην ενθάρρυνση των ατόμων να τηρούν τις προστατευτικές συμπεριφορές. Η κατανόηση της COVID-19, συμπεριλαμβανομένων των τρόπων μετάδοσης, των συμπτωμάτων και των στρατηγικών πρόληψης, είναι ουσιώδης για το κοινό, καθώς η ακριβής και ολοκληρωμένη διάδοση πληροφοριών είναι απαραίτητη για την καταπολέμηση της παραπληροφόρησης και την προώθηση της ευαισθητοποίησης. Η γνώση για την COVID-19 συνδέεται άμεσα με την πιθανότητα τα άτομα να ακολουθούν τις συνιστώμενες προστατευτικές συμπεριφορές, όπως η χρήση μάσκας, η τήρηση κοινωνικών αποστάσεων και το συχνό πλύσιμο των χεριών, καθώς αυτοί που είναι καλά ενημερωμένοι είναι πιο πιθανό να παίρνουν αυτά τα μέτρα σοβαρά. Η σημασία των εκπαιδευτικών εκστρατειών και των μηνυμάτων δημόσιας υγείας στην ενίσχυση της κατανόησης του κοινού για τον ιό είναι επίσης υπογραμμισμένη, ενώ οι εκπαιδευτικές πρωτοβουλίες πρέπει να είναι στοχευμένες και προσαρμοσμένες ώστε να φτάνουν σε ποικίλες πληθυσμιακές ομάδες, εξασφαλίζοντας ότι όλα τα μέλη της κοινότητας έχουν πρόσβαση σε ακριβείς πληροφορίες. Η παραπληροφόρηση και οι μύθοι για την COVID-19 μπορούν να εμποδίσουν τις προστατευτικές συμπεριφορές, και είναι κρίσιμο να αντιμετωπιστούν αυτές οι προκλήσεις με αποτελεσματικές στρατηγικές επικοινωνίας για την προώθηση επιστημονικά τεκμηριωμένων γεγονότων. Η εμπλοκή των ηγετών της κοινότητας και η χρήση διαφόρων καναλιών επικοινωνίας, συμπεριλαμβανομένων των κοινωνικών μέσων, είναι ζωτικής σημασίας για τη διάδοση ακριβών πληροφοριών, καθώς η συμμετοχή της κοινότητας μπορεί να βοηθήσει στην ενίσχυση των θετικών συμπεριφορών και στην υποστήριξη των προσπαθειών δημόσιας υγείας. Οι υπεύθυνοι χάραξης πολιτικής ενθαρρύνονται να δώσουν προτεραιότητα στη διάδοση γνώσεων ως μέρος της στρατηγικής δημόσιας υγείας τους, που περιλαμβάνει επενδύσεις στην επικοινωνιακή υποδομή, υποστήριξη εκπαιδευτικών πρωτοβουλιών και προώθηση της συνεργασίας μεταξύ των υγειονομικών αρχών και της κοινότητας. Η κατανόηση της COVID-19 είναι θεμελιώδης για την προώθηση της τήρησης των προστατευτικών συμπεριφορών, και εξασφαλίζοντας ότι το κοινό είναι καλά ενημερωμένο, οι υγειονομικές αρχές μπορούν να ενισχύσουν τη συμμόρφωση με τα προληπτικά μέτρα, μειώνοντας τελικά τη διάδοση του ιού και μετριάζοντας τις επιπτώσεις του στην κοινωνία[5].

1.3. *Αιτιολογία της ασθένειας και Παράγοντες Κινδύνου*

Η ηλικία είναι ο ισχυρότερος παράγοντας κινδύνου για σοβαρές εκβάσεις της COVID-19. Οι ασθενείς με μία ή περισσότερες από ορισμένες υποκείμενες ιατρικές παθήσεις διατρέχουν επίσης υψηλότερο κίνδυνο. Επιπλέον, το να μην είναι κάποιος εμβολιασμένος ή να μην έχει ενημερωθεί με τις τρέχουσες δόσεις εμβολίων COVID-19 αυξάνει επίσης τον κίνδυνο σοβαρών εκβάσεων της COVID-19.Οι πάροχοι υγείας θα πρέπει να λαμβάνουν υπόψη την ηλικία του ασθενούς, την παρουσία υποκείμενων ιατρικών παθήσεων και άλλων παραγόντων κινδύνου, καθώς και την κατάσταση

εμβολιασμού, για να καθορίσουν τον κίνδυνο σοβαρών εκβάσεων που σχετίζονται με την COVID-19 για κάθε ασθενή[6].

Ακόμα ένας παράγοντας που αυξάνει τον κίνδυνο νόσησης είναι η ισότητα υγείας. Η ισότητα υγείας είναι η κατάσταση κατά την οποία όλοι έχουν μια δίκαιη και ίση ευκαιρία να επιτύχουν το υψηλότερο επίπεδο υγείας τους. Η επίτευξη αυτής της ισότητας απαιτεί συνεχιζόμενες κοινωνικές προσπάθειες για την αντιμετώπιση ιστορικών και σύγχρονων αδικιών, την υπέρβαση οικονομικών, κοινωνικών και άλλων εμποδίων στην υγεία και την περίθαλψη, και την εξάλειψη προλαμβανόμενων ανισοτήτων υγείας. Για να επιτευχθεί η ισότητα υγείας, πρέπει να αλλάξουμε τα συστήματα και τις πολιτικές που έχουν οδηγήσει στις γενεαλογικές αδικίες που δημιουργούν φυλετικές και εθνοτικές ανισότητες υγείας[7].

Μεταξύ πολλών γενετικών διαταραχών που επηρεάζουν τους ανθρώπους από τις πρώτες ημέρες ή χρόνια της ζωής, οι πολυπαραγοντικές ασθένειες έχουν γένεση που επηρεάζεται τόσο από περιβαλλοντικούς παράγοντες όσο και από γονίδια[8].

1.4. Κλινική Παρουσίαση και Συμπτώματα

Είναι σημαντικό να θυμόμαστε ότι τα συμπτώματα της COVID-19 εξελίσσονται διαφορετικά ανάλογα με το άτομο. Τυπικά, τα συμπτώματα εμφανίζονται 2–14 ημέρες μετά την έκθεση, πιο συχνά μετά από 5–6 ημέρες. Την πρώτη εβδομάδα, το αρχικό σύμπτωμα είναι πιθανό να είναι ο πυρετός, που εμφανίζεται σε περίπου 78% των περιπτώσεων. Αυτό ακολουθείται από συμπτώματα όπως βήχας, πονόλαιμος, πόνους στο σώμα και πονοκέφαλοι. Ορισμένα άτομα μπορεί να εμφανίσουν ναυτία και εμετό νωρίτερα από ότι σε άλλες αναπνευστικές λοιμώξεις. Σε σοβαρές περιπτώσεις, μπορεί να απαιτηθεί νοσηλεία και το σύνδρομο οξείας αναπνευστικής ανεπάρκειας, με την ανάγκη για εντατική φροντίδα περίπου 10 ημέρες μετά την εμφάνιση των συμπτωμάτων. Η σειρά και ο τύπος των συμπτωμάτων ποικίλλουν· ορισμένοι μπορεί να έχουν γαστρεντερικά συμπτώματα πριν τον πυρετό ή τον βήχα, ενώ άλλοι μπορεί να μην έχουν καθόλου συμπτώματα. Άλλα κοινά συμπτώματα περιλαμβάνουν ρίγη, κόπωση, απώλεια γεύσης ή όσφρησης, και ρινική συμφόρηση. Σοβαρά συμπτώματα όπως δύσπνοια, σύγχυση, πόνος ή πίεση στο στήθος, ή δυσκολία στην κίνηση ή στην ομιλία απαιτούν άμεση ιατρική φροντίδα. Οι διάφορες παραλλαγές όπως Ομικρον, Άλφα, Βήτα, Γάμμα και Δέλτα προκαλούν παρόμοια συμπτώματα, αλλά μπορεί να παρουσιάζουν συχνότερα συμπτώματα κρυολογήματος όπως πονοκέφαλοι, καταρροή και πονόλαιμος[9].

1.5. Διάγνωση και πρόγνωση

Υπάρχουν δύο τύποι τεστ που μπορούν να βοηθήσουν στη διάγνωση της COVID-19. Τα μοριακά τεστ ανιχνεύουν γενετικό υλικό από τον ιό της COVID-19. Ένα παράδειγμα είναι τα τεστ αλυσιδωτής αντίδρασης πολυμεράσης (PCR), τα οποία είναι πιο ακριβή από τα αντιγονικά τεστ και μπορούν να γίνουν στο σπίτι ή, πιο συχνά, από επαγγελματίες υγείας και να επεξεργαστούν σε εργαστήριο. Τα αντιγονικά τεστ ανιχνεύουν ιικές πρωτεΐνες και είναι γνωστά ως γρήγορα τεστ ή τεστ στο σπίτι. Αν και αξιόπιστα, είναι λιγότερο ακριβή από τα τεστ PCR, ιδιαίτερα σε ασυμπτωματικούς ασθενείς. Αν ένα αντιγονικό τεστ είναι αρνητικό, συνιστάται να επαναληφθεί μετά από 48 ώρες για πιο ακριβές αποτέλεσμα. Αν το αποτέλεσμα είναι θετικό σε οποιοδήποτε από τα δύο τεστ, είναι σχεδόν βέβαιο ότι έχετε COVID-19 και δεν χρειάζεται άλλο τεστ. Αν το αποτέλεσμα του τεστ PCR είναι αρνητικό, πιθανότατα δεν έχετε COVID-19. Σε περίπτωση θετικού αποτελέσματος, χρειάζεται άμεση επικοινωνία με επαγγελματία υγείας για συζήτηση για διαθέσιμες επιλογές[10].

Σε μια μελέτη από τον Bellou et al. περιγράφετε η διεξαγωγή και τα αποτελέσματα 263 μετα-αναλύσεων που εστιάζουν σε διάφορους κινδύνους και παράγοντες που σχετίζονται με την COVID-19. Οι μετα-αναλύσεις εξετάζουν τους κινδύνους θνησιμότητας, εισαγωγής στο νοσοκομείο, εισαγωγής στη μονάδα εντατικής θεραπείας (ΜΕΘ), μηχανικού αερισμού, οξείας νεφρικής βλάβης, φλεβικής θρομβοεμβολής, πνευμονικής εμβολής, οξείας αναπνευστικής δυσχέρειας και βαθιάς φλεβικής θρόμβωσης. Συνολικά, αξιολογήθηκαν 91 μοναδικοί προγνωστικοί παράγοντες, οι οποίοι κατηγοριοποιήθηκαν σε επτά κατηγορίες, όπως βιοδείκτες, συννοσηρότητες, απεικονιστικοί δείκτες, δημογραφικά χαρακτηριστικά, περιβαλλοντικοί παράγοντες, φάρμακα και συμπτώματα ή κλινικά σημεία.
Τα ευρήματα από τις μετα-αναλύσεις έδειξαν ότι ένας μεγάλος αριθμός από αυτές παρουσίασε στατιστικά σημαντικά αποτελέσματα, με αρκετές να περιλαμβάνουν πάνω από 1000 περιστατικά και να έχουν υψηλό βαθμό ακρίβειας. Μερικά από τα σημαντικά ευρήματα περιλαμβάνουν την επίδραση της αποφρακτικής υπνικής άπνοιας και του ιστορικού φλεβικής θρομβοεμβολής στον κίνδυνο νοσηλείας, την επίδραση του γυναικείου φύλου στον κίνδυνο εισαγωγής στη Μονάδα Εντατικής Θεραπείας (ΜΕΘ), και την επίδραση διαφόρων συννοσηροτήτων όπως ο καρκίνος και η Χρόνια Αποφρακτική Πνευμονοπάθεια (ΧΑΠ) στον κίνδυνο θνησιμότητας[11].

1.6. Θεραπευτικές Προσεγγίσεις

Για άτομα που νοσηλεύονται για φροντίδα COVID-19, η φροντίδα παρέχεται βάσει της αντίδρασης του ανοσοποιητικού συστήματος του ατόμου και της ανάγκης για υποστήριξη οξυγόνου.

Το πρόσθετο οξυγόνο μπορεί να χορηγηθεί μέσω σωλήνα στη μύτη, ενώ σε ορισμένα άτομα μπορεί να χρειαστεί τοποθέτηση σωλήνα στην αεροφόρο οδό για μηχανικό αερισμό, ώστε να διοχετευτεί αέρας στους πνεύμονες. Σε πολύ σοβαρές περιπτώσεις, μπορεί να χρησιμοποιηθεί εξωσωματική μεμβρανική οξυγόνωση (ECMO), μια συσκευή που μιμείται τη λειτουργία της καρδιάς και των πνευμόνων.

Για τη θεραπεία της σοβαρής COVID-19, μπορεί να χορηγηθούν φάρμακα όπως η ρεμδεσιβίρη, η μπαρισιτινίμπη (Olumiant), η τοσιλιζουμάμπη (Actemra), ή κορτικοστεροειδή όπως η δεξαμεθαζόνη. Η μπαρισιτινίμπη χορηγείται σε μορφή χαπιού, η τοσιλιζουμάμπη ως ένεση, ενώ η δεξαμεθαζόνη μπορεί να δοθεί είτε σε χάπι είτε με ενδοφλέβια ένεση.

Μια επιπλέον θεραπευτική επιλογή είναι το αναρρωτικό πλάσμα, προερχόμενο από αίμα ατόμων που έχουν αναρρώσει από COVID-19. Το πλάσμα, το οποίο περιέχει αντισώματα του ανοσοποιητικού συστήματος, μπορεί να χρησιμοποιηθεί για τη βοήθεια ατόμων με εξασθενημένο ανοσοποιητικό σύστημα να αναρρώσουν από τη νόσο[10].

Οι αναδυόμενες θεραπευτικές προσεγγίσεις για την αντιμετώπιση του COVID-19 περιλαμβάνουν διάφορες καινοτόμες μεθόδους. Πρώτον, αναπτύσσονται νέα αντιικά φάρμακα που στοχεύουν άμεσα τον ιό SARS-CoV-2, προσφέροντας νέες δυνατότητες αντιμετώπισης. Επίσης, τα μονοκλωνικά αντισώματα χρησιμοποιούνται για την εξουδετέρωση του ιού, παρέχοντας μια εξειδικευμένη θεραπευτική επιλογή. Η ανοσοθεραπεία αποτελεί μια άλλη προσέγγιση που ενισχύει το ανοσοποιητικό σύστημα ώστε να καταπολεμήσει τον ιό πιο αποτελεσματικά. Παράλληλα, τα εμβόλια νέας γενιάς αναπτύσσονται για να προσφέρουν μακροχρόνια προστασία και να αντιμετωπίζουν νέες παραλλαγές του ιού. Επιπλέον, τα αντιφλεγμονώδη φάρμακα χρησιμοποιούνται για τη μείωση της φλεγμονώδους αντίδρασης του οργανισμού, ενώ η γονιδιακή θεραπεία εξετάζεται ως μέθοδος για τη διόρθωση των επιπτώσεων του ιού

στο ανθρώπινο σώμα. Τέλος, οι αναστολείς εισόδου του ιού αποτελούν μια πολλά υποσχόμενη κατηγορία φαρμάκων που εμποδίζουν την είσοδο του ιού στα ανθρώπινα κύτταρα, προσφέροντας έτσι έναν επιπλέον τρόπο προστασίας[12].

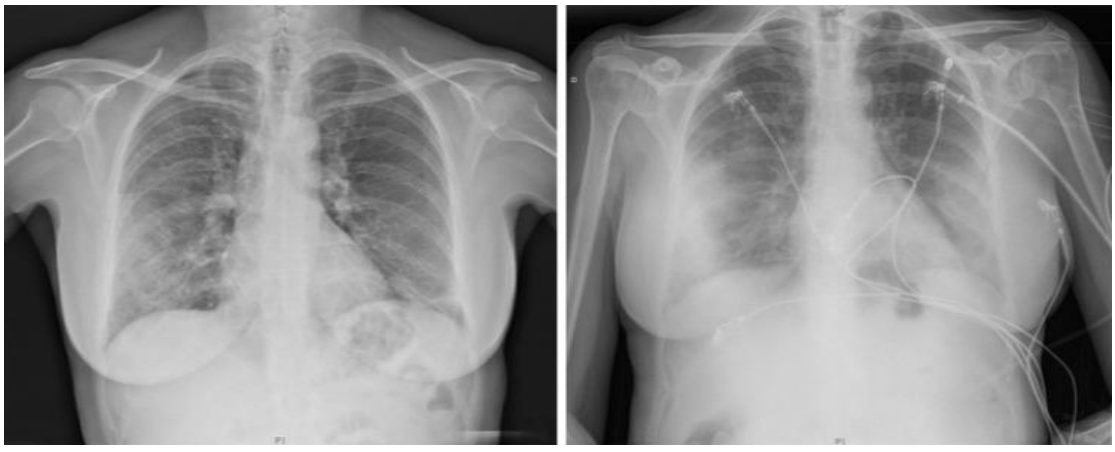
2. Ακτινολογικό Μηχάνημα(Chest X-rays)

Η ακτινογραφία θώρακος είναι μια συχνά χρησιμοποιούμενη μέθοδος λόγω του χαμηλού κόστους και της ευρείας διαθεσιμότητάς της, επιτρέποντας τη μελέτη διάφορων καταστάσεων με απλό και γρήγορο τρόπο. Ορισμένες μελέτες έχουν προτείνει ότι η ακτινογραφία θώρακος είναι μια χρήσιμη μέθοδος τόσο για τη διάγνωση όσο και για την παρακολούθηση της πνευμονίας που προκαλείται από τη μόλυνση SARS-CoV-2, αλλά έχει χαμηλή ευαισθησία στην αναγνώριση των πνευμονικών αλλοιώσεων που προκαλούνται από τη μόλυνση. Η χρήση των ακτινογραφιών θώρακος περιοριζόταν στην παρακολούθηση των ασθενών που νοσηλεύονται στη ΜΕΘ, των οποίων η ευπάθεια τους θα καθιστούσε δύσκολη τη μεταφορά τους για αξονική τομογραφία θώρακος.

Η σοβαρότητα της πνευμονίας COVID-19 δεν μπορεί να προσδιοριστεί από ένα θετικό ρινοφαρυγγικό επίχρισμα SARS-CoV-2. Επομένως, είναι απαραίτητη η διεξαγωγή μιας συμπληρωματικής ακτινολογικής μελέτης. Η προγνωστική αξία των ακτινογραφικών εικόνων που πραγματοποιούνται στα αρχικά στάδια της νόσου δείχνει σημαντική συσχέτιση μεταξύ της εμπλοκής του πνευμονικού παρεγχύματος - που εκτιμάται με ποσοστό των περιοχών που επηρεάζονται από θολερότητα τύπου ground-glass (GGOs) ή πύκνωση - και της σοβαρότητας της νόσου.

Οι πιο συνηθισμένες εκδηλώσεις που παρατηρούνται στις ακτινογραφίες θώρακος των ασθενών με COVID-19 είναι οι περιοχές αδιαφάνειας τύπου "ground-glass" (GGOs), συχνά συνοδευόμενες από δικτυωτές αδιαφάνειες και πνευμονική πύκνωση. Αυτές, όπως και σε άλλες ιογενείς πνευμονίες, είναι συνήθως πολυλοβιακές και διμερείς, επηρεάζοντας κυρίως τους κάτω λοβούς. Ένα από τα πιο χαρακτηριστικά σημάδια της πνευμονίας COVID-19 είναι η περιφερειακή και πολυεστιακή κατανομή των πνευμονικών διηθημάτων.

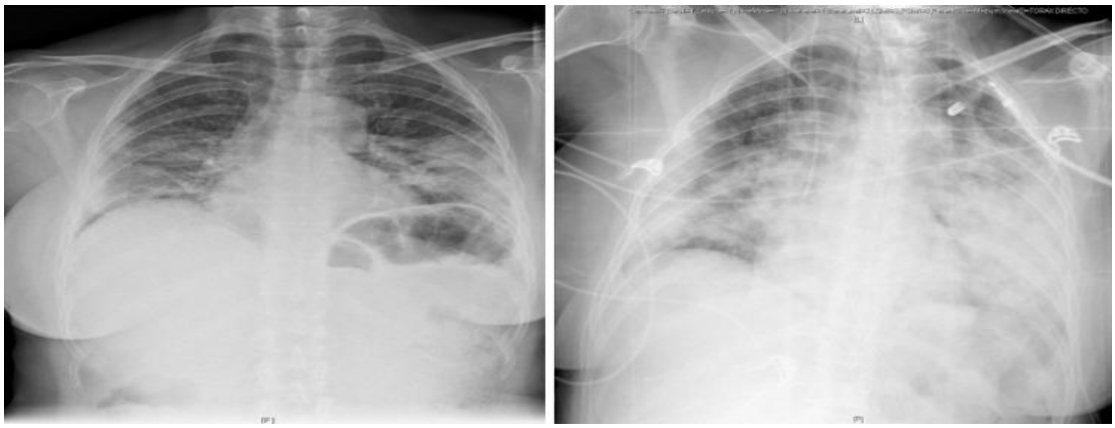
Οι ακτινολογικές αλλοιώσεις τείνουν να εξελίσσονται γρήγορα, σχηματίζοντας ένα μοτίβο, με τη μέγιστη σοβαρότητα και εμπλοκή του πνευμονικού παρεγχύματος να εμφανίζονται συνήθως μεταξύ της 6ης και 12ης ημέρας από την έναρξη των συμπτωμάτων. Η πλευριτική συλλογή είναι εξαιρετικά σπάνια σε ασθενείς με SARS-CoV-2, αλλά αν εμφανιστεί, συνήθως αφορά τα τελικά στάδια της νόσου. Η παρουσία πνευμονικών κοιλοτήτων ή πνευμοθώρακα είναι επίσης ασυνήθιστη, αλλά μπορεί να παρατηρηθεί σε ορισμένες περιπτώσεις COVID-19.[13].



Εικόνα 2.1 [13]

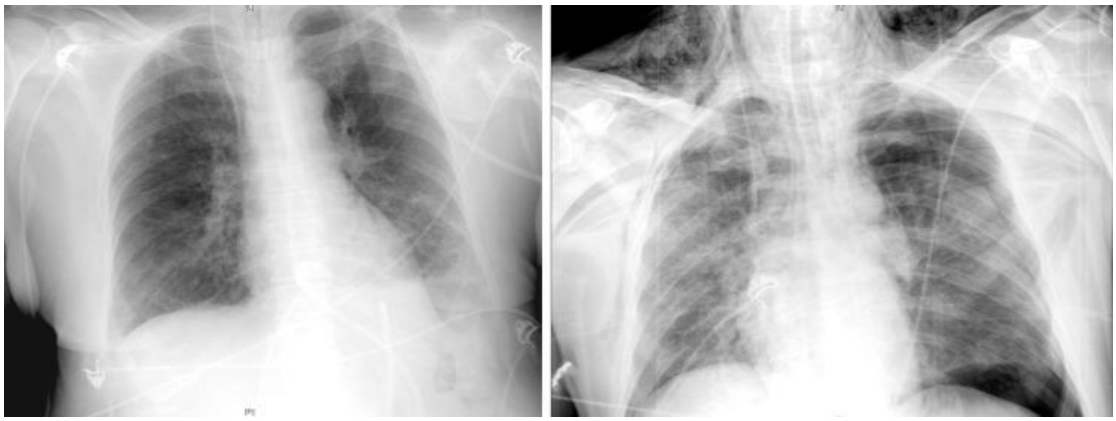
Στην εικόνα 2.1 παρουσιάζονται ευρήματα ακτινογραφίας θώρακος σε γυναίκα 60 ετών με επιβεβαιωμένη πνευμονία από τον ιό SARS-CoV-2 (θετικό τεστ RT-PCR). Ακτινογραφία θώρακος PA (αριστερά) με διάσπαρτες αδιαφάνειες στο μεσαίο και

κάτω δεξί πνεύμονα και στον κάτω αριστερό πνεύμονα. Ακτινογραφία θώρακος AP (δεξιά) με περιφερειακά διατεταγμένες διμερείς πνευμονικές αδιαφάνειες



Εικόνα 2.2 [13]

Στην εικόνα 2.2 παρουσιάζονται ευρήματα ακτινογραφίας θώρακος PA σε γυναίκα 55 ετών με διάφορους βαθμούς πνευμονίας από τη νόσο COVID-19, οριζόμενη από διάχυτες αδιαφάνειες τύπου ground-glass και πύκνωσης, που αφορούν κυρίως την κάτω ζώνη και των δύο πνευμόνων.



Εικόνα 2.3 [13]

Στην εικόνα 2.3 παρουσιάζονται ευρήματα ακτινογραφίας θώρακος AP. Ευρήματα ακτινογραφίας θώρακος AP (αριστερά) σε άνδρα 80 ετών με διμερή πνευμονία COVID-19 και συνδεόμενη αριστερή πλευριτική συλλογή. Ευρήματα ακτινογραφίας θώρακος AP (δεξιά) σε άνδρα 84 ετών με διμερείς κυψελιδικές διηθήσεις, διάχυτα διατεταγμένες και αριστερό πνευμοθώρακα υπό τάση με υποδόριο εμφύσημα.

Τα δημοσιευμένα δεδομένα υποδηλώνουν ότι η ακτινογραφία θώρακος έχει υψηλή χρησιμότητα σε ασθενείς με μόλυνση SARS-CoV-2, ειδικά σε εκείνους με μέτρια έως σοβαρή πνευμονική εμπλοκή και στα προχωρημένα στάδια της νόσου. Επιπλέον, μπορεί να χρησιμεύσει ως εργαλείο πρώτης γραμμής όταν οι πόροι είναι περιορισμένοι, παίζοντας βασικό ρόλο στην παρακολούθηση των ασθενών και στην αξιολόγηση των ενδεχόμενων συνδεόμενων επιπλοκών[13].

3. Παραγωγή/Εξαγωγή Χαρακτηριστικών στην Ιατρική Απεικόνιση

Η παραγωγή χαρακτηριστικών, γνωστή και ως εξαγωγή χαρακτηριστικών, είναι μια θεμελιώδης διαδικασία στην ανάλυση δεδομένων ιατρικής απεικόνισης. Περιλαμβάνει τη μετατροπή των ακατέργαστων δεδομένων απεικόνισης σε ένα σύνολο μετρήσιμων και ποσοτικών χαρακτηριστικών που αποτυπώνουν τα ουσιώδη χαρακτηριστικά της περιοχής ενδιαφέροντος (ROI) μέσα στην εικόνα. Αυτά τα χαρακτηριστικά μπορούν να παρέχουν βαθύτερες γνώσεις για την παθολογία, να ενισχύσουν την ακρίβεια της διάγνωσης και να βοηθήσουν στην πρόβλεψη των αποτελεσμάτων της θεραπείας.

Η διαδικασία εξαγωγής χαρακτηριστικών ξεκινά με την απόκτηση υψηλής ποιότητας ιατρικών εικόνων, οι οποίες συνήθως λαμβάνονται μέσω μεθόδων όπως η μαγνητική τομογραφία (MRI), η αξονική τομογραφία (CT) ή η τομογραφία εκπομπής ποζιτρονίων (PET). Ακολουθώντας την απόκτηση εικόνων, εφαρμόζονται προ-επεξεργαστικά βήματα για την τυποποίηση των εικόνων και τη βελτίωση της ποιότητάς τους. Αυτό μπορεί να περιλαμβάνει μείωση θορύβου, κανονικοποίηση και ευθυγράμμιση των εικόνων.

Στη συνέχεια, η περιοχή ενδιαφέροντος, όπως ένας όγκος ή μια αλλοίωση, εντοπίζεται και διαχωρίζεται από τον περιβάλλοντα ιστό. Αυτός ο διαχωρισμός μπορεί να πραγματοποιηθεί χειροκίνητα από ακτινολόγους ή αυτόματα χρησιμοποιώντας προηγμένους αλγόριθμους επεξεργασίας εικόνας. Μόλις απομονωθεί η περιοχή ενδιαφέροντος, εξάγεται μια ποικιλία χαρακτηριστικών. Αυτά τα χαρακτηριστικά μπορούν να κατηγοριοποιηθούν σε χαρακτηριστικά βασισμένα στο σχήμα, την ένταση και την υφή.

- **Χαρακτηριστικά βασισμένα στο σχήμα** περιγράφουν τις γεωμετρικές ιδιότητες της περιοχής ενδιαφέροντος, όπως το μέγεθος, την περίμετρο, τον όγκο και την επιφάνεια.
- **Χαρακτηριστικά βασισμένα στην ένταση** καταγράφουν την κατανομή των εντάσεων pixel ή voxel μέσα στην περιοχή ενδιαφέροντος, αντανακλώντας τα υποκείμενα χαρακτηριστικά του ιστού.
- **Χαρακτηριστικά βασισμένα στην υφή** αναλύουν τα πρότυπα και τις χωρικές διατάξεις των εντάσεων pixel, παρέχοντας πληροφορίες για την ετερογένεια και την δομική πολυπλοκότητα του ιστού.

Αναλύοντας αυτά τα εξαγόμενα χαρακτηριστικά, οι κλινικοί και οι ερευνητές μπορούν να ανακαλύψουν πρότυπα και συσχετίσεις που μπορεί να μην είναι εμφανείς με οπτική επιθεώρηση. Αυτά τα χαρακτηριστικά μπορούν να χρησιμοποιηθούν για την ανάπτυξη προγνωστικών μοντέλων για τη διάγνωση της νόσου, την πρόγνωση και την ανταπόκριση στη θεραπεία, συμβάλλοντας τελικά στην πιο εξατομικευμένη και αποτελεσματική φροντίδα των ασθενών[14,15].

3.1. Χαρακτηριστικά πρώτης τάξης

Αυτά τα χαρακτηριστικά αναφέρονται στα στατιστικά δεδομένα των αποχρώσεων του γκρι μιας εικόνας. Ορισμένα από τα βασικά χαρακτηριστικά είναι:

1. **Μέση τιμή**

$$\text{mean} = \frac{\sum_i \sum_j g(i,j)}{N}$$

Όπου το g(i,j) είναι η απόχρωση του γκρι στο σημείο (i,j) και το N είναι ο αριθμός των εικονοστοιχείων.

2. **Διακύμανση (std²)**

$$\text{std}^2 = \frac{\sum_i \sum_j (g(i,j) - \text{mean})^2}{N-1}$$

Η διακύμανση μετρά τη μεταβλητότητα των αποχρώσεων σε σχέση με τη μέση τιμή.

3. **Ασυμμετρία (skewness)**

$$s = \frac{1}{N} \frac{\sum_i \sum_j (g(i,j) - \text{mean})^3}{\text{std}^3}$$

Η λοξότητα είναι μια στατιστική παράμετρος που καθορίζει τον βαθμό ασυμμετρίας μιας κατανομής (θετική ή αρνητική) και είναι πολύτιμη όταν τα δεδομένα έχουν ακραίες τιμές. Αν η τιμή της είναι γύρω από το μηδέν (0), σημαίνει ότι η κατανομή είναι συμμετρική. Αρνητικές τιμές (αρνητική ασυμμετρία) σημαίνουν ότι ο αριθμός των εικονοστοιχείων με χαμηλότερες τιμές γκρι στη γκρι κλίμακα υπερβαίνει τη μέση τιμή και σύρει τη μέση τιμή προς την αριστερή πλευρά της κατανομής. Αντίστοιχα, θετικές τιμές (θετική ασυμμετρία) σημαίνουν ότι ο αριθμός των εικονοστοιχείων που έχουν υψηλότερες τιμές γκρι σύρει τη μέση τιμή προς τη δεξιά πλευρά της κατανομής.

4. **Κυρτότητα (kyrtosis)**

$$k = \frac{1}{N} \frac{\sum_i \sum_j (g(i,j) - \text{mean})^4}{\text{std}^4}$$

,όπου std είναι η τυπική απόκλιση και mean είναι η μέση τιμή.

Η κυρτότητα είναι ένα χαρακτηριστικό σχήματος που περιγράφει την αιχμηρότητα της κατανομής σε αναλογία με την κανονική κατανομή και εκτιμά τη συγκρίσιμη κορυφή ή την επίπεδη επιφάνεια μιας κατανομής. Τιμές της κυρτότητας γύρω από το μηδέν (0) ορίζουν την κανονική κατανομή, που ονομάζεται Μεσοκουρτική. Η κανονική κατανομή, γενικά, είναι το πρότυπο/σημείο αναφοράς. Αρνητικές τιμές κυρτότητας περιγράφουν μια Πλατυκουρτική κατανομή, και οι θετικές τιμές αντιπροσωπεύουν μια Λεπτοκουρτική κατανομή.

3.2 Χαρακτηριστικά δεύτερης τάξης

Τα χαρακτηριστικά δεύτερης τάξης στην ανάλυση υφής εικόνας παρέχουν σημαντικές πληροφορίες σχετικά με τη δομή και τις σχέσεις μεταξύ γειτονικών εικονοστοιχείων. Το πακέτο **Pyradiomics** είναι ένα από τα πιο δημοφιλή εργαλεία για την εξαγωγή τέτοιων χαρακτηριστικών, ειδικά για εφαρμογές στην ιατρική απεικόνιση. Εκτός από τα χαρακτηριστικά που βασίζονται στη **GLCM** και τη **GLRLM**, το PyRadiomics υποστηρίζει αρκετές ακόμη μήτρες και χαρακτηριστικά δεύτερης τάξης.

Πιο κάτω δίνονται οι όροι και συμβολισμοί που θα χρησιμοποιήσουμε στις εξισώσεις για κάθε μήτρα.

1. Οριακές πιθανότητες στήλης $p_y(j)$

$$p_y(j) = \sum_{i=1}^{N_g} p(i,j)$$

2. Μέση ένταση γκρίζου επιπέδου του p_x:

$$\mu_x = \sum_{i=1}^{N_g} p_x(i) \cdot i$$

3. Μέση ένταση γκρίζου επιπέδου του p_y:

$$\mu_y = \sum_{j=1}^{N_g} p_y(j) \cdot j$$

4. Τυπική απόκλιση του p_x::

$$\sigma_x$$

5. Τυπική απόκλιση του p_y:

$$\sigma_y$$

6. Ε(έψιλον):

$$\epsilon \approx 2.2 \times 10^{-16}$$

7. Πίνακας συν-εμφάνισης P(i, j) για ένα αυθαίρετο δ και θ:

$$P(i,j)$$

8. Κανονικοποιημένος πίνακας συν-εμφάνισης p(i, j):

$$p(i,j) = \frac{P(i,j)}{\sum P(i,j)}$$

9. Αριθμός διακριτών επιπέδων έντασης στην εικόνα:

$$N_g$$

10. Οριακές πιθανότητες γραμμής $p_\chi(i)$:

$$p_x(i) = \sum_{j=1}^{N_g} p(i,j)$$

11. $p_{x+y}(k)$, όπου $i + j = k$ και $k = 2,3, \dots, 2N_g$:

$$p_{x+y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j)$$

12. $p_{x-y}(k)$, όπου $|i - j| = k$ και $k = 0,1, \dots, N_g - 1$:

$$p_{x-y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j)$$

13. Εντροπία του $p_x H_X$:

$$H_X = -\sum_{i=1}^{N_g} p_x(i) \log_2(p_x(i) + \epsilon)$$

14. Εντροπία του $p_y H_Y$:

$$H_Y = -\sum_{j=1}^{N_g} p_y(j) \log_2\big(p_y(j) + \epsilon\big)$$

15. Εντροπία του $p(i,j) H_{XY}$:

$$H_{XY} = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)\log_2(p(i,j)+\epsilon)$$

16. H_{XY1}:

$$H_{XY1} = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)\log_2\left(p_x(i)p_y(j)+\epsilon\right)$$

17. H_{XY2} :

$$H_{XY2} = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_x(i)p_y(j)\log_2\left(p_x(i)p_y(j)+\epsilon\right)$$

Παρακάτω παρατίθεται μια πλήρης λίστα των χαρακτηριστικών δεύτερης τάξης που υποστηρίζονται από το **PyRadiomics**:

3.2.1. Μήτρα Συνεμφάνισης Γκρι Επιπέδων (GLCM):

Κάθε στοιχείο του πίνακα δείχνει πόσο συχνά ζεύγη εικονοστοιχείων με συγκεκριμένες εντάσεις εμφανίζονται μαζί σε μια συγκεκριμένη απόσταση και κατεύθυνση.

1. Autocorrelation $= \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)\cdot i\cdot j$
2. Cluster Prominence $= \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \left(i+j-\mu_x-\mu_y\right)^4\cdot p(i,j)$
3. Cluster Tendency $= \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \left(i+j-\mu_x-\mu_y\right)^2\cdot p(i,j)$
4. Cluster Shade $= \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \left(i+j-\mu_x-\mu_y\right)^3\cdot p(i,j)$
5. Energy $= \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)^2$
6. Difference Entropy $= -\sum_{i=0}^{N_g-1} p_x(i)\log_2(p_x(i))$
7. Correlation $= \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)\cdot(i-\mu_x)\cdot(j-\mu_y)}{\sigma_x\cdot\sigma_y}$
8. Entropy $= -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)\log_2(p(i,j))$
9. Contrast $= \sum_{n=0}^{N_g-1} n^2\cdot\left(\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)|i-j|=n\right)$
10. IMC1 $= \frac{HXY-HXY1}{\max\{HX,HY\}}$
11. IMC2 $= \sqrt{1-\exp[-2(HXY2-HXY)]}$
12. Inverse Variance $= \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \frac{p(i,j)}{(i-j)^2}$, for $i\neq j$
13. Maximum Probability $= \max(p(i,j))$

3.2.2. Μήτρα Μήκους Εκτέλεσης Γκρι Επιπέδων (GLRLM):

Η **GLRLM** καταγράφει τη συχνότητα εμφάνισης σειρών εικονοστοιχείων με την ίδια τιμή του γκρι σε συγκεκριμένο μήκος και κατεύθυνση.

N_g είναι ο αριθμός των διακριτών επιπέδων έντασης στην εικόνα.
N_r είναι ο αριθμός των διακριτών Run Lengths στην εικόνα.
N_p είναι ο αριθμός των voxel στην εικόνα.
$N_r(\theta)$ είναι ο αριθμός των runs στην εικόνα κατά γωνία θ, ο οποίος είναι ίσος με:

$$\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} P(i,j \mid \theta)\ ,\ \text{όπου}\ 1 \le N_r(\theta) \le N_p.$$

$P(i,j \mid \theta)$ είναι ο πίνακας Run Length για μια αυθαίρετη κατεύθυνση θ.
$p(i,j \mid \theta)$ *είναι ο κανονικοποιημένος πίνακας Run Length, ο οποίος ορίζεται ως:*

$$p(i,j \mid \theta) = \frac{P(i,j \mid \theta)}{N_r(\theta)}$$

Από αυτή τη μήτρα **GLRLM** εξάγονται τα ακόλουθα χαρακτηριστικά:

1. Short Run Emphasis (SRE):
$$SRE = \frac{\sum_{i=1}^{N_g}\ \sum_{j=1}^{N_r}\ \frac{\mathbf{P}(i,j \mid \theta)}{j^2}}{N_r(\theta)}$$
2. Long Run Emphasis (LRE):
$$LRE = \frac{\sum_{i=1}^{N_g}\ \sum_{j=1}^{N_r}\ \mathbf{P}(i,j \mid \theta) j^2}{N_r(\theta)}$$
3. Gray Level Non-Uniformity (GLNU):
$$GLNU = \frac{\sum_{i=1}^{N_g}\ \left(\sum_{j=1}^{N_r}\ \mathbf{P}(i,j \mid \theta)\right)^2}{N_r(\theta)}$$
4. Run Length Non-Uniformity (RLNU):
$$RLNU = \frac{\sum_{j=1}^{N_r}\ \left(\sum_{i=1}^{N_g}\ \mathbf{P}(i,j \mid \theta)\right)^2}{N_r(\theta)}$$
5. Run Percentage (RP):
$$RP = \frac{N_r(\theta)}{N_p}$$
6. Gray Level Variance (GLV):
$$GLV = \sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j \mid \theta)(i-\mu)^2$$
, όπου $\mu = \sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j \mid \theta) i$
7. Run Variance (RV):
$$RV = \sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j \mid \theta)(j-\mu)^2$$
, όπου $\mu = \sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j \mid \theta) i$
8. Run Entropy:
$$RE = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j \mid \theta)\log_2(p(i,j \mid \theta) + \epsilon)$$

3.2.3 Μήτρα Ζωνών Ομοιομορφίας Γκρι Επιπέδων (GLSZM):

Η **GLSZM** καταγράφει τις ομοιόμορφες ζώνες σε μια εικόνα, ανεξάρτητα από την κατεύθυνση. Μια ζώνη ορίζεται ως μια συστάδα συνεχόμενων εικονοστοιχείων με την ίδια τιμή.

Σημείωση: Τα χαρακτηριστικά **GLSZM** είναι τα ίδια με αυτά της **GLRLM** με τη διαφορά ότι αντί να έχουμε RunLengths έχουμε SizeZones.

3.2.4. Μήτρα Εξάρτησης επιπέδου γκρί (GLDM):

Ένα γειτονικό voxel με επίπεδο γκρι j θεωρείται εξαρτώμενο από το κεντρικό voxel με επίπεδο γκρι i αν $|i-j|\leq\alpha$, όπου α είναι μία αυθόρμητη σταθερά. Στον πίνακα εξάρτησης επιπέδων γκρι $\mathbf{P}(i,j)$, το στοιχείο (i,j) περιγράφει πόσες φορές ένα voxel με επίπεδο γκρι i με j εξαρτώμενα voxels στη γειτονιά του εμφανίζεται στην εικόνα.
N_g είναι ο αριθμός των διακριτών επιπέδων έντασης στην εικόνα.
N_d είναι ο αριθμός των διακριτών μεγεθών εξάρτησης στην εικόνα.
N_z είναι ο αριθμός των ζωνών εξάρτησης στην εικόνα, ο οποίος είναι ίσος με:

$$\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}\mathbf{P}(i,j)$$

$\mathbf{P}(i,j)$ είναι ο πίνακας εξάρτησης.
$p(i,j)$ είναι ο κανονικοποιημένος πίνακας εξάρτησης, ο οποίος ορίζεται ως:

$$p(i,j)=\frac{P(i,j)}{N_z}$$

Από αυτή τη μήτρα **GLDM** εξάγονται τα ακόλουθα χαρακτηριστικά:

1. Small Dependence Emphasis (SDE):

$$SDE=\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}\frac{\mathbf{P}(i,j)}{i^2}}{N_z}$$

2. Large Dependence Emphasis (LDE):

$$LDE=\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}\mathbf{P}(i,j)j^2}{N_z}$$

3. Dependence Variance (DV):

$$DV=\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}p(i,j)(j-\mu)^2,\ \text{όπου}\ \mu=\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}jp(i,j)$$

4. Dependence Entropy (DE):

$$\mathrm{DE}=-\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}p(i,j)\log_2(p(i,j)+\varepsilon)$$

3.2.5. Μήτρα Διαφοράς Γειτονικών Γκρι Επιπέδων (NGTDM):

Ένας Πίνακας Διαφοράς Γειτονικών Τόνων του Γκρι ποσοτικοποιεί τη διαφορά μεταξύ μιας τιμής γκρι και της μέσης τιμής γκρι των γειτονικών της voxels (s) σε απόσταση δ.

Από αυτή τη μήτρα **NGTDM** εξάγονται τα ακόλουθα χαρακτηριστικά:

1. Coarseness $=\frac{1}{\sum_{i=1}^{N_g}p_i s_i}$

2. Busyness $= \frac{\sum_{i=1}^{N_g} p_i s_i}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} |ip_i - jp_j|}$, where $p_i \neq 0, p_j \neq 0$
3. Complexity $= \frac{1}{N_{v,p_i}} \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} |i-j| \frac{p_i s_i + p_j s_j}{p_i + p_j}$, where $p_i \neq 0, p_j \neq 0$
4. Strength $= \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (p_i + p_j)(i-j)^2}{\sum_{i=1}^{N_g} s_i}$, where $p_i \neq 0, p_j \neq 0$
5. Contrast $= \left(\frac{1}{N_{g,p}(N_{g,p}-1)} \sum_{i=1}^{N_g} \sum_j \sum_j p_i p_i p_j (i-j)^2\right) \left(\frac{1}{N_{v,p}} \sum_{i=1}^{N_g} s_i\right)$, where $p_i \neq 0, p_j \neq 0$

Ένα παράδειγμα μιας μήτρας NGTDM είναι η πιο κάτω. Στον πίνακα αυτό, οι στήλες αντιπροσωπεύουν τα εξής:

- Η πρώτη στήλη (i) αναφέρεται στις διακριτές τιμές έντασης γκρι (gray levels) που υπάρχουν στην εικόνα.
- Η δεύτερη στήλη (nini) αναφέρει τον αριθμό των voxels με την αντίστοιχη ένταση γκρι i.
- Η τρίτη στήλη (pi) αντιπροσωπεύει την κανονικοποιημένη συχνότητα εμφάνισης των voxels με τιμή έντασης i στην εικόνα. Υπολογίζεται ως:

$$p_i = \frac{n_i}{\sum n_i}$$

- Η τέταρτη στήλη (sisi) αναφέρεται στο άθροισμα των διαφορών μεταξύ κάθε τιμής έντασης ii και του μέσου όρου των γειτονικών voxels της ίδιας τιμής.

i	n_i	p_i	s_i
1	6	0.375	13.35
2	2	0.125	2.00
3	4	0.25	2.63
4	0	0.00	0.00
5	4	0.25	10.075

4. Μηχανική Μάθηση

Η πανδημία COVID-19 έχει δημιουργήσει σημαντικές προκλήσεις στον τομέα της ιατρικής διάγνωσης και θεραπείας. Ένα από τα κύρια εργαλεία για την ανίχνευση και παρακολούθηση της νόσου είναι η απεικόνιση του θώρακα μέσω ακτινογραφιών. Η χρήση της μηχανικής μάθησης στις ιατρικές εικόνες έχει αναδειχθεί ως μια καινοτόμος προσέγγιση για την αυτόματη ανάλυση και διάγνωση ασθενειών, όπως η πνευμονία που προκαλείται από τον ιό SARS-CoV-2.

Η μηχανική μάθηση, και πιο συγκεκριμένα οι αλγόριθμοι βαθιάς μάθησης, μπορούν να εκπαιδευτούν για να αναγνωρίζουν παθολογίες στις ακτινογραφίες θώρακα με μεγάλη ακρίβεια. Αυτές οι τεχνικές αξιοποιούν τεράστια σύνολα δεδομένων από εικόνες, μαθαίνοντας να διακρίνουν τα χαρακτηριστικά γνωρίσματα που σχετίζονται με την COVID-19 πνευμονία. Οι αυτοματοποιημένες λύσεις που βασίζονται στη μηχανική μάθηση έχουν τη δυνατότητα να βελτιώσουν τη διάγνωση, μειώνοντας το χρόνο και το κόστος, ενώ ταυτόχρονα παρέχουν αξιόπιστα αποτελέσματα που μπορούν να βοηθήσουν τους ιατρούς στη λήψη αποφάσεων.

Στην παρούσα μελέτη, θα εξετάσουμε την εφαρμογή συγκεκριμένων τεχνικών απεικόνισης για την ανίχνευση της COVID-19 πνευμονίας μέσω ακτινογραφιών θώρακα, αξιοποιώντας μεθόδους μηχανικής μάθησης για τη βελτίωση της ακρίβειας και της αποδοτικότητας της διάγνωσης.

4.1. Εισαγωγή στη Μηχανική Μάθηση

Η Μηχανική Μάθηση, όπως την περιέγραψε ο Arthur Samuel, είναι "η μελέτη που επιτρέπει στους υπολογιστές να μαθαίνουν χωρίς να προγραμματίζονται με ακρίβεια". Στην εργασία του Alan Turing (1950), παρουσιάστηκε ένα μοντέλο για την τεχνητή νοημοσύνη, όπου μια μηχανή θεωρείται ευφυής αν μπορεί να απαντά με τρόπο που να μην διακρίνεται από τις ανθρώπινες αντιδράσεις. Η Μηχανική Μάθηση αποτελεί κομμάτι της τεχνητής νοημοσύνης, καθώς οι υπολογιστές αποκτούν γνώσεις από δεδομένα του παρελθόντος και κάνουν προβλέψεις για το μέλλον. Η απόδοση αυτών των συστημάτων πρέπει να είναι συγκρίσιμη με αυτή ενός ανθρώπου. Ένας πιο τεχνικός ορισμός δόθηκε από τον Tom M. Mitchell το 1997: "Ένα πρόγραμμα υπολογιστή θεωρείται ότι μαθαίνει από την εμπειρία E σχετικά με μια κατηγορία εργασιών T και τις μετρήσεις απόδοσης P, αν η απόδοσή του στις εργασίες T, όπως μετριέται με το P, βελτιώνεται χάρη στην εμπειρία E". Ένα παράδειγμα φαίνεται πιο κάτω:

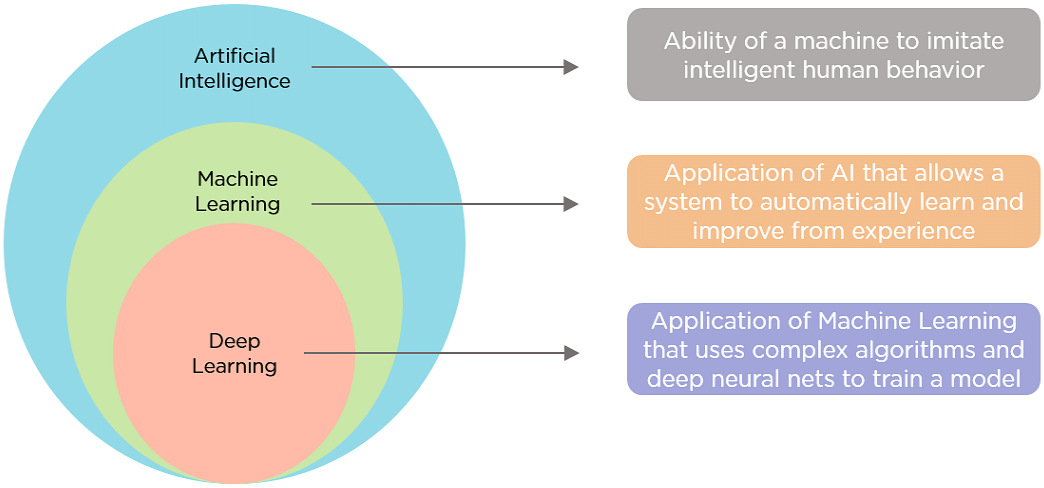
```
A handwriting recognition learning problem:

Task T: recognizing and classifying handwritten words within images
Performance measure P: percent of words correctly classified,
accuracy
Training experience E: a data-set of handwritten words with given
classifications
```

Εικόνα 4.1

Για να εκτελέσει την εργασία T, το σύστημα μαθαίνει από το παρεχόμενο σύνολο δεδομένων. Ένα σύνολο δεδομένων είναι μια συλλογή από πολλά παραδείγματα. Ένα παράδειγμα είναι μια συλλογή χαρακτηριστικών[16].

4.1.1.Διαφορές Τεχνικής νοημοσύνης, Μηχανικής μάθησης και Βαθειάς μάθησης

Οι τρεις όροι συχνά χρησιμοποιούνται εναλλακτικά, αλλά δεν αναφέρονται ακριβώς στα ίδια πράγματα. Πιο κάτω απεικονίζονται οι θεμελιώδεις διαφορές μεταξύ της τεχνητής νοημοσύνης, της μηχανικής μάθησης και της βαθιάς μάθησης.



Εικόνα 4.2: Διαφορές Τεχνικής νοημοσύνης, Μηχανικής μάθησης και Βαθειάς μάθησης [17]

Η Τεχνητή Νοημοσύνη είναι η έννοια της δημιουργίας έξυπνων ευφυών μηχανών. Η Μηχανική Μάθηση είναι ένα υποσύνολο της τεχνητής νοημοσύνης που βοηθά στην κατασκευή εφαρμογών με βάση την τεχνητή νοημοσύνη. Η Βαθιά Μάθηση είναι ένα υποσύνολο της μηχανικής μάθησης που χρησιμοποιεί τεράστιους όγκους δεδομένων και πολύπλοκους αλγόριθμους για να εκπαιδεύσει ένα μοντέλο[17]

4.1.2. Εφαρμογές Μηχανικής Μάθησης

Η μηχανική μάθηση έχει ένα ευρύ φάσμα εφαρμογών που επηρεάζουν σημαντικά διάφορους τομείς. Μερικές από τις αξιοσημείωτες εφαρμογές περιλαμβάνουν την αναγνώριση εικόνας, όπου οι μηχανές εντοπίζουν αντικείμενα ή χαρακτηριστικά μέσα σε εικόνες, την αναγνώριση ομιλίας, που επιτρέπει τη μετατροπή της προφορικής γλώσσας σε κείμενο, και τη διάγνωση ιατρικών καταστάσεων, βοηθώντας στην αναγνώριση ασθενειών από ιατρικά δεδομένα. Επιπλέον, η μηχανική μάθηση ενισχύει τα συστήματα σύστασης, που προτείνουν προϊόντα ή περιεχόμενο με βάση τη συμπεριφορά των χρηστών, και την προγνωστική συντήρηση, που προβλέπει βλάβες εξοπλισμού πριν αυτές συμβούν. Βελτιώνει επίσης τις χρηματοοικονομικές υπηρεσίες μέσω της ανίχνευσης απάτης και της διαχείρισης κινδύνων, βελτιώνει τη μεταφορά με αυτοκινούμενα αυτοκίνητα και βελτιστοποιεί το μάρκετινγκ μέσω της τμηματοποίησης και στόχευσης πελατών. Επιπλέον, η επεξεργασία φυσικής γλώσσας επιτρέπει στις μηχανές να κατανοούν και να δημιουργούν ανθρώπινη γλώσσα, ενώ τα αυτόνομα ρομπότ και οι προσωπικοί βοηθοί, όπως η Siri ή η Alexa, αναδεικνύουν τις δυνατότητές της στη ρομποτική και τις καθημερινές ανθρώπινες αλληλεπιδράσεις[18].

4.2 Είδη Μηχανικής Μάθησης

Η Μηχανική Μάθηση χωρίζεται συνήθως σε τρεις κύριες κατηγορίες: Εποπτευόμενη Μάθηση, Μη Εποπτευόμενη Μάθηση και Ενισχυτική Μάθηση.

1. Εποπτευόμενη Μάθηση

Στην εποπτευόμενη μάθηση, το σύστημα εκπαιδεύεται με δεδομένα που συνοδεύονται από ετικέτες ή στόχους, επιτρέποντας στον αλγόριθμο να συνδέει τα χαρακτηριστικά με τις σωστές εξόδους. Ο στόχος είναι να μάθει το μοντέλο να κάνει προβλέψεις για νέα δεδομένα. Οι δύο πιο διαδεδομένες μορφές εποπτευόμενης μάθησης είναι η ταξινόμηση(classification) και η παλινδρόμηση(regression).

- Στην ταξινόμηση, η μηχανή εκπαιδεύεται να προβλέπει διακριτές κατηγορίες. Για παράδειγμα, μπορεί να προβλέψει αν η αξία μιας μετοχής θα αυξηθεί ή θα μειωθεί, ή αν ένα άρθρο ανήκει στην κατηγορία της πολιτικής ή της ψυχαγωγίας.

- Στην παλινδρόμηση, η μηχανή προβλέπει μια συνεχή τιμή, όπως οι πωλήσεις ενός προϊόντος ή το μισθό για μια συγκεκριμένη θέση εργασίας, βασισμένη σε περιγραφικά χαρακτηριστικά.

2. Μη Εποπτευόμενη Μάθηση

Στη μη εποπτευόμενη μάθηση, το σύστημα μαθαίνει από δεδομένα που δεν φέρουν ετικέτες ή προκαθορισμένες κατηγορίες. Ο στόχος είναι να εντοπιστούν μοτίβα ή σχέσεις μεταξύ των δεδομένων. Ένα συνηθισμένο παράδειγμα μη εποπτευόμενης μάθησης είναι η **ομαδοποίηση**, όπου παρόμοια παραδείγματα ομαδοποιούνται με βάση κοινά χαρακτηριστικά, χωρίς να υπάρχει προκαθορισμένος στόχος.

3. Ενισχυτική Μάθηση

Η ενισχυτική μάθηση βασίζεται στην έννοια της ανταμοιβής και της τιμωρίας, καθώς ένας πράκτορας αλληλεπιδρά με ένα περιβάλλον για να επιτύχει έναν στόχο. Ο πράκτορας μαθαίνει ποιες ενέργειες οδηγούν στη μέγιστη απόδοση μέσω πολλών βημάτων, λαμβάνοντας θετικά ή αρνητικά σήματα ανάδρασης (σήματα ενίσχυσης). Για παράδειγμα, ένας αλγόριθμος μπορεί να προσπαθεί να μεγιστοποιήσει τους πόντους σε ένα παιχνίδι, κάνοντας τις καλύτερες δυνατές κινήσεις.[16].

4.3 Μεθόδοι μείωσης/επιλογής χαρακτηριστικών

4.3.1 PCA

Η ανάλυση κύριων συνιστωσών (PCA) είναι μια τεχνική μείωσης διαστάσεων που αποσκοπεί στη συμπύκνωση μεγάλων συνόλων δεδομένων, διατηρώντας τις βασικές πληροφορίες. Αυτή η μέθοδος λειτουργεί μέσω πέντε βασικών βημάτων:

1. Κανονικοποίηση των αρχικών μεταβλητών ώστε να έχουν συγκρίσιμα μεγέθη.

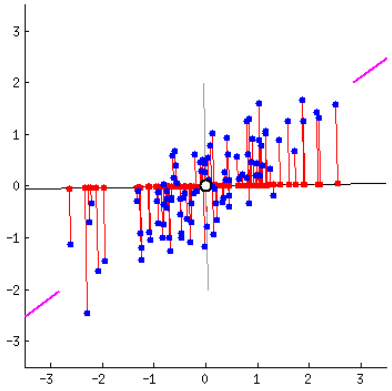
2. Υπολογισμός του πίνακα συνδιακύμανσης για την κατανόηση των συσχετίσεων μεταξύ των μεταβλητών.

3. Εύρεση των ιδιοτιμών και ιδιοδιανυσμάτων του πίνακα συνδιακύμανσης για τον εντοπισμό των κύριων συνιστωσών.

4. Επιλογή των σημαντικότερων κύριων συνιστωσών με βάση την αντίστοιχη διακύμανσή τους.

5. Μετασχηματισμός των δεδομένων σύμφωνα με τις κύριες συνιστώσες.

Το PCA μετατρέπει τις αρχικές μεταβλητές σε νέες μεταβλητές (κύριες συνιστώσες), οι οποίες είναι ασυσχέτιστες και περιέχουν τη μέγιστη δυνατή πληροφορία. Η πρώτη κύρια συνιστώσα εξηγεί το μεγαλύτερο μέρος της διακύμανσης, ενώ η κάθε επόμενη προσθέτει την υπόλοιπη πληροφορία.



Εικόνα 4.3 [19]

Στην εικόνα 4.3 το διάγραμμα διασποράς του συνόλου δεδομένων δείχνει την πρώτη κύρια συνιστώσα, είναι περίπου η γραμμή που ταιριάζει με τα μοβ σημάδια επειδή περνά από την αρχή των αξόνων και είναι η γραμμή στην οποία η προβολή των σημείων (κόκκινες τελείες) είναι η πιο εξαπλωμένη. Είναι η γραμμή που μεγιστοποιεί τη διακύμανση (τον μέσο όρο των τετραγωνικών αποστάσεων από τα προβεβλημένα σημεία (κόκκινες τελείες) μέχρι την αρχή των αξόνων)

Αν και η μείωση διαστάσεων συνεπάγεται απώλεια ακρίβειας, η PCA προσφέρει απλοποίηση των δεδομένων, καθιστώντας την ανάλυση πιο εύκολη και ταχύτερη. Γεωμετρικά, οι κύριες συνιστώσες είναι οι άξονες που εξηγούν τη μεγαλύτερη διακύμανση στα δεδομένα, γεγονός που αυξάνει την πληροφορία που περιέχεται στις πρώτες κύριες συνιστώσες.[19].

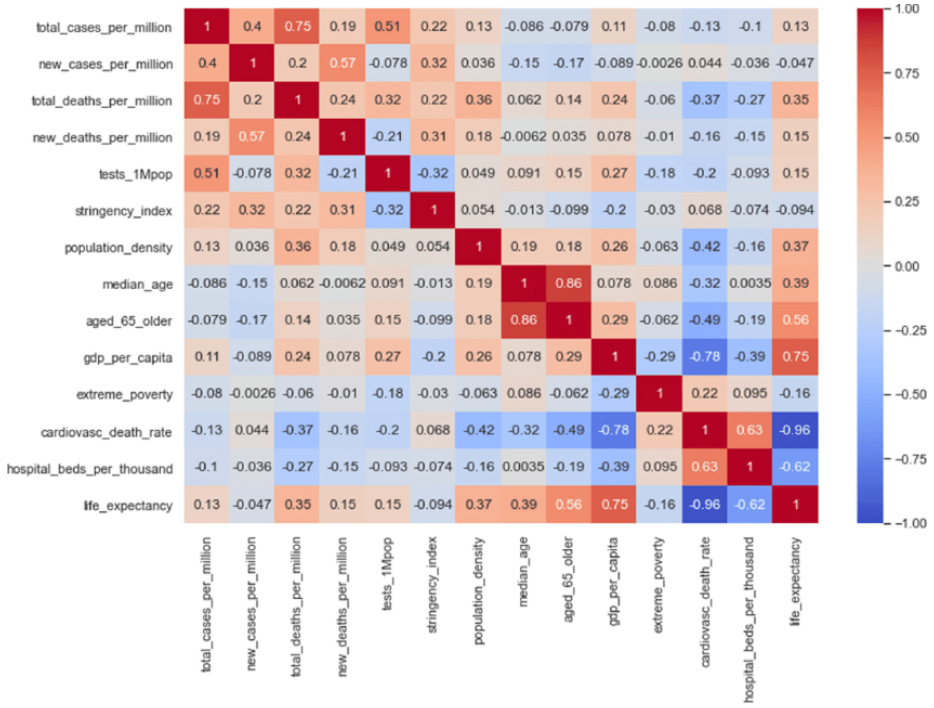
4.3.2 RFE

Ο RFE (Recursive Feature Elimination) είναι ένας αλγόριθμος επιλογής χαρακτηριστικών τύπου περιτυλίγματος (wrapper-type). Αυτό σημαίνει ότι ενσωματώνει έναν αλγόριθμο μηχανικής μάθησης ως πυρήνα του, ο οποίος χρησιμοποιείται για να καθοδηγήσει τη διαδικασία επιλογής χαρακτηριστικών. Αυτό διαφέρει από τις μεθόδους επιλογής φίλτρου (filter-based), όπου κάθε χαρακτηριστικό αξιολογείται ανεξάρτητα και επιλέγονται εκείνα με τις καλύτερες βαθμολογίες.

Το RFE ξεκινά με όλα τα χαρακτηριστικά του συνόλου δεδομένων και αφαιρεί σταδιακά τα λιγότερο σημαντικά, μέχρι να παραμείνει ο επιθυμητός αριθμός χαρακτηριστικών. Η διαδικασία αυτή περιλαμβάνει την εκπαίδευση του αλγορίθμου μηχανικής μάθησης για την κατάταξη των χαρακτηριστικών, την απομάκρυνση των λιγότερο σημαντικών και την επανεκπαίδευση του μοντέλου. Αυτό επαναλαμβάνεται έως ότου φτάσει στον καθορισμένο αριθμό επιλεγμένων χαρακτηριστικών.[20].

4.3.3 Συντελεστής Συσχέτισης

Η συσχέτιση είναι ένα μέτρο της γραμμικής σχέσης μεταξύ δύο ή περισσότερων μεταβλητών. Μέσω της συσχέτισης, μπορούμε να προβλέψουμε μία μεταβλητή από την άλλη. Η λογική πίσω από τη χρήση της συσχέτισης για την επιλογή χαρακτηριστικών είναι ότι οι καλές μεταβλητές έχουν υψηλή συσχέτιση με τον στόχο. Επιπλέον, οι μεταβλητές θα πρέπει να έχουν συσχέτιση με τον στόχο, αλλά να είναι ασυσχέτιστες μεταξύ τους[21].



Εικόνα 4.4 [22]

Η εικόνα 4.4 απεικονίζει τη συσχέτιση μεταξύ χαρακτηριστικών με τη μορφή χάρτη θερμοκρασίας. Κάθε κελί του πίνακα δείχνει τον συσχετισμό ανάμεσα σε δύο χαρακτηριστικά, κυμαινόμενο από -1 (αρνητική σχέση) έως 1 (θετική σχέση). Τα μπλε και τα κόκκινα χρώματα υποδηλώνουν αρνητικούς και θετικούς συσχετισμούς αντίστοιχα, ενώ τα ανοιχτόχρωμα κελιά αντιπροσωπεύουν συσχετίσεις κοντά στο μηδέν. Η διαγώνιος του πίνακα, συνήθως με τιμές 1, αναπαριστά τον συσχετισμό χαρακτηριστικού με τον εαυτό του

4.4 Βασικοί Αλγόριθμοι Μηχανικής Μάθησης

4.4.1 Ελάχιστης απόστασης

Ο Ταξινομιτής ελάχιστης απόστασης (Minimum Distance Classifier, MDC) είναι μία από τις πιο απλές μεθόδους ταξινόμησης. Βασίζεται στη μέτρηση της απόστασης μεταξύ ενός δείγματος και των κεντρικών σημείων (μέσων τιμών) των διαφορετικών κλάσεων. Κάθε νέο δείγμα ταξινομείται στην κλάση της οποίας το κεντρικό σημείο βρίσκεται πιο κοντά, συνήθως με βάση την Ευκλείδεια απόσταση. Ο MDC έχει το πλεονέκτημα της χαμηλής υπολογιστικής πολυπλοκότητας, αλλά μπορεί να μην είναι τόσο ακριβής σε δεδομένα όπου οι κλάσεις δεν διαχωρίζονται γραμμικά ή παρουσιάζουν μεγάλες επικαλύψεις. Η ακρίβεια του εξαρτάται σε μεγάλο βαθμό από τη σωστή επιλογή χαρακτηριστικών που διαχωρίζουν επαρκώς τις κλάσεις.

4.4.2 Κ-πλησιέστεροι γείτονες – K nearest Neighbours (KNN)

Ο αλγόριθμος k-Nearest Neighbors (k-πλησιέστεροι γείτονες, KNN) είναι μια απλή αλλά ισχυρή μέθοδος μη παραμετρικής ταξινόμησης. Η βασική ιδέα του KNN είναι να ταξινομεί ένα δείγμα βάσει των κλάσεων των k πλησιέστερων γειτόνων του στο χαρακτηριστικό χώρο, που καθορίζεται συνήθως από την Ευκλείδεια απόσταση. Το k είναι μια παράμετρος που ορίζει πόσοι γείτονες θα ληφθούν υπόψη και μπορεί να επιλεγεί μέσω διασταύρωσης επικύρωσης. Ο KNN είναι ιδανικός για δεδομένα όπου οι κλάσεις είναι διαχωρισμένες τοπικά, αλλά μπορεί να υποφέρει από υψηλό υπολογιστικό κόστος για μεγάλα δεδομένα, καθώς απαιτεί τον υπολογισμό της απόστασης για κάθε δείγμα στο σύνολο εκπαίδευσης[23].

4.4.3 Bayesian

Ο Bayesian classifier (Μπεϋζιανός ταξινομητής) είναι μια στατιστική μέθοδος ταξινόμησης που βασίζεται στο θεώρημα του Bayes. Η βασική αρχή του ταξινομητή αυτού είναι η πιθανότητα. Για ένα νέο δείγμα, υπολογίζεται η πιθανότητα να ανήκει σε κάθε κλάση, λαμβάνοντας υπόψη τις παρατηρήσεις του και την εκ των προτέρων πιθανότητα της κλάσης (prior probability). Το μοντέλο επιλέγει την κλάση με την υψηλότερη πιθανότητα (posterior probability). Ένας ειδικός τύπος Bayesian classifier είναι ο Naive Bayes, ο οποίος υποθέτει ανεξαρτησία μεταξύ των χαρακτηριστικών, κάτι που καθιστά τον υπολογισμό των πιθανοτήτων απλούστερο και γρήγορο. Ο Bayesian classifier είναι αποτελεσματικός όταν οι υποθέσεις του ισχύουν, αλλά μπορεί να επηρεαστεί από την ακρίβεια των εκτιμήσεων των πιθανοτήτων σε πιο σύνθετα ή μη γραμμικά δεδομένα[24].

4.4.4 Linear Discriminant Analysis(LDA)

Η γραμμική διακριτική ανάλυση (Linear Discriminant Analysis, LDA) είναι μια μέθοδος στατιστικής ταξινόμησης που χρησιμοποιείται για την εύρεση της γραμμικής συνάρτησης που μεγιστοποιεί τον διαχωρισμό μεταξύ διαφορετικών κλάσεων. Ο στόχος του LDA είναι να μειώσει τη διάσταση του δεδομένου προβλήματος, διατηρώντας ταυτόχρονα τη διακριτική πληροφορία. Για κάθε δείγμα, το LDA προβάλλει τα δεδομένα σε έναν γραμμικό χώρο έτσι ώστε να ελαχιστοποιεί τη διασπορά εντός των κλάσεων και να μεγιστοποιεί τη διασπορά μεταξύ των κλάσεων. Είναι αποτελεσματικό όταν οι κλάσεις διαχωρίζονται γραμμικά και οι κατανομές τους είναι κανονικές (Gaussian) με ίδια διασπορά. Το LDA χρησιμοποιείται ευρέως σε εφαρμογές όπως στην αναγνώριση προσώπου και στη βιοϊατρική ταξινόμηση[25].

4.4.5 Λογιστική Παλινδρόμηση(Logistic Regressor)

Σε ορισμένα προβλήματα, η μεταβλητή απόκρισης δεν κατανέμεται κανονικά. Για παράδειγμα, μια ρίψη νομίσματος μπορεί να οδηγήσει σε δύο αποτελέσματα: κορώνα ή γράμματα. Η κατανομή Bernoulli περιγράφει την πιθανότητα κατανομής μιας τυχαίας μεταβλητής που μπορεί να πάρει την θετική περίπτωση με πιθανότητα P ή την αρνητική περίπτωση με πιθανότητα 1-P. Αν η μεταβλητή απόκρισης αντιπροσωπεύει μια πιθανότητα, πρέπει να περιορίζεται στο εύρος {0,1}.

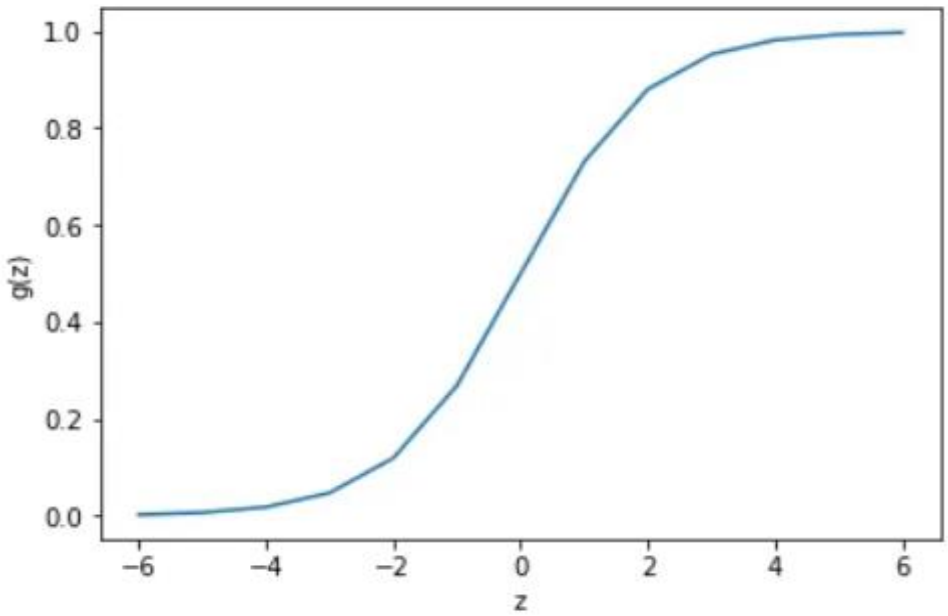
Στη λογιστική παλινδρόμηση, η μεταβλητή απόκρισης περιγράφει την πιθανότητα ότι το αποτέλεσμα είναι η θετική περίπτωση. Αν η μεταβλητή απόκρισης είναι ίση ή υπερβαίνει ένα κατώφλι διάκρισης, προβλέπεται η θετική κλάση. Διαφορετικά, προβλέπεται η αρνητική κλάση.

Η μεταβλητή απόκρισης μοντελοποιείται ως συνάρτηση ενός γραμμικού συνδυασμού των μεταβλητών εισόδου χρησιμοποιώντας τη λογιστική συνάρτηση.

Δεδομένου ότι η υπόθεση $\hat{y}$ πρέπει να ικανοποιεί $0 \leq \hat{y} \leq 1$, αυτό μπορεί να επιτευχθεί με την εισαγωγή της λογιστικής συνάρτησης ή της "Σιγμοειδούς Συνάρτησης".

$$g(z) = \frac{1}{1 + e^{-z}}$$

Η συνάρτηση g(z) χαρτογραφεί οποιονδήποτε πραγματικό αριθμό στο διάστημα (0, 1), καθιστώντας τη χρήσιμη για τη μετατροπή μιας αυθαίρετης αξίας συνάρτησης σε μια συνάρτηση που είναι καλύτερα προσαρμοσμένη για ταξινόμηση. Ακολουθεί ένα διάγραμμα της τιμής της σιγμοειδούς συνάρτησης για το εύρος {-6,6}:



Εικόνα 4.5:Γραφική αναπαράσταση Σιγμοειδούς Συνάρτησης

[16]

Τώρα, επιστρέφοντας στο πρόβλημα της λογιστικής παλινδρόμησης, ας υποθέσουμε ότι το z είναι μια γραμμική συνάρτηση μιας μοναδικής εξηγητικής μεταβλητής x. Μπορούμε τότε να εκφράσουμε το z ως εξής:

$$z = w_0 + w_1 x$$

Και η λογιστική συνάρτηση μπορεί τώρα να γραφτεί ως εξής:

$$g(x) = \frac{1}{1 + e^{-(w_0 + w_y, x)}}$$

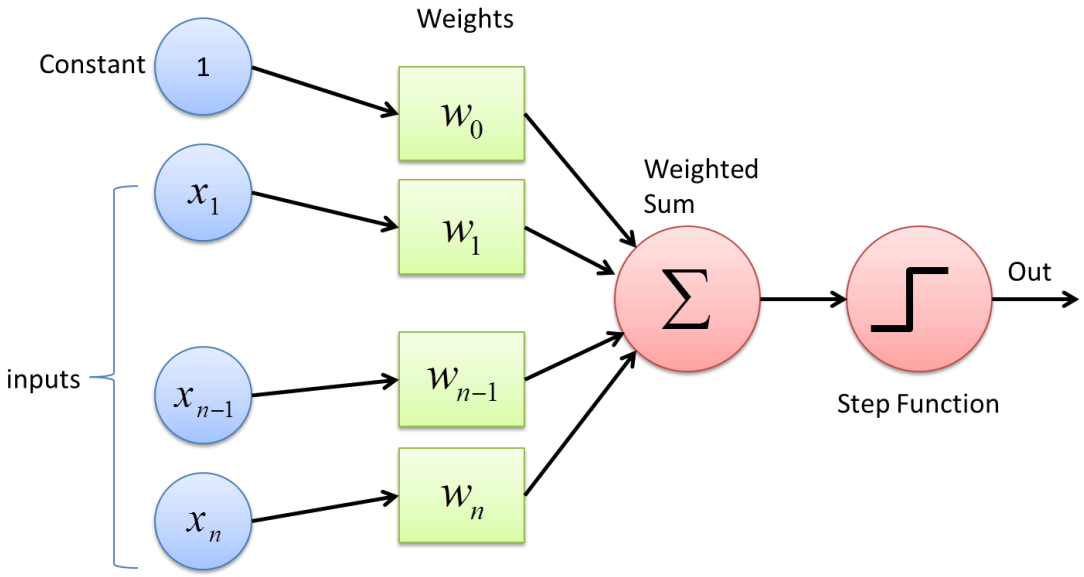
Σημειώστε ότι το g(x) ερμηνεύεται ως η πιθανότητα της εξαρτημένης μεταβλητής. g(x) = 0.7, μας δίνει μια πιθανότητα 70% ότι το αποτέλεσμα μας είναι 1. Η πιθανότητα ότι η πρόβλεψή μας είναι 0 είναι απλώς το συμπλήρωμα της πιθανότητας ότι είναι 1 (π.χ. αν η πιθανότητα ότι είναι 1 είναι 70%, τότε η πιθανότητα ότι είναι 0 είναι 30%).

Η είσοδος στη σιγμοειδή συνάρτηση 'g' δεν χρειάζεται να είναι γραμμική συνάρτηση. Μπορεί κάλλιστα να είναι κύκλος ή οποιοδήποτε σχήμα[16].

$$z = (w_0 + w_1 x_1^2 + w_2 x_2^2)$$

4.4.6 Perceptron

Ο perceptron είναι ένας από τους πρώτους αλγόριθμους μάθησης που χρησιμοποιήθηκαν στην τεχνητή νοημοσύνη, και πιο συγκεκριμένα στη μηχανική μάθηση. Είναι ένας γραμμικός ταξινομητής, ο οποίος μαθαίνει να διαχωρίζει δεδομένα σε δύο κατηγορίες χρησιμοποιώντας μια απλή γραμμική συνάρτηση απόφασης. Ο perceptron ενημερώνει το βάρος των εισόδων μέσω επαναληπτικών βημάτων μάθησης, κατά τα οποία συγκρίνει τις προβλέψεις του με τις πραγματικές ετικέτες των δεδομένων εκπαίδευσης. Αν οι προβλέψεις είναι λανθασμένες, τα βάρη προσαρμόζονται αναλόγως. Ο perceptron είναι κατάλληλος για προβλήματα γραμμικά διαχωρίσιμα, αλλά δεν μπορεί να λύσει πιο περίπλοκα, μη γραμμικά προβλήματα, τα οποία απαιτούν πιο σύνθετα μοντέλα, όπως τον Multi-Layer Perceptron (MLP)[26].



Εικόνα 4.6 Αλγόριθμος Perceptron

[27]

Τα βασικά μέρη του Perceptron είναι:

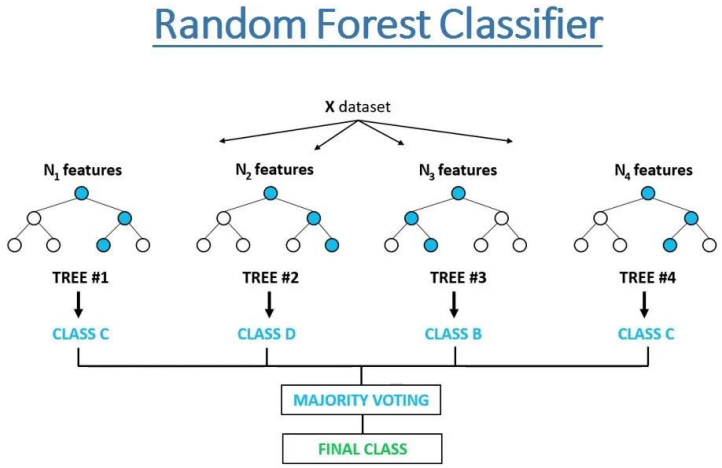
1. **Επίπεδο Εισόδου**: Το επίπεδο εισόδου αποτελείται από έναν ή περισσότερους νευρώνες εισόδου, οι οποίοι λαμβάνουν σήματα εισόδου από τον εξωτερικό κόσμο ή από άλλα επίπεδα του νευρωνικού δικτύου.
2. **Βάρη**: Κάθε νευρώνας εισόδου συνδέεται με ένα βάρος, το οποίο αντιπροσωπεύει τη δύναμη της σύνδεσης μεταξύ του νευρώνα εισόδου και του νευρώνα εξόδου.
3. **Bias**: Ένας όρος μεροληψίας προστίθεται στο επίπεδο εισόδου για να παρέχει στον perceptron επιπλέον ευελιξία στη μοντελοποίηση πολύπλοκων προτύπων στα δεδομένα εισόδου.
4. **Συνάρτηση Ενεργοποίησης**: Η συνάρτηση ενεργοποίησης καθορίζει την έξοδο του perceptron βάσει του σταθμισμένου αθροίσματος των εισόδων και του όρου μεροληψίας. Συνηθισμένες συναρτήσεις ενεργοποίησης που χρησιμοποιούνται στους perceptrons περιλαμβάνουν τη συνάρτηση σκαλοπατιού (step function), τη συνάρτηση sigmoid και τη συνάρτηση ReLU.
5. **Έξοδος**: Η έξοδος του perceptron είναι μια δυαδική τιμή, είτε 0 είτε 1, η οποία υποδεικνύει την κλάση ή την κατηγορία στην οποία ανήκουν τα δεδομένα εισόδου.
6. **Αλγόριθμος Εκπαίδευσης**: Ο perceptron εκπαιδεύεται συνήθως χρησιμοποιώντας έναν αλγόριθμο εποπτευόμενης μάθησης, όπως ο αλγόριθμος μάθησης perceptron ή η μέθοδος backpropagation. Κατά τη διάρκεια της εκπαίδευσης, τα βάρη και οι τα bias του perceptron προσαρμόζονται ώστε να ελαχιστοποιηθεί το σφάλμα μεταξύ της προβλεπόμενης εξόδου και της πραγματικής εξόδου για ένα σύνολο εκπαίδευσης[27].

4.4.2 Τυχαία Δάση(Random Forest)

Το τυχαίο δάσος είναι ένας εποπτευόμενος αλγόριθμος μηχανικής μάθησης, ο οποίος βασίζεται σε πολλαπλά δέντρα αποφάσεων. Αποτελεί μια δημοφιλή τεχνική που χρησιμοποιείται τόσο σε προβλήματα παλινδρόμησης όσο και ταξινόμησης. Ο αλγόριθμος αυτός δημιουργεί ένα "δάσος" από δέντρα αποφάσεων, το οποίο εκπαιδεύεται μέσω της μεθόδου bagging (bootstrap aggregating), μια τεχνική που βελτιώνει την ακρίβεια των αλγορίθμων συνδυάζοντας πολλές προβλέψεις.

Στην πράξη, το τυχαίο δάσος προβλέπει το αποτέλεσμα βασισμένο στις προβλέψεις που κάνουν τα επιμέρους δέντρα αποφάσεων, είτε υπολογίζοντας το μέσο όρο (για παλινδρόμηση) είτε ψηφίζοντας για την πιο συχνή πρόβλεψη (για ταξινόμηση). Όσο περισσότερα δέντρα περιλαμβάνονται, τόσο αυξάνεται η ακρίβεια των προβλέψεων.

Το τυχαίο δάσος έχει το πλεονέκτημα ότι μειώνει την υπερπροσαρμογή (overfitting), η οποία είναι συχνό πρόβλημα στα μεμονωμένα δέντρα αποφάσεων, ενώ παράλληλα αυξάνει την ακρίβεια και διαχειρίζεται αποτελεσματικά τα ελλιπή δεδομένα. Επιπλέον, μπορεί να παράγει αξιόπιστα αποτελέσματα χωρίς εκτεταμένη ρύθμιση υπερ-παραμέτρων. Σημαντικό στοιχείο είναι ότι σε κάθε σημείο διάσπασης ενός δέντρου, επιλέγεται ένα τυχαίο υποσύνολο χαρακτηριστικών, κάτι που συμβάλλει στη μείωση της υπερπροσαρμογής.

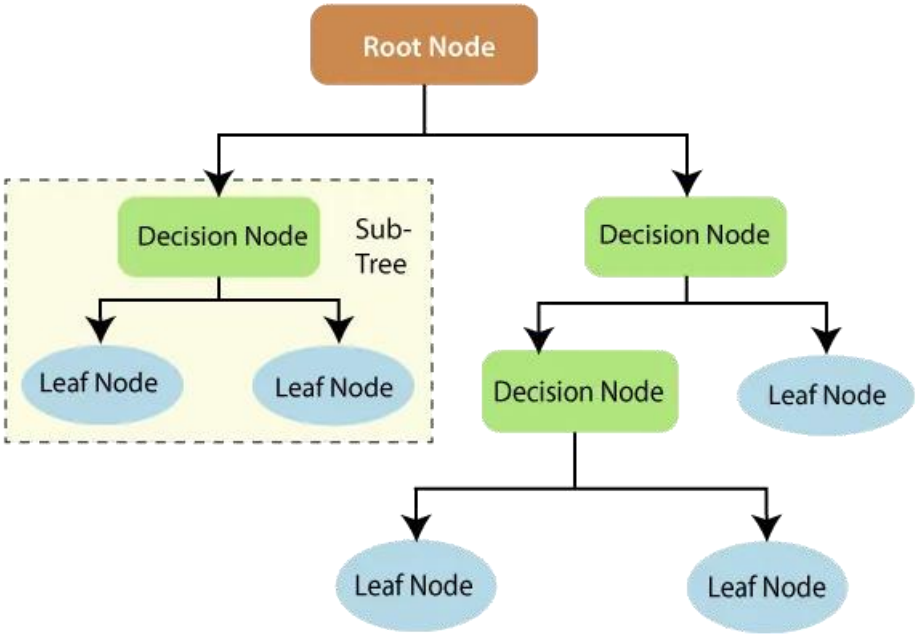


Εικόνα 4.7 Αλγόριθμος τυχαίων δασών

[28]

4.4.3 Δέντρα αποφάσεων

Ένα δέντρο αποφάσεων αποτελείται από τρία βασικά μέρη: τους κόμβους απόφασης, τα φύλλα κόμβων και τον κόμβο ρίζας. Ο αλγόριθμος δέντρου αποφάσεων διαχωρίζει το σύνολο των δεδομένων σε κλάδους, οι οποίοι συνεχίζουν να διαχωρίζονται μέχρι να φτάσουν σε κόμβους φύλλων, οι οποίοι δεν διαχωρίζονται περαιτέρω. Οι κόμβοι απόφασης αντιπροσωπεύουν τα χαρακτηριστικά που χρησιμοποιούνται για την πρόβλεψη του αποτελέσματος, ενώ οι κόμβοι φύλλων είναι τα τελικά αποτελέσματα[28].

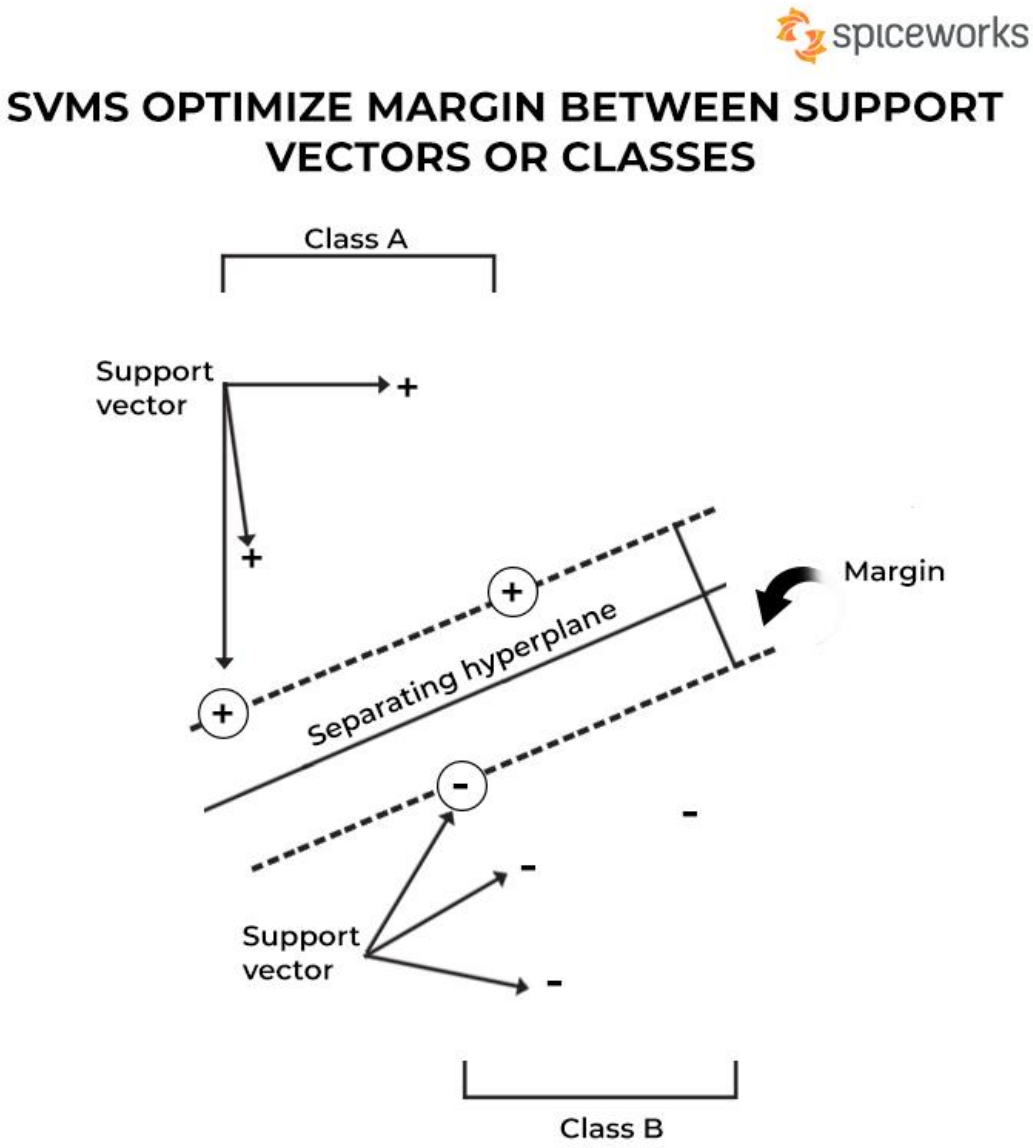


Εικόνα 4.8 Αλγόριθμος δένδρου απόφασης [28]

4.4.4 Υποστηρικτικές μηχανές διανυσμάτων (Support Vector Machines)

Η μηχανή διανυσμάτων υποστήριξης (SVM) είναι ένας αλγόριθμος μηχανικής μάθησης που χρησιμοποιεί εποπτευόμενα μοντέλα για την επίλυση προβλημάτων ταξινόμησης, παλινδρόμησης και ανίχνευσης ανωμαλιών. Η SVM επιτυγχάνει αυτό το στόχο προσδιορίζοντας βέλτιστους μετασχηματισμούς στα δεδομένα που βοηθούν στον καθορισμό ορίων μεταξύ σημείων δεδομένων, σύμφωνα με τις κατηγορίες, τις ετικέτες ή τα αποτελέσματα που έχουν οριστεί εκ των προτέρων.

Η βασική αρχή της SVM είναι να βρει ένα υπερεπίπεδο που διαχωρίζει με σαφήνεια τα σημεία δεδομένων διαφορετικών κατηγοριών. Αυτό το υπερεπίπεδο τοποθετείται με τέτοιο τρόπο ώστε να μεγιστοποιείται το περιθώριο μεταξύ των διακριτών κατηγοριών, δηλαδή, η απόσταση μεταξύ των πλησιέστερων σημείων από διαφορετικές κατηγορίες.



Εικόνα 4.9 Μηχανές Υποστήριξης Διανυσμάτων (SVMs)

[29]

Το περιθώριο σε έναν αλγόριθμο SVM αναφέρεται στο μεγαλύτερο πλάτος της ζώνης που εκτείνεται παράλληλα με το υπερεπίπεδο, χωρίς να περιλαμβάνει τα διανύσματα υποστήριξης. Αυτά τα υπερεπίπεδα είναι πιο εύκολα να καθοριστούν σε γραμμικά διαχωρίσιμα προβλήματα. Ωστόσο, σε πιο σύνθετα ή πραγματικά σενάρια, ο αλγόριθμος SVM προσπαθεί να μεγιστοποιήσει το περιθώριο, με αποτέλεσμα να γίνονται μερικές φορές λανθασμένες ταξινομήσεις για μεμονωμένα σημεία δεδομένων.

Στο μαθηματικό πλαίσιο, το SVM περιλαμβάνει μια σειρά αλγορίθμων μηχανικής μάθησης που χρησιμοποιούν συναρτήσεις πυρήνα για να μετασχηματίσουν τα χαρακτηριστικά των δεδομένων. Οι συναρτήσεις αυτές επιτρέπουν τη χαρτογράφηση των δεδομένων σε υψηλότερες διαστάσεις, διευκολύνοντας τον διαχωρισμό των σημείων δεδομένων σε πιο περίπλοκα, μη γραμμικά προβλήματα. Η χρήση αυτών των πυρήνων απλοποιεί τον διαχωρισμό των συνόλων δεδομένων, προσθέτοντας νέες διαστάσεις για την καλύτερη ανάλυση των σημείων.

Τύποι SVM

2. **Γραμμικό SVM:** Ένας γραμμικός ταξινομητής SVM χρησιμοποιείται για σύνολα δεδομένων που μπορούν να διαχωριστούν με μια ευθεία γραμμή. Αυτά τα δεδομένα ονομάζονται γραμμικά διαχωρίσιμα. Στην περίπτωση αυτή, ο αλγόριθμος βρίσκει το υπερεπίπεδο που χωρίζει τα δεδομένα με τον μέγιστο δυνατό τρόπο.
3. **Μη γραμμικό ή SVM με πυρήνα(kernel):** Για δεδομένα που δεν μπορούν να διαχωριστούν γραμμικά, χρησιμοποιείται η μέθοδος των πυρήνων. Αυτός ο τύπος SVM εφαρμόζει μη γραμμικούς μετασχηματισμούς, προσθέτοντας επιπλέον διαστάσεις για να διαχωρίσει καλύτερα τα δεδομένα. Οι μη γραμμικοί ταξινομητές SVM χρησιμοποιούν υπερεπίπεδα που επιτυγχάνουν πιο σύνθετους διαχωρισμούς[29].

4.5 Διαδικασία Ανάπτυξης Μοντέλων Μηχανικής Μάθησης

4.5.1 Συλλογή δεδομένων και προεπεξεργασία

Για τους επιστήμονες δεδομένων, το αρχικό βήμα στη διαδικασία ανάπτυξης της μηχανικής μάθησης περιλαμβάνει τη συλλογή και την προετοιμασία των δεδομένων για την εκπαίδευση του μοντέλου. Συλλέγουν δεδομένα από διάφορες πηγές και ακολουθεί η κανονικοποίηση των δεδομένων. Η διαδικασία συλλογής δεδομένων είναι κρίσιμη, καθώς η ποιότητα και η ποσότητα των δεδομένων που συλλέγονται επηρεάζουν σημαντικά την επιτυχία του μοντέλου.

Κατά τη φάση προεπεξεργασίας των δεδομένων, οι επιστήμονες δεδομένων επικεντρώνονται στον εντοπισμό και τη διόρθωση των ελλιπών δεδομένων και στην αφαίρεση των άσχετων δεδομένων. Η ετικετοποίηση των δεδομένων γίνεται επίσης σε αυτό το στάδιο για να διευκολυνθεί η διαδικασία της μηχανικής μάθησης. Εργασίες καθαρισμού δεδομένων(data cleansing), όπως η αντικατάσταση λανθασμένων ή ελλιπών τιμών, η απαλοιφή επαναλαμβανόμενων τιμών(deduplication) και η επαύξηση δεδομένων(data augmentation), επίσης εκτελούνται. Παρά το χρόνο και την προσπάθεια που απαιτείται για την προετοιμασία των δεδομένων, αυτό το βήμα είναι ζωτικής σημασίας δεδομένης της εξάρτησης των μοντέλων μηχανικής μάθησης από ακριβή και ολοκληρωμένα δεδομένα[30].

4.5.2 Διαίρεση δεδομένων σε σύνολα εκπαίδευσης και δοκιμής

Η διαίρεση των δεδομένων χωρίζει ένα σύνολο δεδομένων σε τρεις κύριες υποομάδες: το σύνολο εκπαίδευσης, το σύνολο επικύρωσης και το σύνολο δοκιμών.

Το **σύνολο εκπαίδευσης (training set)** είναι ένα σύνολο δεδομένων που χρησιμοποιείται για να εκπαιδεύσει ένα μοντέλο μηχανικής μάθησης, δίνοντάς του τη δυνατότητα να μάθει τα χαρακτηριστικά και τα μοτίβα των δεδομένων. Η ποικιλία δεδομένων στο σύνολο εκπαίδευσης είναι σημαντική, ώστε το μοντέλο να εκπαιδευτεί καλά και να είναι ικανό να αντιμετωπίσει διάφορες συνθήκες που μπορεί να συναντήσει μελλοντικά.

Το **σύνολο επικύρωσης (validation set)** χρησιμοποιείται για την αξιολόγηση της απόδοσης του μοντέλου κατά τη διάρκεια της εκπαίδευσης. Είναι εργαλείο για την προσαρμογή των παραμέτρων του μοντέλου και αποτρέπει την υπερπροσαρμογή, δηλαδή την ικανότητα του μοντέλου να αποδίδει καλά στα δεδομένα εκπαίδευσης αλλά

όχι σε νέα, αόρατα δεδομένα. Το σύνολο επικύρωσης παρέχει ανατροφοδότηση στο μοντέλο μετά από κάθε εποχή εκπαίδευσης (epoch).

Το **σύνολο δοκιμών (testing set)** χρησιμοποιείται μετά την ολοκλήρωση της εκπαίδευσης για να δοκιμαστεί το μοντέλο σε νέα δεδομένα, που δεν έχει δει ποτέ πριν. Με αυτό τον τρόπο γίνεται η τελική αξιολόγηση της απόδοσης του μοντέλου.

Η διαίρεση του συνόλου δεδομένων σε **εκπαίδευση, επικύρωση και δοκιμή** βοηθά στην εξισορρόπηση μεταξύ του **bias** και της **διακύμανσης (variance)** του μοντέλου. Το σύνολο εκπαίδευσης επιτρέπει στο μοντέλο να αναγνωρίσει μοτίβα, ενώ τα σύνολα επικύρωσης και δοκιμής αξιολογούν την ικανότητά του να γενικεύει σε νέα δεδομένα.

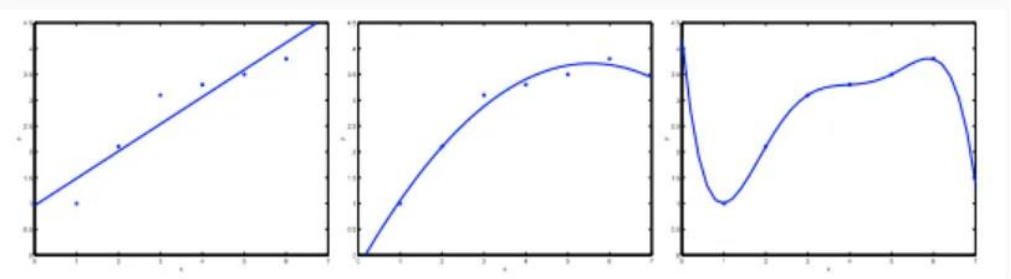
Μια πιο προηγμένη τεχνική είναι η **k-fold cross-validation**, η οποία χωρίζει το σύνολο δεδομένων σε **k**υποσύνολα. Το μοντέλο εκπαιδεύεται σε **k-1** υποσύνολα και επικυρώνεται στο ένα υπόλοιπο. Η διαδικασία αυτή επαναλαμβάνεται **k** φορές, και η απόδοση του μοντέλου εκτιμάται με βάση τον μέσο όρο όλων των επαναλήψεων, προσφέροντας μια πιο σταθερή και ακριβή αξιολόγηση της γενικότητας του μοντέλου[31].

4.5.3 Υπερπροσαρμογή(overfitting)

Η υπερπροσαρμογή (overfitting) συμβαίνει όταν ένα μοντέλο γίνεται υπερβολικά πολύπλοκο και προσαρμόζεται στενά στα δεδομένα εκπαίδευσης, ακόμη και καταγράφοντας το θόρυβο ή τα ασυνήθιστα μοτίβα. Αυτό μειώνει την ικανότητά του να γενικεύει σε νέα δεδομένα. Η εικόνα δείχνει τρία σενάρια:

1. Στην **αριστερή εικόνα**, βλέπουμε την **ανεπαρκή προσαρμογή (underfitting)**, όπου το μοντέλο είναι πολύ απλό (γραμμικό) και δεν καταφέρνει να προσαρμοστεί στα δεδομένα, οδηγώντας σε χαμηλή ακρίβεια τόσο στα δεδομένα εκπαίδευσης όσο και στα νέα δεδομένα.
2. Στη **μεσαία εικόνα**, με την προσθήκη ενός τετραγωνικού όρου (x^2), το μοντέλο βελτιώνεται και **προσαρμόζεται καλύτερα** στα δεδομένα, βρίσκοντας μια ισορροπία μεταξύ πολυπλοκότητας και απόδοσης.
3. Στην **δεξιά εικόνα**, η χρήση ενός πολυωνύμου 5ης τάξης προκαλεί **υπερπροσαρμογή**. Το μοντέλο καταφέρνει να ταιριάζει στα δεδομένα εκπαίδευσης σχεδόν τέλεια, αλλά χάνει την ικανότητά του να γενικεύει καλά για νέα, αόρατα δεδομένα.

Αυτή η αντίθεση δείχνει την ανάγκη για ένα μοντέλο που ισορροπεί μεταξύ του **bias** και της **διακύμανσης (variance)**, αποφεύγοντας τόσο την ανεπαρκή όσο και την υπερβολική προσαρμογή(overfitting) [16].



Εικόνα 4.10 Υποπροσαρμογή(underfitting), κατάλληλη προσαρμογή και υπερπροσαρμογή(overfitting)

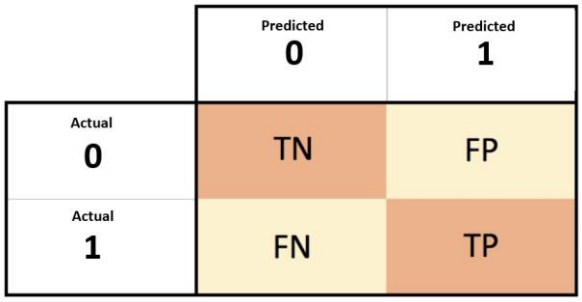
[16]

Βελτιστοποίηση Υπερπαραμέτρων του Μοντέλου: Η παρακολούθηση ενός μοντέλου περιλαμβάνει την προσαρμογή των υπερπαραμέτρων(hyperparameters) για την επίτευξη απόδοσης. Αυτή η διαδικασία απαιτεί επαναληπτικές προσαρμογές με βάση τη συμπεριφορά του μοντέλου, η οποία γίνεται με ένα ξεχωριστό σύνολο επικύρωσης.

Τυχαιοποίηση στο διαχωρισμό δεδομένων: Η τυχαιοποίηση είναι απαραίτητη στη μηχανική μάθηση, διασφαλίζοντας μη μεροληπτικά σύνολα εκπαίδευσης, επικύρωσης και δοκιμών. Η τυχαία ανακατάταξη του συνόλου δεδομένων πριν από την κατάτμηση ελαχιστοποιεί τον κίνδυνο εισαγωγής μοτίβων που σχετίζονται με τη σειρά των δεδομένων. Αυτό αποτρέπει τα μοντέλα από το να μαθαίνουν θορυβώδη δεδομένα με βάση τη διάταξη. Η τυχαιοποίηση ενισχύει την ικανότητα γενίκευσης των μοντέλων, κάνοντάς τα ανθεκτικά σε διάφορες κατανομές δεδομένων. Προστατεύει επίσης από πιθανές μεροληψίες, διασφαλίζοντας ότι κάθε υποσύνολο αντικατοπτρίζει την ποικιλομορφία που υπάρχει στο συνολικό σύνολο δεδομένων[31].

4.5.4 Μετρικές αξιολόγισης

Σημαντικό κομμάτι της μηχανικής μάθησης, αποτελούν οι μετρικές αξιολόγησης. Χρησιμοποιώντας τις μετρικές αυτές, μπορούμε να διακρίνουμε την επιτυχία που έχει κάποιο μοντέλο μηχανικής μάθησης και αν αυτό επαρκεί ως προς τις απαιτήσεις που τίθενται. Παράλληλα, καθίσταται εφικτή η σύγκριση μεταξύ των μοντέλων. Μία μέθοδος οπτικοποίησης των αποτελεσμάτων της κατηγοριοποίησης, αποτελούν τα Confusion Matrices. Σε ένα πρόβλημα δυαδικής κατηγοριοποίησης, το Confusion Matrix αποτελεί ένα πίνακα 2x2, στον οποίο φαίνονται το πλήθος των True Positive (αληθώς θετικά - TP), True Negative (αληθώς αρνητικά - TN), False Positive (ψευδώς θετικά - FP) και False Negative (ψευδώς αρνητικά – FN). Στο επόμενο σχήμα φαίνεται η δομή ενός πίνακα σύγχισης, σε ένα πρόβλημα κατηγοριοποίησης με κατηγορίες y = {0, 1}:



Εικόνα 4.11

Με βάση τα στοιχεία αυτά, μπορούμε να υπολογίσουμε κάποιες μετρικές, με τις οποίες δίνεται η δυνατότητα να καταλάβουμε την ακρίβεια της κατηγοριοποίησης.

Το Accuracy Score , είναι ο λόγος των ορθών προβλέψεων δια του συνόλου των στοιχείων:

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

Σε ένα σύνολο δεδομένων στο οποίο τα TN είναι ελάχιστα, το accuracy score πολύ πιθανόν να είναι ψηλό, αν ο αλγόριθμος κατηγοριοποιήσει όλα τα δεδομένα ως θετικά. Στην περίπτωση που αποτελεί σημαντικό σφάλμα η λάθος ταξινόμηση των αρνητικών

στοιχείων, το accuracy score δεν αντιπροσωπεύει την επιθυμητή ακρίβεια του μοντέλου.
Για το λόγο αυτό, υπολογίζονται οι μετρικές Precision και Recall. Σε ένα πρόβλημα κατηγοριοποίησης, αν έχουμε precision 1.0, σημαίνει ότι κάθε στοιχείο που κατηγοριοποιήθηκε στην κατηγορία Y όντως ανήκει σε αυτή την κατηγορία, χωρίς να λαμβάνονται υπόψιν τα στοιχεία της κατηγορίας Y που κατηγοριοποιήθηκαν λανθασμένα. Το precision, υπολογίζεται με την ακόλουθη συνάρτηση:

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

Αν έχουμε recall 1.0, σημαίνει ότι κάθε στοιχείο της κατηγορίας Y κατηγοριοποιήθηκε σωστά, χωρίς να λαμβάνονται υπόψιν το πλήθος των στοιχείων της άλλης κατηγορίας που κατηγοριοποιήθηκαν εσφαλμένα στην κατηγορία Y.

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

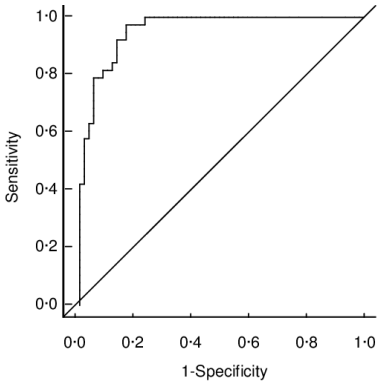
Λόγω του ότι οι δύο αυτές μετρικές αρκετές φορές έχουν αντίστροφη σχέση, συνήθως υπολογίζεται το F1-Score, το οποίο αποτελεί τον αρμονικό μέσο μεταξύ Precision και Recall, και υπολογίζεται με την συνάρτηση:

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

Σημαντική μετρική αξιολόγησης, είναι και το Receiver Operator Characteristics (ROC Curve) και το AUC (Area Under the Curve). Για την επεξήγηση του ROC και AUC, πρέπει αρχικά να αναφερθούμε στο Fall Out ή False Positive Rate (FPR) το οποίο ορίζεται ως:

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

Πριν να οριστεί η τελική κατηγοριοποίηση από τον εκάστοτε αλγόριθμο, αρχικά υπολογίζεται η πιθανότητα να ανήκει σε κάποια από τις κατηγορίες το στοιχείο προς κατηγοριοποίηση. Στη συνέχεια το αποτέλεσμα αυτό συγκρίνεται με το ορισμένο Threshold Value, και έπειτα το στοιχείο κατηγοριοποιείται ανάλογα. Κατά την δημιουργία της καμπύλης ROC, λαμβάνονται οι πιθανότητες κατηγοριοποίησης και όχι η τελική απόφαση του αλγόριθμου, και ουσιαστικά δοκιμάζονται τα αποτελέσματα με διάφορα Threshold Values. Για κάθε ένα από αυτά, υπολογίζεται το TPR (Recall) και το FPR, και έπειτα τα στοιχεία αυτά παρουσιάζονται ως μια καμπύλη με άξονες τους FPR και TPR. Ως αποτέλεσμα, γίνεται μια απεικόνιση της ακρίβειας του αλγόριθμου σε ένα ευρύ φάσμα από Threshold Values, πράγμα το οποίο θα ήταν αδύνατον με την χρήση των Confusion Matrices.
Με την δημιουργία της καμπύλης ROC, είναι στη συνέχεια εύκολο να υπολογιστεί το AUC, το οποίο αποτελεί το εμβαδόν που περικλείεται από την καμπύλη, έτσι ώστε να είναι άμεση η σύγκριση μεταξύ διαφορετικών καμπύλων ROC[32].



Εικόνα 4.12 Παράδειγμα Καμπύλης ROC

5. Μεθοδολογία

Το κεφάλαιο εμβαθύνει στην περιγραφή των δεδομένων. Εδώ, περιγράφονται λεπτομερώς οι πηγές, τα χαρακτηριστικά και οι μέθοδοι προεπεξεργασίας του συνόλου δεδομένων ακτινολογικών εικόνων. Αυτή η ενότητα υπογραμμίζει την ποικιλομορφία και την πολυπλοκότητα των δεδομένων, μαζί με τα συγκεκριμένα βήματα προεπεξεργασίας που έγιναν για την τυποποίηση και προετοιμασία των δεδομένων για αποτελεσματική εκπαίδευση μοντέλων.

Η επόμενη ενότητα για την αρχιτεκτονική του μοντέλου προσφέρει μια ολοκληρωμένη ματιά στη δομή των μοντέλων μηχανικής μάθησης. Εξηγεί το σκεπτικό πίσω από την επιλογή κάθε μοντέλου, τις αρχιτεκτονικές του αποχρώσεις και τις συγκεκριμένες διαμορφώσεις που το καθιστούν κατάλληλο για το έργο της ταξινόμησης εικόνων. Αυτό το μέρος είναι κρίσιμο για την κατανόηση των θεωρητικών θεμελίων του σχεδιασμού του μοντέλου.

Στην εκπαίδευση του μοντέλου, το κεφάλαιο προχωρά στις πρακτικές πτυχές της εφαρμογής των μοντέλων μηχανικής μάθησης. Καλύπτει τη ρύθμιση περιβάλλοντος, τις διαδικασίες εκπαίδευσης, τον συντονισμό υπερπαραμέτρων και τους αλγόριθμους που χρησιμοποιούνται. Αυτή η ενότητα είναι ιδιαίτερα σημαντική για την κατανόηση του τρόπου με τον οποίο το μοντέλο μαθαίνει από το σύνολο δεδομένων και πώς βελτιστοποιείται η απόδοσή του.

Η ενότητα της επικύρωση του μοντέλου εξετάζει τις τεχνικές και τις μετρήσεις που χρησιμοποιούνται για την επικύρωση του. Εξηγεί πώς δοκιμάστηκε και βελτιώθηκε η απόδοση του μοντέλου κατά τη διαδικασία ανάπτυξης, διασφαλίζοντας την αξιοπιστία και τη στιβαρότητά του.

Η προτελευταία ενότητα, Αξιολόγηση Μοντέλου, παρουσιάζει τα κριτήρια και τις μεθόδους που χρησιμοποιήθηκαν για την αξιολόγηση του τελικού μοντέλου. Εξηγεί τις διάφορες μετρήσεις που χρησιμοποιούνται για την αξιολόγηση της απόδοσης του μοντέλου, όπως η ακρίβεια, η ανάκληση, η βαθμολογία F1 και άλλα. Αυτή η ενότητα είναι κρίσιμη για την κατανόηση του πόσο καλά αποδίδει το μοντέλο σε πρακτικά σενάρια.

Τέλος, το κεφάλαιο πραγματεύεται τους περιορισμούς και τις προκλήσεις που συναντήθηκαν κατά τη διάρκεια της μελέτης. Συζητά με ειλικρίνεια τους περιορισμούς του μοντέλου και του συνόλου δεδομένων, καθώς και τα εμπόδια που συναντήθηκαν κατά τη διάρκεια της έρευνας. Αυτή η ενότητα παρέχει μια ειλικρινή αξιολόγηση των προκλήσεων της μελέτης, προσφέροντας μια ισορροπημένη προοπτική.

5.1. Εξοπλισμός και λογισμικό

Για την παρούσα διπλωματική εργασία χρησιμοποιήθηκε πάνω σε φορητό υπολογιστή με ενσωματωμένα Windows 22H2, με διαθέσιμη μνήμη τυχαίας προσπέλασης (RAM) 8 GB και επεξεργαστή (CPU) AMD Ryzen 3 2200U.Το περιβάλλον λογισμικού, που υποστηρίζεται από την Python και βασικές βιβλιοθήκες όπως, η NumPy, Scipy και Matplotlib, που παρείχε μια ισχυρή πλατφόρμα για την υλοποίηση μοντέλων.Το περιβάλλον όπου πραγματοποιήθηκε η εργασία είναι το Visual Studio Code.

5.2.Χαρακτηριστικά των Δεδομένων

Το σύνολο δεδομένων που στηρίζει αυτήν την εργασία αποτελείται από προεπεξεργασμένες ακτινολογικές εικόνες και έχει δημοσιευθεί στην ιστοσελίδα Kaggle. Η Radiological Society of North America (RSNA®) απευθύνθηκε στην κοινότητα μηχανικής μάθησης της Kaggle και συνεργάστηκε με τα Εθνικά Ινστιτούτα Υγείας (NIH) των ΗΠΑ, την Εταιρεία Θωρακοπνευμονικής Ακτινολογίας (STR)[33]

και την MD.ai για την ανάπτυξη ενός πλούσιου συνόλου δεδομένων για αυτή την πρόκληση[34].

Πίνακας 5.1

2018 RSNA Pneumonia Detection Challenge Dataset Description	
Imaging Modality	X-ray Preferred name: digital radiography RadLex ID: RID10351
Number of Images	30,000 frontal view chest radiographs from the 112,000-image public National Institutes of Health (NIH) CXR8 dataset ● 16,248 posteroanterior views ● 13,752 anteroposterior views ● Test: 4,527
Imaging file and structure set format	Portable Network Graphics images were converted into Digital Imaging and Communications in Medicine format, and patient sex, patient age, and projection (anteroposterior or posteroanterior) were added to the Digitial Imaging and Communications in Medicine tags
Annotation Pattern	● Whole study label ● Whole image (2D) label ● 2D ROI(s)
Annotation methodology and structure	Method of annotation ● Manual Annotation output ● Bounding boxes Annotation software ● md.ai Storage, Portability, Interoperability ● RSNA Website ZIP file
Common data elements	PDE339-Pneumonia Detection Element Details for Pneumonia Detection Name: Pneumonia Detection Definition: Detection of pneumonia Question: Pneumonia Detection ValuesValue References Enumerated (exactly 1 value): ● 0 Unknown ● 1 Pneumonia present ● 2 Pneumonia absent
Data use agreement/licensing	● Open licensing ● Non-commercial purpose ● References to dataset ● Terms
Imaging file and structure set format	DICOM - metadata/tags (based on individual task)
Image Characteristics	Resolution ● Original ● Downsampled

2018 RSNA Pneumonia Detection Challenge Dataset Description	
	Pre-processing ● Standard normalization ● Histogram normalization ● Other Burned-in PHI ● No ● Removed
Labeler demographics	● 18 radiologists from 16 different institutions, including 12 chest radiologists from the STR Specialty ● Mean of 10.6 years of experience (age range, 3-35 years) ● Scope of annotation (e.g., multi-institutional)
Reference	Shih G, et al. Augmenting the National Institutes of Health Chest Radiograph Dataset with Expert Annotations of Possible Pneumonia https://pubs.rsna.org/doi/10.1148/ryai.2019180041 [35]

Στον πίνακα 5.1 παρουσιάζονται τα αποτελέσματα της ανίχνευσης πνευμονίας με τη χρήση ακτινογραφιών θώρακα από το σύνολο δεδομένων RSNA Pneumonia Detection Challenge. Ο πίνακας περιλαμβάνει πληροφορίες για τις προβολές των ακτινογραφιών (πρόσθιο-οπίσθιες και οπίσθιο-πρόσθιες), τον αριθμό των εικόνων που χρησιμοποιήθηκαν, καθώς και τα αποτελέσματα της ανίχνευσης πνευμονίας, όπως καθορίστηκαν από τους ακτινολόγους με την εφαρμογή μεθοδολογιών ανάλυσης εικόνας και προεπεξεργασίας.

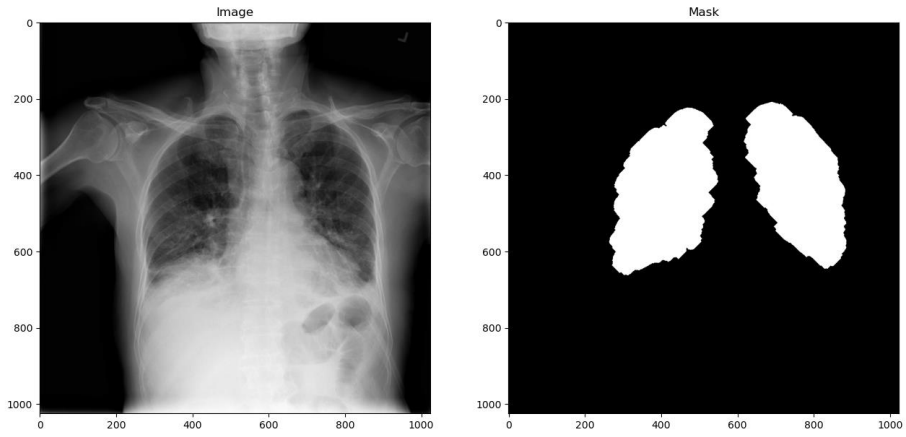
Για αυτή τη μελέτη, χρησιμοποιήσαμε δύο σύνολα δεδομένων. Το πρώτο σύνολο δεδομένων δημιουργήθηκε με αυτοματοποιημένη δημιουργία μασκών χρησιμοποιώντας τη μέθοδο κατωφλίωσης[36] για την επιλογή του βέλτιστου κατωφλίου (-320 HU) για κάθε εικόνα, προκειμένου να δημιουργηθεί η δυαδική μάσκα. Πριν από την εφαρμογή του κατωφλίου, εφαρμόσαμε τη μέθοδο ισοστάθμισης του ιστογράμματος, ώστε τα όρια των πνευμόνων να είναι πιο ευδιάκριτα, καθώς οι ασθενείς παρουσίαζαν υψηλότερα επίπεδα γκρι και περισσότερα τεχνητά αντικείμενα (καρδιακοί βηματοδότες, καθετήρες, σωλήνες θώρακος κ.λπ.). Μετά την εφαρμογή της κατωφλίωσης, χρησιμοποιήσαμε τη συνάρτηση skimage.measure.regionprops για την ανίχνευση περιοχών. Ταξινομήσαμε τις περιοχές και επιλέξαμε τις δύο μεγαλύτερες (δύο πνεύμονες) με έκταση άνω των 20000 pixels.

Στη συνέχεια, εφαρμόσαμε τη συνάρτηση clear_border για να αφαιρέσουμε περιοχές που συνδέονται με τα όρια της εικόνας, ώστε να μειωθούν τυχόν ψευδή θετικά που μπορεί να προκληθούν από τα τεχνητά αντικείμενα. Δεν χρησιμοποιήσαμε τη μέθοδο fill_holes διότι δεν θέλαμε να συμπεριλάβουμε τα τεχνητά αντικείμενα που αναφέρθηκαν προηγουμένως. Τελικά, καταλήξαμε σε 4130 μάσκες για τους ασθενείς και 8029 για τους υγιείς. Τελικά αφαιρέσαμε δεδομένα από τους υγιείς έτσι ώστε να έχουμε 4130 μάσκες σε κάθε κατηγορία. Έπειτα από κάποιες δοκιμές των μοντέλων και στατιστικές αναλύσεις t-test συμπεράναμε ότι υπάρχουν πολλά δεδομένα που απέχουν πολλές τυπικές αποκλίσεις από την μέση τιμή(outliers) τα οποία μολύνουν τα

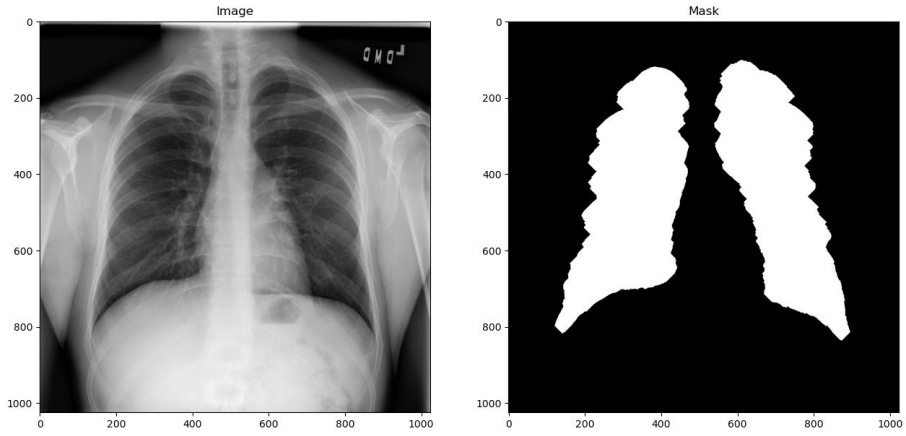
μοντέλα μας. Συνεπώς αφαιρέσαμε αυτούς τους outliers με z-score πάνω από 1.6. Τελικά καταλήξαμε με 1466 δεδομένα για τους ασθενείς και 1441για τους υγιείς.

Για το δεύτερο σύνολο δεδομένων, δημιουργήσαμε ένα γραφικό περιβάλλον διεπαφής (GUI) για τη χειροκίνητη επιλογή και αποθήκευση των μασκών. Για κάθε περίπτωση, φορτώσαμε την εικόνα και χρησιμοποιώντας το εργαλείο PolygonSelector από το matplotlib.widgets, επιλέξαμε και αποθηκεύσαμε χειροκίνητα τις μάσκες σε ξεχωριστά αρχεία για τις δύο κατηγορίες. Επειδή η διαδικασία ήταν πολύ χρονοβόρα, επιλέξαμε μόνο 650 μάσκες για κάθε κατηγορία.

Και στα δύο σύνολα δεδομένων, οι μάσκες και οι αντίστοιχες εικόνες, μαζί με το patient_id, χρησιμοποιήθηκαν ως είσοδος στο radiomics.featureextractor για την εξαγωγή των χαρακτηριστικών, τα οποία αποθηκεύτηκαν σε δύο αρχεία CSV, των ασθενών και των φυσιολογικών.



Εικόνα 5.1 Ακτινογραφία θώρακος ασθενή με πνευμονία και η αντίστοιχη μάσκα πνευμόνων



Εικόνα 5.2 Ακτινογραφία θώρακος φυσιολογικού πνεύμονα και η αντίστοιχη μάσκα πνευμόνων

5.3 Εξαγωγή χαρακτηριστικών

Αφού αποθηκεύσαμε όλες τις εικόνες και τις αντίστοιχες μάσκες τους με το μοναδικό αναγνωριστικό (ID) από τις δύο κατηγορίες σε ξεχωριστά αρχεία, προχωρήσαμε στη διαδικασία εξαγωγής των χαρακτηριστικών που περιγράφονται στο θεωρητικό υπόβαθρο. Κατά τη διαδικασία αυτή, κάθε εικόνα και η αντίστοιχη μάσκα εισάγονται στον εξαγωγέα χαρακτηριστικών της βιβλιοθήκης Pyradiomics για κάθε κατηγορία ξεχωριστά, και τα εξαγόμενα χαρακτηριστικά αποθηκεύονται σε ένα ενιαίο αρχείο CSV.

Με επιτυχία καταμετρήσαμε 89 διακριτά χαρακτηριστικά από τις εικόνες. Σημειώνεται ότι δεν χρησιμοποιήθηκαν χαρακτηριστικά σχήματος, διότι δεν παρέχουν χρήσιμη πληροφορία για την ανάλυση μας (το σχήμα του πνεύμονα είναι το ίδιο τόσο στους ασθενείς όσο και στους υγιείς). Όλες οι εικόνες που χρησιμοποιήθηκαν είναι σε κλίμακα του γκρι, γεγονός που επιτρέπει την εστίαση στην ένταση των εικονοστοιχείων και την εξαγωγή σχετικών στατιστικών.

Συγκεκριμένα, τα χαρακτηριστικά που εξήχθησαν περιλαμβάνουν:

- **Χαρακτηριστικά Πρώτης Τάξης Στατιστικών** (17): Περιγράφουν βασικές στατιστικές της κατανομής της έντασης των εικονοστοιχείων, όπως η μέση τιμή, η διασπορά, η εντροπία κ.ά.
- **Μήτρα Συνεμφάνισης Επιπέδου Γκρι (GLCM)** (21): Αποτυπώνει τη συσχέτιση μεταξύ των γειτονικών εικονοστοιχείων σε διαφορετικές κατευθύνσεις και αποστάσεις.
- **Μήτρα Μήκους Σειράς Επιπέδου Γκρι (GLRLM)** (16): Περιγράφει τη διάρκεια ή την έκταση των ομοιογενών περιοχών σε μια εικόνα.
- **Μήτρα Ζώνης Μεγέθους Επιπέδου Γκρι (GLSZM)** (16): Αξιολογεί τη συχνότητα και την κατανομή των ζωνών παρόμοιας έντασης.
- **Μήτρα Διαφοράς Γειτονικών Αποχρώσεων του Γκρι (NGTDM)** (5): Περιγράφει την ομοιογένεια των γειτονικών εικονοστοιχείων σε σχέση με τον τοπικό μέσο όρο.
- **Μήτρα Εξάρτησης Επιπέδου Γκρι (GLDM)** (14): Αποτυπώνει την εξάρτηση των εικονοστοιχείων από τους γείτονές τους, με βάση συγκεκριμένα κατώφλια έντασης.

Τα παραπάνω χαρακτηριστικά εξήχθησαν από 2160 εικόνες για κάθε κατηγορία (ασθενείς και υγιείς) και αποθηκεύτηκαν σε δύο ξεχωριστά αρχεία CSV (Comma Separated Values). Αυτή η οργανωμένη αποθήκευση επιτρέπει την εύκολη ανάλυση και την επεξεργασία των δεδομένων σε μεταγενέστερα στάδια της μελέτης. Η προσεκτική επιλογή και εξαγωγή αυτών των χαρακτηριστικών αποτελεί το θεμέλιο για την ακρίβεια των μοντέλων μηχανικής μάθησης που θα χρησιμοποιηθούν για την κατηγοριοποίηση και ανάλυση των εικόνων.

5.4 Στατιστικές Μεθόδοι - Μείωση Χαρακτηριστικών

Στη συνέχεια, σχεδιάσαμε το επιβλεπόμενο σύστημα μηχανικής μάθησης το οποίο αποτελείται από τα εξής βήματα:

1. Κανονικοποίηση δεδομένων.
2. Μείωση χαρακτηριστικών (χρησιμοποιώντας στατιστικές μεθόδους και άλλες μεθόδους μείωσης χαρακτηριστικών).
3. Στη συνέχεια, επιλέγουμε τα καλύτερα/πιο κατάλληλα χαρακτηριστικά από τη μειωμένη ομάδα χαρακτηριστικών και τα χρησιμοποιούμε.

4. Σχεδιάσαμε το σύστημα μηχανικής μάθησης (επιλογή ταξινομητή και μέθοδος αξιολόγησης) και αξιολογήσαμε την ακρίβειά του για να βρούμε τον καλύτερο συνδυασμό χαρακτηριστικών.

5.4.1 Κανονικοποίηση δεδομένων

Η κανονικοποίηση των δεδομένων είναι απαραίτητη, διότι το εύρος τιμών διαφέρει πολύ μεταξύ των χαρακτηριστικών (π.χ η μέση τιμή των τιμών έντασης των εικονοστοιχείων μιας περιοχής της εικόνας, μπορεί να έχει τιμές που κυμαίνονται από 0 έως 255, και ένα δεύτερο χαρακτηριστικό, όπως η τυπική απόκλιση των τιμών έντασης των εικονοστοιχείων, μπορεί να κυμαίνεται από 0 έως 7. Λαμβάνοντας υπόψη ότι το σύστημα σχεδιάζεται από τον συνδυασμό διάφορων χαρακτηριστικών, είναι κατανοητό ότι οι τιμές των χαρακτηριστικών πρέπει να κανονικοποιηθούν για να συμμετάσχουν και/ή να επηρεάσουν τη διαδικασία λήψης αποφάσεων του συστήματος ισότιμα.

Για την κανονικοποίηση, χρησιμοποιήσαμε τη μέθοδο preprocessing.scale από τη βιβλιοθήκη sklearn, η οποία μετασχηματίζει τα δεδομένα μας έτσι ώστε να έχουν μέση τιμή 0 και τυπική απόκλιση 1.

$$Z = \frac{x - \mu}{\sigma}$$

,όπου Z είναι η νέα τιμή του δεδομένου, χ η τιμή πριν τον μετασχηματισμό, μ η μέση τιμή και σ τυπική απόκλιση.

Η κανονικοποίηση πραγματοποιήθηκε σε όλα τα χαρακτηριστικά, διότι όλα τα δεδομένα που χρησιμοποιήθηκαν είναι συνεχή και δεν περιλαμβάνουν κατηγορικά δεδομένα. Σε περιπτώσεις όπου έχουμε κατηγορικά δεδομένα, όπως το φύλο, που μπορεί να αντιπροσωπεύεται με τιμές όπως "άνδρας" ή "γυναίκα", η κανονικοποίηση δεν είναι κατάλληλη μέθοδος επεξεργασίας. Αυτά τα δεδομένα συνήθως απαιτούν διαφορετικές προσεγγίσεις, όπως η τεχνική του one-hot encoding, η οποία μετατρέπει κατηγορικές τιμές σε δυαδικά χαρακτηριστικά. Για παράδειγμα, το χαρακτηριστικό "φύλο" μπορεί να κωδικοποιηθεί ως δύο ξεχωριστά δυαδικά χαρακτηριστικά (π.χ., "άνδρας" = 1, "γυναίκα" = 0).

Επιπλέον, ορισμένα αριθμητικά χαρακτηριστικά, όπως η "ηλικία", μπορεί να μην ωφεληθούν από την κανονικοποίηση με τον ίδιο τρόπο που ωφελούνται άλλα χαρακτηριστικά. Η ηλικία είναι ένα χαρακτηριστικό που μετράται σε ακέραιες τιμές και αντιπροσωπεύει μια διακριτή μεταβλητή, όπου οι δεκαδικές τιμές δεν έχουν ουσιαστικό νόημα. Για παράδειγμα, η κανονικοποίηση της ηλικίας μπορεί να οδηγήσει σε μη διαισθητικές τιμές που δεν αντιπροσωπεύουν πραγματικές ηλικίες.

Ωστόσο, στο πλαίσιο της μελέτης μας, όλα τα δεδομένα που χρησιμοποιήθηκαν ήταν συνεχή, προερχόμενα από χαρακτηριστικά εικόνας (όπως η ένταση των εικονοστοιχείων και η τυπική απόκλιση). Αυτά τα συνεχή δεδομένα επωφελούνται άμεσα από την κανονικοποίηση, καθώς η διαδικασία αυτή εξασφαλίζει ότι όλα τα χαρακτηριστικά συνεισφέρουν ισότιμα στη διαδικασία ανάλυσης και λήψης αποφάσεων. Δεδομένου ότι η βιβλιοθήκη pyradiomics περιέχει μόνο συνεχή δεδομένα, δεν υπήρχε ανάγκη να αποκλείσουμε κάποιο χαρακτηριστικό από την κανονικοποίηση.

Η κανονικοποίηση δεν εφαρμόστηκε στον ταξινομητή Random Forest, καθώς αυτός ο ταξινομητής δεν απαιτεί κανονικοποίηση των δεδομένων. Αυτό συμβαίνει διότι η διαδικασία λήψης αποφάσεων στον Random Forest βασίζεται σε διαδοχικά splits των δεδομένων με βάση συγκεκριμένα κατώφλια (thresholds), τα οποία καθορίζονται ανεξάρτητα από την κλίμακα των δεδομένων. Συγκεκριμένα, κάθε δέντρο στο δάσος

(forest) επιλέγει τα χαρακτηριστικά και τα αντίστοιχα κατώφλια με τέτοιο τρόπο ώστε να μεγιστοποιείται η διαχωριστική ικανότητα μεταξύ των κλάσεων. Επειδή αυτά τα κατώφλια δεν εξαρτώνται από το εύρος ή τη διασπορά των χαρακτηριστικών, η κανονικοποίηση δεν επηρεάζει την απόδοση του ταξινομητή. Ως αποτέλεσμα, η χρήση μη κανονικοποιημένων δεδομένων στον Random Forest επιτρέπει την απλοποίηση της διαδικασίας προεπεξεργασίας χωρίς να μειώνεται η ακρίβεια του μοντέλου.

Αντιθέτως, ταξινομητές όπως ο Support Vector Machine (SVM) επηρεάζονται σημαντικά από την κανονικοποίηση των δεδομένων. Αυτό συμβαίνει επειδή ο SVM χρησιμοποιεί μεθόδους όπως η ελαχιστοποίηση της απόστασης μεταξύ των διαχωριστικών υπερεπιπέδων και των δεδομένων εκπαίδευσης, οι οποίες εξαρτώνται απόλυτα από την κλίμακα των χαρακτηριστικών. Χωρίς κανονικοποίηση, χαρακτηριστικά με μεγαλύτερο εύρος τιμών θα μπορούσαν να κυριαρχήσουν στη διαδικασία εκμάθησης, οδηγώντας σε λιγότερο αποτελεσματικά μοντέλα. Έτσι, για την ορθή λειτουργία του SVM και άλλων παρόμοιων ταξινομητών, η κανονικοποίηση είναι απαραίτητη για να εξασφαλιστεί ότι όλα τα χαρακτηριστικά θα συνεισφέρουν ισότιμα στην απόφαση του μοντέλου.

5.4.2 Μείωση χαρακτηριστικών

PCA

Υπολογίσαμε τη μήτρα συνεμφάνισης των κανονικοποιημένων δεδομένων για να κατανοήσουμε τη σχέση διακύμανσης μεταξύ διαφορετικών χαρακτηριστικών. Υπολογίσαμε τα ιδιοδιανύσματα και τις ιδιοτιμές της μήτρας συνεμφάνισης χρησιμοποιώντας τη βιβλιοθήκη sklearn.decomposition. Τα ιδιοδιανύσματα αντιπροσωπεύουν την κατεύθυνση της μέγιστης διακύμανσης, ενώ οι ιδιοτιμές δείχνουν το μέγεθος της διακύμανσης σε αυτές τις κατευθύνσεις. Επιλέξαμε τις κύριες συνιστώσες που κατέγραφαν ένα σημαντικό ποσοστό 80% της συνολικής διακύμανσης. Στη συνέχεια έγινε προβολή των δεδομένων στις κύριες συνιστώσες, μειώνοντας τις διαστάσεις του χώρου. Η λογική πίσω από την επιλογή αυτής της μεθόδου της ανάλυσης κύριων συνιστωσών (PCA) είναι ότι είναι αποτελεσματική στη μείωση των διαστάσεων του χώρου χαρακτηριστικών διατηρώντας τα ουσιώδη πρότυπα και τις τάσεις στα δεδομένα. Αυτό βοηθά στην επιτάχυνση της διαδικασίας εκπαίδευσης.

RFE

Χρησιμοποιήσαμε ένα μοντέλο SVM με γραμμικό πυρήνα ως τον βασικό μοντέλο για το RFE, δεδομένης της αποτελεσματικότητάς του στη διαχείριση γραμμικών σχέσεων και την αξιολόγηση της σημαντικότητας των χαρακτηριστικών. Ξεκινώντας με όλα τα χαρακτηριστικά, το RFE εκπαίδευσε το μοντέλο και κατάταξε τα χαρακτηριστικά κατά σημαντικότητα. Το χαρακτηριστικό/χαρακτηριστικά με τη μικρότερη σημαντικότητα αφαιρέθηκαν και το μοντέλο εκπαιδεύτηκε ξανά. Αυτή η διαδικασία επαναλήφθηκε μέχρι να διατηρηθεί ένας προκαθορισμένος αριθμός χαρακτηριστικών, εξασφαλίζοντας ότι χρησιμοποιήθηκαν μόνο τα πιο σημαντικά χαρακτηριστικά για την εκπαίδευση του μοντέλου. Η λογική πίσω από την επιλογή της RFE είναι ότι συνδυάζει τα οφέλη των μεθόδων wrapper (με την αξιολόγηση της απόδοσης του μοντέλου), με τις embedded μεθόδους (χρησιμοποιώντας τη σημασία των χαρακτηριστικών που είναι ειδική για το μοντέλο). Είναι ιδιαίτερα χρήσιμο όταν το σύνολο δεδομένων μπορεί να περιέχει πλεόνασμα χαρακτηριστικών.

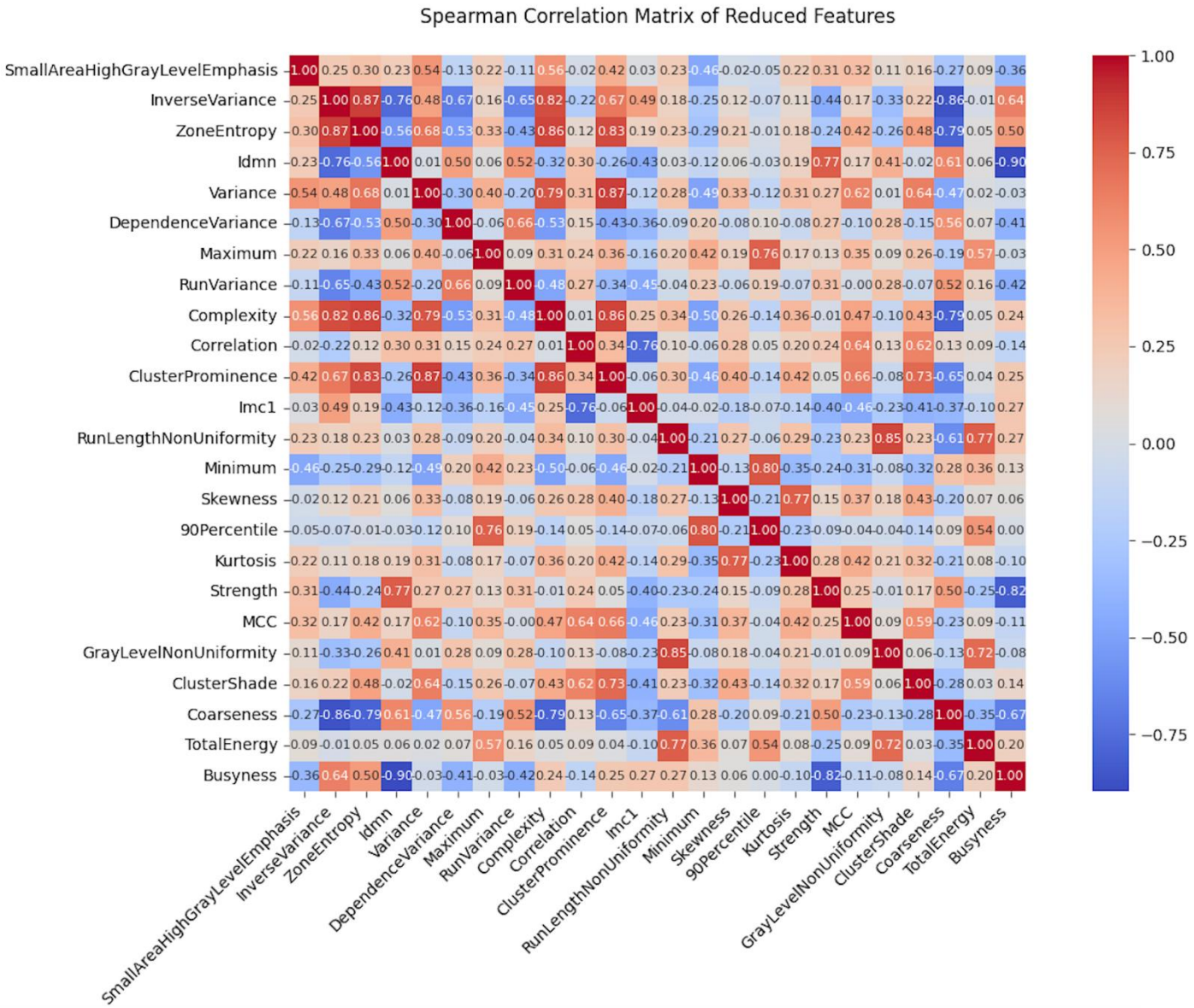
Μήτρα συσχέτισης(spearman correlation)

Υπολογίσαμε τη μήτρα συσχέτισης μεταξύ των χαρακτηριστικών. Τα χαρακτηριστικά με υψηλή συσχέτιση με άλλα χαρακτηριστικά (>0.2) απομακρύνθηκαν. Αυτό βοήθησε στο να διασφαλιστεί ότι κάθε χαρακτηριστικό προσέφερε μοναδικές πληροφορίες στο μοντέλο. Επιλέχθηκαν μόνο τα χαρακτηριστικά που έδειξαν χαμηλή αλληλοσυσχέτιση

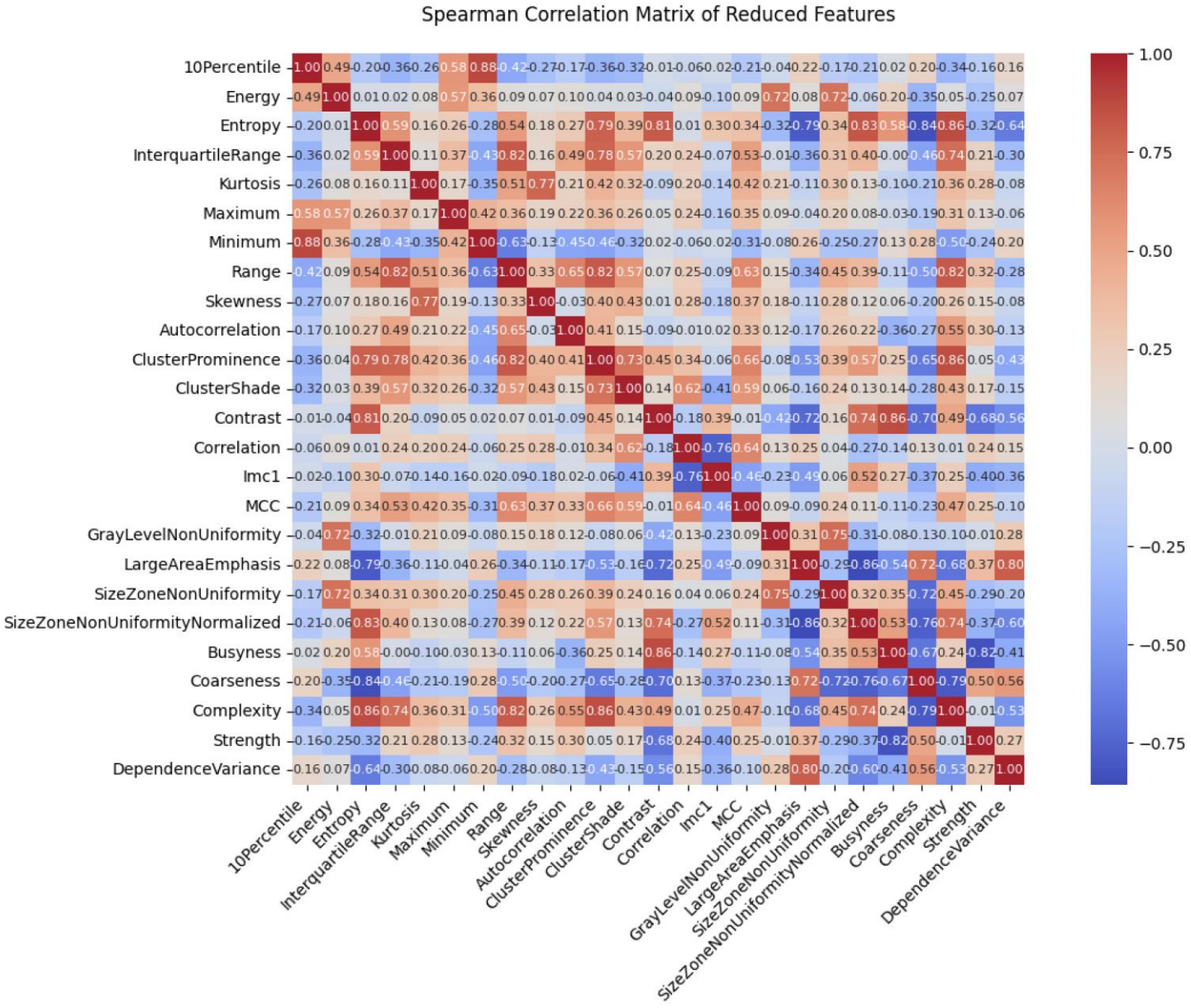
μεταξύ τους για περαιτέρω ανάλυση. Η επιλογή με βάση τη συσχέτιση είναι άμεση και αποτελεσματική στον εντοπισμό πλεοναζόντων χαρακτηριστικών που δεν προσφέρουν επιπρόσθετες πληροφορίες. Αυτό βοηθά στη βελτίωση της ερμηνευσιμότητας του μοντέλου και μειώνει τον κίνδυνο υπερπροσαρμογής.

Συνδιαστικές μεθόδοι

Δοκιμάσαμε δυνδιασμό της μεθοδου PCA και Spearman Corellation. Αρχικά φορτώνουμε τα δεδομένα από δύο ξεχωριστά αρχεία CSV, που περιέχουν τα χαρακτηριστικά για τις ομάδες ελέγχου και ασθενών αντίστοιχα. Συνδυάζουμε τα δεδομένα δημιουργώντας ένα ενιαίο σύνολο και αφαιρούμε τις πρώτες δύο στήλες(labels και patient_id). Στη συνέχεια, χρησιμοποιούμε Ανάλυση Κύριων Συνιστωσών (PCA) για να αναλύσουμε τα δεδομένα μας σε κύριες συνιστώσες που εξηγούν την περισσότερη διασπορά(variance) στα δεδομένα και εξάγουμε τα βάρη των κύριων συνιστωσών(loadings) όπως φαίνεται στους πίνακες 6.5 και 6.6. Μελετάμε τις απόλυτες τιμές των βαρών για να αξιολογήσουμε τη σημασία κάθε χαρακτηριστικού. Κατατάσσουμε τα χαρακτηριστικά σε αριθμό σημαντικότητας. Και στις δύο περιπτώσεις χρησιμοποιήθηκε Spearman Correlation. Στη μια περίπτωση δεν εφαρμόσαμε PCA. Οι πιο κάτω γραφικές απεικονίζουν τις δύο μεθόδους και τα τελικά χαρακτηριστικά.



Εικόνα 5.3 Spearman Correlation με PCA



Εικόνα 5.4 Spearman Correlation χωρίς PCA

6. RESULTS

Οι πιο κάτω πίνακες δείχνουν την μέση τιμή και τυπική απόκλιση των μέτρων αξιολόγησης για κάθε μέθοδο μείωσης χαρακτηριστικών και τον αντίστοιχο ταξινομητή.

Πίνακας 6.1

Όλα τα χαρακτηριστικά				
Classifier	**Sensitivity(%)**	**Specificity(%)**	**Accuracy(%)**	**F1 Score(%)**
MDC	76.99 ± 0.54	69.78 ± 0.57	73.42 ± 0.36	74.50 ± 0.36
KNN	85.66 ± 0.90	73.33 ± 0.81	79.55 ± 0.62	80.86 ± 0.60
Bayesian	83.53 ± 0.24	71.02 ± 0.23	77.33 ± 0.18	78.80 ± 0.18
LDA	85.59 ± 0.65	80.91 ± 0.68	83.27 ± 0.33	83.77 ± 0.33
Logistic Regressor	85.46 ± 0.52	80.42 ± 0.50	82.96 ± 0.17	83.50 ± 0.19
Perceptron	76.71 ± 2.78	76.94 ± 1.10	76.82 ± 1.46	76.93 ± 1.73
SVM(linear)	86.41 ± 0.44	80.69 ± 0.40	83.57 ± 0.25	84.14 ± 0.25
SVM(poly)	85.58 ± 0.60	80.50 ± 0.53	83.06 ± 0.26	83.59 ± 0.28
SVM(rbf)	85.50 ± 0.34	80.62 ± 0.47	83.08 ± 0.32	83.60 ± 0.30
SVM(sigmoid)	70.42 ± 0.72	69.67 ± 0.90	70.05 ± 0.53	70.34 ± 0.52
Random Forest	80.76 ± 0.85	80.24 ± 0.77	80.50 ± 0.72	80.69 ± 0.73
CART	77.28 ± 0.84	77.42 ± 0.76	77.35 ± 0.32	77.48 ± 0.38

Πίνακας 6.2

Μέθοδος μείωσης χαρακτηριστικών: Spearman Correlation				
Classifier	**Sensitivity(%)**	**Specificity(%)**	**Accuracy(%)**	**F1 Score(%)**
MDC	76.95 ± 0.49	69.88 ± 0.28	73.45 ± 0.24	74.51 ± 0.28
KNN	85.57 ± 0.82	72.66 ± 0.67	79.17 ± 0.58	80.55 ± 0.57
Bayesian	83.53 ± 0.33	71.12 ± 0.42	77.38 ± 0.31	78.83 ± 0.28
LDA	85.55 ± 0.70	80.87 ± 0.65	83.23 ± 0.26	83.73 ± 0.28
Logistic Regressor	85.42 ± 0.34	80.35 ± 0.56	82.91 ± 0.31	83.44 ± 0.28
Perceptron	77.14 ± 1.88	77.51 ± 1.77	77.32 ± 1.30	77.42 ± 1.34
SVM(linear)	86.26 ± 0.41	80.91 ± 0.55	83.61 ± 0.28	84.14 ± 0.26
SVM(poly)	86.38 ± 0.36	78.59 ± 0.61	82.52 ± 0.27	83.29 ± 0.23
SVM(rbf)	85.86 ± 0.69	80.24 ± 0.55	83.07 ± 0.27	83.65 ± 0.30
SVM(sigmoid)	70.71 ± 0.89	69.63 ± 0.51	70.18 ± 0.36	70.51 ± 0.49
Random Forest	80.81 ± 0.59	80.28 ± 0.65	80.55 ± 0.31	80.73 ± 0.32
CART	76.75 ± 1.44	77.36 ± 0.64	77.05 ± 0.57	77.12 ± 0.76

Πίνακας 6.3

Μέθοδος μείωσης χαρακτηριστικών: PCA				
Classifier	**Sensitivity(%)**	**Specificity(%)**	**Accuracy(%)**	**F1 Score(%)**
MDC	77.05% ± 0.31	69.94% ± 0.67	73.53% ± 0.39	74.59% ± 0.33
KNN	85.43% ± 0.62	73.27% ± 0.96	79.40% ± 0.62	80.71% ± 0.55
Bayesian	83.57% ± 0.40	70.85% ± 0.33	77.26% ± 0.28	78.75% ± 0.27
LDA	85.53% ± 0.56	80.90% ± 0.54	83.24% ± 0.37	83.73% ± 0.36
Logistic Regressor	85.61% ± 0.62	80.39% ± 0.43	83.02% ± 0.26	83.57% ± 0.30
Perceptron	76.96% ± 1.96	77.20% ± 1.68	77.08% ± 1.44	77.20% ± 1.50
SVM(linear)	86.21% ± 0.36	80.94% ± 0.58	83.60% ± 0.30	84.13% ± 0.27
SVM(poly)	86.35% ± 0.40	78.38% ± 0.34	82.40% ± 0.22	83.19% ± 0.22
SVM(rbf)	85.64% ± 0.44	80.31% ± 0.64	83.00% ± 0.26	83.56% ± 0.24
SVM(sigmoid)	70.74% ± 0.60	69.61% ± 0.20	70.18% ± 0.29	70.52% ± 0.37
Random Forest	80.68% ± 0.54	80.44% ± 0.54	80.56% ± 0.27	80.72% ± 0.28
CART	77.48% ± 0.53	77.39% ± 0.80	77.44% ± 0.57	77.60% ± 0.54

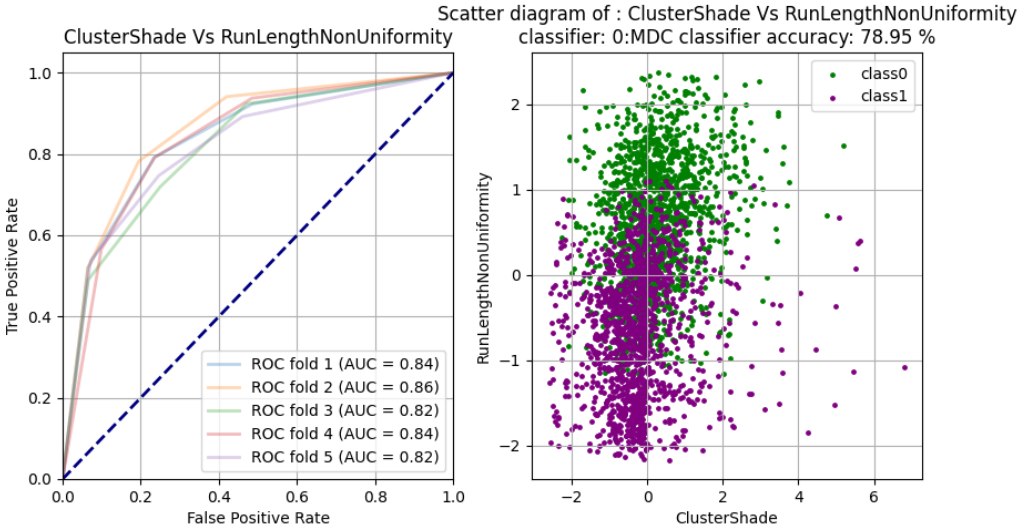
Πίνακας 6.4

Μέθοδος μείωσης χαρακτηριστικών: RFE				
Classifier	**Sensitivity(%)**	**Specificity(%)**	**Accuracy(%)**	**F1 Score(%)**
MDC	77.14% ± 0.63	70.21% ± 0.63	73.70% ± 0.43	74.74% ± 0.43
KNN	85.27% ± 0.58	72.85% ± 0.61	79.12% ± 0.46	80.46% ± 0.43
Bayesian	83.47% ± 0.41	70.87% ± 0.25	77.23% ± 0.28	78.71% ± 0.28
LDA	85.40% ± 0.40	80.80% ± 0.31	83.12% ± 0.25	83.61% ± 0.26
Logistic Regressor	85.63% ± 0.31	80.28% ± 0.35	82.98% ± 0.25	83.53% ± 0.24
Perceptron	77.91% ± 2.27	75.70% ± 2.40	76.81% ± 2.00	77.21% ± 1.96
SVM(linear)	86.28% ± 0.40	80.99% ± 0.37	83.66% ± 0.31	84.19% ± 0.31
SVM(poly)	86.19% ± 0.91	78.40% ± 0.25	82.33% ± 0.47	83.10% ± 0.52
SVM(rbf)	85.95% ± 0.50	80.16% ± 0.68	83.08% ± 0.29	83.67% ± 0.27
SVM(sigmoid)	70.38% ± 0.73	69.49% ± 0.99	69.93% ± 0.52	70.25% ± 0.49
Random Forest	81.12% ± 0.98	79.65% ± 0.57	80.39% ± 0.54	80.67% ± 0.60
CART	76.94% ± 0.79	76.80% ± 0.98	76.87% ± 0.64	77.04% ± 0.63

Για τα αποτελέσματα μας εφαρμόσαμε brute force μέθοδο για δοκιμή όλων των συνδυασμών χαρακτηριστικών, για δύο και τρεις διαστάσεις και διαλέξαμε αυτά με την καλύτερη ακρίβεια και ROC_AUC. Την τεχνική αυτή την εφαρμόσαμε και στα δύο σύνολα δεδομένων με όλους τους ταξινομητές.

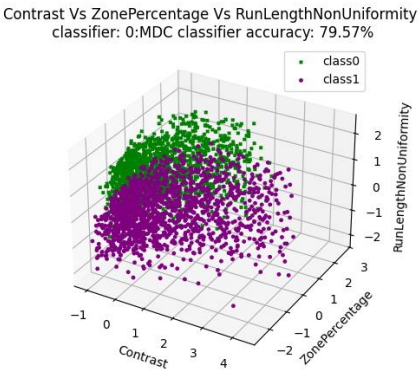
Πρώτο σύνολο δεδομένων

MDC



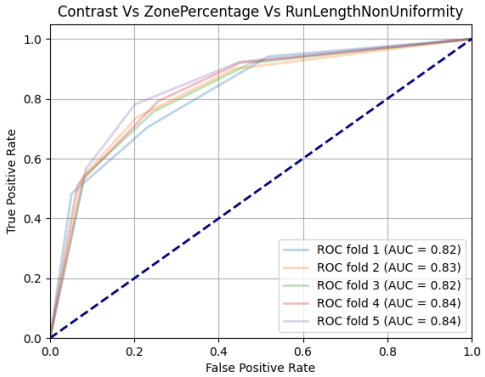
Εικόνα 6.1.1

Evaluation Metrics_2d:
sensitivity: 78.85
specificity: 79.04
accuracy: 78.95
f1_score: 79.07



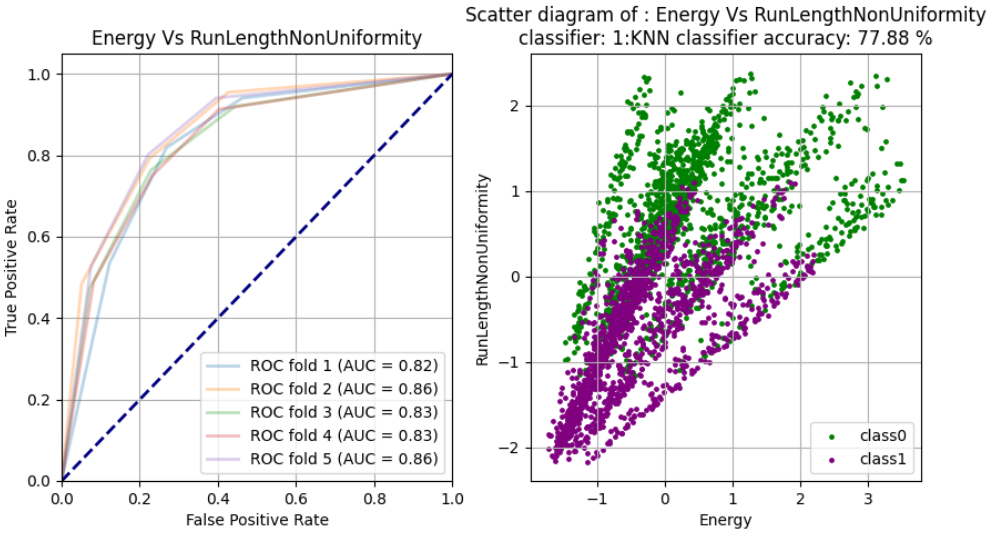
Εικόνα 6.1.2

Evaluation Metrics_3d
sensitivity: 81.04
specificity: 78.07
accuracy: 79.57
f1_score: 80.00



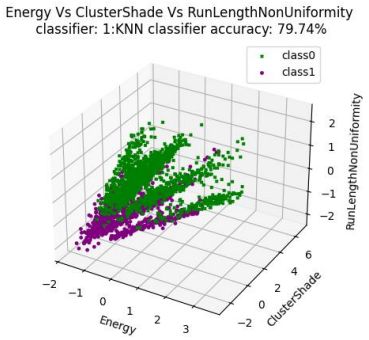
Εικόνα 6.1.3

KNN



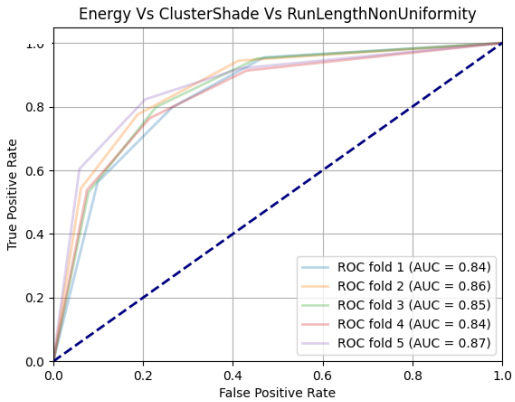
Εικόνα 6.2.1

Evaluation Metrics_2d:
sensitivity: 75.72
specificity: 80.08
accuracy: 77.88
f1_score: 77.54



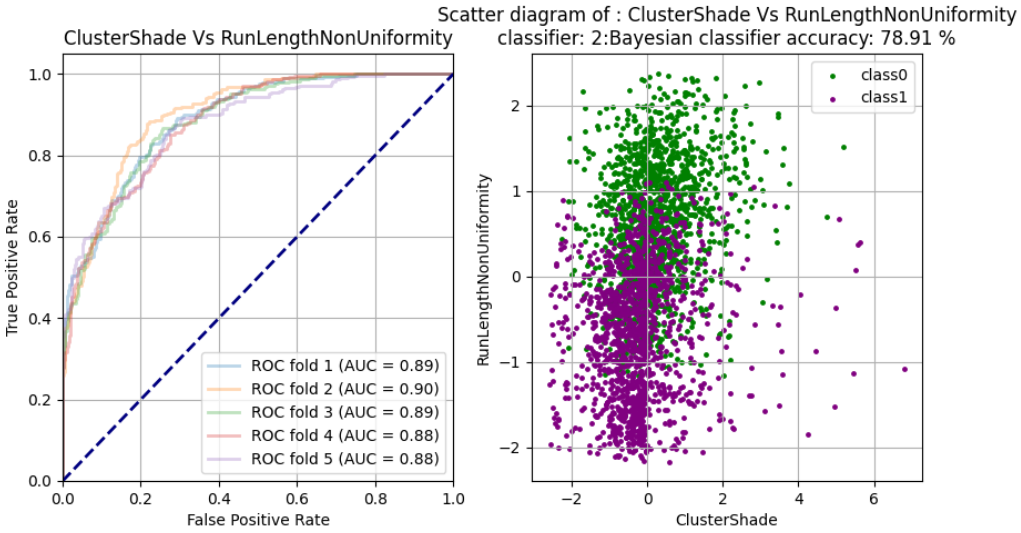
Evaluation Metrics_3d:
sensitivity: 78.79
specificity: 80.71
accuracy: 79.74
f1_score: 79.68

Εικόνα 6.2.2



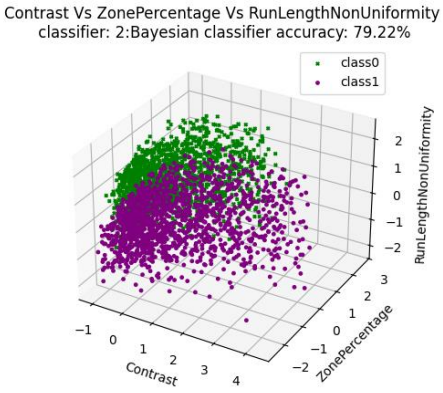
Εικόνα 6.2.3

Bayesian



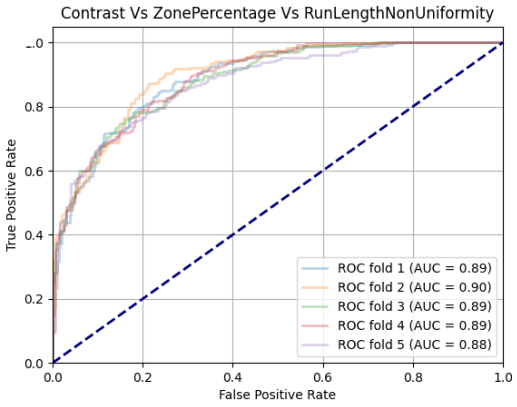
Εικόνα 6.3.1

Evaluation Metrics_2d:
sensitivity: 80.08
specificity: 77.72
accuracy: 78.91
f1_score: 79.30



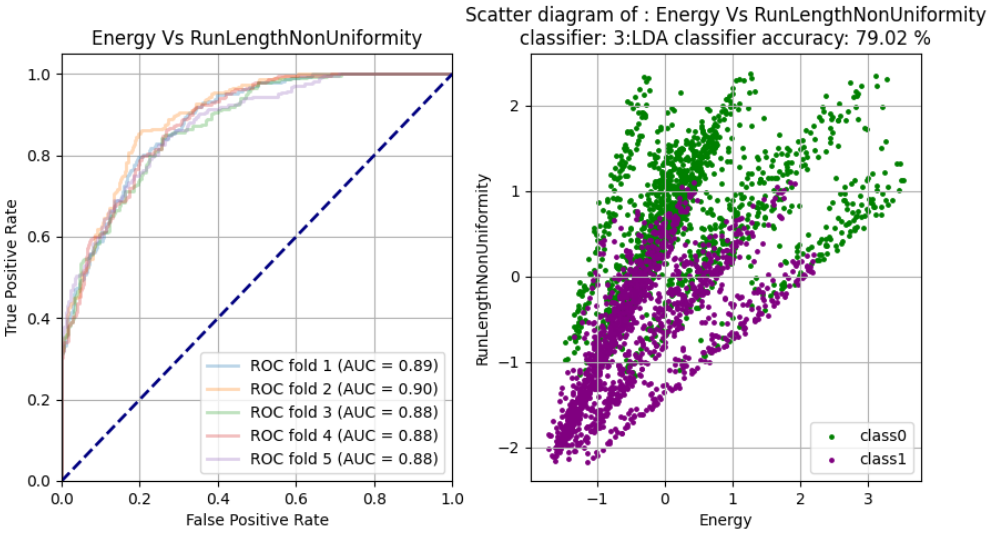
Evaluation Metrics:
sensitivity: 81.51
specificity: 76.89
accuracy: 79.22
f1_score: 79.83

Εικόνα 6.3.2



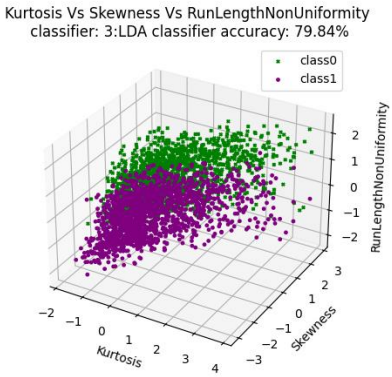
Εικόνα 6.3.3

LDA



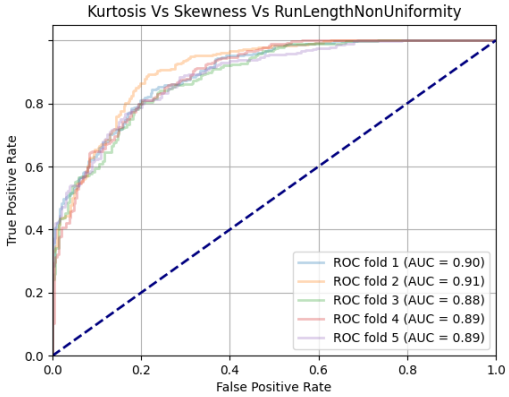
Εικόνα 6.4.1

Evaluation Metrics_2d:
sensitivity: 79.54
specificity: 78.49
accuracy: 79.02
f1_score: 79.27



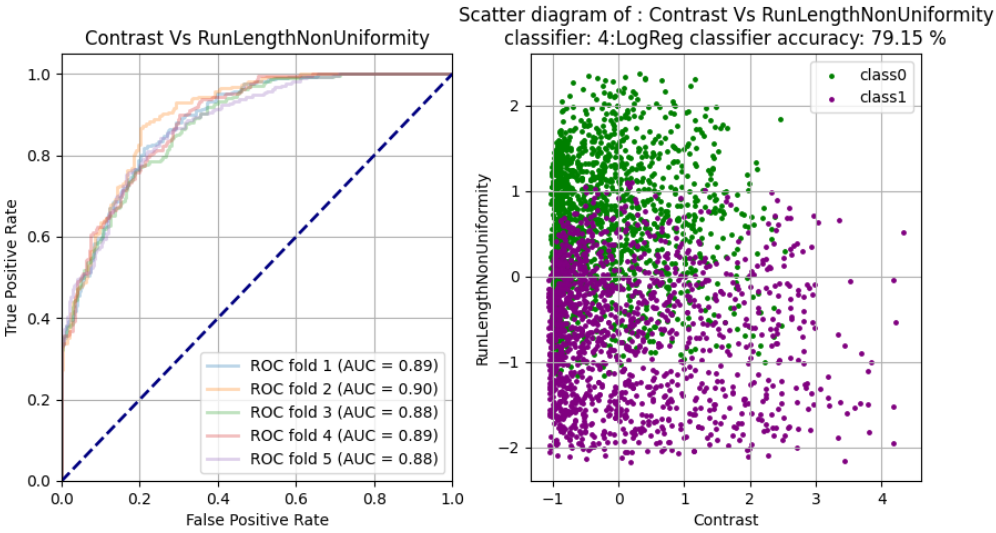
Εικόνα 6.4.2

Evaluation Metrics_3d:
sensitivity: 81.17
specificity: 78.49
accuracy: 79.84
f1_score: 80.24



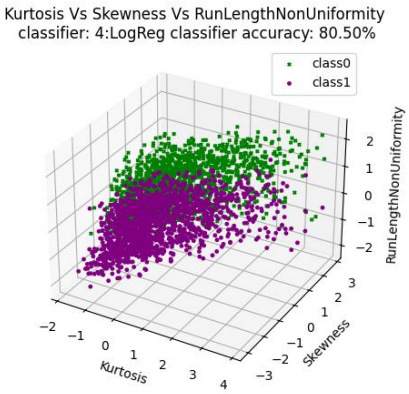
Εικόνα 6.4.3

Logistic Regressor



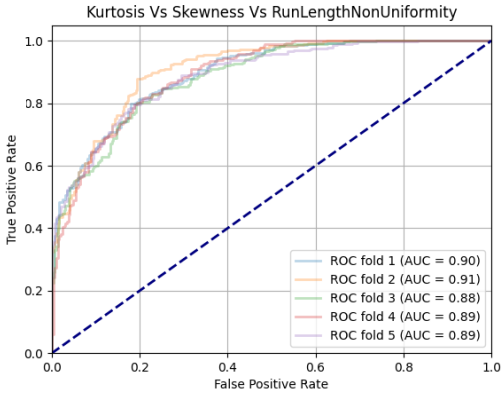
Εικόνα 6.5.1

Evaluation Metrics_2d:
sensitivity: 79.54
specificity: 78.76
accuracy: 79.15
f1_score: 79.37

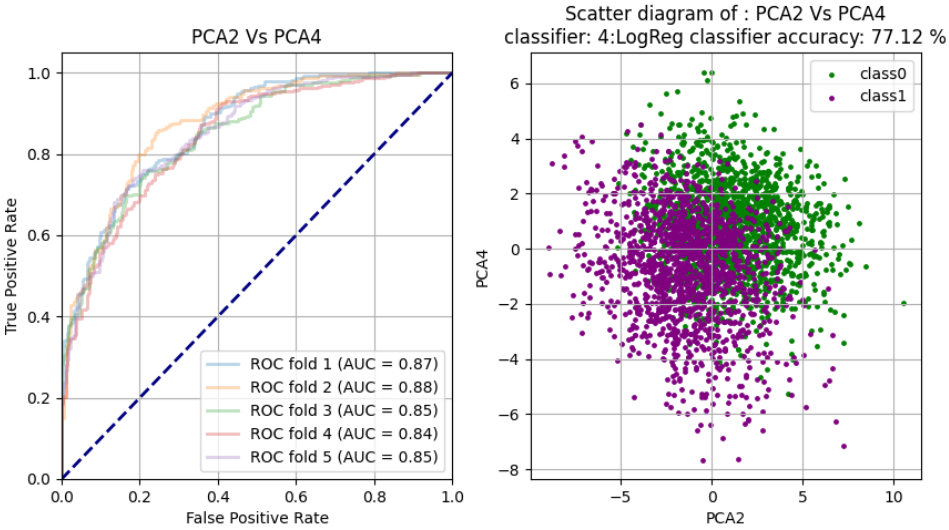


Εικόνα 6.5.2

Evaluation Metrics_3d:
sensitivity: 80.63
specificity: 80.36
accuracy: 80.50
f1_score: 80.66

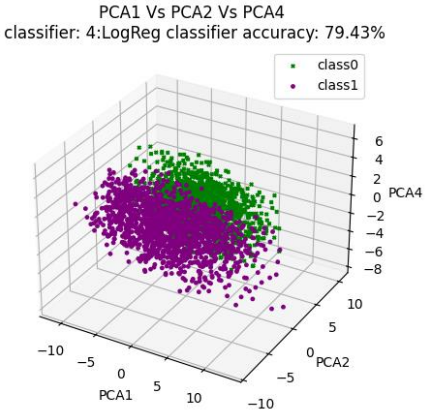


Εικόνα 6.5.3



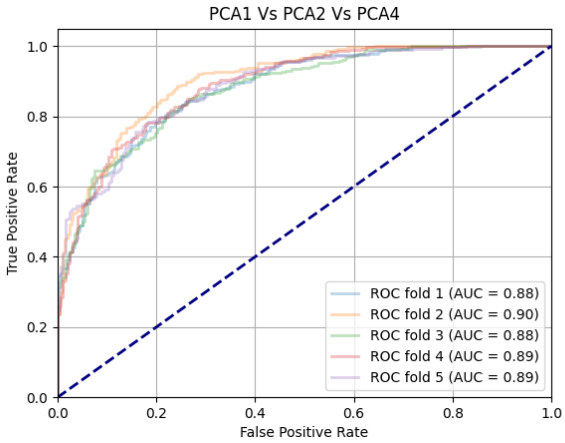
Εικόνα 6.5.4

Evaluation Metrics_2d:
sensitivity: 77.76
specificity: 76.47
accuracy: 77.12
f1_score: 77.42



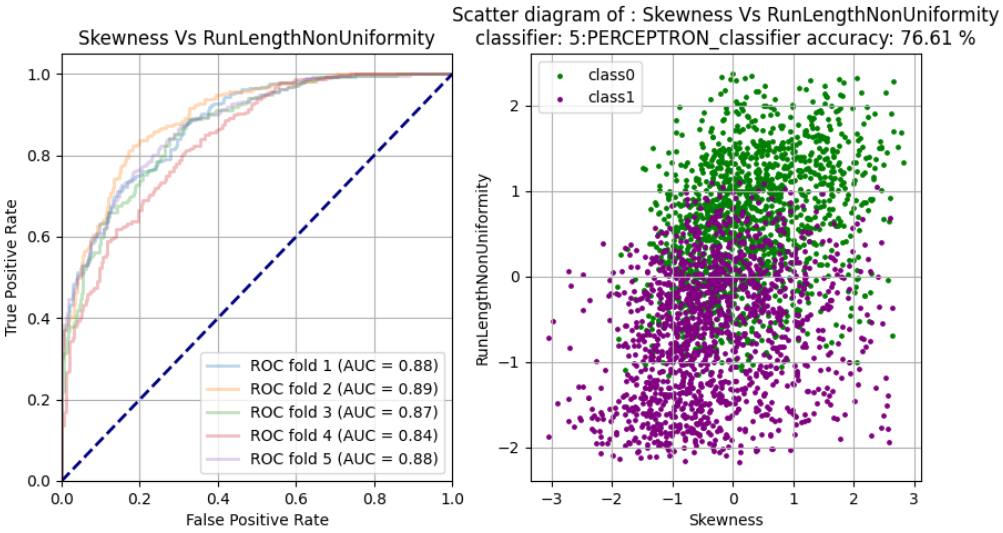
Evaluation Metrics_3d:
sensitivity: 80.08
specificity: 78.76
accuracy: 79.43
f1_score: 79.70

Εικόνα 6.5.5



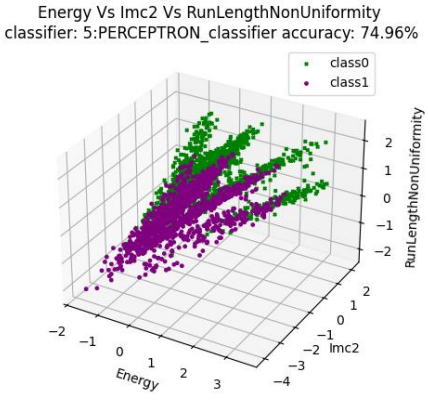
Εικόνα 6.5.6

Perceptron



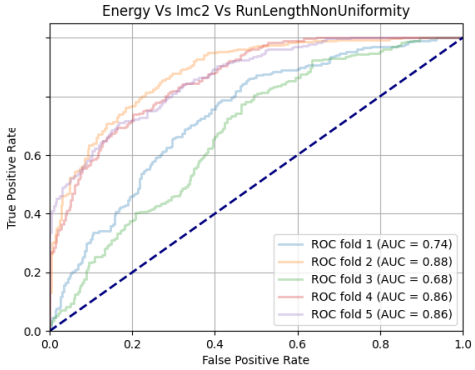
Εικόνα 6.6.1

Evaluation Metrics_2d:
sensitivity: 77.42
specificity: 75.78
accuracy: 76.61
f1_score: 76.95



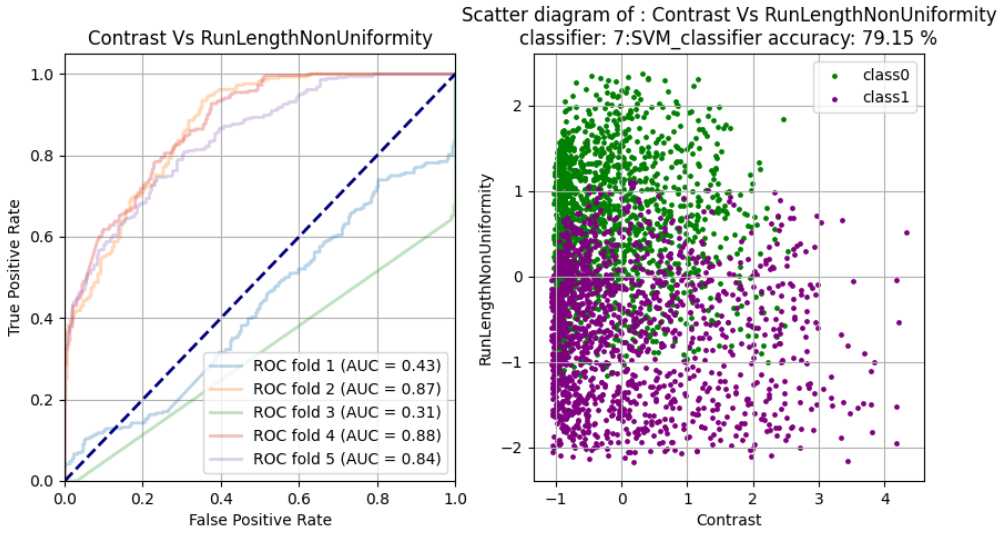
Εικόνα 6.6.1

Evaluation Metrics_3d:
sensitivity: 73.53
specificity: 76.41
accuracy: 74.96
f1_score: 74.76



Εικόνα 6.6.3

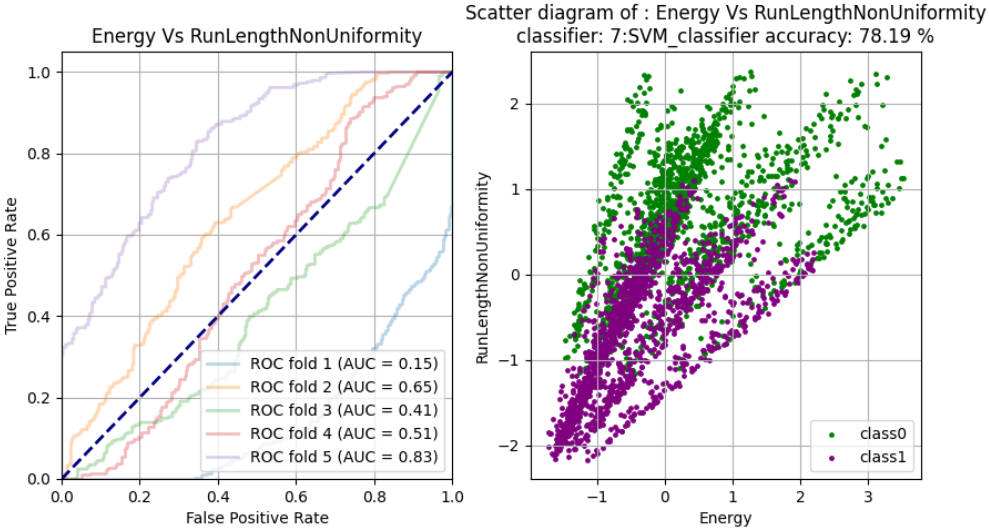
SVM(linear kernel)



Εικόνα 6.7.1

Evaluation Metrics:
sensitivity: 79.60
specificity: 78.70
accuracy: 79.15
f1_score: 79.39

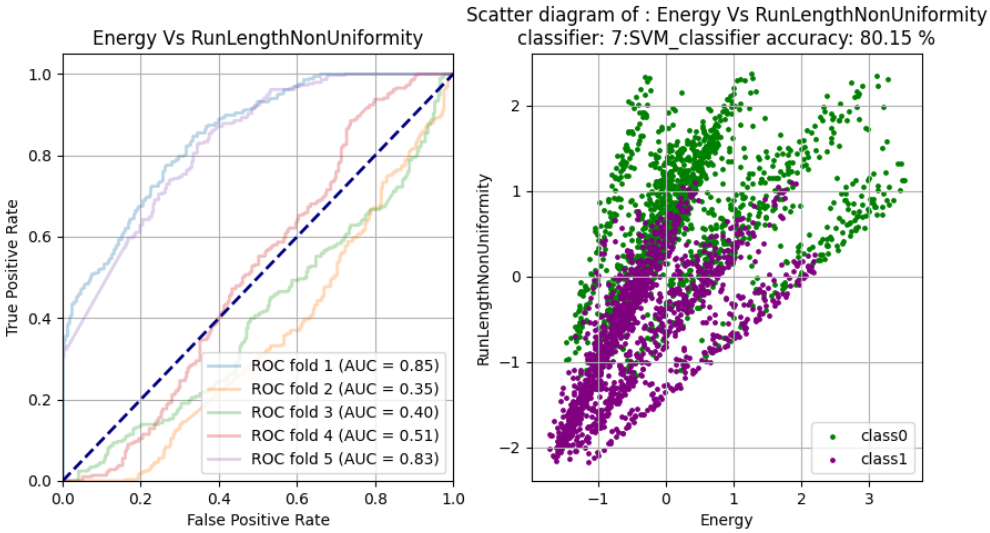
SVM(polynomial kernel)



Εικόνα 6.7.2

Evaluation Metrics:
sensitivity: 72.24
specificity: 84.25
accuracy: 78.19
f1_score: 76.96

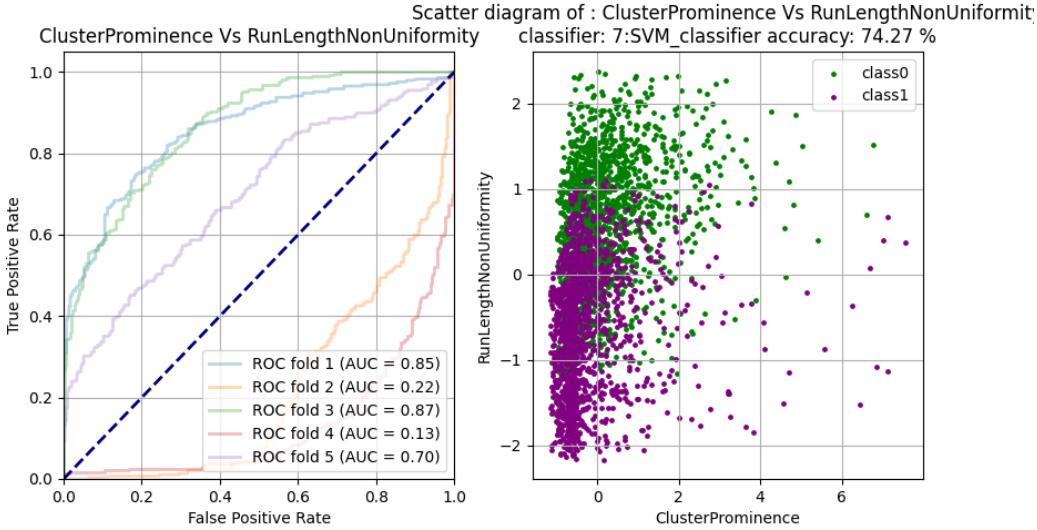
SVM(rbf)



Εικόνα 6.7.3

Evaluation Metrics:
sensitivity: 76.94
specificity: 83.41
accuracy: 80.15
f1_score: 79.63

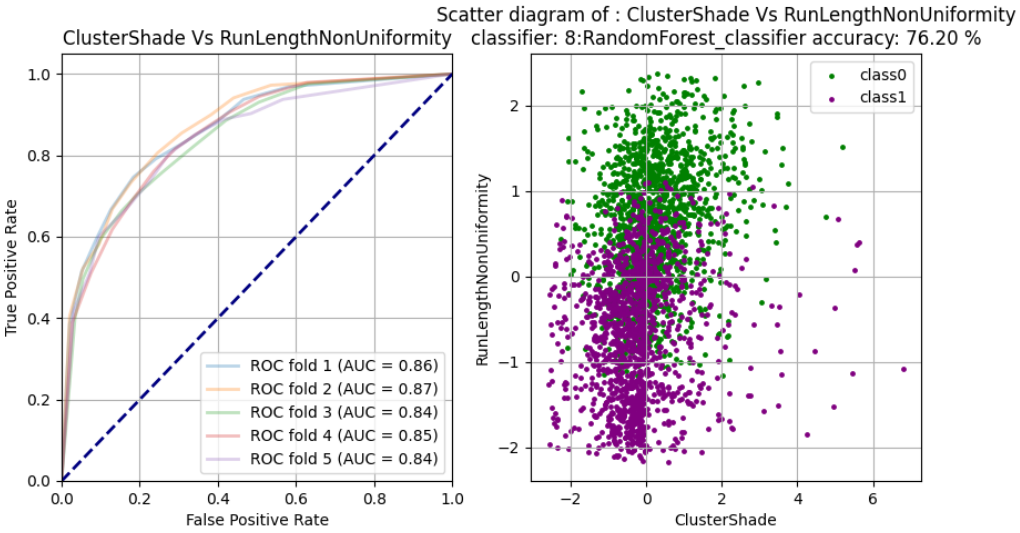
SVM(sigmoid)



Εικόνα 6.7.4

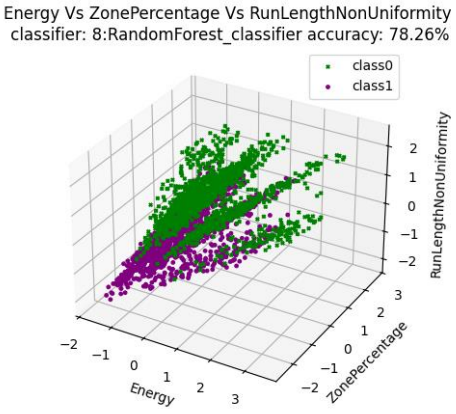
Evaluation Metrics:
sensitivity: 72.17
specificity: 76.41
accuracy: 74.27
f1_score: 73.88

Random Forest



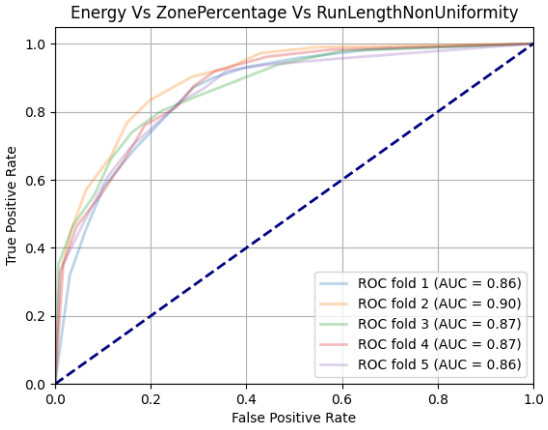
Εικόνα 6.8.1

Evaluation Metrics_2d:
sensitivity: 76.26
specificity: 76.13
accuracy: 76.20
f1_score: 76.37



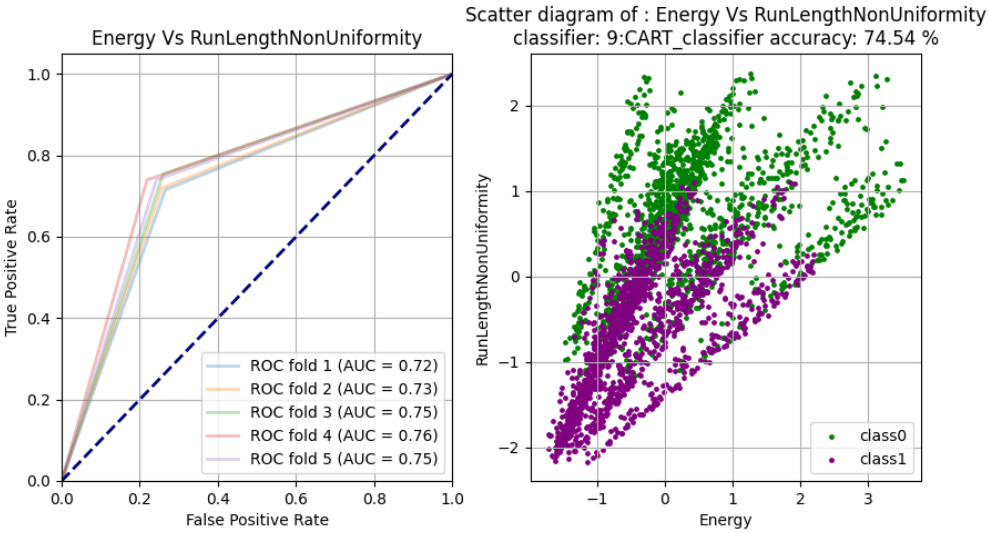
Εικόνα 6.8.2

Evaluation Metrics_3d:
sensitivity: 77.49
specificity: 79.04
accuracy: 78.26
f1_score: 78.24



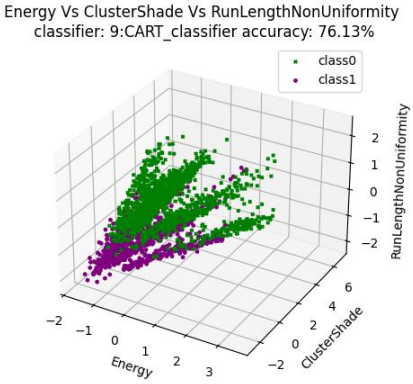
Εικόνα 6.8.3

CART



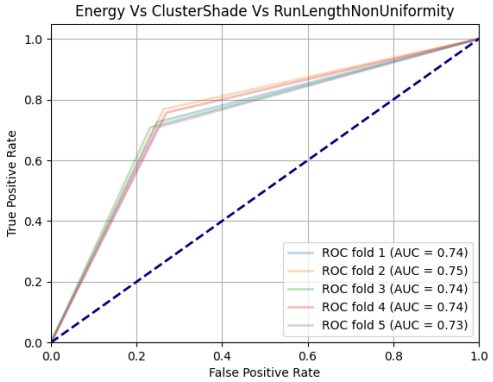
Εικόνα 6.9.1

Evaluation Metrics_2d:
sensitivity: 74.69
specificity: 74.39
accuracy: 74.54
f1_score: 74.74



Evaluation Metrics_3d:
sensitivity: 77.69
specificity: 74.53
accuracy: 76.13
f1_score: 76.65

Εικόνα 6.9.2



Εικόνα 6.9.3

Στο πρώτο σύνολο δεδομένων, σαν μία δεύτερη μέθοδος, εφαρμόστηκε μείωση χαρακτηριστικών μέσω της μεθόδου **PCA**. Με τη χρήση αυτής της μεθόδου, βρέθηκαν τα **PCA2** και **PCA4** ως ο καλύτερος συνδυασμός χαρακτηριστικών σχεδόν σε όλους τους ταξινομητές. Τα **loadings** αντικατοπτρίζουν τη συσχέτιση ή τη συνεισφορά κάθε αρχικού χαρακτηριστικού στην κύρια συνιστώσα. Όσο υψηλότερη είναι η απόλυτη τιμή του loading για ένα συγκεκριμένο χαρακτηριστικό, τόσο περισσότερο αυτό το χαρακτηριστικό συνεισφέρει στην κύρια συνιστώσα.Τα χαρακτηριστικά με τη μεγαλύτερη συνεισφορά φαίνονται στους πίνακες 6.5 και 6.6.

Πίνακας 6.5

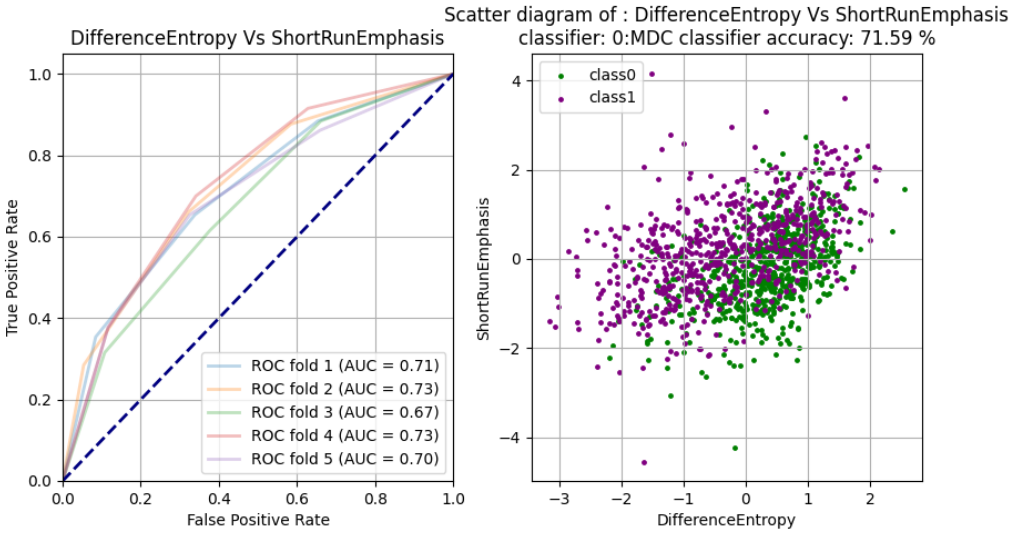
Κύρια χαρακτηριστικά που συνεισφέρουν στο PCA2:	
Χαρακτηριστικό	**Loading**
ClusterShade	0.255137
Correlation	0.253827
Skewness	0.211642
Imc1	0.203828
MCC	0.202172
ClusterProminence	0.201247
Variance	0.200914
RunLengthNonUniformity	0.200576
Range	0.197621
DependenceNonUniformity	0.196910

Πίνακας 6.6

Κύρια χαρακτηριστικά που συνεισφέρουν στο PCA4	
Χαρακτηριστικό	**Loading**
GrayLevelNonUniformity	0.351829
DependenceNonUniformity	0.348146
RunLengthNonUniformity	0.347256
SizeZoneNonUniformity	0.345587
TotalEnergy	0.219633
Energy	0.201247
Correlation	0.188351
ClusterShade	0.166859
Imc2	0.157407
Strength	0.156157

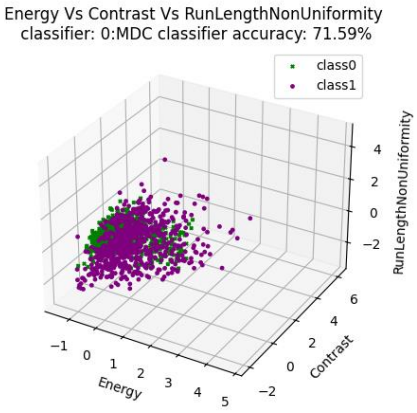
Δεύτερο σύνολο δεδομένων

MDC



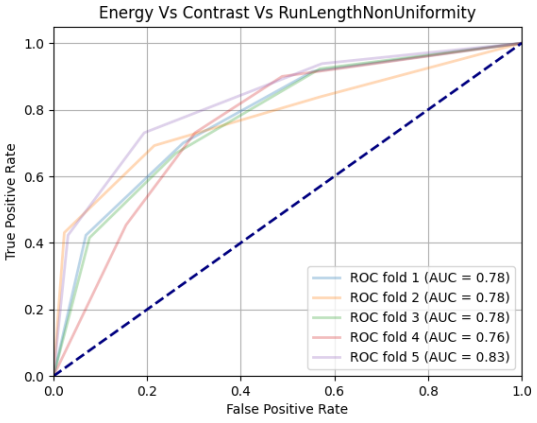
Εικόνα 6.10.1

Evaluation Metrics_2d:
sensitivity: 76.70
specificity: 68.46
accuracy: 72.57
f1_score: 73.63



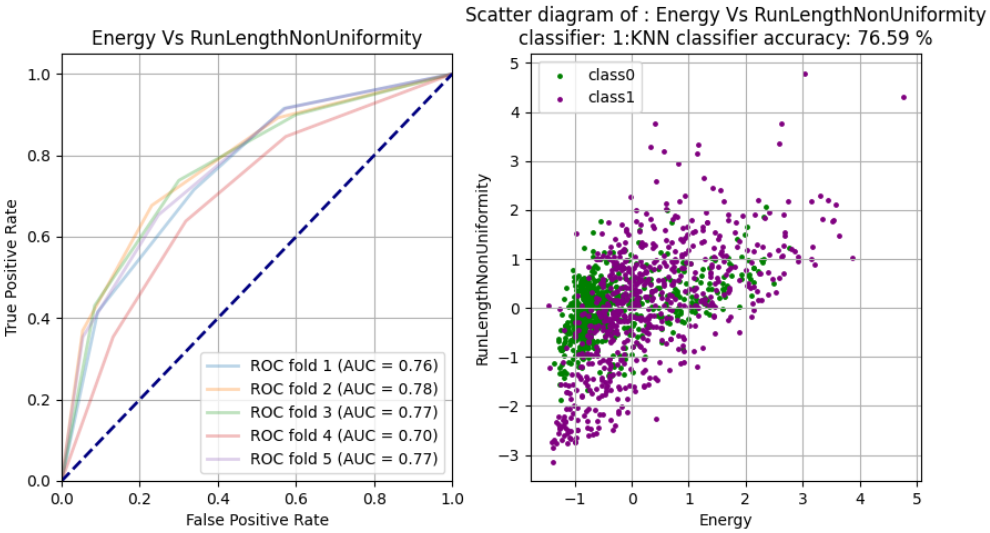
Εικόνα 6.10.2

Evaluation Metrics_3d:
sensitivity: 78.09
specificity: 67.08
accuracy: 72.57
f1_score: 73.98



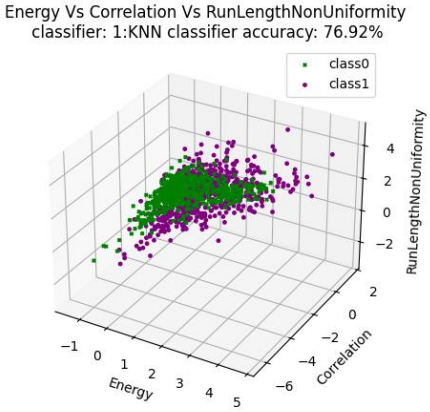
Εικόνα 6.10.3

KNN



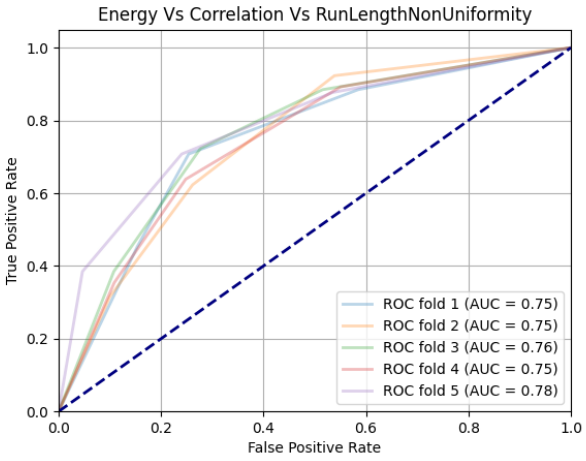
Εικόνα 6.11.1

Evaluation Metrics_2d:
sensitivity: 84.10
specificity: 67.23
accuracy: 75.65
f1_score: 77.52



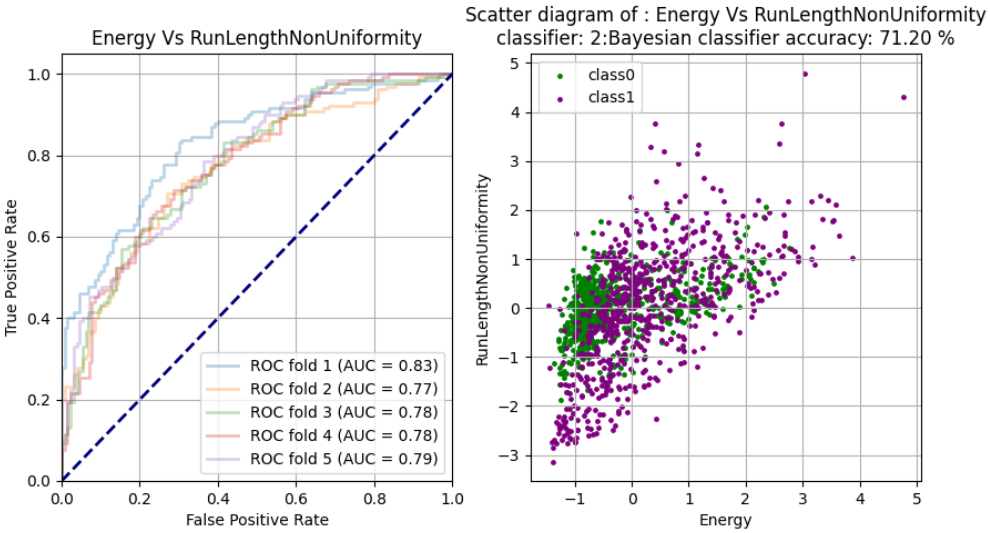
Εικόνα 6.11.2

Evaluation Metrics_3d:
sensitivity: 83.18
specificity: 71.23
accuracy: 77.20
f1_score: 78.46



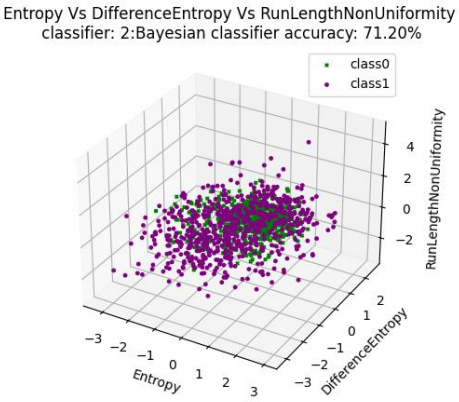
Εικόνα 6.11.3

Bayesian



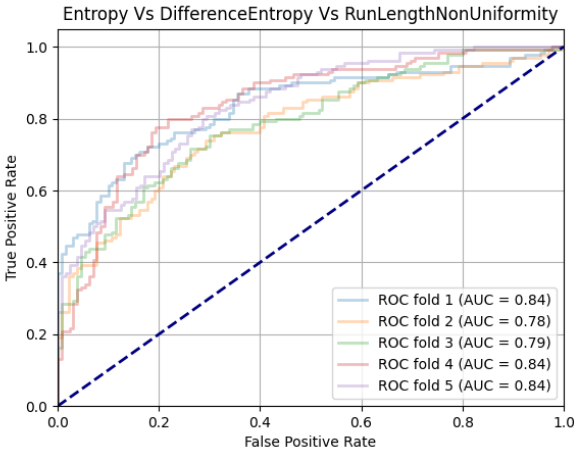
Εικόνα 6.12.1

Evaluation Metrics_2d
sensitivity: 78.86
specificity: 62.46
accuracy: 70.65
f1_score: 72.84



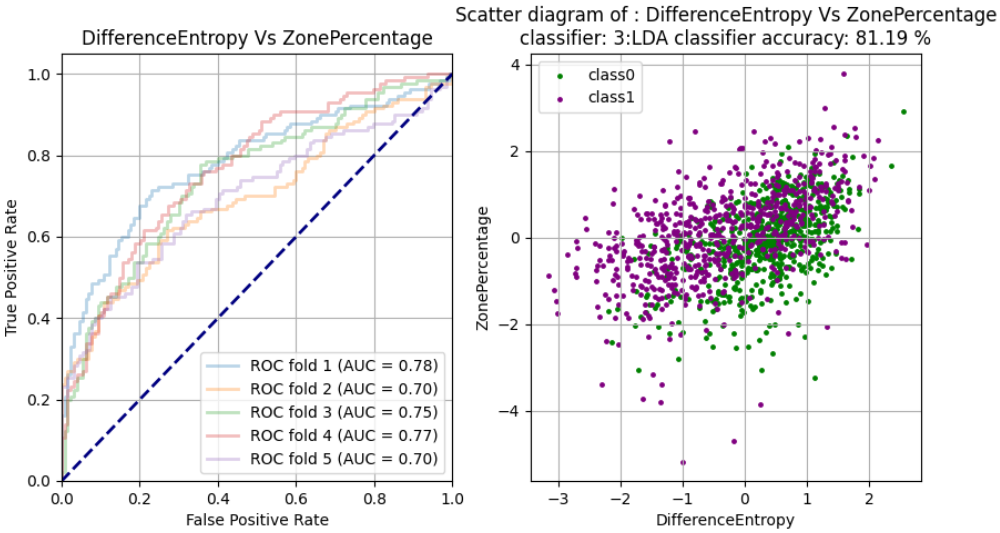
Εικόνα 6.12.2

Evaluation Metrics_3d:
sensitivity: 79.17
specificity: 62.62
accuracy: 70.88
f1_score: 73.08



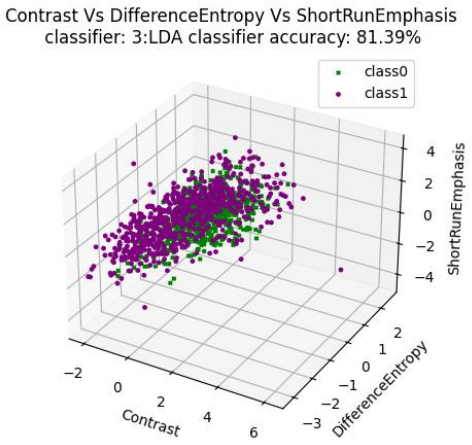
Εικόνα 6.12.3

LDA



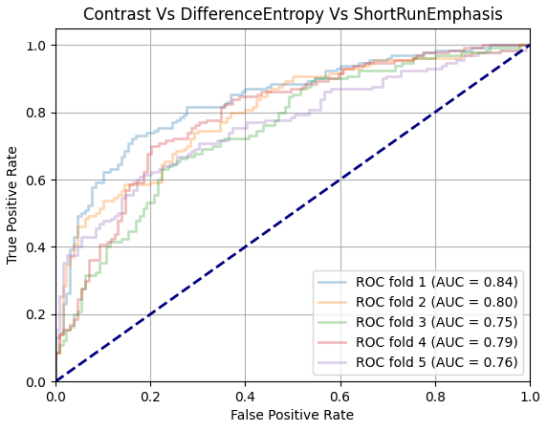
Εικόνα 6.13.1

Evaluation Metrics_2d:
sensitivity: 85.34
specificity: 75.69
accuracy: 80.51
f1_score: 81.38



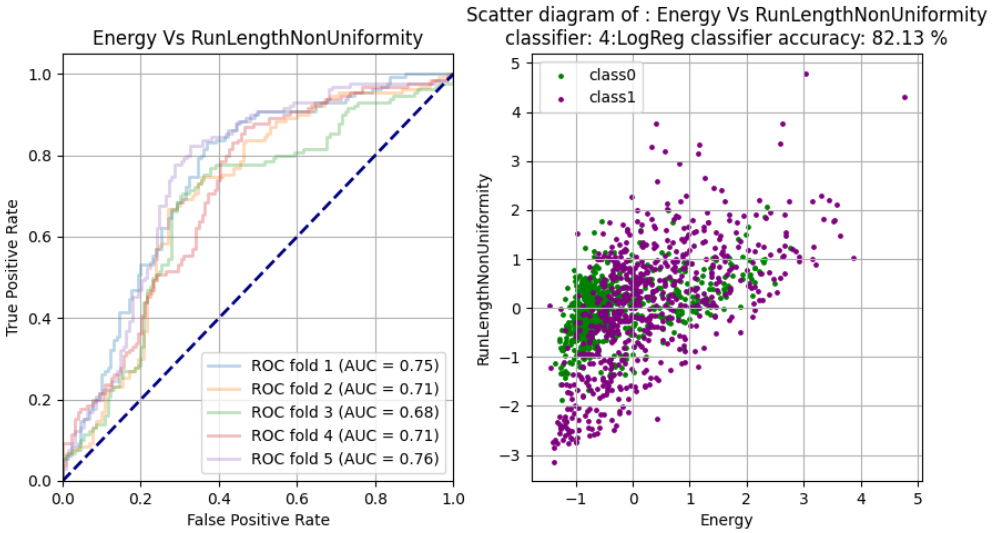
Εικόνα 6.13.2

Evaluation Metrics_3d
sensitivity: 86.88
specificity: 77.54
accuracy: 82.20
f1_score: 82.98



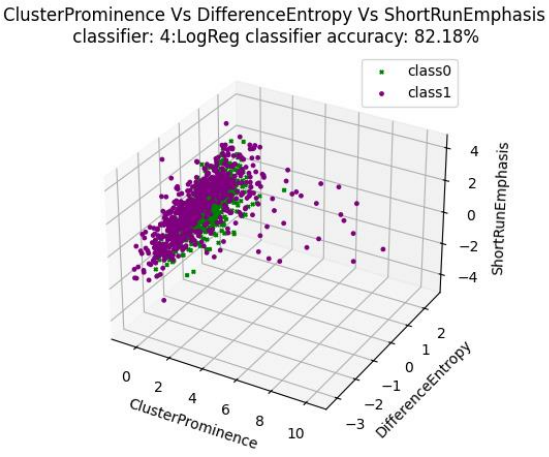
Εικόνα 6.13.3

Logistic Regressor



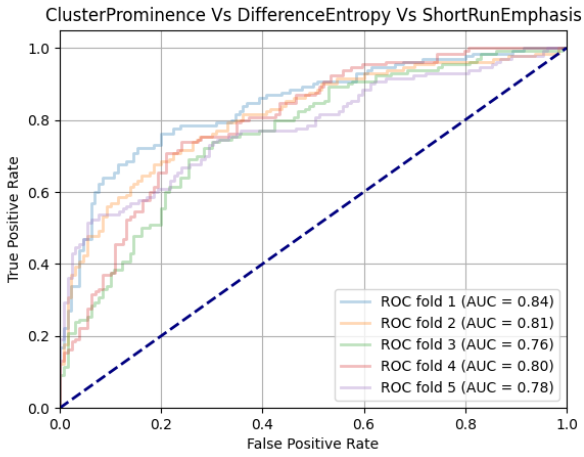
Εικόνα 6.14.1

Evaluation Metrics_2d:
sensitivity: 82.72
specificity: 80.15
accuracy: 81.43
f1_score: 81.65



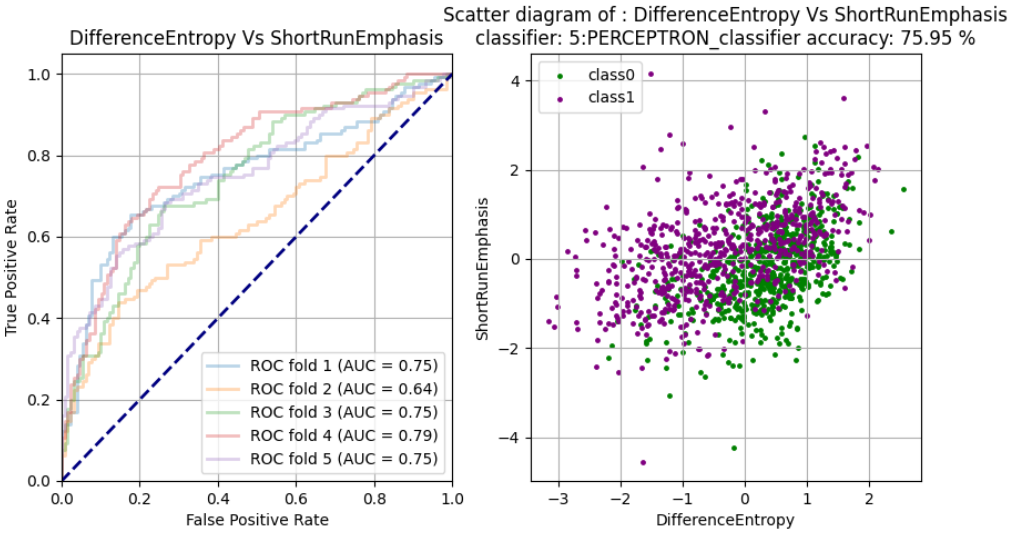
Εικόνα 6.14.2

Evaluation Metrics_3d:
sensitivity: 83.80
specificity: 80.31
accuracy: 82.05
f1_score: 82.34



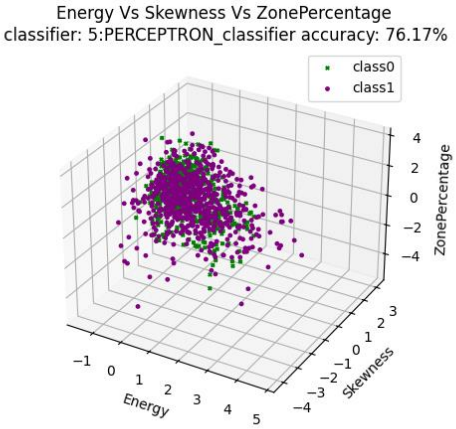
Εικόνα 6.14.3

Perceptron



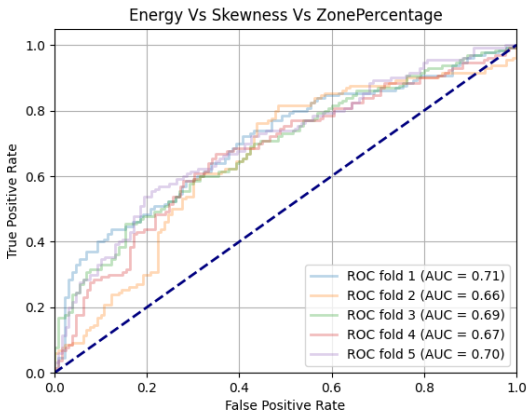
Εικόνα 6.15.1

Evaluation Metrics_2d:
sensitivity: 75.62
specificity: 74.77
accuracy: 75.19
f1_score: 75.27



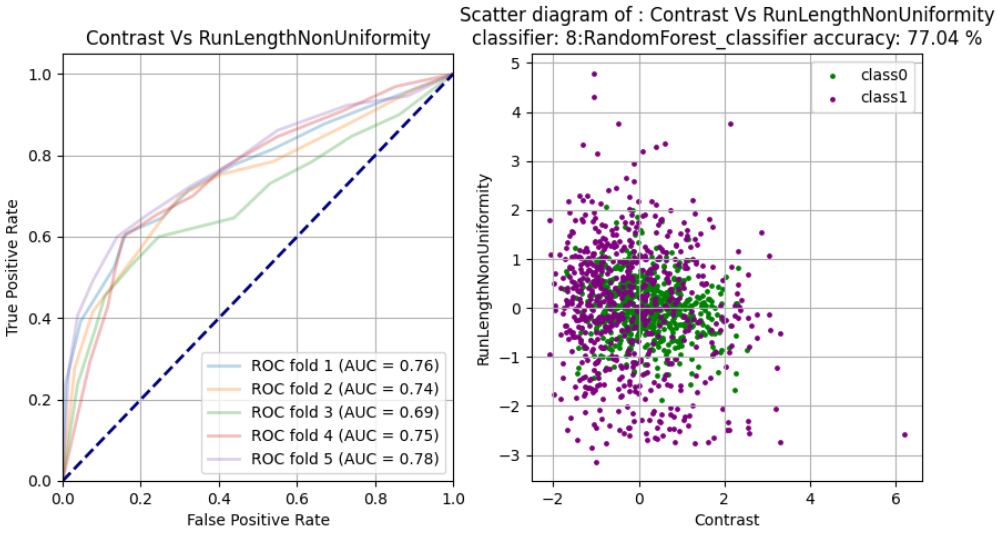
Evaluation Metrics_3d:
sensitivity: 76.23
specificity: 72.92
accuracy: 74.58
f1_score: 74.96

Εικόνα 6.15.2



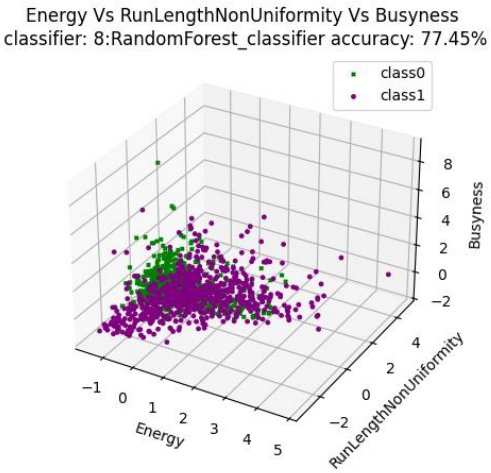
Εικόνα 6.15.3

Random Forest



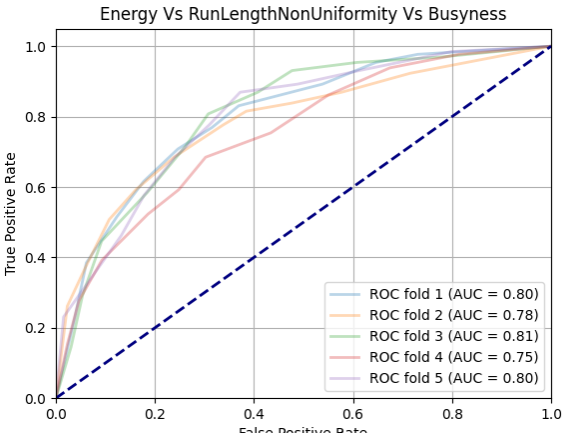
Εικόνα 6.16.1

Evaluation Metrics_2d:
sensitivity: 77.01
specificity: 76.31
accuracy: 76.66
f1_score: 76.71



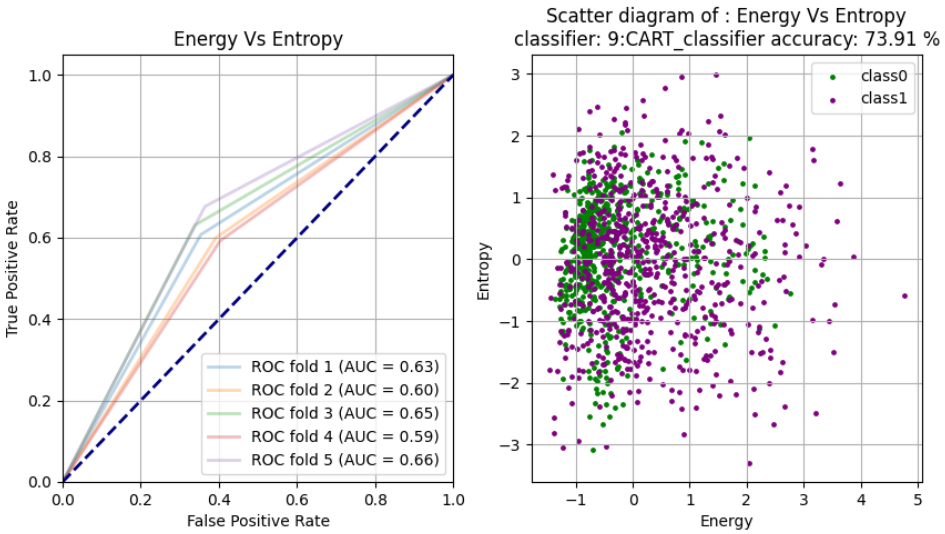
Εικόνα 6.16.2

Evaluation Metrics_3d:
sensitivity: 79.48
specificity: 78.31
accuracy: 78.89
f1_score: 78.99



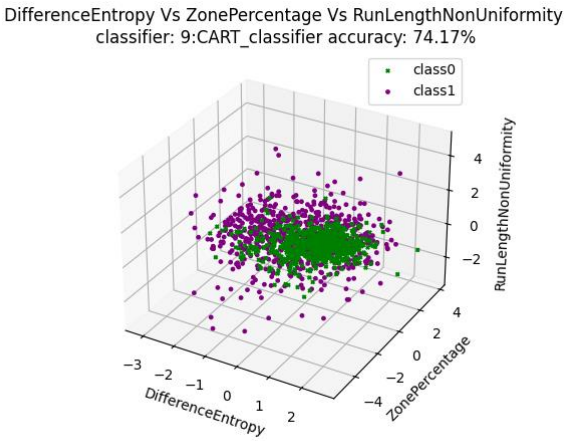
Εικόνα 6.16.3

CART



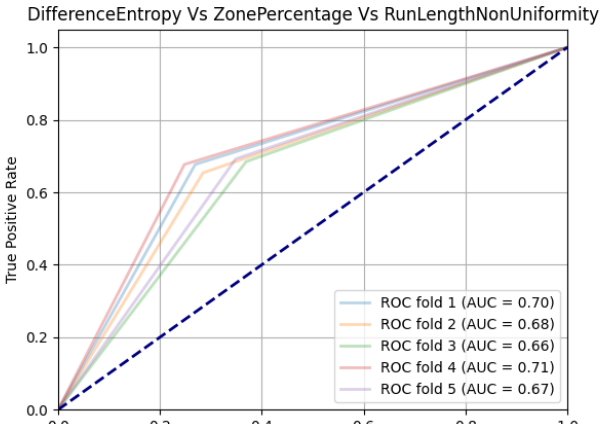
Εικόνα 6.17.1

Evaluation Metrics_2d:
sensitivity: 73.92
specificity: 74.62
accuracy: 74.27
f1_score: 74.15



Εικόνα 6.17.2

Evaluation Metrics_3d:
sensitivity: 74.54
specificity: 74.77
accuracy: 74.65
f1_score: 74.59



Εικόνα 6.17.3

7. Συζήτηση αποτελεσμάτων και συμπεράσματα

7.1 Εκπαίδευση μοντέλων σε όλα τα χαρακτηριστικά(πρώτο σύνολο δεδομένων)

Τα αποτελέσματα των ταξινομητών που στους μετά από 10 επαναλήψεις παρουσιάζουν στατιστικές μετρήσεις για την ευαισθησία, την ειδικότητα, την ακρίβεια και το F1 score, τόσο για το σύνολο των χαρακτηριστικών όσο και για τις διαφορετικές μεθόδους μείωσης χαρακτηριστικών (Spearman Correlation, PCA, RFE). Παρακάτω αναλύονται τα βασικά ευρήματα για κάθε σύνολο δεδομένων.

1. Χωρίς Μείωση Χαρακτηριστικών

Από τον πίνακα 6.1 παρατηρούμε πως ο *SVM (linear)* ξεχώρισε με την υψηλότερη ακρίβεια (83.57% ± 0.25) και F1 score (84.14% ± 0.25), δείχνοντας συνεπή απόδοση. Ο *SVM (rbf)* και ο *SVM (poly)* είχαν επίσης εξαιρετικές επιδόσεις με παρόμοια ακρίβεια (83.08% ± 0.32 και 83.06% ± 0.26 αντίστοιχα). Ο *Random Forest* είχε ελαφρώς χαμηλότερες αποδόσεις, αλλά ήταν σταθερός με ακρίβεια 80.50% ± 0.72 και F1 score 80.69% ± 0.73. Αντίθετα, ο *SVM (sigmoid)* παρουσίασε χαμηλότερη απόδοση, με ακρίβεια 70.05% ± 0.53 και F1 score 70.34% ± 0.52. Συνολικά, οι γραμμικοί ταξινομητές SVM και οι μη γραμμικοί SVM με πυρήνα rbf φαίνεται να είναι οι πιο αποτελεσματικοί, με σταθερά υψηλά επίπεδα απόδοσης.

2.Μείωση Χαρακτηριστικών με Συσχέτιση Spearman

Στον πίνακα 6.2 με τη χρήση της συσχέτισης Spearman, οι ταξινομητές *SVM (rbf)* και *Random Forest* είχαν και πάλι υψηλές επιδόσεις, με ακρίβεια 82.77% και 82.87% αντίστοιχα. Ο *SVM (linear)* επίσης σημείωσε αξιόλογη ακρίβεια 82.70% και F1 score 83.15%. Οι *Logistic Regressor* και *LDA* είχαν συγκρίσιμα αποτελέσματα με ακρίβεια 82.32% και 82.46%, ενώ ο *SVM (poly)* σημείωσε την υψηλότερη ευαισθησία (86.77%), αλλά η ειδικότητά του ήταν χαμηλότερη σε σχέση με άλλα μοντέλα. Ο *CART* σημείωσε ακρίβεια 81.29%, με ελαφρώς χαμηλότερη ειδικότητα (77.79%). Γενικά, η συσχέτιση Spearman βελτίωσε ελαφρώς την ακρίβεια για τα περισσότερα μοντέλα, χωρίς σημαντικές θυσίες στην ειδικότητα.

3. Μείωση Χαρακτηριστικών με PCA

Στον πίνακα 6.3 στην περίπτωση του PCA, ο *SVM (linear)* παρουσίασε και πάλι την υψηλότερη απόδοση με ακρίβεια 83.60% ± 0.30 και F1 score 84.13% ± 0.27. Ο *SVM (rbf)* ακολούθησε στενά με ακρίβεια 83.00% ± 0.26 και F1 score 83.56% ± 0.24. Οι *KNN* και *Bayesian* παρουσίασαν καλές επιδόσεις με ακρίβειες 79.40% ± 0.62 και 77.26% ± 0.28 αντίστοιχα. Ο *SVM (sigmoid)* είχε και πάλι τη χαμηλότερη απόδοση, με ακρίβεια 70.18% ± 0.29 και F1 score 70.52% ± 0.37. Το PCA φαίνεται να έχει μικρή επίδραση στη βελτίωση των ταξινομητών, καθώς οι γραμμικοί και μη γραμμικοί SVM παραμένουν οι καλύτεροι σε όρους ακρίβειας και F1 score.

4. Μείωση Χαρακτηριστικών με RFE

Στον πίνακα 6.4 η χρήση του RFE οδήγησε και πάλι σε εξαιρετικές επιδόσεις για τον *SVM (linear)*, ο οποίος σημείωσε ακρίβεια 83.66% ± 0.31 και F1 score 84.19% ± 0.31. Οι *SVM (rbf)* και *SVM (poly)* ήταν επίσης αποδοτικοί, με ακρίβειες 83.08% ± 0.29 και 82.33% ± 0.47 αντίστοιχα. Ο *Random Forest* διατήρησε σταθερές επιδόσεις με ακρίβεια 80.39% ± 0.54 και F1 score 80.67% ± 0.60. Ο *SVM (sigmoid)* εξακολούθησε να έχει τη χαμηλότερη απόδοση, με ακρίβεια 69.93% ± 0.52. Συνολικά, το RFE φαίνεται να ενισχύει την απόδοση των πιο ισχυρών ταξινομητών, όπως οι SVM, διατηρώντας υψηλές επιδόσεις ευαισθησίας και ειδικότητας.

Συμπεράσματα
Οι ταξινομητές *SVM (linear)* και *SVM (rbf)* διατήρησαν σταθερά υψηλές επιδόσεις ανεξαρτήτως της χρήσης ή μη μεθόδων μείωσης χαρακτηριστικών. Οι μέθοδοι μείωσης χαρακτηριστικών, όπως το Spearman, το PCA και το RFE, βελτίωσαν ελαφρώς την απόδοση κάποιων ταξινομητών, αλλά δεν άλλαξαν δραματικά την τάση των SVM να παραμένουν οι καλύτεροι για την ταξινόμηση. Ο *Random Forest* παρουσίασε σταθερή απόδοση, ενώ ο *SVM (sigmoid)* αποδείχθηκε ότι ήταν λιγότερο αποτελεσματικός σε όλες τις περιπτώσεις.

7.2. Εκπαίδευση μοντέλων σε συνδυασμούς χαρακτηριστικών ανά δύο και ανά τρία

Πρώτο σύνολο δεδομένων

Στην παρούσα μελέτη, το χαρακτηριστικό RunLengthNonUniformity αναδείχθηκε ως το καλύτερο χαρακτηριστικό για τη διάκριση των ασθενών με πνευμονία από τους υγιείς μάρτυρες. Συγκεκριμένα, το χαρακτηριστικό αυτό αναδείχθηκε ως το πιο σημαντικό σε όλες τις αναλύσεις που πραγματοποιήθηκαν, τόσο σε δισδιάστατες όσο και σε τρισδιάστατες γραφικές παραστάσεις για κάθε ταξινομητή. Η σημασία του RunLengthNonUniformity τονίζεται από το γεγονός ότι ήταν το πιο διακριτό χαρακτηριστικό σε όλες τις αναπαραστάσεις δεδομένων, γεγονός που το καθιστά ιδανικό για τον διαχωρισμό των ομάδων.

Σε γενικές γραμμές, τα χαρακτηριστικά που σχετίζονται με το Run Length, όπως GrayLevelNonUniformity,DependenceNonUniformity,RunLengthNonUniformity,SizeZoneNonUniformity, αποδείχθηκαν τα πλέον αποτελεσματικά για την εκτέλεση αυτής της ταξινόμησης. Αυτά τα χαρακτηριστικά, αποτυπώνουν την ομοιομορφία των ακολουθιών γκρίζων επιπέδων μέσα σε μία εικόνα. Συγκεκριμένα, το RunLengthNonUniformity μετρά το βαθμό μη ομοιομορφίας των Run-Lengths αυτών των ακολουθιών. Όσο περισσότερο υγρό συγκεντρώνεται στους πνεύμονες των ασθενών, τόσο πιο ομοιόμορφη γίνεται η υφή των πνευμονικών ιστών, όπως παρατηρείται σε ακτινογραφίες ασθενών με πνευμονία. Αυτό σημαίνει ότι οι ασθενείς παρουσιάζουν χαμηλότερες τιμές μη ομοιομορφίας, καθώς οι ανωμαλίες στον πνευμονικό ιστό μειώνονται λόγω της συγκέντρωσης υγρού.

Αν και τα χαρακτηριστικά Correlation και ClusterShade επίσης είχαν καλές επιδόσεις, δεν ήταν τόσο αποδοτικά όσο τα χαρακτηριστικά Run Length που αναφέρθηκαν προηγουμένως. Το Correlation μετρά την γραμμική εξάρτηση μεταξύ των επιπέδων του γκρι, αποτυπώνοντας την ομοιογένεια ή την κανονικότητα στις σχέσεις μεταξύ των εικονοστοιχείων. Από την άλλη, το ClusterShade είναι ένα μέτρο του Skewness και της ομοιομορφίας του GLCM. Μια υψηλότερη τιμή της συνεπάγεται μεγαλύτερη ασυμμετρία ως προς τη μέση τιμή. Χρησιμοποιείται για να ανιχνεύει περιοχές ανωμαλιών ή συγκεντρώσεων ανωμαλιών στις εικόνες. Παρόλο που αυτά τα χαρακτηριστικά συνεισέφεραν στη διάκριση, δεν κατάφεραν να φτάσουν την απόδοση των άλλων χαρακτηριστικών GLCM.

Συνοψίζοντας, τα χαρακτηριστικά GLCM, με πιο σημαντικό το RunLengthNonUniformity, αποδείχθηκαν τα καλύτερα για την ταξινόμηση της πνευμονίας. Η ικανότητά τους να αποτυπώνουν την μη ομοιομορφία και τις ανωμαλίες στην υφή των εικόνων τους δίνει ένα σαφές πλεονέκτημα στη διάκριση των ασθενών με πνευμονία από τους υγιείς.

Η επιλογή αυτών των χαρακτηριστικών μπορεί να εξηγηθεί από το γεγονός ότι η COVID-19 πνευμονία επηρεάζει σημαντικά την υφή του πνεύμονα, προκαλώντας διαταραχές στη φυσιολογική ομοιογένεια των δομών. Οι βλάβες που προκαλεί ο ιός στον πνεύμονα, όπως η ίνωση και οι GGO προκαλούν αλλαγές που ανιχνεύονται μέσω αυτών των χαρακτηριστικών, επιτρέποντας στους ταξινομητές να διακρίνουν με μεγαλύτερη ακρίβεια τις παθολογικές από τις φυσιολογικές περιοχές. Οι διαφορές στην κατανομή αυτών των χαρακτηριστικών στα δεδομένα επιτρέπουν την καλύτερη ταξινόμηση.

Από το πείραμα PCA, τα χαρακτηριστικά που συνέβαλαν περισσότερο στη PCA2 ήταν το **Cluster Shade (**0.255137) και **Correlation (**0.253827) . Επιπλέον, τα χαρακτηριστικά **Gray Level Non-Uniformity** (0.351829), **Dependence Non-Uniformity** (0.348146), **Run Length Non-Uniformity**(0.347256) και **Size Zone Non-Uniformity** (0.345587) παρουσίασαν σημαντική διαφορά από τα υπόλοιπα, υπογραμμίζοντας τον κυρίαρχο ρόλο τους στη διακύμανση που αποτυπώθηκε από αυτά τα συστατικά.
Ωστόσο, λόγω της υψηλής συσχέτισης που παρατηρήθηκε μεταξύ των, επιλέξαμε μόνο το **Run Length Non-Uniformity** για τα μοντέλα μας, ώστε να αποφύγουμε την επανάληψη και προβλήματα πολυδιάστατης συσχέτισης.
Επιπλέον, τα χαρακτηριστικά **Energy** (0.219633) και **Imc** (Μέτρο Πληροφορίας Συσχέτισης) έδειξαν επίσης σημαντική συμβολή, επιτρέποντάς μας να βελτιστοποιήσουμε το σύνολο των χαρακτηριστικών χωρίς να χάσουμε σημαντικές πληροφορίες.

Δεύτερο σύνολο δεδομένων

Οι γραφικές παραστάσεις δείχνουν ότι τα κορυφαία χαρακτηριστικά που διακρίνουν τους ασθενείς με COVID-19 πνευμονία από τα άτομα ελέγχου είναι το **Energy**, το **RunLengthNonUniformity**, η **DifferenceEntropy** και το **ZonePercentage**. Αυτά τα χαρακτηριστικά επισημαίνονται λόγω της απόδοσής τους σε εργασίες ταξινόμησης, όπως φαίνεται από τις καμπύλες ROC και τα διαγράμματα διασποράς, και συνδυάζονται με μοντέλα μηχανικής μάθησης.

Το πρώτο βασικό χαρακτηριστικό είναι η **Ενέργεια**, η οποία αντιπροσωπεύει το άθροισμα των τετραγώνων των τιμών έντασης μέσα στην περιοχή ενδιαφέροντος. Μια υψηλότερη τιμή υποδεικνύει μεγαλύτερη ομοιομορφία στις εντάσεις των voxel. Οι πνευμονικοί ιστοί που επηρεάζονται από τον COVID-19 μπορεί να εμφανίζουν χαρακτηριστικά μοτίβα μεγαλύτερης ενέργειας λόγω αλλαγών στην ομοιογένεια του ιστού που προκαλούνται από φλεγμονή ή βλάβη. Το RunLengthNonUniformity αντικατοπτρίζει τη συνέπεια των διαδρομών γκρι επιπέδου προς μια συγκεκριμένη κατεύθυνση. Ποσοτικοποιεί τις ανωμαλίες στην υφή, που θα μπορούσαν να είναι σημαντικές για την ανίχνευση των χαρακτηριστικών αλλαγών στον ιστό λόγω της COVID-19 πνευμονίας.

Η **DifferenceEntropy** είναι ένα άλλο σημαντικό χαρακτηριστικό, το οποίο απεικονίζει την τυχαιότητα ή την πολυπλοκότητα της εικόνας, συγκεκριμένα τονίζοντας τις διαφορές έντασης ανάμεσα σε ζεύγη εικονοστοιχείων. Η αυξημένη πολυπλοκότητα θα μπορούσε να σηματοδοτεί ανωμαλίες που προκαλούνται από τη μόλυνση. Το **ZonePercentage** σχετίζεται με το μέγεθος των ομοιογενών ζωνών στην εικόνα. Χαμηλότερα ποσοστά θα μπορούσαν να υποδεικνύουν την αύξηση του μεγέθους των περιοχών που μολύνονται από την ασθένεια.

Οι καμπύλες ROC που προκύπτουν από πέντε πτυχές (cross-validation) δείχνουν τιμές AUC που κυμαίνονται από 0.70 έως 0.90, γεγονός που υπογραμμίζει τη σημασία αυτών των χαρακτηριστικών για τη διάκριση μεταξύ των δύο ομάδων.

Ένα επιπλέον σημαντικό χαρακτηριστικό, είναι το **ShortRunEmphasis**. Πρόκειται για ένα χαρακτηριστικό που μετρά την κατανομή των μικρών μηκών διαδρομής. Αυτή η μέτρηση είναι ευαίσθητη σε λεπτές υφές και μπορεί επίσης να είναι σημαντική για τον εντοπισμό αλλαγών που σχετίζονται με τον COVID-19, καθώς οι μικρές διαδρομές μπορεί να αυξάνονται σε ιστούς που έχουν επηρεαστεί από την πνευμονία. Η SRE αντικατοπτρίζει συνήθως περιοχές με επαναλαμβανόμενα, λεπτά πρότυπα που παρατηρούνται στους πνεύμονες των μολυσμένων ασθενών.

Μελλοντικές βελτιώσεις

Για μελλοντική εργασία, θα μπορούσαμε να συμπεριλάβουμε μια κατηγορία που αφορά την ήπια πνευμονία και να τη συγκρίνουμε με τα δεδομένα από το πρώτο μας σύνολο για να εξετάσουμε εάν μοιάζει με την ήπια μορφή της πνευμονίας.

Επιπλέον, θα μπορούσαμε να προσπαθήσουμε να κατηγοριοποιήσουμε την πνευμονία COVID-19 σε σύγκριση με άλλες μορφές πνευμονίας, ώστε να εξετάσουμε αν μπορούν αυτά τα μοντέλα να ανιχνεύσουν διαφορές που δεν φαίνονται με γυμνό μάτι.

Επίσης, θα μπορούσαμε να χρησιμοποιήσουμε μεθόδους βαθιάς μάθησης (deep learning) για τη δημιουργία μασκών (mask generation) και την εκπαίδευση των μοντέλων.

Μια πρόσθετη βελτίωση θα μπορούσε να είναι η εφαρμογή τεχνικών ενίσχυσης δεδομένων (data augmentation) για τη βελτίωση της γενίκευσης του μοντέλου μας. Παράλληλα, θα ήταν χρήσιμο να αξιολογήσουμε την απόδοση των μοντέλων σε μεγαλύτερα και πιο ετερογενή σύνολα δεδομένων για την καλύτερη κατανόηση της απόδοσης τους σε πραγματικές κλινικές συνθήκες.

Αναφορές - Πηγές

[1] Anon Lung Anatomy *Physiopedia*

[2] Shi Y, Wang G, Cai X, Deng J, Zheng L, Zhu H, Zheng M, Yang B and Chen Z 2020 An overview of COVID-19 *J. Zhejiang Univ. Sci. B* **21** 343–60

[3] Anon Coronavirus disease (COVID-19)

[4] Doernberg S B, Holubar M, Jain V, Weng Y, Lu D, Bollyky J B, Sample H, Huang B, Craik C S, Desai M, Rutherford G W and Maldonado Y 2022 Incidence and Prevalence of Coronavirus Disease 2019 Within a Healthcare Worker Cohort During the First Year of the Severe Acute Respiratory Syndrome Coronavirus 2 Pandemic *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am.* **75** 1573–84

[5] Miller L M S, Gee P M and Katz R A 2021 The Importance of Understanding COVID-19: The Role of Knowledge in Promoting Adherence to Protective Behaviors *Front. Public Health* **9** 581497

[6] CDC 2020 Healthcare Workers *Cent. Dis. Control Prev.*

[7] Anon 2023 What is Health Equity? | Health Equity | CDC

[8] van de Weijer M P, Pelt D H M, de Vries L P, Huider F, van der Zee M D, Helmer Q, Ligthart L, Willemsen G, Boomsma D I, de Geus E and Bartels M 2022 Genetic and environmental influences on quality of life: The COVID-19 pandemic as a natural experiment *Genes Brain Behav.* **21** e12796

[9] Anon 2021 COVID-19 symptoms: Timeline and progression

[10] Anon Coronavirus disease 2019 (COVID-19) - Diagnosis and treatment - Mayo Clinic

[11] Bellou V, Tzoulaki I, van Smeden M, Moons K G M, Evangelou E and Belbasis L 2022 Prognostic factors for adverse outcomes in patients with COVID-19: a field-wide systematic review and meta-analysis *Eur. Respir. J.* **59** 2002964

[12] Vivekanandhan K, Shanmugam P, Barabadi H, Arumugam V, Daniel Raj Daniel Paul Raj D, Sivasubramanian M, Ramasamy S, Anand K, Boomi P, Chandrasekaran B, Arokiyaraj S and Saravanan M 2021 Emerging Therapeutic Approaches to Combat COVID-19: Present Status and Future Perspectives *Front. Mol. Biosci.* **8**

[13] Churruca M, Martínez-Besteiro E, Couñago F and Landete P 2021 COVID-19 pneumonia: A review of typical radiological characteristics *World J. Radiol.* **13** 327–43

[14] Gillies R J, Kinahan P E and Hricak H 2016 Radiomics: Images Are More than Pictures, They Are Data *Radiology* **278** 563–77

[15] Kumar V, Gu Y, Basu S, Berglund A, Eschrich S A, Schabath M B, Forster K, Aerts H J W L, Dekker A, Fenstermacher D, Goldgof D B, Hall L O, Lambin P, Balagurunathan Y, Gatenby R A and Gillies R J 2012 QIN "Radiomics: The Process and the Challenges" *Magn. Reson. Imaging* **30** 1234–48

[16] Nabi J 2019 Machine Learning —Fundamentals *Medium*

[17] Anon Differences Between AI vs. Machine Learning vs. Deep Learning | Simplilearn *Simplilearn.com*

[18] Anon Applications of Machine Learning - Javatpoint

[19] Anon Principal Component Analysis (PCA) Explained *Built In*

[20] Brownlee J 2020 Recursive Feature Elimination (RFE) for Feature Selection in Python *MachineLearningMastery.com*

[21] Gupta A 2020 Feature Selection Techniques in Machine Learning *Anal. Vidhya*

[22] Anon Fig. 1. Heatmap plot representing the correlation matrix between... *ResearchGate*

[23] Peterson L E 2009 K-nearest neighbor *Scholarpedia* **4** 1883

[24] Zhang H The Optimality of Naive Bayes

[25] Anon 2023 What Is Linear Discriminant Analysis? | IBM
[26] Anon Neural Networks - A Comprehensive Foundation - Simon Haykin.pdf
[27] Anon 2021 What is Perceptron? A Beginners Guide for 2024 *Simplilearn.com*
[28] Khushaktov M F 2023 Introduction Random Forest Classification By Example *Medium*
[29] Anon All You Need to Know About Support Vector Machines *Spiceworks Inc*
[30] Anon Machine Learning Development Process: From Data Collection to Model Deployment
[31] Khanna S 2023 A Comprehensive Guide to Train-Test-Validation Split in 2024 *Anal. Vidhya*
[32] Anon What is A Confusion Matrix in Machine Learning? The Model Evaluation Tool Explained
[33] Anon Society of Thoracic Radiology
[34] Anon RSNA Pneumonia Detection Challenge
[35] Shih G, Wu C C, Halabi S S, Kohli M D, Prevedello L M, Cook T S, Sharma A, Amorosa J K, Arteaga V, Galperin-Aizenberg M, Gill R R, Godoy M C B, Hobbs S, Jeudy J, Laroia A, Shah P N, Vummidi D, Yaddanapudi K and Stein A 2019 Augmenting the National Institutes of Health Chest Radiograph Dataset with Expert Annotations of Possible Pneumonia *Radiol. Artif. Intell.* **1** e180041
[36] Anon youtube_channel/Python Tutorial Series/image_processing1.ipynb at main · lukepolson/youtube_channel *GitHub*