Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομέας Συστημάτων Μετάδοσης Πληροφορίας και Τεχνολογίας Υλικών

Διερεύνηση και Ταξινόμηση Σταδίων Καρκίνου του Μαστού βάσει Γονιδιακών Δεδομένων

Διπλωματικη Εργασια

της

ΑΓΓΕΛΙΚΗΣ ΕΜΜΑΝΟΥΕΛΑΣ ΣΥΡΡΗ

Επιβλέπων: Γεώργιος Ματσόπουλος
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2024

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομέας Συστημάτων Μετάδοσης Πληροφορίας και Τεχνολογίας Υλικών

Διερεύνηση και Ταξινόμηση Σταδίων Καρκίνου του Μαστού βάσει Γονιδιακών Δεδομένων

Διπλωματικη Εργασια

της

ΑΓΓΕΛΙΚΗΣ ΕΜΜΑΝΟΥΕΛΑΣ ΣΥΡΡΗ

Επιβλέπων: Γεώργιος Ματσόπουλος
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 15η Ιουλίου 2024.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Γεώργιος Ματσόπουλος	Αθανάσιος Δ. Παναγόπουλος	Παναγιώτης Τσανάκας
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2024

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομέας Συστημάτων Μετάδοσης Πληροφορίας και Τεχνολογίας Υλικών

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ενυπογράφως ότι είμαι αποκλειστικός συγγραφέας της παρούσας Πτυχιακής Εργασίας, για την ολοκλήρωση της οποίας κάθε βοήθεια είναι πλήρως αναγνωρισμένη και αναφέρεται λεπτομερώς στην εργασία αυτή. Έχω αναφέρει πλήρως και με σαφείς αναφορές, όλες τις πηγές χρήσης δεδομένων, απόψεων, θέσεων και προτάσεων, ιδεών και λεκτικών αναφορών, είτε κατά κυριολεξία είτε βάσει επιστημονικής παράφρασης. Αναλαμβάνω την προσωπική και ατομική ευθύνη ότι σε περίπτωση αποτυχίας στην υλοποίηση των ανωτέρω δηλωθέντων στοιχείων, είμαι υπόλογος έναντι λογοκλοπής, γεγονός που σημαίνει αποτυχία στην Πτυχιακή μου Εργασία και κατά συνέπεια αποτυχία απόκτησης του Τίτλου Σπουδών, πέραν των λοιπών συνεπειών του νόμου περί πνευματικών δικαιωμάτων. Δηλώνω, συνεπώς, ότι αυτή η Πτυχιακή Εργασία προετοιμάστηκε και ολοκληρώθηκε από εμένα προσωπικά και αποκλειστικά και ότι, αναλαμβάνω πλήρως όλες τις συνέπειες του νόμου στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής άλλης πνευματικής ιδιοκτησίας.

(Υπογραφή)

Αγγελική Εμμανουέλα Συρρή
Διπλωματούχα Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Ιούλιος 2024

Περίληψη

Η σταδιοποίηση του καρκίνου είναι η διαδικασία προσδιορισμού της ανάπτυξης και της εξάπλωσής του καρκίνου σε όλο το σώμα, και χρησιμεύει ως κρίσιμη προϋπόθεση στην κλινική πρακτική για τον σχεδιασμό της θεραπείας και την αξιολόγηση της πρόγνωσης. Οι ασθενείς που ταξινομούνται στο ίδιο στάδιο έχουν συνήθως παρόμοιες προγνώσεις και επωφελούνται από παρόμοια πλάνα θεραπείας. Παρά τις υπάρχουσες μεθόδους ανίχνευσης σταδίου καρκίνου, παρουσιάζονται περιορισμοί που καθιστούν αναγκαία την εξερεύνηση καινοτόμων προσεγγίσεων. Η παρούσα διπλωματική εργασία διερευνά την ανάπτυξη ενός μοντέλου βασισμένου σε γονιδιακά δεδομένα για την ταξινόμηση των παθολογικών σταδίων των ασθενών. Η εστίαση είναι στη διαφοροποίηση μεταξύ των σταδίων *I, II* και *III* του καρκίνου του μαστού. Στο πλαίσιο αυτό, εξετάστηκαν πέντε διαφορετικές μεθοδολογίες για την επίλυση του προβλήματος, χρησιμοποιώντας τόσο κλασσικούς αλγόριθμους μηχανικής μάθησης όσο και τεχνικές νευρωνικών δικτύων, τύπου Πολλαπλών Επιπέδων Perceptrons (MLP). Τα έμφυτα προβλήματα των βιοϊατρικών δεδομένων, και ειδικότερα των δεδομένων γονιδιακής έκφρασης, περιλαμβάνουν την υψηλή διαστασιμότητα και την άνιση κατανομή των κλάσεων. Για την αντιμετώπιση αυτών των προκλήσεων, χρησιμοποιούνται τεχνικές επιλογής χαρακτηριστικών, όπως η μετατροπή των γονιδιακών δεδομένων σε βιολογικά μονοπάτια με ανάλυση εμπλουτισμού συνόλων γονιδίων (GSE), καθώς και η χρήση των κλινικών μεταδεδομένων των ασθενών για τη μείωση της ετερογένειας εντός του συνόλου. Επιπλέον, για τον εμπλουτισμό του συνόλου εκπαίδευσης, εφαρμόζονται τεχνικές δημιουργίας συνθετικών δεδομένων. Παράλληλα, μια άλλη μέθοδος αφορά τον μετασχηματισμό του προβλήματος ταξινόμησης σταδίων σε πρόβλημα ταξινόμησης μεταβάσεων σταδίων. Τα αποτελέσματα της μελέτης υπογραμμίζουν τις εγγενείς προκλήσεις και τους περιορισμούς στην επίτευξη του επιθυμητού επιπέδου ακρίβειας και αξιοπιστίας στην ταξινόμηση σταδίων βάσει γονιδιωματικών δεδομένων. Παρά τις δυσκολίες, η προτεινόμενη μεθοδολογία μπορεί να επεκταθεί σε άλλα είδη καρκίνων που παρουσιάζουν μικρότερο βαθμό ετερογένειας σε σχέση με τον καρκίνο του μαστού. Συνοψίζοντας, η παρούσα εργασία προτείνει ένα σύνολο μοντέλων πρόγνωσης σταδίων του καρκίνου του μαστού που βασίζονται σε γονιδιακά δεδομένα και παράγωγα τους, αξιολογώντας την αποτελεσματικότητα και τους περιορισμούς διαφορετικών μεθοδολογιών και τεχνολογιών, και παρέχει μια βάση για μελλοντική έρευνα και βελτιώσεις των αλγορίθμων και των τεχνικών επιλογής χαρακτηριστικών.

Λέξεις Κλειδιά

Καρκίνος του Μαστού, Μηχανική Μάθηση, Νευρωνικά Δίκτυα, Εξέλιξη της νόσου, Σταδιοποίηση, Ταξινόμηση

Abstract

Cancer staging is the process of determining the growth and spread of cancer throughout the body. It serves as a critical prerequisite in clinical practice for treatment planning and prognosis assessment. Patients in the same stage share similar prognosis and hence, benefit from similar treatment plans. Despite existing methods for cancer staging detection, limitations persist, necessitating the exploration of innovative approaches. This thesis investigates the development of a model based on genomic data for classification of patients' pathological stages, focusing on distinguishing between stages I, II, and III of breast cancer. To address this problem, five different methodologies were examined, utilizing both classical machine learning algorithms and neural network techniques, such as Multi-Layer Perceptrons (MLP). The inherent challenges of biomedical data, particularly gene expression data, include high dimensionality and imbalanced class distribution. To tackle these challenges, feature engineering techniques are employed, such as transforming genomic data into biological pathways using Gene Set Enrichment Analysis (GSEA), and leveraging patients' clinical metadata to reduce heterogeneity within the dataset. Additionally, synthetic data generation techniques are applied to augment the training set. Another method tested involves the transformation of the problem into a stage transition classification task to refine the focus on changes between consecutive stages. The study's results highlight the inherent challenges and limitations in achieving the desired level of accuracy and reliability in stage classification based on genomic data. Despite the difficulties, the proposed methodology can be extended to other cancer types with lower heterogeneity compared to breast cancer. In summary, this thesis proposes a set of models for breast cancer staging classification based on genomic data and derived biological insights. It evaluates the effectiveness and limitations of various methodologies and technologies, offering a basis for future research and potential advancements in the field.

Keywords

Breast Cancer, Machine Learning, Neural Networks, Disease Progression, Staging, Classification

Στους γονείς μου, Έθελ και Γιώργο

Ευχαριστίες

Θα ήθελα να εκφράσω τις ειλικρινείς μου ευχαριστίες σε όσους συνέβαλαν στην ολοκλήρωση αυτής της διπλωματικής εργασίας και με υποστήριξαν κατά τη διάρκεια των σπουδών μου. Πρωτίστως, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα της διπλωματικής μου εργασίας κ. Γιώργο Ματσόπουλο, ο οποίος με ενθάρρυνε και ενέπνευσε να ασχοληθώ με τον τομέα της Βιοϊατρικής. Επίσης, θα ήθελα να ευχαριστήσω την κα. Ουρανία Πετροπούλου, η οποία συνέβαλε καθοριστικά στην πορεία της διπλωματικής και έκανε κάθε ανυπέρβλητο πρόβλημα να μοιάζει εύκολο. Θερμές ευχαριστίες απευθύνω και στον κ. Γιώργο Παλιούρα, για την εμπιστοσύνη που μου έδειξε, αναθέτοντάς μου ένα τόσο ενδιαφέρον και επίκαιρο θέμα στο Εργαστηρίο Τεχνολογίας Γνώσεων και Λογισμικού (SKEL | The AI lab) του Ινστιτούτου Πληροφορικής και Τηλεπικοινωνιών, Εθνικού Κέντρου Έρευνας Φυσικών Επιστημών «ΔΗΜΟΚΡΙΤΟΣ». Η επιτυχής ολοκλήρωση της διπλωματικής εργασίας οφείλεται σε μεγάλο βαθμό στην συνεργασία την οποία ανέπτυξαν μαζί μου, επιστήμονες και ερευνητές τους οποίους και ευχαριστώ θερμά. Ειδικότερα, ευχαριστώ πολύ για την συνεργασία τον κ. Νίκο Κατζούρη και τους ερευνητές Νικόλα και Βασίλη Μαγγίνα, οι οποίοι με καθοδήγησαν και με στήριξαν καθ' όλη την διάρκεια της ερευνητικής αυτής προσπάθειας. Βαθιά ευγνωμοσύνη θέλω να εκφράσω και προς τους φίλους μου που με στήριξαν σε κάθε βήμα της πορείας μου, από τα σχολικά μου χρόνια μέχρι και σήμερα. Η Κωνσταντίνα, η Ήλια, η Χριστιάνα και η Κατερίνα στάθηκαν δίπλα μου σε κάθε δύσκολη στιγμή, προσφέροντας ανεκτίμητη υποστήριξη. Η Δέσποινα, ο Δημήτρης, η Σοφία, η Άντα, η Χριστίνα και η Εύα γέμισαν τα φοιτητικά μου χρόνια με χαρά και αξέχαστες στιγμές, κάνοντάς τα πιο φωτεινά και όμορφα με τη φιλία και την αγάπη τους. Τέλος, θα ήθελα να εκφράσω την ευγνωμοσύνη μου προς την οικογένεια μου, τον αδελφό μου, Αλέξη, και ιδιαίτερα τους γονείς μου, Έθελ και Γιώργο, στους οποίους οφείλω όσα έχω καταφέρει μέχρι στιγμής, για την αμέριστη αγάπη και υποστήριξή τους που μου έχουν προσφέρει σε κάθε στάδιο της ζωής μου.

Αθήνα, Ιούνιος 2024

Αγγελική Εμμανουέλα Συρρή

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Γενικά

Ο καρκίνος του μαστού αποτελεί μια από τις κύριες αιτίες θανάτου μεταξύ γυναικών ανά τον κόσμο. Ευθύνεται για 670.000 θανάτους παγκοσμίως για το έτος 2022 [10] και το έτος 2024 αναμένεται ότι περίπου 310.720 γυναίκες θα διαγνωστούν με κακοήθη νεοπλασία στον μαστό μόνο στην Αμερική [11]. Ο αντίκτυπος της νόσου γίνεται ακόμα πιο αισθητός, αφού στατιστικά στοιχεία δείχνουν πως 1 στις 8 γυναίκες θα διαγνωστεί με καρκίνο του μαστού κάποια στιγμή στη ζωή της [10]. Τα υψηλά αυτά ποσοστά υπογραμμίζουν την επείγουσα ανάγκη για αποτελεσματικές διαγνωστικές και θεραπευτικές μεθόδους.

Μία από τις μεγαλύτερες προκλήσεις για την αντιμετώπιση του καρκίνου του μαστού είναι η ετερογένεια του. Ο καρκίνος του μαστού δεν είναι μια ομοιόμορφη νόσος, αλλά μια πολύπλοκη ομάδα ασθενειών με διάφορους ιστολογικούς υποτύπους, γενετικές μεταλλάξεις και μοριακά προφίλ. Αυτή η μεταβλητότητα οδηγεί σε διαφορές στον τρόπο με τον οποίο η νόσος εξελίσσεται και ανταποκρίνεται στη θεραπεία, καθιστώντας την μια μεγάλη πρόκληση για τους κλινικούς γιατρούς. Κατά συνέπεια, η ιατρική ακριβείας (Precision Medicine) έχει αναδειχθεί ως μια κρίσιμη προσέγγιση για την προσαρμογή των σχεδίων θεραπείας με βάση τα μεμονωμένα χαρακτηριστικά του ασθενούς, ενισχύοντας την πιθανότητα επιτυχούς καταπολέμησης της νόσου [12].

Παραδοσιακά, η μαστογραφία και η διαδερμική βιοψία αποτελούν τις κεντρικές διαγνωστικές τεχνικές για τον καρκίνο του μαστού. Η μαστογραφία, μια απεικονιστική μέθοδος και η διαδερμική βιοψία, μια ελάχιστα επεμβατική διαδικασία που εξάγει κύτταρα για εξέταση, χρησιμοποιούνται ευρέως λόγω της προσβασιμότητας και του σχετικά χαμηλού κόστους τους. Ωστόσο, αυτές οι τεχνικές έχουν περιορισμούς. Η μαστογραφία, αν και είναι χρήσιμη, μπορεί να μην αναγνωρίσει ορισμένους τύπους όγκων, ιδιαίτερα σε πυκνό ιστό μαστού, και η βιοψία μπορεί μερικές φορές να αποφέρει ασαφή αποτελέσματα λόγω ανεπαρκούς κυτταρικού υλικού. Αυτοί οι διαγνωστικοί περιορισμοί υπογραμμίζουν την ανάγκη για βελτιωμένες μεθόδους για την εξασφάλιση ακριβούς και έγκαιρης ανίχνευσης [13].

Επιπλέον, η έγκαιρη διάγνωση του καρκίνου του μαστού είναι κρίσιμη, καθώς αυξάνει σημαντικά τις πιθανότητες επιτυχούς θεραπείας και επιβίωσης του ασθενή. Όταν ο καρκίνος του μαστού διαγνωστεί σε πρώιμο στάδιο, το ποσοστό πενταετούς επιβίωσης για ασθενείς με καρκίνο του μαστού μπορεί να ξεπεράσει το 90% [14]. Αντίθετα, η διάγνωση σε τελευταίο στάδιο συσχετίζεται συχνά με χαμηλότερα ποσοστά επιβίωσης και πιο επιθετικές μεθόδους

θεραπείας. Ως εκ τούτου, η ενίσχυση των δυνατοτήτων έγκαιρης ανίχνευσης αποτελεί πρωταρχικό στόχο στην καταπολέμηση του καρκίνου του μαστού.

Παράλληλα, οι ραγδαίες εξελίξεις στον τομέα της τεχνητής νοημοσύνης σε συνδυασμό με τις εξελίξεις στη βιοϊατρική απεικόνιση και την μοντελοποίηση έχουν συμβάλει καθοριστικά στην ανάπτυξη της εξατομικευμένης ιατρικής για την διάγνωση και την θεραπεία του καρκίνου του μαστού. Οι αλγόριθμοι μηχανικής μάθησης, ένας κλάδος της τεχνητής νοημοσύνης, διαπρέπουν στην ανάλυση τεράστιων ποσοτήτων ιατρικών δεδομένων, στον εντοπισμό περίπλοκων προτύπων και στην πρόβλεψη των αποτελεσμάτων της νόσου με υψηλή ακρίβεια, προσφέροντας πρωτοφανή υποστήριξη στη μάχη κατά του καρκίνου του μαστού [15].

Με αυτόν τον τρόπο και αξιοποιώντας δεδομένα ασθενών, συμπεριλαμβανομένων γενετικών, μοριακών και απεικονιστικών πληροφοριών, οι υγειονομικοί μπορούν να σχεδιάσουν προσαρμοσμένα σχέδια θεραπείας, τα οποία είναι πιο αποτελεσματικά και λιγότερο επεμβατικά. Αυτή η προσέγγιση όχι μόνο βελτιώνει τα αποτελέσματα των ασθενών αλλά και ελαχιστοποιεί τις περιττές θεραπείες και τις σχετικές παρενέργειες.

Συμπερασματικά, η αντιμετώπιση της νόσου του καρκίνου του μαστού μέσω βελτιωμένων διαγνωστικών και θεραπευτικών στρατηγικών είναι υψίστης σημασίας. Η ετερογένεια της νόσου, σε συνδυασμό με τους περιορισμούς των παραδοσιακών διαγνωστικών μεθόδων, καθιστά αναγκαία τη μετάβαση προς την ιατρική ακριβείας. Η τεχνολογική πρόοδος είναι ζωτικής σημασίας για την επίτευξη αυτών των στόχων. Αγκαλιάζοντας αυτές τις καινοτομίες, είναι δυνατό να ενισχυθεί η έγκαιρη ανίχνευση και η προσαρμογή των θεραπειών σε μεμονωμένους ασθενείς και τελικά να μειωθεί η θνησιμότητα και η νοσηρότητα που σχετίζεται με τον καρκίνο του μαστού.

1.2 Αντικείμενο της Εργασίας

Η σταδιοποίηση του καρκίνου του μαστού είναι ένα κρίσιμο βήμα για τον αποτελεσματικό σχεδιασμό θεραπείας και διάγνωσης ενός ασθενούς. Επί του παρόντος, οι πιο κοινές μέθοδοι για την ανίχνευση σταδίων καρκίνου του μαστού περιλαμβάνουν συνδυασμό κλινικής εξέτασης, απεικονιστικών τεχνικών όπως η μαστογραφία και η μαγνητική τομογραφία και η παθολογική αξιολόγηση μέσω βιοψιών. Παρά την πρόοδο σε αυτά τα διαγνωστικά εργαλεία, ο ακριβής προσδιορισμός του σταδίου του καρκίνου του μαστού παραμένει ένα σύνθετο και απαιτητικό έργο για τους ειδικούς.

Μία από τις κύριες δυσκολίες στην ανίχνευση σταδίου προκύπτει από την ετερογενή φύση του καρκίνου του μαστού. Η ασθένεια περιλαμβάνει μια μεγάλη ποικιλία υποτύπων, ο καθένας με ξεχωριστά γενετικά και μοριακά χαρακτηριστικά. Αυτή η ποικιλομορφία περιπλέκει τη διαδικασία σταδιοποίησης, καθώς διαφορετικοί υποτύποι μπορεί να παρουσιάζουν παρόμοια κλινικά χαρακτηριστικά, ταυτόχρονα όμως να ποικίλλουν σημαντικά ως προς την πρόγνωση και την ανταπόκρισή τους στη θεραπεία. Κατά συνέπεια, οι γιατροί αντιμετωπίζουν συχνά προκλήσεις στη διάκριση μεταξύ πρώιμων και προχωρημένων σταδίων της νόσου, οδηγώντας σε πιθανές καθυστερήσεις στην έναρξη της κατάλληλης θεραπείας ή σε περιττές επιθετικές παρεμβάσεις.

Δεδομένων αυτών των προκλήσεων, η ανάπτυξη μιας μεθόδου για την ταξινόμηση σταδίων με βάση γονιδιωματικά δεδομένα έχει μεγάλη σημασία. Τα γονιδιωματικά δεδομένα

παρέχουν μια ολοκληρωμένη εικόνα για το μοριακό προφίλ του καρκίνου του μαστού. Αξιοποιώντας αυτές τις πληροφορίες, είναι δυνατό να αναπτυχθούν πιο ακριβή και αξιόπιστα προγνωστικά μοντέλα που μπορούν να βελτιώσουν την ακρίβεια της σταδιοποίησης του καρκίνου του μαστού.

Επομένως, η παρούσα διπλωματική εργασία διερευνά τις τρέχουσες δυνατότητες δημιουργίας ενός τέτοιου προγνωστικού μοντέλου. Η έρευνα επικεντρώνεται στην αξιολόγηση διαφόρων τεχνικών μηχανικής χαρακτηριστικών για τη μείωση της διάστασης των γονιδιωματικών δεδομένων και την επιλογή των πιο συναφών χαρακτηριστικών. Αυτό το βήμα είναι κρίσιμο, καθώς τα γονιδιωματικά δεδομένα είναι συνήθως υψηλών διαστάσεων και ο εντοπισμός των πιο ωφέλιμων χαρακτηριστικών μπορεί να βελτιώσει σημαντικά την απόδοση του μοντέλου.

Μετά τη διαδικασία προεπεξεργασίας των δεδομένων η έρευνα αναπτύσσει και αξιολογεί πέντε διαφορετικές μεθοδολογίες χρησιμοποιώντας τεχνικές μηχανικής μάθησης (Machine Learning, ML) και βαθιάς μάθησης (Deep Learning, DL). Κάθε μοντέλο επικυρώνεται για την αξιολόγηση της απόδοσής του στην ταξινόμηση σταδίων καρκίνου του μαστού με βάση τα γονιδιωματικά χαρακτηριστικά. Η συγκριτική ανάλυση των μεθοδολογιών παρέχει πληροφορίες για τα πλεονεκτήματα και τους περιορισμούς τους, επισημαίνοντας τις πιο υποσχόμενες προσεγγίσεις για μελλοντική ανάπτυξη και βελτίωση.

1.3 Συνεισφορά

Η παρούσα διπλωματική εργασία αποτελεί μια εκτενή μελέτη αφιερωμένη στην ανάπτυξη ενός προηγμένου προγνωστικού μοντέλου ικανού να διακρίνει με ακρίβεια τα στάδια του καρκίνου του μαστού χρησιμοποιώντας γονιδιωματικά δεδομένα. Ο πρωταρχικός στόχος είναι η αξιοποίηση των δυνατοτήτων των μεθοδολογιών αιχμής στους τομείς ML και DL, όπου μέσω της χρήσης δεδομένων γονιδιακής έκφρασης και παράγωγων τους τα υποψήφια μοντέλα θα είναι σε θέση να διακρίνουν τα σταδία του καρκίνου του μαστού.

Για να επιτευχθεί αυτό, έχει δοκιμαστεί και αξιολογηθεί ένα ολοκληρωμένο σύνολο τεχνικών ML και DL. Αυτές οι μεθοδολογίες περιλαμβάνουν κλασσικούς αλγόριθμους, όπως Gradient Boosting, Support Vector Machines (SVM), Decision Trees (DT) και μεθόδους συνόλου όπως Random Forests (RF), καθώς και πιο εξελιγμένες αρχιτεκτονικές βαθιάς μάθησης, όπως Πολλαπλών Επιπέδων Perceptron (MLP). Επιπλέον, διερευνήθηκαν τεχνικές, όπως η δημιουργία συνθετικών δεδομένων, αλλά και ο μετασχηματισμός του ζητήματος σε πρόβλημα δυαδικής ταξινόμησης για τη βελτίωση της απόδοσης του μοντέλου.

Παρά την αναλυτική προσέγγιση, τα αποτελέσματα αποκάλυψαν αρκετές προκλήσεις και περιορισμούς στην επίτευξη του επιθυμητού επιπέδου ακρίβειας και αξιοπιστίας. Αυτή η έρευνα συμβάλλει στο πεδίο παρέχοντας πολύτιμες γνώσεις σχετικά με αυτές τις προκλήσεις και εντοπίζοντας τομείς για περαιτέρω βελτίωση.

Συγκεκριμένα, η εργασία αυτή συμβάλλει στην:

- **Επισήμανση των διαγνωστικών προκλήσεων και των δυνατοτήτων των γονιδιακών δεδομένων να παρέχουν πληροφορίες για την εξέλιξη της νόσου**: Η μελέτη υπογραμμίζει τις εγγενείς δυσκολίες στην ταξινόμηση σταδίων καρκίνου του μαστού

χρησιμοποιώντας γονιδιακά δεδομένα. Παρέχει μια ανάλυση για το πού υπολείπονται τα κλασσικά μοντέλα ML και τα μοντέλα DL, προσφέροντας μια ρεαλιστική αξιολόγηση των τρεχουσών δυνατοτήτων και περιορισμών.

- **Αξιολόγηση Μεθόδων**: Η παρούσα εργασία προσφέρει μια κριτική αξιολόγηση των αποτελεσμάτων των μεθόδων ML και DL, παρατηρώντας την απόδοσή τους στο πλαίσιο της σταδιοποίησης του καρκίνου του μαστού. Η συγκριτική ανάλυση εμπλουτίζει την κατανόηση του τρόπου με τον οποίο διαφορετικοί αλγόριθμοι χειρίζονται τα γενετικά δεδομένα και τις συγκεκριμένες προκλήσεις που αντιμετωπίζουν.

- **Ενίσχυση Μεθοδολογικών Προσεγγίσεων**: Με τον πειραματισμό με συνθετικά δεδομένα και τον μετασχηματισμό του προβλήματος πολλαπλών κλάσεων σε δυαδικό, αυτή η έρευνα διερευνά εναλλακτικές στρατηγικές για τη βελτίωση των αποτελεσμάτων του μοντέλου. Αν και αυτές οι τεχνικές δεν κατάφεραν να σχηματίσουν ισχυρά προγνωστικά μοντέλα, συνέβαλαν στην βελτίωση των αναπτυσσόμενων μοντέλων και παρέχουν τη βάση για επαναληπτικό πειραματισμό και βελτιστοποίηση.

- **Διευκόλυνση περαιτέρω έρευνας**: Οι γνώσεις που αποκτήθηκαν από αυτήν τη μελέτη ανοίγουν το δρόμο για μελλοντικές ερευνητικές προσπάθειες. Καταγράφοντας τόσο τις επιτυχίες όσο και τις ελλείψεις, αυτή η εργασία ενθαρρύνει τις μετέπειτα μελέτες να βασιστούν σε αυτά τα ευρήματα, να εξερευνήσουν νέες μεθοδολογίες και να βελτιώσουν τους υπάρχοντες αλγόριθμους για την καλύτερη αντιμετώπιση της πολυπλοκότητας της σταδιοποίησης του καρκίνου του μαστού.

1.4 Δομή και Οργάνωση

Η παρούσα έρευνα είναι οργανωμένη σε επτά κεφάλαια, καθένα από τα οποία καλύπτει ένα σημαντικό μέρος της μελέτης:

- Στο **Κεφάλαιο 2** παρατίθεται το υπόβαθρο για την κατανόηση του καρκίνου του μαστού συμπεριλαμβανομένης της ανατομίας του μαστού, της βιολογίας και της ετερογένειας του καρκίνου, της σταδιοποίησης της νόσου, καθώς και της γονιδιωματικής ανάλυσης και μοντελοποίησης της εξέλιξής του. Ακόμη, περιέχεται μια εκτεταμένη βιβλιογραφική ανασκόπηση αναφορικά με την έρευνα που έχει γίνει στον τομέα του καρκίνου του μαστού με χρήση γονιδιακών δεδομένων.

- Στο **Κεφάλαιο 3** αναλύεται το θεωρητικό υπόβαθρο της Τεχνητής Νοημοσύνης, καλύπτοντας αλγόριθμους Μηχανικής και Βαθιάς Μάθησης, μετρικές απόδοσης, καθώς και τη διαχείριση μη ισορροπημένων και μεγάλων διαστάσεων δεδομένων.

- Στο **Κεφάλαιο 4** περιγράφονται αναλυτικά τα δεδομένα που χρησιμοποιήθηκαν στην παρούσα έρευνα, η επεξεργασία τους, καθώς και οι μέθοδοι που χρησιμοποιήθηκαν για την πρόβλεψη του καρκίνου του μαστού βάσει γονιδιακής έκφρασης.

- Στο **Κεφάλαιο 5** παραβάλλεται ένα σύνολο μεθοδολογιών που δοκιμάστηκαν για την υλοποίηση υποψήφιων μοντέλων πρόβλεψης.

- Στο **Κεφάλαιο 6** παρουσιάζονται και σχολιάζονται αναλυτικά τα αποτελέσματα των μεθοδολογιών που δοκιμάστηκαν.
- Στο **Κεφάλαιο 7** συνοψίζονται τα συμπεράσματα της διπλωματικής εργασίας και προτείνονται κατευθύνσεις για μελλοντική έρευνα.

Κεφάλαιο 2

Καρκίνος του Μαστού και Ανάλυση Γονιδιωματικών Δεδομένων

Στο κεφάλαιο αυτό ορίζονται οι βασικές έννοιες γύρω από τον καρκίνο του μαστού. Ακολούθως, παρατίθεται το σύστημα σταδιοποίησης Tumor-Node-Metastasis (TNM) και περιγράφεται το γονιδίωμα, η γονιδιακή έκφραση, οι βιοδείκτες και ο ρόλος που έχουν στην εξέλιξη και καταγραφή της νόσου. Στο τέλος του κεφαλαίου, παρουσιάζεται εκτεταμένη βιβλιογραφική ανασκόπηση αναφορικά με την έρευνα που έχει γίνει στον τομέα μοντελοποίησης του καρκίνου του μαστού με χρήση γονιδιακών δεδομένων.

2.1 Καρκίνος του Μαστού και Σταδιοποίηση

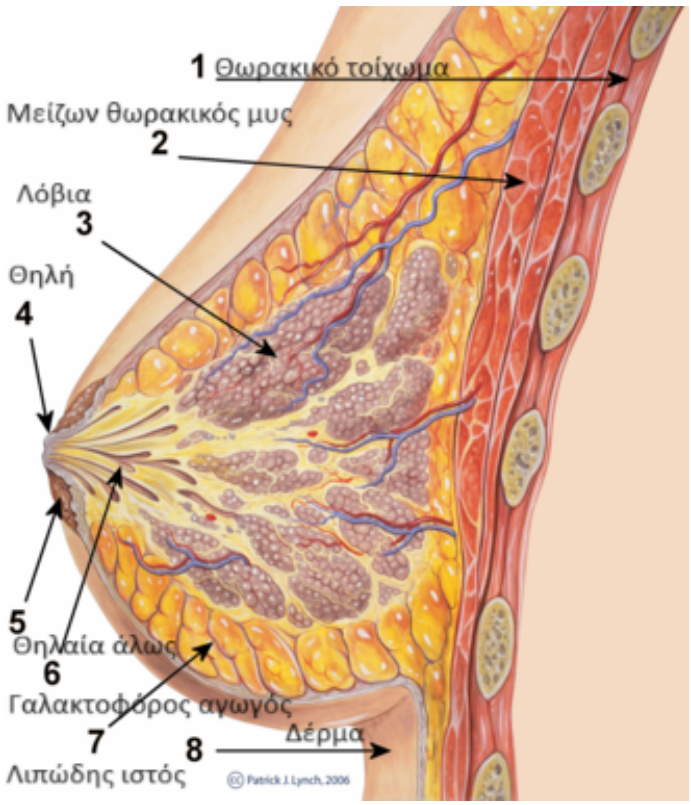
Η παρούσα έρευνα επικεντρώνεται στον καρκίνο του μαστού στις γυναίκες, καθώς αυτός είναι ο πιο συχνά αναφερόμενος στη βιβλιογραφία [16]. Σε αυτήν την ενότητα, θα παρατεθεί η ανατομία του μαστού και θα μελετηθεί η ανάπτυξη του καρκίνου σε αυτήν την περιοχή. Στη συνέχεια, θα παρουσιαστεί το σύστημα σταδιοποίησης TNM, το οποίο χρησιμοποιείται ευρέως στη διάγνωση του καρκίνου του μαστού και αποτελεί την κοινή γλώσσα επικοινωνίας μεταξύ των επαγγελματιών υγείας για τον χαρακτηρισμό των σταδίων.

2.1.1 Ανατομία του Μαστού

Η διερεύνηση και η εξοικείωση με την ανατομία του μαστού είναι ιδιαίτερα σημαντική για την κατανόηση των νόσων που αφορούν το συγκεκριμένο όργανο, αλλά και την ανάλυση της εξέλιξης τους. Ο μαστός είναι ένα σύνθετο όργανο που αποτελείται κυρίως από αδενικό, λιπώδη και συνδετικό ιστό. Βρίσκεται στο μπροστινό μέρος του θωρακικού τοιχώματος εκατέρωθεν του οστού του μαστού ή του στέρνου με συνδέσμους. Στηρίζεται στον μείζονα θωρακικό μυ και εκτείνεται από τη δεύτερη έως την έκτη πλευρά και από το στέρνο έως τη μέση μασχαλιαία γραμμή.

Κάθε μαστός αποτελείται από περίπου 10-20 απλούς αδένες. Ο αδένας είναι ένα ζωικό όργανο που παράγει κάποια ουσία, η οποία είναι χρήσιμη για τον οργανισμό. Ανάλογα με το που εκκρίνουν τις ουσίες που παράγουν οι αδένες διακρίνονται σε ενδοκρινείς, εξωκρινείς και μικτούς. Έτσι, ο μαστός, του οποίου παράγωγο είναι το πρωτόγαλα, εντάσσεται στους εξωκρινής αδένες.

Ο μαστός υφίσταται σημαντικές αλλαγές καθ' όλη τη διάρκεια της ζωής μιας γυναίκας, ιδίως κατά την εφηβεία, την εγκυμοσύνη και τη γαλουχία. Ο αδενικός ιστός του μαστού είναι οργανωμένος σε 15 έως 20 λοβούς (lobes), καθένας από τους οποίους αποτελείται από λοβίδια (lobules) που περιέχουν συστάδες γαλακτοπαραγωγών αδένων που ονομάζονται κυψελίδες. Αυτά τα λοβίδια συνδέονται με γαλακτοφόρους πόρους ή κανάλια (ducts), οι οποίοι μεταφέρουν το γάλα από τις κυψελίδες στη θηλή κατά τη διάρκεια της γαλουχίας. Το σύστημα των πόρων είναι ιδιαίτερα διακλαδισμένο, σχηματίζοντας ένα δίκτυο σε όλο τον ιστό του μαστού. Τους πόρους και τα λοβίδια περιβάλλει υποστηρικτικός συνδετικός ιστός γνωστός ως στρώμα, ο οποίος περιέχει αιμοφόρα αγγεία, λεμφαγγεία και νεύρα. Οι αγωγοί καταλήγουν στη θηλή [17].



Σχήμα 2.1: *Ανατομία μαστού*

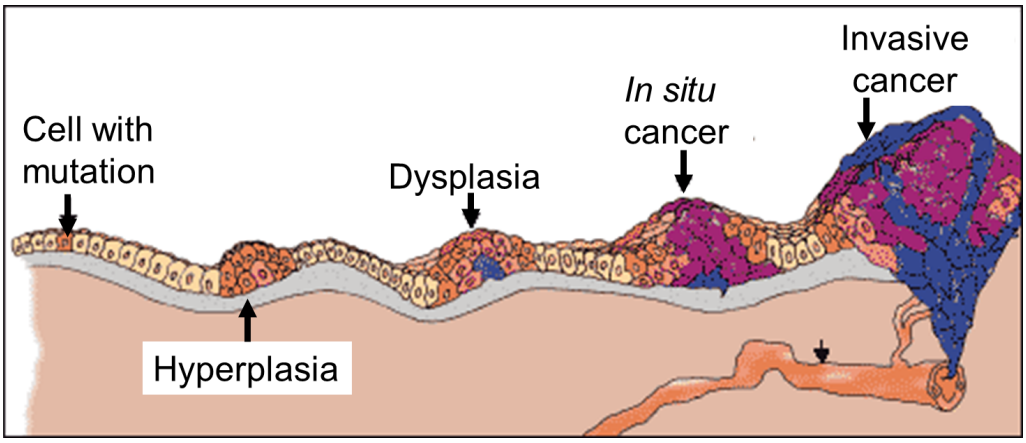
Ο μαστός ανταποκρίνεται σε μια πολύπλοκη αλληλεπίδραση ορμονών που προκαλούν την ανάπτυξη, και την παραγωγή γάλακτος. Οι τρεις κύριες ορμόνες που επηρεάζουν τον μαστό είναι τα οιστρογόνα, η προγεστερόνη και η προλακτίνη, οι οποίες προκαλούν αλλαγές στον αδενικό ιστό του μαστού και της μήτρας κατά τη διάρκεια του εμμηνορροϊκού κύκλου.

Σημαντικό είναι επίσης πως ο μαστός δεν έχει μυϊκό ιστό. Ένα στρώμα λίπους περιβάλλει τους αδένες και εκτείνεται σε όλο τον μαστό [18]. Η ποσότητα του λιπώδους ιστού στο μαστό ποικίλλει μεταξύ των ατόμων. Η θηλή, που βρίσκεται στο κέντρο της θηλαίας άλω, περιβάλλεται από λείες μυϊκές ίνες, γνωστές ως σύμπλεγμα θηλής-θηλαίας άλω, οι οποίες συστέλλονται κατά τη διάρκεια του θηλασμού για να διευκολύνουν την έκκριση του γάλακτος.

Συνολικά, η περίπλοκη ανατομία του μαστού επιτρέπει τις πρωταρχικές λειτουργίες του, δηλαδή την παραγωγή γάλακτος και τον θηλασμό.

2.1.2 Κατανόηση του Καρκίνου: Βιολογία και Ετερογένεια

Ως καρκίνο ορίζουμε μια ομάδα περισσότερων από 100 ασθενειών που αναπτύσσονται με την πάροδο του χρόνου και αφορούν την ανεξέλεγκτη διαίρεση των κυττάρων του σώματος. Ο καρκίνος ξεκινά από ένα κύτταρο, το οποίο ξεφεύγει από τους φυσιολογικούς περιορισμούς της κυτταρικής διαίρεσης και πολλαπλασιάζεται με δικούς του ρυθμούς λόγω κάποιας μετάλλαξης. Όλα τα κύτταρα που παράγονται από την διαίρεση αυτού του πρώτου, προγονικού κυττάρου, οι απόγονοι του δηλαδή, εμφανίζουν επίσης ακατάλληλο πολλαπλασιασμό. Έτσι καταλήγει να υπάρχει στον οργανισμό μια μάζα κυττάρων ή ένας όγκος που απαρτίζεται από αυτά τα μη φυσιολογικά κύτταρα, που έχουν ως κοινό χαρακτηριστικό τον ακανόνιστο πολλαπλασιασμό τους. Για να αναπτυχθεί ένας όγκος αναγκαία είναι και η ύπαρξη και άλλων μεταλλάξεων που θα συμβάλλουν στην δημιουργία ανώμαλων κυττάρων. Έτσι, τα κύτταρα αυτά και οι απόγονοι τους καταλήγουν να είναι μη φυσιολογικοί τόσο στην ανάπτυξη όσο και στην εμφάνιση τους. Το πλήθος των μεταλλάξεων που απαιτείται για τη δημιουργία των όγκων δεν είναι σταθερό και ούτε κοινό σε κάθε περίπτωση [1].



Σχήμα 2.2: *Στάδια ανάπτυξης καρκίνου [1]*

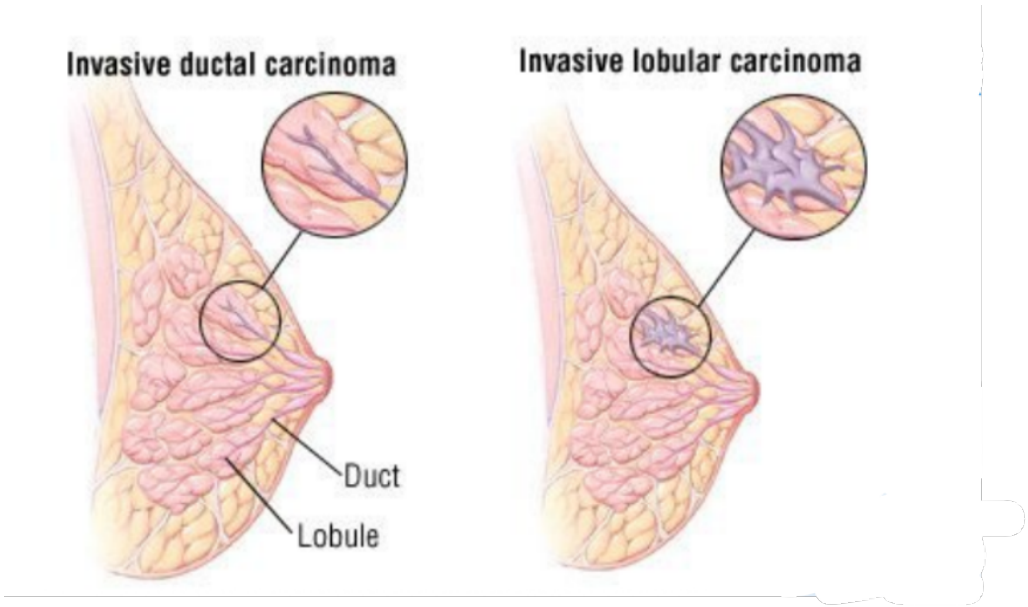
Η μάζα που δημιουργείται μπορεί είτε να παραμείνει εντός του ίδιου ιστού και άρα γίνεται λόγος για **μη διηθητικό καρκίνο** (in situ) ή μπορεί να αρχίσει να εισβάλει σε κοντινούς ιστούς, όπου τότε γίνεται αναφορά για **διηθητικό καρκίνο**. Ο διηθητικός όγκος είναι κακοήθης και τα κύτταρα που αποβάλλονται στο αίμα ή τη λέμφο από έναν κακοήθη όγκο είναι πιθανό να δημιουργήσουν νέους όγκους (μεταστάσεις) σε όλο το σώμα. Οι όγκοι απειλούν τη ζωή ενός ατόμου, όταν η ανάπτυξή τους διαταράσσει τους ιστούς και τα όργανα που απαιτούνται για την επιβίωση.

Ο καρκίνος μπορεί να αναπτυχθεί σχεδόν σε οποιονδήποτε ιστό του σώματος και κάθε τύπος καρκίνου έχει μοναδικά χαρακτηριστικά, ωστόσο ο τρόπος με τον οποίο εξελίσσεται ο καρκίνος, όπως περιγράψαμε παραπάνω, είναι αρκετά παρόμοιος σε όλες τις μορφές της νόσου.

Αντικείμενο της παρούσας διπλωματικής είναι η μελέτη του καρκίνου που αναπτύσσεται στον μαστό. Αυτό μπορεί να αφορά είτε τον έναν είτε και τους δύο μαστούς. Ο καρκίνος του μαστού μπορεί να αναπτυχθεί σε διαφορετικά σημεία του μαστού και από εκεί προκύπτουν οι διαφορετικοί ιστολογικοί τύποι.

Ο καρκίνος του μαστού συναντάται πιο συχνά:

- στους γαλακτοφόρους πόρους και ο ιστολογικός τύπος εκεί ονομάζεται πορογενής καρκίνος (ductal carcinoma) και,
- στα λοβίδια, από όπου ξεκινά ο λοβιακός καρκίνος (lobular carcinoma).



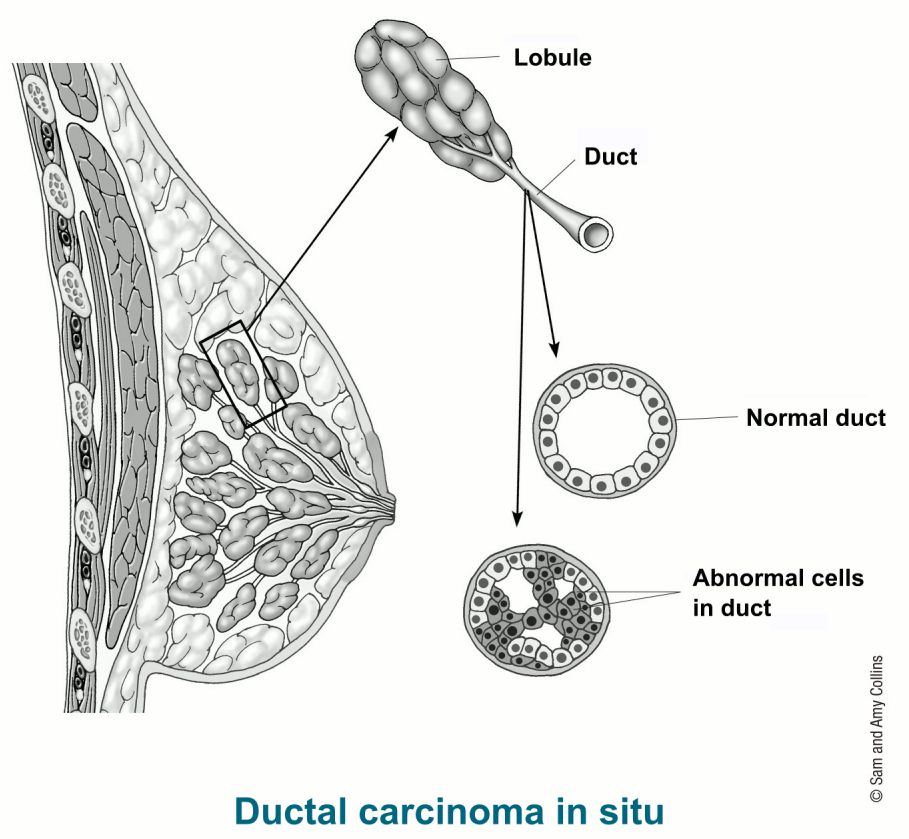
Σχήμα 2.3: *Παθολογία Καρκίνου του Μαστού: Πορογενές και Λοβιακό καρκίνωμα*

Ο καρκίνος του μαστού χαρακτηρίζεται και περαιτέρω ανάλογα με την εξάπλωση του ή μη. Παρακάτω εντοπίζονται και κατηγοριοποιούνται οι κύριες ομάδες τύπων καρκίνων λαμβάνοντας υπόψη και τους προαναφερθέντες τύπους, σημεία έναρξης ουσιαστικά της νόσου.

Μη διηθητικό καρκίνωμα

Περίπου 1 στους 5 νέους καρκίνους του μαστού είναι μη διηθητικός πορογενής καρκίνος (DCIS) και πρόκειται για έναν μη επεμβατικό ή προ-επεμβατικό καρκίνο του μαστού. Αυτό σημαίνει ότι τα κύτταρα που βρίσκονται στους πόρους έχουν μετατραπεί σε καρκινικά κύτταρα, αλλά δεν έχουν εξαπλωθεί μέσω των τοιχωμάτων των πόρων στον κοντινό ιστό του μαστού. Επειδή το DCIS δεν έχει εξαπλωθεί στον ιστό του μαστού γύρω του, δεν μπορεί να κάνει μεταστάσεις πέρα από το μαστό σε άλλα μέρη του σώματος. Ωστόσο, το DCIS μπορεί μερικές φορές να μετατραπεί σε διηθητικό καρκίνο [19]. Τότε, ο καρκίνος έχει εξαπλωθεί από τον πόρο στον κοντινό ιστό και από εκεί μπορεί να οδηγήσει σε μετάσταση σε άλλα μέρη του σώματος.

Ο μη διηθητικός λοβιακός καρκίνος (LCIS) είναι ένας τύπος in situ καρκινώματος του μαστού. Ωστόσο, δεν είναι όλες οι αλλοιώσεις, διογκώσεις ή σκληρύνσεις καρκινικές και κακοήθεις. Ενώ το DCIS θεωρείται προ-καρκίνος, δεν είναι σαφές αν το LCIS είναι προ-καρκίνος ή αν είναι απλώς ένας γενικός παράγοντας κινδύνου για την ανάπτυξη καρκίνου του μαστού. Αυτό προκύπτει από παρατηρήσεις, όπου χωρίς θεραπεία, σε ελάχιστες περιπτώσεις το LCIS έχει εξελιχθεί σε διηθητικό καρκίνο. Επειδή δεν είναι σαφές αν το LCIS είναι προκαρκινικό συχνά γίνεται αναφορά σε αυτόν ως λοβιακή νεοπλασία [20].



Σχήμα 2.4: *Μη διηθητικό πορογενές καρκίνωμα [2]*

Διηθητικό καρκίνωμα

Οι καρκίνοι του μαστού που έχουν εξαπλωθεί στον περιβάλλοντα ιστό του μαστού είναι γνωστοί ως διηθητικοί καρκίνοι.

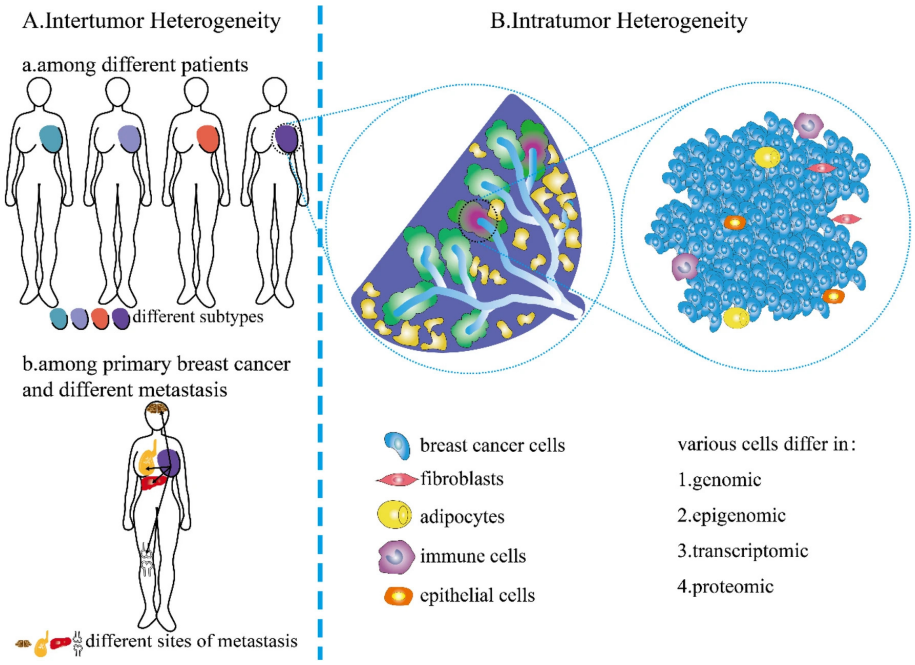
Ο διηθητικός πορογενής καρκίνος (IDC) είναι ο πιο κοινός τύπος καρκίνου του μαστού. Περίπου 4 στους 5 διηθητικούς καρκίνους του μαστού είναι καρκινώματα των πόρων (IDC) [2]. Το διηθητικό πορογενές καρκίνωμα ξεκινά από τα κύτταρα που βρίσκονται στο γαλακτοφόρο πόρο του μαστού. Από εκεί, ο καρκίνος διαπερνά το τοίχωμα του πόρου και αναπτύσσεται στους κοντινούς ιστούς του μαστού. Σε αυτό το σημείο, μπορεί να είναι σε θέση να εξαπλωθεί (δηλαδή να κάνει μετάσταση) σε άλλα μέρη του σώματος μέσω του λεμφικού συστήματος και της κυκλοφορίας του αίματος.

Περίπου 1 στους 5 διηθητικούς καρκίνους του μαστού είναι διηθητικό λοβιακό καρκίνωμα (ILC). Το ILC ξεκινά από τους αδένες του μαστού που παράγουν γάλα (λοβίδια). Όπως και το IDC, μπορεί να εξαπλωθεί σε άλλα μέρη του σώματος. Το διηθητικό λοβιακό καρκίνωμα μπορεί να είναι πιο δύσκολο να εντοπιστεί με τη φυσική εξέταση και την απεικόνιση, όπως η μαστογραφία, από το διηθητικό καρκίνωμα του πόρου. Σε σύγκριση με άλλα είδη διηθητικού καρκινώματος, είναι πιο πιθανό να προσβάλει και τους δύο μαστούς. Περίπου 1 στις 5 γυναίκες με ILC ενδέχεται να έχουν καρκίνο και στους δύο μαστούς κατά τη στιγμή της διάγνωσης.

Η παραπάνω ταξινόμηση του καρκίνου του μαστού σε διηθητικό και μη διηθητικό, ενώ χρήσιμη, συχνά δεν επαρκεί για μια ολοκληρωμένη διάγνωση και κατάλληλη θεραπεία. Ο

καρκίνος του μαστού θεωρείται μια από τις πιο ετερογενείς μορφές καρκίνου στις γυναίκες.

Η ετερογένεια της νόσου δύναται να εντοπιστεί τόσο εντός του όγκου (intratumor heterogeneity), που αναφέρεται στις διαφορές που εμφανίζει ο καρκινικός ιστός στον ίδιο όγκο στον ίδιο ασθενή, όσο και μεταξύ διαφορετικών όγκων ή μεταστάσεων μεταξύ διαφορετικών ασθενών (intertumor heterogeneity) [3]. Αυτές οι διαφορές πηγάζουν από τη γενετική, επιγενετική και πρωτεϊνική ποικιλομορφία των καρκινικών κυττάρων, δημιουργώντας προκλήσεις στη διάγνωση και την εύρεση κατάλληλης θεραπείας. Επιπλέον, ο καρκίνος είναι μια δυναμική οντότητα, η οποία μεταβάλλεται όσο ο όγκος εξελίσσεται, γεγονός που επιβεβαιώνει και την ανάγκη για εξατομικευμένες ιατρικές προσεγγίσεις.



Σχήμα 2.5: *Μορφές ετερογένειας καρκίνου στον μαστό [3]*

Η ετερογένεια **εντός του όγκου** μπορεί να παρατηρηθεί μεταξύ διαφορετικών περιοχών του ίδιου όγκου, μεταξύ του πρωτοπαθούς όγκου και των μεταστατικών περιοχών, ακόμη και μεταξύ διαφορετικών μεταστάσεων. Αυτές οι διαφορές μπορεί να περιλαμβάνουν πτυχές της γενετικής, επιγενετικής και πρωτεϊνικής σύνθεσης του όγκου, δημιουργώντας προκλήσεις στη διάγνωση και τη θεραπεία.

Η ετερογένεια **μεταξύ των όγκων** αντιστοιχεί στις διαφορές του ίδιου τύπου καρκίνου μεταξύ διαφορετικών ασθενών. Συνήθως συνδέεται με τους διάφορους υποτύπους του καρκίνου του μαστού, την ανταπόκρισή τους στη θεραπεία και τα αποτελέσματά τους. Ο καρκίνος του μαστού ταξινομείται σε διάφορες ομάδες με βάση παράγοντες όπως οι ορμονικοί υποδοχείς (ER και PR), ο δείκτης Ki67 και η έκφραση HER2. Αυτές οι ομάδες σχηματίζουν τους διαφορετικούς υποτύπους καρκίνου μαστού που συναντώνται. Σε αυτούς συμπεριλαμβάνονται οι Luminal-A, Luminal B-, Luminal B+, HER2+, Basal και Triple-negative (TNBC). Κάθε υποομάδα έχει τη δική της πρόγνωση και ανταποκρίνεται διαφορετικά σε συγκεκριμένες θεραπείες [21].

2.1.3 Σταδιοποίηση

Η σταδιοποίηση της νόσου αποτελεί κρίσιμο σημείο τόσο για την πρόγνωση όσο και για τη θεραπεία του καρκίνου. Ο προσδιορισμός σταδίου λειτουργεί συμπληρωματικά μαζί με άλλους βιολογικούς παράγοντες για να το πετύχει αυτό. Αποτελεί έναν κοινό κώδικα επικοινωνίας μεταξύ των επαγγελματιών υγείας, επιτρέποντας τον προσδιορισμό της πορείας και εξέλιξης της νόσου για κάθε ασθενή. Προκειμένου να επιτευχθεί ακριβέστερη εκτίμηση της επέκτασης της νόσου και να διαμορφωθεί η βέλτιστη θεραπευτική προσέγγιση, έχουν αναπτυχθεί διάφορα συστήματα σταδιοποίησης. Αυτά βασίζονται στην ανατομική έκταση της νόσου, την τάση για επέκταση, καθώς και στα ιστολογικά και κλινικά χαρακτηριστικά.

Ένα ευρέως χρησιμοποιούμενο σύστημα σταδιοποίησης είναι το σύστημα TNM. Στο ακρώνυμο αυτό το *T* αντιπροσωπεύει το μέγεθος πρωτοπαθούς όγκου, το *N* αναφέρεται ποσοτικά στην προσβολή μασχαλιαίων λεμφαδένων και το *M* αναπαριστά την παρουσία ή όχι μεταστάσεων στον ασθενή. Το σύστημα αυτό δημιουργήθηκε πριν από 60 χρόνια και εξελίσσεται συνεχώς για να ανταπεξέλθει στις κλινικές αλλαγές. Για τον καρκίνο του μαστού, το σύστημα TNM περιλαμβάνει 5 στάδια (0, IA, IB, IIA, IIB, IIIA, IIIB, IIIC, IV) [4].

Η διαδικασία κατηγοριοποίησης ενός όγκου σε ένα στάδιο μπορεί να γίνει είτε κλινικά είτε παθολογοανατομικά. Η κλινική σταδιοποίηση περιλαμβάνει φυσική εξέταση, με προσεκτική επιθεώρηση και ψηλάφηση του δέρματος, του μαστικού αδένα και των λεμφαδένων (μασχαλιαίων, υπερκλείδιων και τραχηλικών), απεικόνιση και παθολογοανατομική εξέταση του μαστού ή άλλων ιστών, ανάλογα με την περίπτωση, για να τεθεί η διάγνωση του καρκινώματος του μαστού. Η παθολογοανατομική σταδιοποίηση περιλαμβάνει όλα τα δεδομένα που αποκομίστηκαν από την κλινική σταδιοποίηση, καθώς και δεδομένα από τη χειρουργική διερεύνηση και εκτομή, καθώς και την παθολογική εξέταση του αρχικού καρκινώματος, των περιφερειακών λεμφαδένων και των μεταστατικών περιοχών αν υπάρχουν. Συνεπώς, μεγαλύτερη ακρίβεια στην εύρεση σταδίου του καρκίνου προσφέρει η παθολογοανατομική σταδιοποίηση.

Σύστημα TNM – Ταξινόμηση του όγκου (T)

Η συνιστώσα "*T*" αξιολογεί το μέγεθος και την έκταση του πρωτοπαθούς όγκου εντός του μαστού. Οι όγκοι κατηγοριοποιούνται σε ένα φάσμα από *T*0 έως *T*4, το οποίο αντικατοπτρίζει το αυξανόμενο μέγεθος και τη συμμετοχή των παρακείμενων ιστών. Παράγοντες όπως η εισβολή στις γύρω δομές και η διείσδυση μέσω του δέρματος ή του θωρακικού τοιχώματος λαμβάνονται υπόψη για τον καθορισμό του σταδίου *T*, προσφέροντας πληροφορίες για την τοπική εξάπλωση του καρκίνου. Ο καθορισμός της κατηγορίας γίνεται είτε με βάσει την κλινική εξέταση είτε παθολογοανατομικά. Σαφώς, η παθολογοανατομική εξέταση παρουσιάζει μεγαλύτερη ακρίβεια.

Σύστημα TNM - Ταξινόμηση των επιχώριων λεμφαδένων (N)

Η συνιστώσα *N* υποδηλώνει τη συμμετοχή των λεμφαδένων. Οι λεμφαδένες διαδραματίζουν καθοριστικό ρόλο στο ανοσοποιητικό σύστημα του οργανισμού και λειτουργούν ως πύλη για την εξάπλωση των καρκινικών κυττάρων. Το σύστημα σταδιοποίησης *N* αξιολογεί

κατά πόσο ο καρκίνος έχει διεισδύσει σε αυτούς, με ταξινομήσεις που κυμαίνονται από $N0$ (που υποδηλώνει μη συμμετοχή λεμφαδένων) έως $N3$ (που υποδηλώνει εκτεταμένη εξάπλωση σε πολλούς κόμβους). Η αξιολόγηση περιλαμβάνει τον αριθμό των προσβεβλημένων κόμβων και το μέγεθός τους, βοηθώντας στην κατανόηση της πιθανότητας περιφερειακής μετάστασης.

Σύστημα TNM – Ταξινόμηση της μεταστατικής νόσου (M)

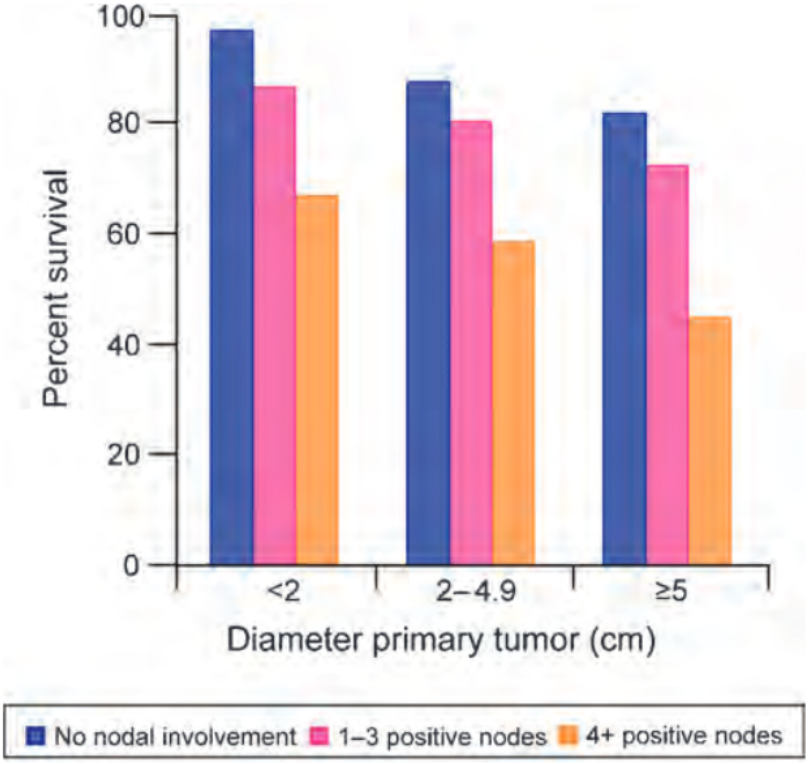
Τέλος, η συνιστώσα "M", που αντιπροσωπεύει τη μετάσταση, εξετάζει εάν ο καρκίνος έχει εξαπλωθεί σε απομακρυσμένες περιοχές πέραν του μαστού και των κοντινών λεμφαδένων. Η μετάσταση, ιδίως σε ζωτικά όργανα όπως οι πνεύμονες, το ήπαρ, τα οστά ή ο εγκέφαλος, επηρεάζει σημαντικά την πρόγνωση και τις στρατηγικές θεραπείας. Το σύστημα σταδιοποίησης M απλοποιεί αυτή την αξιολόγηση σε δύο κατηγορίες: $M0$ (απουσία απομακρυσμένων μεταστάσεων) και $M1$ (παρουσία απομακρυσμένων μεταστάσεων). Ο προσδιορισμός της μεταστατικής εξάπλωσης βοηθά στον καθορισμό της κατάλληλης θεραπευτικής προσέγγισης, με έμφαση τόσο στον τοπικό έλεγχο όσο και στη συστηματική αντιμετώπιση.

Η σταδιοποίηση του καρκίνου του μαστού σύμφωνα με την American Joint Committee on Cancer [4] διαμορφώνεται ως εξής:

Πίνακας 2.1: *Ανατομικά στάδια [4]*

ΣΤΑΔΙΟ	**T**	**N**	**M**
Stage 0	Tis	N0	M0
Stage IA	T1*	N0	M0
Stage IB	T0	N1mi	M0
	T1*	N1mi	M0
Stage IIA	T0	N1**	M0
	T1*	N1**	M0
	T2	N0	M0
Stage IIB	T2	N1	M0
	T3	N0	M0
Stage IIIA	T0	N2	M0
	T1*	N2	M0
	T2	N2	M0
	T3	N1	M0
	T3	N2	M0
Stage IIIB	T4	N0	M0
	T4	N1	M0
	T4	N2	M0
Stage IIIC	Any T	N3	M0
Stage IV	Any T	Any N	M1

Η σημασία της σταδιοποίησης στην πρόγνωση και στη θεραπεία αποτυπώνεται και στο παρακάτω διάγραμμα. Συγκεκριμένα, παρατηρούμε στο Σχήμα 2.6 τα ποσοστά επιβίωσης για 211.645 διαγνωσμένες υποθέσεις καρκίνου του μαστού τη χρονική περίοδο 2001-2002. Γίνεται ξεκάθαρο πως η έγκαιρη πρόγνωση και η αντιμετώπιση του καρκίνου του μαστού σε πρώιμα στάδια της εξέλιξης του οδηγεί σε καλύτερες προγνωστικές προβλέψεις για την



Σχήμα 2.6: *Ποσοστό επιβίωσης 5 χρόνων με βάση το μέγεθος του αρχικού όγκου και το πλήθος των λεμφαδένων [4]*

επιβίωση του ασθενούς.

2.2 Γονιδιωματική Ανάλυση Δεδομένων

2.2.1 Από το κύτταρο στα γονίδια

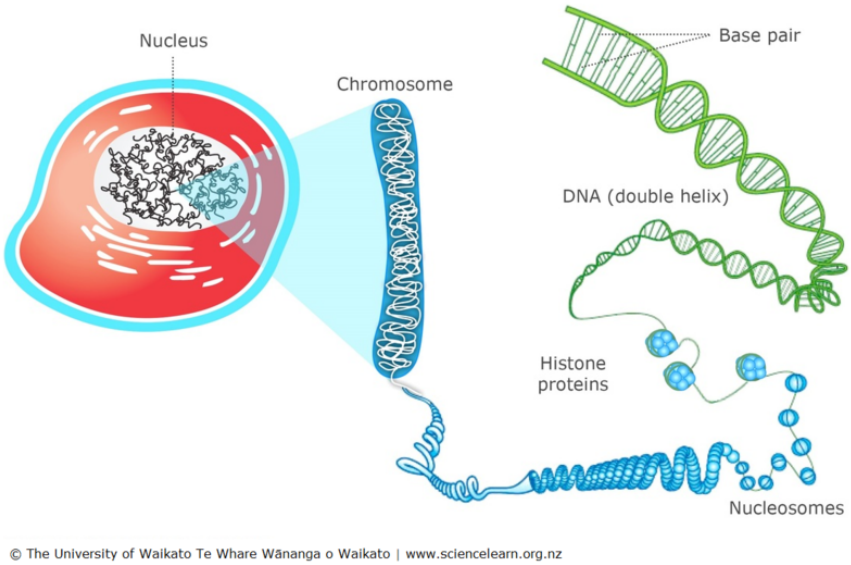
Κάθε ζωντανός οργανισμός αποτελείται από κύτταρα που συνεργάζονται για τη διατήρηση των βιολογικών λειτουργιών. Το κύτταρο κατέχει έναν εξαιρετικά σημαντικό ρόλο και αποτελεί το βασικό δομικό στοιχείο των έμβιων όντων [22]. Συγκεκριμένα, το ανθρώπινο σώμα περιέχει περισσότερα από 30 τρισεκατομμύρια κύτταρα.

Τα βασικά δομικά στοιχεία του κυττάρου είναι η κυτταρική μεμβράνη, το κυτταρόπλασμα και ο πυρήνας. Η κυτταρική μεμβράνη περιβάλλει το κύτταρο και ελέγχει τις ουσίες που εισέρχονται και εξέρχονται από το κύτταρο. Το εσωτερικό του κυττάρου περιέχει διάφορα μικροσκοπικά οργανίδια, όπως το σύμπλεγμα Golgi, τα μιτοχόνδρια, και το ενδοπλασματικό δίκτυο. Τα οργανίδια αυτά βρίσκονται μέσα στο κυτταρόπλασμα, δηλαδή το υγρό στο εσωτερικό του κυττάρου. Στο κυτταρόπλασμα λαμβάνουν χώρα οι περισσότερες χημικές αντιδράσεις και εκεί παράγονται οι περισσότερες πρωτεΐνες. Κάθε οργανίδιο συμβάλλει στην εκτέλεση συγκεκριμένων λειτουργιών εντός του κυτταρόπλασματος. Ανάμεσα στα οργανίδια ξεχωρίζει ο πυρήνας ως το μεγαλύτερο οργανίδιο μέσα στο κύτταρο και αποτελεί το κέντρο της κυτταρικής δραστηριότητας [23]. Στον πυρήνα συγκεντρώνεται το DNA, η γενετική πληροφορία του κυττάρου, και από εκεί εκτελούνται οι διεργασίες της μετάγγισης της γενετικής πληροφορίας για την παραγωγή πρωτεϊνών, που αποτελούν τα κυριότερα δομικά και λειτουργικά στοιχεία του κυττάρου. Είναι επίσης το σημείο όπου παράγεται το μεγαλύτερο μέρος του RNA [24].

Ο πυρήνας, μια εξειδικευμένη δομή που απαντάται στα περισσότερα κύτταρα, ελέγχει και ρυθμίζει τις δραστηριότητες του κυττάρου, όπως την ανάπτυξη και τον μεταβολισμό. Επιπλέον, και σημαντικότερο περιέχει το DNA. Η δομή του DNA είναι απλή, τόσο απλή που στις αρχές του 1940 οι βιολόγοι ήταν δύσπιστοι σχετικά με το πως το DNA θα μπορούσε να περιέχει όλη την γενετική πληροφορία. Πρόκειται για ένα μακρύ πολυμερές που συγκροτείται από αζωτούχες-πρωτεϊνικές βάσεις, φωσφορικές ρίζες και ένα σάκχαρο με πέντε άτομα άνθρακα. Οι 4 βάσεις - αδενίνη (A), γουανίνη (G), κυτοσίνη (C) και θυμίνη (T) - σχηματίζουν ζεύγη μεταξύ τους (A με T και G με C). Μια βάση, ένα σάκχαρο και ένα φωσφορικό άλας σχηματίζουν ένα νουκλεοτίδιο. Αυτά στη συνέχεια ενώνονται σειριακά και προκύπτουν οι πολυνουκλεοτιδικές αλυσίδες. Το DNA αποτελείται από δύο από αυτές τις πολυνουκλεοτιδικές αλυσίδες, σχηματίζοντας έναν διπλό έλικα. Οι δύο αλυσίδες είναι αντιπαράλληλες μεταξύ τους και οι δεσμοί υδρογόνου μεταξύ των τμημάτων βάσεων των νουκλεοτιδίων τις συγκρατούν [25]. Είναι η σειρά ή η αλληλουχία αυτών των ζευγών βάσεων που παρέχει τις πληροφορίες που απαιτούνται για την ανάπτυξη και την εξέλιξη του σώματός μας.

Το μακρομόριο αυτό έχει μέγεθος περίπου δύο μέτρα, και κάθε αλυσίδα έχει πλάτος λιγότερο μικρότερο από ένα εκατομμυριοστό του εκατοστού. Όμως ο πυρήνας έχει μέγεθος μόλις $6\mu m$ σε διάμετρο. Η δύσκολη αυτή διεργασία συγκέντρωσης του DNA γίνεται με την βοήθεια ειδικών πρωτεϊνών, που ενώνονται με το DNA και το διπλώνουν. Οι δομές που συγκεντρώνεται το DNA ονομάζονται χρωμοσώματα. Τα χρωμοσώματα αποτελούνται από αλυσίδες του DNA μαζί με ιστονικές πρωτεΐνες. Με εξαίρεση κάποια συγκεκριμένα κύτταρα, όπως τους γαμέτες ή τα ερυθροκύτταρα, κάθε κύτταρο περιέχει δύο αντίγραφα από κάθε

χρωμόσωμα [25].



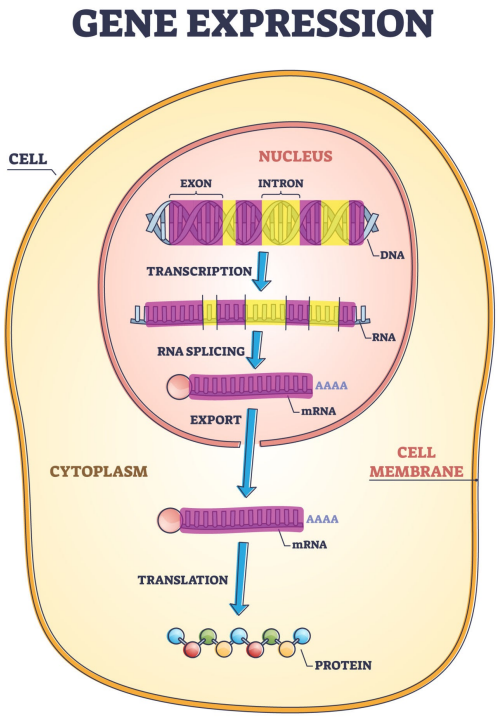
Σχήμα 2.7: *Σχηματική αναπαράσταση της συσχετίσης του πυρήνα του κυττάρου, του χρωμοσώματος, και των γονιδίων*

Η πιο σημαντική λειτουργία του DNA είναι η αποθήκευση και μεταβίβαση γενετικής πληροφορίας μέσω των γονιδίων. Τα γονίδια είναι τμήματα του DNA που βρίσκονται μέσα στα χρωμοσώματα. Διακρίνονται σε κωδικοποιητικά (coding) και μη κωδικοποιητικά (non-coding). Τα κωδικοποιητικά γονίδια περιέχουν τις οδηγίες για τη δημιουργία πρωτεϊνών, ενώ τα μη κωδικοποιητικά γονίδια παράγουν μόρια RNA ή συμμετέχουν σε άλλες λειτουργικές διαδικασίες.

Η ικανότητα των γονιδίων να παράγουν τελικά προϊόντα, όπως πρωτεΐνες ή μη κωδικοποιητικό RNA, συνοψίζεται στην διαδικασία της γονιδιακής έκφρασης. Κατά τη γονιδιακή έκφραση, οι πληροφορίες που περιέχει το γονίδιο χρησιμοποιούνται για τη σύνθεση των απαραίτητων βιολογικών μορίων.

Η γονιδιακή έκφραση αποτελείται από δύο βασικά στάδια. Αρχικά, εξειδικευμένες κυτταρικές δομές διαβάζουν το γονίδιο και χρησιμοποιούν τις πληροφορίες του για να παράγουν ένα μοριακό μήνυμα με τη μορφή ενός μορίου mRNA (αγγελιοφόρο ριβονουκλεϊκό οξύ) - η διαδικασία αυτή είναι γνωστή ως μεταγραφή (transcription). Στη συνέχεια, το μόριο mRNA μετακινείται από τον πυρήνα στο κυτταρόπλασμα του κυττάρου. Ένα ριβόσωμα διαβάζει το μήνυμα και παράγει μια πρωτεΐνη που ταιριάζει ακριβώς με τις οδηγίες που κωδικοποιούνται στο γονίδιο - η διαδικασία αυτή ονομάζεται μετάφραση (translation).

Σε κάθε κύτταρο του ανθρώπινου οργανισμού συναντάται το ίδιο DNA και άρα και τα ίδια γονίδια. Παρ' όλα αυτά υπάρχουν πολλοί και διαφορετικοί τύποι κυττάρων σε ένα πολυκυτταρικό οργανισμό, οι οποίοι διαφέρουν τόσο σε δομή όσο και σε λειτουργία. Οι τύποι κυττάρων διαφοροποιούνται μεταξύ τους επειδή συνθέτουν και συσσωρεύουν διαφορετικά σύνολα RNA και πρωτεϊνών μορίων. Επιπλέον, οι αναλύσεις των RNAs δείχνουν ότι, ανά πάσα στιγμή, ένα τυπικό ανθρώπινο κύτταρο εκφράζει το 30-60% των περίπου 25.000 γονιδίων του σε κάποιο σημαντικό επίπεδο. Υπάρχουν περίπου 20.000 γονίδια που κωδικοποιούν πρωτεΐνες και περίπου 5.000 που κωδικοποιούν μη κωδικοποιημένα RNA γονίδια



Σχήμα 2.8: *Η διαδικασία της γονιδιακής έκφρασης*

στον άνθρωπο. Επομένως, οι διάφοροι τύποι κυττάρων εκφράζουν, δηλαδή ενεργοποιούν ή απενεργοποιούν, διαφορετικά γονίδια, γεγονός που οδηγεί στη μεγάλη ποικιλία κυτταρικών δομών και λειτουργιών.

Η γονιδιακή έκφραση είναι μια διαδικασία που επιτρέπει την μελέτη, αλλά και την ερμηνεία των διαφορών που υπάρχουν στους κυτταρικούς οργανισμούς. Αν και εντυπωσιακές διαφορές παρουσιάζονται ήδη από τα πρωτεϊνοκωδικοποιητικά RNAs (mRNAs) σε εξειδικευμένους κυτταρικούς τύπους, δεν είναι κατάλληλα για να αναδείξουν το πλήρες φάσμα των διαφορών στο τελικό πρότυπο παραγωγής πρωτεϊνών. Αυτό προκύπτει, καθώς πέρα από το RNA, υπάρχουν και άλλοι παράγοντες, που ρυθμίζουν την γονιδιακή έκφραση [26].

Η γονιδιακή έκφραση, η οποία είναι καθοριστική για τον σχηματισμό πρωτεϊνών, στην διαδικασία σχηματισμού της από το DNA προς το RNA ρυθμίζεται και επηρεάζεται από διάφορους παράγοντες που συναντά σε διαφορετικά στάδια αυτής της πορείας [25]. Το κύτταρο μπορεί να ελέγξει τις πρωτεΐνες που παράγει:

1. από το πότε και πόσο συχνά μεταγράφεται ένα συγκεκριμένο γονίδιο (μεταγραφικός έλεγχος - transcriptional control)

2. ελέγχοντας την ωρίμανση και την επεξεργασία του RNA μεταγράφων (έλεγχος RNAs-επεξεργασίας - RNA processing control)

3. επιλέγοντας ποια ολοκληρωμένα mRNAs εξάγονται από τον πυρήνα στο κυτταρόλυμα και καθορίζοντας το σημείο στο κυτταρόλυμα που βρίσκονται (έλεγχος της μεταφοράς και του εντοπισμού του RNA)

4. επιλέγοντας ποια mRNAs στο κυτταρόπλασμα θα μεταφραστούν από τα ριβοσώματα (μεταφραστικός έλεγχος - translational control)

5. την επιλεκτική αποσταθεροποίηση ορισμένων μορίων mRNA στο κυτταρόπλασμα (mRNA αποικοδόμησης - mRNA degradation control)

6. μέσω της επιλεκτικής αποικοδόμησης συγκεκριμένων πρωτεϊνικών μορίων (έλεγχος αποικοδόμησης πρωτεϊνών - protein degradation control)

7. μέσω της ενεργοποίησης, απενεργοποίησης ή εντοπισμού συγκεκριμένων πρωτεϊνικών μορίων (έλεγχος πρωτεϊνικής δραστηριότητας)

Οι διαφορές στη γονιδιακή έκφραση μεταξύ των κυτταρικών τύπων αποκαλύπτονται επομένως πληρέστερα μέσω μεθόδων που εμφανίζουν άμεσα τα επίπεδα των πρωτεϊνών, μαζί με τις μετα-μεταφραστικές τροποποιήσεις τους. Ένας από τους τρόπους με τους οποίους καταφέρνουμε να έχουμε το μεταγράφωμα του κυττάρου είναι μέσω της μεθόδου single-cell RNA sequencing. Μέσω της συγκεκριμένης μεθόδου παρέχεται ένα στιγμιότυπο των αλληλουχιών RNA σε κάθε δείγμα και πρόκειται αυτή τη στιγμή για ένα από τα πιο ισχυρά εργαλεία διαθέσιμα για την ανάλυση της γονιδιακής έκφρασης.

2.2.2 Βιοδείκτες

Σύμφωνα με το Εθνικό Ινστιτούτο Καρκίνου, ένας βιοδείκτης είναι "ένα βιολογικό μόριο που βρίσκεται στο αίμα, σε άλλα σωματικά υγρά ή σε ιστούς και αποτελεί ένδειξη μιας φυσιολογικής ή μη φυσιολογικής διαδικασίας ή μιας κατάστασης ή ασθένειας", όπως ο καρκίνος [27]. Οι βιοδείκτες συνήθως διαφοροποιούν έναν πάσχοντα ασθενή από ένα άτομο χωρίς ασθένεια. Οι μεταβολές μπορεί να οφείλονται σε διάφορους παράγοντες, συμπεριλαμβανομένων των μεταλλάξεων της γεννητικής σειράς ή των σωματικών μεταλλάξεων, των μεταγραφικών αλλαγών και των μετα-μεταφραστικών τροποποιήσεων.

Οι βιοδείκτες περιλαμβάνουν ένα φάσμα βιολογικών μορίων, από πρωτεΐνες (π.χ. ένα ένζυμο ή έναν υποδοχέα), νουκλεϊκά οξέα (π.χ. ένα microRNA ή άλλο μη κωδικοποιητικό RNA), αντισώματα και πεπτίδια μέχρι και υπογραφές γονιδιακής έκφρασης. Ακόμα, οι βιοδείκτες μπορεί να είναι κληρονομικοί και να ανιχνεύονται ως παραλλαγές αλληλουχίας στο DNA ή μπορεί να είναι σωματικοί και να προσδιορίζονται ως μεταλλάξεις στο DNA που προέρχεται από ιστό όγκου.

Οι βιοδείκτες διαδραματίζουν σημαντικό ρόλο σε εφαρμογές στην ογκολογία για αξιολόγηση κινδύνου, διάγνωση, πρόγνωση, αλλά και πρόβλεψη για ανταπόκριση στη θεραπεία. Η ραγδαία αύξηση γνώσεων σχετικά με την βιολογία του καρκίνου, αλλά και τα τεχνολογικά επιτεύγματα που έχουν σημειωθεί τα τελευταία χρόνια στην μοριακή βιολογία, έχουν οδηγήσει στην σχεδόν καθημερινή δημοσίευση μελετών σχετικά με βιοδείκτες του καρκίνου.

Πιο συγκεκριμένα, οι βιοδείκτες είναι ζωτικής σημασίας για τη διαχείριση διηθητικών καρκινωμάτων. Βοηθούν στην έγκαιρη ανίχνευση και διάγνωση της νόσου, η οποία είναι απαραίτητη για τη βελτίωση των αποτελεσμάτων της θεραπείας. Συγκεκριμένοι βιοδείκτες, όπως οι ορμονικοί υποδοχείς (ER και PR) και ο HER2, χρησιμοποιούνται για την ταξινόμηση

των υποτύπων του καρκίνου του μαστού και την καθοδήγηση των θεραπευτικών αποφάσεων. Για παράδειγμα, η παρουσία ορμονικών υποδοχέων είναι προγνωστικός δείκτης καλής ανταπόκρισης στην ορμονοθεραπεία, η οποία μπορεί να οδηγήσει σε μεγαλύτερα ποσοστά επιβίωσης και καλύτερη συνολική πρόγνωση [28].

Οι γονιδιωματικοί βιοδείκτες για τον καρκίνου του μαστού μπορούν να δείξουν την προδιάθεση ενός οργανισμού για την ανάπτυξη καρκίνου. Μεταλλάξεις στα γονίδια BRCA1/2 είναι από τις κύριες αιτίες για κληρονομικό καρκίνο του μαστού [29].

Ένας άλλος σημαντικός γονιδιωματικός βιοδείκτης είναι το πάνελ γονιδίων PAM50, το οποίο έχει αποδειχθεί ότι μπορεί να λειτουργήσει ως η υπογραφή του κάθε υποτύπου του καρκίνου του μαστού [30]. Το PAM50 αξιολογεί δηλαδή την έκφραση 50 γονιδίων και μπορεί να κατηγοριοποιήσει τον καρκίνο του μαστού σε τέσσερις κύριους υποτύπους: Luminal A, Luminal B, HER2-enriched, και Basal-like. Αυτή η κατηγοριοποίηση είναι κρίσιμη για την κατανόηση της βιολογίας του καρκίνου και για τη λήψη αποφάσεων σχετικά με τη θεραπεία, καθώς διαφορετικοί υποτύποι ανταποκρίνονται με διαφορετικό τρόπο στις θεραπείες. Επιπλέον, το PAM50 μπορεί να παρέχει προβλέψεις για την έκβαση της νόσου, την πιθανότητα υποτροπής και να βοηθήσει στην αξιολόγηση της ανάγκης υποβολής σε χημειοθεραπεία ασθενών με ER-θετικό/ HER2-αρνητικό καρκίνο του μαστού [31].

2.3 Γονιδιωματικά Δεδομένα και Μοντελοποίηση της Εξέλιξης του Καρκίνου του Μαστού

Η κατανόηση των βιολογικών μηχανισμών που διέπουν τον καρκίνο του μαστού είναι ζωτικής σημασίας για την ανάπτυξη αποτελεσματικών διαγνωστικών και προγνωστικών εργαλείων. Η ανάλυση της ανατομίας και φυσιολογίας του μαστού, των υποτύπων της νόσου, του συστήματος ταξινόμησης TNM, καθώς και των λειτουργιών των κυττάρων και ιδιαίτερα της γονιδιακής έκφρασης, συμβάλλει στην αξιοποίηση αυτής της βιολογικής γνώσης για την ανάπτυξη μοντέλων για την ταξινόμηση του σταδίου του καρκίνου του μαστού.

Τα γονιδιωματικά δεδομένα, ιδίως τα προφίλ γονιδιακής έκφρασης, προσφέρουν ένα λεπτομερές μοριακό πορτρέτο του καρκίνου του μαστού. Αυτά τα προφίλ μπορούν να αποκαλύψουν τα επίπεδα δραστηριότητας χιλιάδων γονιδίων ταυτόχρονα, παρέχοντας πληροφορίες για τις υποκείμενες βιολογικές διεργασίες που οδηγούν στην ανάπτυξη και την εξέλιξη του όγκου. Αναλύοντας δεδομένα γονιδιακής έκφρασης, οι ερευνητές μπορούν να εντοπίσουν συγκεκριμένα γονίδια και μονοπάτια που εκφράζονται διαφορετικά σε διάφορα στάδια του καρκίνου του μαστού.

2.4 Σχετική έρευνα

Οι ραγδαίες εξελίξεις στις γονιδιωματικές τεχνολογίες και τις υπολογιστικές μεθόδους έχουν φέρει επανάσταση στην έρευνα για τον καρκίνο, προσφέροντας νέες ευκαιρίες για την εξέλιξη της ιατρικής ακριβείας [32]. Η ανάπτυξη προγνωστικών μοντέλων για τη σταδιοποίηση του καρκίνου του μαστού χρησιμοποιώντας γονιδιωματικά δεδομένα είναι ένα σημαντικό άλμα προς την κατεύθυνση κατανόησης και θεραπείας αυτής της νόσου. Σε αυτήν την ενότητα

πραγματοποιείται μια ανασκόπηση σχετικά με την έρευνα που έχει γίνει στην μοντελοποίηση της εξέλιξης της νόσου, και συγκεκριμένα στους τομείς καθορισμού τύπου, σταδίου και κλινικών προβλέψεων του καρκίνου του μαστού χρησιμοποιώντας γονιδιακά και βιολογικά δεδομένα.

Η χρήση των τεχνικών Μηχανικής Μάθησης (ML) και Βαθιάς Μάθησης (DL) στην ταξινόμηση του καρκίνου έχει συγκεντρώσει μεγάλη προσοχή. Αυτές οι μέθοδοι αξιοποιούν τα δεδομένα γονιδιακής έκφρασης για τη διάκριση μεταξύ διαφορετικών τύπων και σταδίων καρκίνου, προσφέροντας δυνατότητες για πιο ακριβείς και εξατομικευμένες διαγνώσεις. Ένα πλήθος μελετών έχουν εξερευνήσει διάφορα μοντέλα ML και DL, από παραδοσιακούς αλγόριθμους έως εξελιγμένα νευρωνικά δίκτυα. Εκτεταμένη αναφορά και ανάλυση πάνω στις μεθόδους μηχανικής και βαθιάς μάθησης που χρησιμοποιήθηκαν στις έρευνες που αναφέρονται στη συνέχεια θα παρουσιαστεί στο Κεφάλαιο 3, καθώς είναι σημαντικό να αναγνωρίσουμε την αυξανόμενη σημασία της τεχνητής νοημοσύνης στην προώθηση της κατανόησης της εξέλιξης της νόσου.

Μηχανική Μάθηση

Χάρη στην πρόοδο τεχνολογιών Next Generation Sequencing - οι οποίες προσδιορίζουν τη σειρά των νουκλεοτιδίων σε ολόκληρα γονιδιώματα - όπως το single-cell RNA sequencing (scRNA-seq), και την ραγδαία ανάπτυξη του κλάδου της Βιοϊατρικής, βρισκόμαστε στην εποχή των μεγάλων δεδομένων της επιστήμης αυτής. Από τεχνικές, όπως το scRNA-seq παράγονται δεδομένα γονιδιακής έκφρασης, τα οποία επιτρέπουν στους ερευνητές να προσδιορίσουν ποια γονίδια είναι εκφρασμένα μέσα σε ένα ετερογενές δείγμα, στο επίπεδο ενός κυττάρου, και σε τι ποσότητες.

Η χρήση προσεγγίσεων που βασίζονται στη μηχανική μάθηση έχει καταστεί απαραίτητη για την καλύτερη κατανόηση των βιολογικών μηχανισμών, όπως το πώς οι παραλλαγές γονιδίων μπορούν να οδηγήσουν σε φαινοτυπικές αλλαγές [33]. Αυτές οι προσεγγίσεις αξιοποιούν δεδομένα γονιδιακής έκφρασης για την ανάπτυξη μοντέλων που μπορούν να προβλέψουν την εξέλιξη και την πρόγνωση του καρκίνου, συμβάλλοντας σημαντικά στην ανάπτυξη της εξατομικευμένης ιατρικής.

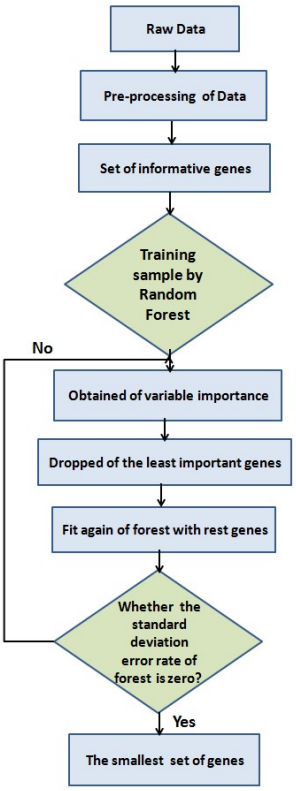
Ωστόσο, η χρήση εργαλείων τεχνητής νοημοσύνης (ΤΝ) και μηχανικής μάθησης για την ανάλυση αυτών των δεδομένων αντιμετωπίζει σημαντικές προκλήσεις, όπως είναι η διαχείριση της υψηλής διαστασιμότητας (high-dimensionality) των δεδομένων και η επεκτασιμότητα (scalability) [34].

Διάφορες μέθοδοι ML έχουν χρησιμοποιηθεί στην ανάλυση γονιδιακής έκφρασης για την διάγνωση τύπου καρκίνου του μαστού και την παροχή πληροφοριών για πιθανές θεραπείες. Συμβατικοί αλγόριθμοι μηχανικής μάθησης, όπως Υποστήριξη Διανυσματικών Μηχανών (SVM), k-Nearest Neighbor (kNN), Naïve Bayes (NB), Random Forest (RF) και σχετικές μέθοδοι έχουν αξιοποιηθεί σε ένα σύνολο ερευνών για πρώιμη ανίχνευση καρκίνου του μαστού.

Η μελέτη των Ram κ.α [35] χρησιμοποίησε τον αλγόριθμο RF για να ταξινομήσει δεδομένα καρκίνου και να εντοπίσει βιοδείκτες από δεδομένα έκφρασης γονιδίων για τρεις τύπους καρκίνου - παχέος εντέρου, προστάτη και λευχαιμίας. Ο αλγόριθμος RF εφαρ-

μόστηκε σε προεπεξεργασμένα σύνολα δεδομένων που περιείχαν 2000 εκφρασμένα γονίδια και αξιολογήθηκε χρησιμοποιώντας μετρικές, όπως Accuracy, Precision και Sensitivity.

Οι Ram κ.α κατέληξαν στην επιλογή των 2000 αυτών γονιδίων, υποσύνολο των οποίων λειτουργεί και ως βιοδείκτης, μέσω μιας αλγοριθμικά επαναληπτικής προσέγγισης. Σε κάθε επανάληψη, υπολογίζεται η σημασία του κάθε γονιδίου, δηλαδή το κατά πόσο ένα χαρακτηριστικό συμβάλει στην βελτίωση της επίδοσης του μοντέλου και ύστερα αφαιρείται το χαρακτηριστικό εκείνο με την χαμηλότερη τιμή από την εκπαίδευση του μοντέλου μέσω του αλγορίθμου RF. Αυτή η διαδικασία συνεχίζεται έως ότου η ελάχιστη τυπική απόκλιση των ποσοστών σφάλματος σε όλα τα δάση μηδενιστεί, υποδεικνύοντας ότι έχει εντοπιστεί το μικρότερο σύνολο γονιδίων.



Σχήμα 2.9: *Μέθοδος των West κ.α για την ταξινόμηση όγκων καρκίνου του μαστού [5]*

Επίσης, οι West κ.α [5] πρότειναν μια προσέγγιση για τη δημιουργία μιας στατιστικής μεθόδου που πραγματοποιεί ταξινόμηση όγκων του μαστού με βάση δεδομένα γονιδιακής έκφρασης. Οι συγγραφείς χρησιμοποίησαν αρχικά ένα μοντέλο δυαδικής παλινδρόμησης σε συνδυασμό με την τεχνική διάσπασης ιδιόμορφων τιμών (Singular Value Decomposition) για να εκτιμήσουν τις παραμέτρους παλινδρόμησης και τις πιθανότητες ταξινόμησης τόσο για δείγματα εκπαίδευσης όσο και για δείγματα επικύρωσης. Για να αντιμετωπίσουν το πρόβλημα της αξιολόγησης των αβεβαιοτήτων που είναι εγγενείς σε οποιοδήποτε μοντέλο πρόβλεψης, πραγματοποιείται διασταυρωμένη επικύρωση κάθε φορά.

Για την μείωση της διαστασιμότητας και του θορύβου των δεδομένων εντοπίστηκαν τα 100 γονίδια, τα οποία είναι πιο άμεσα συσχετισμένα με το αποτέλεσμα της ταξινόμησης. Την παραπάνω τεχνική την εφάρμοσαν σε όλα τα δείγματα του συνόλου δεδομένων μέσω της τεχνικής "hold-one-out", ώστε να λάβουν μια όσο γίνεται πιο πραγματική προγνωστική

αξιολόγηση. Αυτό οδήγησε στην δημιουργία τόσων υποσυνόλων όσο και το πλήθος των συνολικών δειγμάτων. Αν και μεταξύ τους αυτά τα υποσύνολα παρουσίαζαν σημαντικές επικαλύψεις, το εύρημα αυτό αποτυπώνει τη μεταβλητότητα των δειγμάτων και την εγγενή ετερογένεια στα προφίλ γονιδιακής έκφρασης.

Στην έρευνα των West κ.α [5] χρησιμοποιήθηκε η συσχέτιση των γονιδίων, ως παράγοντας για την επιλογή των τελικών χαρακτηριστικών. Η τεχνικές που αποσκοπούν στην επιλογή χαρακτηριστικών ονομάζονται feature engineering και χρησιμοποιούνται συχνά στην ML.

Η έρευνα των Zhang κ.α. [36] πραγματοποιεί ταξινόμηση με χρήση SVM και χρησιμοποιεί μια τεχνική feature engineering που οδηγεί σε αναδρομική εξάλειψη χαρακτηριστικών (RFE) και βελτιστοποίηση παραμέτρων (PO). Ως εκ τούτου ονομάστηκε SVM–RFE–PO. Αυτή η προσέγγιση χρησιμοποιεί αναζήτηση πλέγματος και μερική βελτιστοποίηση σμήνους (Swarm Optimization) για την επιλογή χαρακτηριστικών, σε συνδυασμό με έναν γενετικό αλγόριθμο για την διαδικασία επιλογής χαρακτηριστικών. Στη συνέχεια, χρησιμοποιεί το βέλτιστο σύνολο σημαντικών χαρακτηριστικών για την εκπαίδευση ενός μοντέλου SVM για την ταξινόμηση του καρκίνου του μαστού σε διάφορους υποτύπους του.

Στόχος των μεθόδων επιλογής χαρακτηριστικών είναι η μείωση της διαστασιμότητας του συνόλου δεδομένων κρατώντας τα γονίδια εκείνα που αναπαριστούν καλύτερα το σύνολο δεδομένων σε χαμηλότερη διάσταση. Πέρα από αυτό έχει παρατηρηθεί πως τα υποσύνολα γονιδίων που χαρακτηρίζονται ως βιοδείκτες είναι ασταθή ανάμεσα σε διαφορετικά σύνολα δεδομένων. Όπως φάνηκε και παραπάνω αυτό μπορεί να συμβαίνει ακόμα και μέσα στο ίδιο σύνολο δεδομένων [5], ωστόσο συναντάται και σε άλλες περιπτώσεις. Για παράδειγμα, σύνολα βιοδεικτών που προσδιορίζονται από ανεξάρτητες μελέτες σπάνια εμφανίζουν ουσιαστική επικάλυψη [37, 38]. Αυτό επιβεβαιώνεται και στην περίπτωση των van't Veer κ.α. [38] και των Wang κ.α [37], οι οποίοι προσδιόρισαν σύνολα γονιδίων μεγέθους 70 και 76 αντίστοιχα για να διακρίνουν τον μεταστατικό από τον μη μεταστατικό καρκίνο του μαστού, όμως τα δύο σύνολα είχαν επικάλυψη μόνο 3 γονιδίων [39].

Για την επίλυση του πρώτου εκ των παραπάνω προβλημάτων, δηλαδή της υψηλής διαστασιμότητας και στο πλαίσιο ανάπτυξης μεθόδων επιλογής χαρακτηριστικών, προτείνεται η ανάλυση μονοπατιών (Pathway Analysis). Η ανάλυση μονοπατιών είναι μια ευρέως χρησιμοποιούμενη τεχνική για την εξαγωγή βιολογικού νοήματος από δεδομένα γονιδιακής έκφρασης. Μια ευρέως διαδεδομένη μέθοδος που χρησιμοποιείται για την ερμηνεία και ανάλυση γονιδιακών δεδομένων είναι η Gene Set Enrichment Analysis (GSEA) [40].

Οι Kim κ.α. [41] χρησιμοποιούν την ανάλυση μονοπατιών και ταυτόχρονα προτείνουν μια λύση και στο πρόβλημα εύρεσης σταθερών και αναπαραγωγίσιμων γονιδίων. Η μέθοδος ταξινόμησης υποτύπων καρκίνου που χρησιμοποιούν συνδυάζει χαρακτηριστικά σε επίπεδο γονιδίου και σε επίπεδο μονοπατιού. Οι συγγραφείς προτείνουν τρεις νέες μεθόδους - GLEG (GSEA-Leading-Edge-Genes), GPF (GSEA-enriched-Pathway-Features) και SPF (SVM-Pathway-Features) - που χρησιμοποιούν έναν συνδυασμό της ανάλυσης GSE και του αλγόριθμου SVM για τον εντοπισμό σταθερών και αναπαραγωγίσιμων βιοδεικτών για την ταξινόμηση του καρκίνου.

Οι συγγραφείς δοκίμασαν τις μεθόδους τους σε δύο σύνολα δεδομένων καρκίνου - μετρώντας την πιθανότητα επιβίωσης ασθενών με καρκίνο των ωοθηκών και την πιθανότητα μετάστασης καρκίνου του μαστού - και συνέκριναν την απόδοση με παραδοσιακές μεθόδους

που βασίζονται σε γονίδια και μονοπάτια. Διαπίστωσαν ότι οι μέθοδοι που βασίζονται σε μονοπάτια (GLEG, GPF) γενικά ξεπερνάνε την απόδοση των μεθόδων που βασίζονται σε γονίδια όσον αφορά την ακρίβεια ταξινόμησης και την αναπαραγωγιμότητα των αναγνωρισμένων βιοδεικτών στα σύνολα δεδομένων.

Ταυτόχρονα τα δεδομένα γονιδιακής πέρα από το πρόβλημα της υψηλής διαστασιμότητας, αντιμετωπίζουν και προβλήματα μη-ισορροπημένων κλάσεων. Αυτό συμβαίνει, καθώς κάποιες διαγνώσεις είναι πιο συχνές από άλλες. Το φαινόμενο αυτό οδηγεί τους ερευνητές στην χρήση τεχνικών για δημιουργία συνθετικών δεδομένων.

Συγκεκριμένα, οι Roy κ.α [42] χρησιμοποιούν τεχνικές δημιουργίας συνθετικών δεδομένων για να διακρίνουν μεταξύ πρώιμων και όψιμων σταδίων διηθητικού πορογενή καρκίνου του μαστού με χρήση δεδομένων γονιδιακής έκφρασης. Το ζήτημα μη ισορροπημένων κατηγοριών στα δεδομένα επιλύθηκε με την χρήση τεχνικών δημιουργίας συνθετικών δεδομένων. Συγκεκριμένα, εδώ χρησιμοποιήθηκε η τεχνική Synthetic Minority Over-sampling Technique (SMOTE). Αυτό είχε ως αποτέλεσμα την αύξηση της ορθότητας στο 89% στο σύνολο επικύρωσης.

Παράλληλα εφάρμοσαν διάφορες μεθόδους επιλογής χαρακτηριστικών, όπως Recursive Feature Elimination (RFE), Randomized LASSO και Random Forest για να εντοπίσουν τα πιο σχετικά γονίδια για την εκπαίδευση των μοντέλων. Διαφορετικοί αλγόριθμοι μηχανικής μάθησης όπως Random Forest, Naive Bayes, Linear SVM, Logistic Regression και Decision Tree εκπαιδεύτηκαν και αξιολογήθηκαν χρησιμοποιώντας διασταυρωμένη επικύρωση. Στην έρευνα τους ανέπτυξαν δύο μοντέλα με διαφορετικές εισόδους. Το πρώτο χρησιμοποίησε ολόκληρο το σύνολο δεδομένων γονιδιακής έκφρασης για να εκπαιδεύσει το μοντέλο ταξινόμησης, ενώ το δεύτερο επικεντρώθηκε σε ένα υποσύνολο γονιδίων που είναι γνωστό ότι σχετίζεται με την εξέλιξη της νόσου του καρκίνου του μαστού.

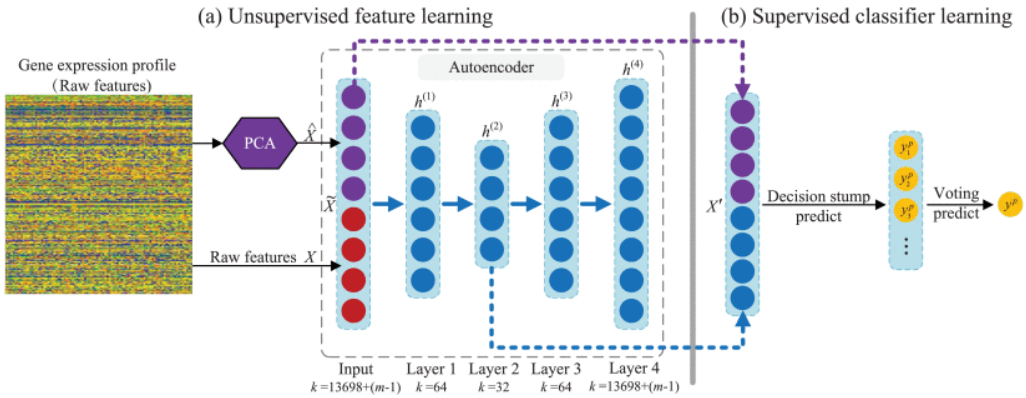
Γενικά, οι αλγόριθμοι ML έχουν αποδειχθεί ένα ισχυρό εργαλείο για τον εντοπισμό δυσδιάκριτων μοτίβων σε πολύπλοκα και υψηλών διαστάσεων δεδομένα σε πολλές εφαρμογές. Ωστόσο, η απόδοση των συμβατικών αλγορίθμων ML εξαρτάται σε μεγάλο βαθμό από την ποιότητα των παρεχόμενων χαρακτηριστικών. Ως εκ τούτου, η απόδοσή τους βασίζεται στην αποτελεσματικότητα των συνοδευτικών μεθόδων επιλογής χαρακτηριστικών.

Βαθιά Μάθηση

Οι μέθοδοι που βασίζονται σε βαθιά μάθηση χρησιμοποιούν τεχνητά νευρωνικά δίκτυα (NN) με πολλαπλά επίπεδα μονάδων επεξεργασίας για αναπαραστάσεις μαθησιακών δεδομένων. Αυτές οι μέθοδοι μπορούν να μάθουν ιεραρχικές αναπαραστάσεις σε δεδομένα υψηλών διαστάσεων, κάτι που αποτελεί βασικό πλεονέκτημα σε σύγκριση με τους συμβατικούς αλγόριθμους ML.

Το MLP είναι μια αρχιτεκτονική νευρωνικών δικτύων με πλήρως συνδεδεμένα επίπεδα, όπου κάθε νευρώνας σε ένα κρυφό στρώμα συνδέεται με όλους τους άλλους νευρώνες στα γειτονικά στρώματα.

Για παράδειγμα, οι Zhang κ.α. [6] πρότειναν ένα πλαίσιο εκμάθησης χαρακτηριστικών χωρίς επίβλεψη για τον εντοπισμό διαφορετικών ιδιοτήτων από τα προφίλ γονιδιακής έκφρασης συνδυάζοντας έναν αλγόριθμο ανάλυσης κύριων συνιστωσών (PCA) και ένα μοντέλο MLP



Σχήμα 2.10: *Διάγραμμα ροής της προτεινόμενης μεθοδολογίας [6]. Η ταξινόμηση αποτελείται από δύο φάσεις: (α) εκμάθηση χαρακτηριστικών χωρίς επίβλεψη, (β) εκμάθηση με επίβλεψη ταξινομητή.*

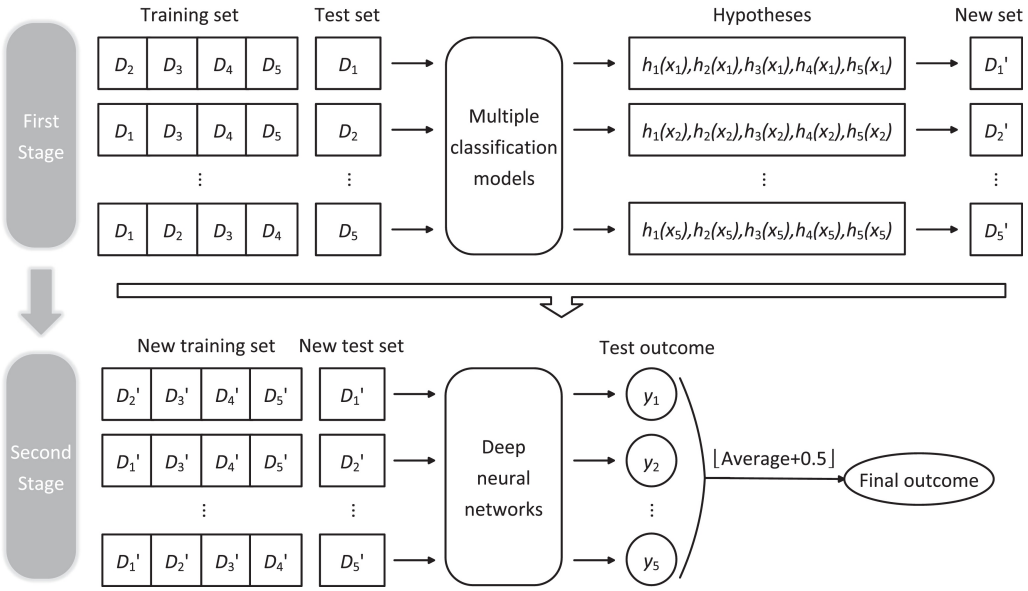
αυτοκωδικοποιητή. Ένας ταξινομητής συνόλου (ensemble classifier) που βασίζεται στον αλγόριθμο AdaBoost με το όνομα PCA-AEAda χρησιμοποιήθηκε για την πρόβλεψη κλινικών αποτελεσμάτων στον καρκίνο του μαστού και συγκεκριμένα για την πρόβλεψη μετάστασης της νόσου.

Επίσης, οι Gao κ.α. [43] πρότειναν την μέθοδο Deep Cancer Subtype Classification (DeepCC) για επιβλεπόμενη ταξινόμηση του υποτύπων καρκίνου του μαστού. Η προσέγγιση αυτή βασίζεται στην ανάλυση λειτουργικών φασμάτων που υποδεικνύουν τις δραστηριότητες των βιολογικών μονοπατιών. Στην έρευνα τους πραγματοποίησαν ανάλυση εμπλουτισμού για κάθε δείγμα και εκπαίδευσαν ένα πολυστρωματικό νευρωνικό δίκτυο για να αντικαταστήσει χαρακτηριστικά σχεδιασμένα με το χέρι. Οι συγγραφείς πέτυχαν ισορροπημένη ακρίβεια μεγαλύτερη από 90% στην ταξινόμηση του καρκίνου του μαστού. Ταυτόχρονα, επειδή το DeepCC εκπαιδεύεται στα λειτουργικά φάσματα, τα οποία είναι άμεσα συσχετισμένα με βιολογικές λειτουργίες, το τελικό μοντέλο είναι και ερμηνεύσιμο.

Συνελικτικά Νευρωνικά Δίκτυα

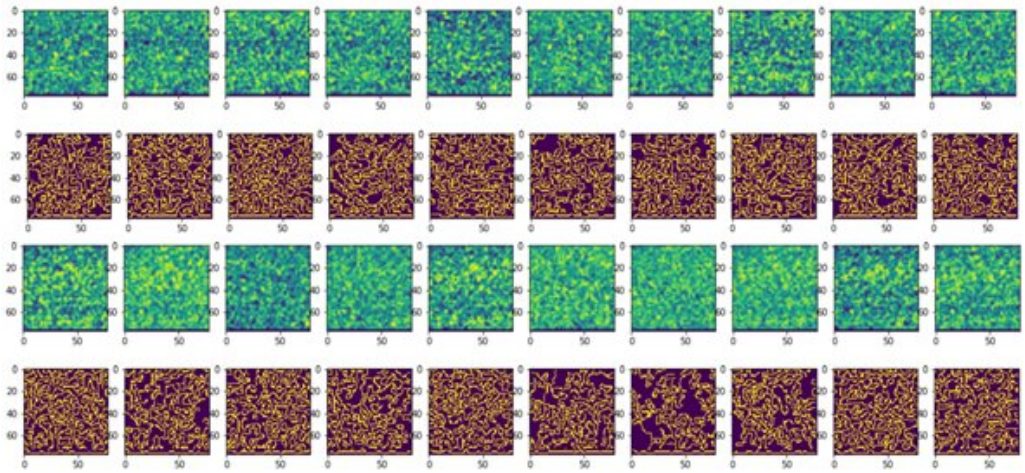
Τα συνελικτικά νευρωνικά δίκτυα (CNN) είναι αρχιτεκτονικές βαθιάς μάθησης που αρχικά σχεδιάστηκαν κυρίως για ανάλυση και επεξεργασία εικόνας. Τα CNN χρησιμοποιούν συνελικτικά φίλτρα για να μαθαίνουν αυτόματα ιεραρχίες χωρικών χαρακτηριστικών στα δεδομένα εισόδου. Οι αρχιτεκτονικές δικτύου χρησιμοποιούν έναν συνδυασμό στοιβαγμένων συνελικτικών και pooling επιπέδων (χρησιμοποιούνται συχνά πρόσθετα επίπεδα κανονικοποιήσης, όπως επίπεδα Batch Normalization ή τα επίπεδα Dropout).

Στην ανάλυση γονιδιακής έκφρασης, οι Xiao κ.α. [7] παρουσίασαν μια μέθοδο συνόλου βασισμένη στην αρχιτεκτονική των CNN, η οποία εφαρμόστηκε σε τρία σύνολα δεδομένων RNA-Seq τριών ειδών καρκίνων, συμπεριλαμβανομένου του αδενοκαρκινώματος πνεύμονα, του αδενοκαρκινώματος Στομάχου και του Διηθητικού Καρκίνου του Μαστού, και επιτεύχθηκε ακρίβεια 98%. Τα δεδομένα εισόδου έχουν φιλτραριστεί και έχουν επιλεχθεί συγκεκριμένα γονίδια μέσω διαφορικής ανάλυσης. Αφού πραγματοποιηθεί η διαδικασία της προεπεξεργασίας τους, εφαρμόζεται η τεχνική συνόλου multi-model, όπου οι προβλέψεις ενός συνόλου ταξινομητών, k-Nearest-Neighbor, Support Vector Machines, Decision Trees,



Σχήμα 2.11: *Τεχνική Multi-Model βαθιάς μάθησης [7]*

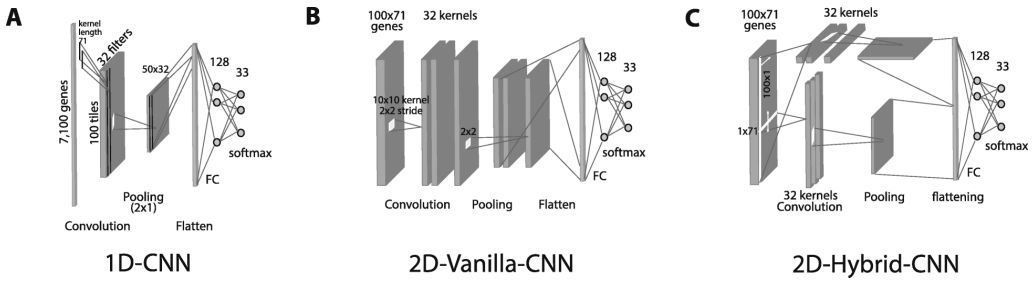
Random Forests, και Gradient Boosting Decision Trees, δίνονται ως είσοδοι σε ένα μοντέλο δεύτερου σταδίου (CNN). Το δεύτερο μοντέλο εκπαιδεύεται να συνδυάζει τις προβλέψεις του πρώτου, ώστε να καταλήξει στο βέλτιστο αποτέλεσμα. Στο Σχήμα 2.11 δίνεται και η αρχιτεκτονική του μοντέλου αυτού.



Σχήμα 2.12: *Δείγματα δεδομένων RNA-sequencing μετασχηματισμένα σε 2D εικόνες [8]*

Σε άλλες σχετικές ερευνητικές εργασίες [8], οι συγγραφείς χρησιμοποίησαν μοντέλα CNN για να ταξινομήσουν τους τύπους όγκων ενσωματώνοντας τα υψηλών διαστάσεων δεδομένα RNA-Seq σε εικόνες 2D. Κατά συνέπεια, η αρχιτεκτονική του CNN για την ταξινόμηση του καρκίνου του μαστού χρησιμοποιώντας δεδομένα γονιδιακής έκφρασης μετασχηματίσαι τα δεδομένα γονιδιακής έκφρασης σε 2D εικόνες που προτάθηκαν από τους Elbashir κ.α [8] και πέτυχαν ακρίβεια 98,76%.

Η έρευνα των Mostavi κ.α [9] περιλαμβάνει την δοκιμή τριών μοντέλων CNN (1D-CNN, 2D-Vanilla-CNN και 2DHybrid-CNN, 2.13) που εκπαιδεύτηκαν και δοκιμάστηκαν χρησιμοποιώντας προφίλ γονιδιακής έκφρασης από 10.340 δείγματα 33 διαφορετικών τύπων καρ-



Σχήμα 2.13: *Σχεδιασμός των 3 μοντέλων CNN από [9]*

κίνου. Στόχος μέσω αυτών των αρχιτεκτονικών ήταν η διερεύνηση και η ανάπτυξη ενός μοντέλου ταξινόμησης των υποτύπων του καρκίνου του μαστού. Στην συνέχεια πρόσθεσαν σε αυτά τα μοντέλα και δείγματα που αντιστοιχούσαν σε φυσιολογικό ιστό και εξέτασαν τις επιπτώσεις στην επίδοση του μοντέλου. Παρατήρησαν ότι η απόδοση του μοντέλου μειώνεται ελαφρώς, ωστόσο σε σύγκριση με άλλα μοντέλα καταφέρνει να έχει αρκετά καλή ακρίβεια 88,42%.

Κεφάλαιο 3

Θεωρητικό Υπόβαθρο

Η Τεχνητή Νοημοσύνη (AI) αποτελεί έναν από τους πιο ταχέως αναπτυσσόμενους και πολυσυζητημένους τομείς της επιστήμης των υπολογιστών και της μηχανικής τα τελευταία χρόνια. Κεντρικό ρόλο σε αυτή την ανάπτυξη παίζουν δύο σημαντικοί κλάδοι της: η Μηχανική Μάθηση και η Βαθιά Μάθηση. Η πρώτη ασχολείται με αλγορίθμους που επιτρέπουν στους υπολογιστές να μαθαίνουν από δεδομένα, ενώ η δεύτερη αξιοποιεί νευρωνικά δίκτυα για την ανάλυση πολύπλοκων και μεγάλων συνόλων δεδομένων. Αυτό το κεφάλαιο εξετάζει τις βασικές αρχές, τους κύριους αλγόριθμους, τις μεθόδους αξιολόγησης της απόδοσης, και τις τεχνικές βελτιστοποίησης αυτών των τεχνολογιών.

3.1 Μηχανική Μάθηση

Η ML εστιάζει στην ανάπτυξη αλγορίθμων και στατιστικών μοντέλων που επιτρέπουν στους υπολογιστές να εκτελούν συγκεκριμένες εργασίες βασιζόμενοι σε μοτίβα και συμπεράσματα από δεδομένα χωρίς να είναι ρητά προγραμματισμένα για αυτά [44]. Οι αλγόριθμοι εκπαίδευσης μηχανικής μάθησης χρησιμοποιούν ένα σύνολο από δεδομένα ώστε να βελτιώσουν την απόδοση τους στο ζητούμενο πρόβλημα, με την βοήθεια μεθόδων στατιστικής και μαθηματικής μοντελοποίησης. Η χρησιμότητα τους έγκειται στο ότι μπορούν να μάθουν να λύνουν προβλήματα τα οποία οι άνθρωποι δεν μπορούν (ή μπορούν με μεγάλη δυσκολία) να λύσουν με τις μηχανιστικές μεθόδους των κλασικών αλγορίθμων, για λόγους πολυπλοκότητας ή μεγάλης απαιτούμενης λεπτομέρειας.

Σε αυτήν την υποενότητα, αναλύονται οι βασικές αρχές και τεχνικές της ML. Αρχικά, εξετάζεται η διαδικασία της ML, τα είδη των αλγορίθμων και φυσικά οι ίδιοι οι αλγόριθμοι. Στη συνέχεια, η εστίαση μετατοπίζεται στην διαδικασία εκπαίδευσης και αξιολόγησης των μοντέλων, όπου περιγράφονται τα κριτήρια και οι μετρικές που χρησιμοποιούνται για την εκτίμηση της απόδοσής.

3.1.1 Εισαγωγή στην Μηχανική Μάθηση

Η διαδικασία της μηχανικής μάθησης μπορεί να συνοψιστεί απλοποιημένα στα παρακάτω βασικά βήματα:

1. **Συλλογή Δεδομένων**: Συγκέντρωση και οργάνωση των δεδομένων που θα χρησιμοποιηθούν για την εκπαίδευση του μοντέλου.

2. **Επεξεργασία Δεδομένων**: Καθαρισμός (διαχείριση μηδενικών δεδομένων, ακραίων τιμών) και μετατροπή των δεδομένων (κανονικοποίηση, κλιμάκωση), ώστε να βρίσκονται σε κατάλληλη μορφή για την εκπαίδευση του μοντέλου.

3. **Επιλογή Μοντέλου**: Επιλογή του κατάλληλου αλγορίθμου ή του συνδυασμού αλγορίθμων για την επίλυση του προβλήματος.

4. **Εκπαίδευση Μοντέλου**: Χρήση των δεδομένων για την εκπαίδευση του μοντέλου, προσαρμόζοντας τις παραμέτρους του, ώστε να βελτιώσει την απόδοσή του.

5. **Αξιολόγηση Μοντέλου**: Εκτίμηση της απόδοσης του μοντέλου χρησιμοποιώντας μετρικές αξιολόγησης και, ενδεχομένως, βελτιστοποίηση του μοντέλου.

6. **Εφαρμογή Μοντέλου**: Χρήση του εκπαιδευμένου μοντέλου σε νέα δεδομένα για την πρόβλεψη ή την ταξινόμηση.

Οι τεχνικές μηχανικής μάθησης χωρίζονται σε διάφορες κατηγορίες, ανάλογα με το είδος των προβλημάτων που προσπαθούν να λύσουν. Οι κυριότερες κατηγορίες περιλαμβάνουν:

- **Επιβλεπόμενη Μάθηση (Supervised Learning)**: Χρησιμοποιώντας ετικετοποιημένα δεδομένα, δηλαδή δείγματα, όπου κάθε είσοδος x ακολουθείται από μια έξοδο y, στόχος της επιβλεπόμενης μάθησης είναι η εκμάθηση μιας συνάρτησης, η οποία μπορεί να αντιστοιχίσει μια είσοδο σε μια έξοδο [45]. Αφού το μοντέλο εκπαιδευτεί στα δεδομένα εκπαίδευσης, τότε προκύπτει μια συνάρτηση. Μαθηματικά, δεδομένου ενός συνόλου εκπαίδευσης $(x_i, y_i)_{i=1}^n$, όπου το x_i αντιπροσωπεύει την είσοδο και το y_i την αντίστοιχη έξοδο, οι αλγόριθμοι μάθησης με επίβλεψη προσπαθούν να βρουν μια συνάρτηση f έτσι ώστε $f(x_i) \approx y_i$. Δύο σημαντικές υποκατηγορίες επιβλεπόμενης μάθησης αποτελούν η ταξινόμηση (classification) και η παλινδρόμηση (regression).

  - Ταξινόμηση: Το πρόβλημα ανάθεσης των δεδομένων σε διαφορετικές διακριτές κατηγορίες που ονομάζονται κλάσεις. Η συγκεκριμένη υποκατηγορία είναι άρρηκτα συνδεδεμένη με την παρούσα εργασία και θα αναλυθεί περαιτέρω και στην συνέχεια.
  - Παλινδρόμηση: Το πρόβλημα εκτίμησης μιας συνεχούς ποσότητας.

- **Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning)** : Σε αντίθεση με την επιβλεπόμενη μάθηση, στους αλγορίθμους μάθησης χωρίς επίβλεψη δεν δίνονται ετικέτες στα δεδομένα εισόδου [46]. Αν και ο στόχος της μη επιβλεπόμενης μάθησης ποικίλει, μπορεί να συνοψιστεί στην ικανότητα των μοντέλων να ανακαλύπτουν μοτίβα, τα οποία εμφανίζονται στα δεδομένα χωρίς καμία εξωτερική επισήμανση. Ο πρωταρχικός στόχος είναι η μοντελοποίηση των δεδομένων, προκειμένου να μάθουμε περισσότερα για αυτά. Σημαντικές τεχνικές μάθησης χωρίς επίβλεψη περιλαμβάνουν την ομαδοποίηση (clustering) και τη μείωση διαστάσεων (dimensionality reduction).

  Η κωδικοποίηση των δεδομένων σε έναν διαφορετικό χώρο μικρότερης διάστασης από τον αρχικό μπορεί να συμβεί για πολλούς λόγους, όπως οικονομίας χώρου, οπτικοποίησης δεδομένων, ή και σαν ενδιάμεσο βήμα για την επίλυση κάποιου άλλου

προβλήματος. Σε περίπτωση εφαρμογής μείωσης διαστάσεων επιθυμητό είναι κάθε διάσταση, δηλαδή κάθε χαρακτηριστικό του προβλήματος, να αντιστοιχεί σε κάποια ερμηνεύσιμη περιγραφή του δεδομένου. Αρκετά δημοφιλής μέθοδοι μείωσης διαστασιμότητας είναι η Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis - PCA) και η t-distributed Stochastic Neighbor Embedding - t-SNE.

- **Ενισχυτική Μάθηση (Reinforcement Learning)**: Περιλαμβάνει την εκπαίδευση ενός αλγορίθμου μέσω ενός συστήματος επιβράβευσης και τιμωρίας, όπου ο αλγόριθμος μαθαίνει να λαμβάνει αποφάσεις με βάση τις ανταμοιβές που δέχεται από τις ενέργειές του. Το πρόβλημα μοντελοποιείται ως μια διαδικασία απόφασης Markov (MDP) με καταστάσεις S, ενέργειες A και ανταμοιβές R, με στόχο την εύρεση μιας πολιτικής $\pi : S \rightarrow A$ που μεγιστοποιεί την αναμενόμενη αθροιστική ανταμοιβή [47].

3.1.2 Αλγόριθμοι Μηχανικής Μάθησης

Αναπόσπαστο εργαλείο για την ανάπτυξη προγνωστικών μοντέλων είναι οι αλγόριθμοι μηχανικής μάθησης. Χάρη στην λειτουργικότητα τους, τα μοντέλα έχουν την δυνατότητα να μαθαίνουν από τα δεδομένα και να προβλέπουν ή να ταξινομούν νέες, άγνωστες πληροφορίες. Σε αυτήν την ενότητα, θα εξετάσουμε πιο αναλυτικά κάποιους από τους βασικούς αλγορίθμους που χρησιμοποιούνται συχνά στην βιβλιογραφία.

Στο παρόν κεφάλαιο θα δώσουμε έμφαση σε αλγορίθμους επιβλεπόμενης μηχανικής μάθησης και συγκεκριμένα σε αλγορίθμους ταξινόμησης. Στόχος της ταξινόμησης είναι η πρόβλεψη της κατηγορίας ή της ομάδας στην οποία ανήκει ένα δείγμα δεδομένων. Δίνονται δεδομένα εκπαίδευσης με ετικέτες (labels) που καθορίζουν την κατηγορία κάθε δείγματος. Στο τέλος το μοντέλο θα πρέπει να μπορεί να προβλέψει σωστά τις κατηγορίες νέων, μη ετικετοποιημένων, άγνωστων δειγμάτων.

Decision Trees

Η βασική ιδέα πίσω από τα δέντρα απόφασης (Decision Trees) είναι η χρήση μιας δενδροειδούς δομής για τη λήψη αποφάσεων [48]. Πρόκειται για ένα κατευθυνόμενο δέντρο, το οποίο ξεκινάει από έναν κόμβο "ρίζα", χωρίς εισερχόμενες ακμές και αποτελεί το σημείο εκκίνησης της διαδικασίας λήψης αποφάσεων. Όλοι οι κόμβοι που ακολουθούν αυτόν τον πρώτο ονομάζονται εσωτερικοί και έχουν ακριβώς μια εισερχόμενη ακμή. Κάθε εσωτερικός κόμβος αντιπροσωπεύει μια δοκιμή του αλγορίθμου σε ένα χαρακτηριστικό, που οδηγεί σε περαιτέρω διακλάδωση με βάση το αποτέλεσμα που θα έχει αυτή η δοκιμή. Τα φύλλα του δέντρου αποτελούν τους κόμβους απόφασης ή τερματισμού και δεν διακλαδίζονται περαιτέρω.

Τα δέντρα αποφάσεων παίρνουν αποφάσεις περνώντας από τη ρίζα σε έναν κόμβο φύλλου με βάση τα αποτελέσματα των δοκιμών σε κάθε εσωτερικό κόμβο. Η διαδρομή που ακολουθείται καθορίζεται από τις τιμές των χαρακτηριστικών των δεδομένων εισόδου. Η τελική απόφαση ή ταξινόμηση δίνεται από τον κόμβο φύλλων που φτάνει στο τέλος αυτής της διαδρομής.

Στόχος της χρήσης των δέντρων αποφάσεων στην διαδικασία της μηχανικής μάθησης

είναι μετά την διαδικασία εκμάθησης να κατασκευαστεί ένα δέντρο απόφασης, το οποίο προσεγγίζει τον ιδανικό ταξινομητή, δηλαδή καταφέρνει να διαχωρίζει σχεδόν τα δεδομένα. Παράλληλα, στόχος των δέντρων απόφασης - αν και υπολογιστικά αρκετά δύσκολο - είναι να είναι όσο πιο μικρά γίνεται.

$$h \leftarrow DecisionTree.train(<x_i, y_i>)_{i=1}^{n}$$

$$minimal(h)$$

Γίνεται φανερό ότι το κύριο ζήτημα γύρω από την ανάπτυξη και υλοποίηση των δέντρων απόφασης είναι η επιλογή του χαρακτηριστικού εκείνου που θα οδηγήσει στον διαχωρισμό ενός κόμβου, με τέτοιον τρόπο, ώστε να αυξήσουμε την πιθανότητα να οδηγηθούμε σε μικρότερο δέντρο. Υπάρχουν διάφοροι αλγόριθμοι που υλοποιούν τεχνικές για τον υπολογισμό της επιλογής ενός χαρακτηριστικού απόφασης, με στόχο την αύξηση του κέρδους πληροφορίας που λαμβάνουμε.

1. Εντροπία: Αποτελεί ένα μέτρο αταξίας σε έναν κόμβο. Το μέγιστο επίπεδο εντροπίας έχει τιμή 1, ενώ το ελάχιστο έχει τιμή 0. Έτσι, τα φύλλα που ανήκουν σε μια κλάση θα έχουν εντροπία 0, ενώ, η εντροπία για έναν κόμβο όπου οι κλάσεις διαιρούνται ισομερώς θα είναι 1. Υπολογίζεται από τον ακόλουθο μαθηματικό τύπο:

   $$E = -\sum_{i=1}^{n} p_i * log_2(pi)$$

   όπου p_i η πιθανότητα επιλογής δείγματος που ανήκει στην κλάση i.

2. Gini Index: Μετρά την πιθανότητα λανθασμένης ταξινόμησης ενός δείγματος αν του αποδοθεί τυχαία μια ετικέτα. Όσο χαμηλότερος είναι ο δείκτης Gini, τόσο μικρότερη είναι η πιθανότητα εσφαλμένης ταξινόμησης. Ο τύπος για τον υπολογισμό του Gini Index είναι:

   $$Gini = 1 - \sum_{i=1}^{j} P(i)^2$$

   όπου το j συμβολίζει το πλήθος των κλάσεων του προβλήματος και το $P(i)$ είναι το πλήθος των δειγμάτων της κλάσης i προς το συνολικό πλήθος δειγμάτων.

Ένα από τα βασικά χαρακτηριστικά που έχουν οδηγήσει στην ευρεία χρήση των δέντρων απόφασης είναι η ανοιχτή δομή τους. Ο αλγόριθμος είναι τύπου white box, δηλαδή μπορούμε να δούμε κάθε δέντρο που έχει φτιάξει και να το οπτικοποιήσουμε. Με αυτόν τον τρόπο επικρατεί μια διαφάνεια και μπορεί ένας ερευνητής να παρατηρήσει τα πιο σημαντικά χαρακτηριστικά.

Random Forests

Παρατηρούμε στα Decision Trees πως μικρές διαφορές στο σύνολο δεδομένων, μπορούν να κατασκευαστούν δέντρα τα οποία μεν έχουν ελαφρά διαφορετική δομή, παρουσιάζουν δε σημαντικά διαφορετικές προβλέψεις. Χρησιμοποιώντας όμως τμήματα του συνόλου δεδομένων, είναι δυνατό να κατασκευαστούν πολλά και διαφορετικά δέντρα αποφάσεων και εν

τέλει να ληφθεί μια πρόβλεψη με βάση την πλειοψηφία των ψήφων από τα επιμέρους δέντρα αποφάσεων. Λειτουργούν δηλαδή συνδυαστικά (ensemble methods) και αυτή είναι και η βασική αρχή των Random Forests [49].

Μια από τις πιο γνωστές συνδυαστικές μεθόδους είναι το Bagging (Bootstrap aggregating), που προτάθηκε από τον Leo Breiman το 1996, όπου δημιουργούνται k τυχαία υποσύνολα του συνόλου δεδομένων που χρησιμοποιούνται για την εκπαίδευση των μοντέλων.

Ο αλγόριθμος των τυχαίων δασών είναι επέκταση της bagging μεθόδου και χρησιμοποιεί επιπρόσθετα τυχαία επιλογή χαρακτηριστικών για να δημιουργήσει ένα ασυσχέτιστο δάσος από δέντρα απόφασης. Η ενσωμάτωση της τυχαιότητας των χαρακτηριστικών χρησιμεύει στην παραγωγή ενός τυχαίου υποσυνόλου αυτών, το οποίο θα τροφοδοτηθεί σε κάθε δέντρο ώστε να εξασφαλιστεί όσο το δυνατόν μικρότερη συσχέτιση μεταξύ των επιμέρους δέντρων του δάσους.

$$y = \underset{c}{\operatorname{argmax}} \sum_{i=1}^{k} I(y_i = c)$$

Για κάθε ένα από τα k διαφορετικά υποσύνολα δεδομένων κατασκευάζεται ένα δέντρο απόφασης. Ύστερα, συνδυάζοντας τα αποτελέσματα των k δέντρων αποφάσεων, από την πλειοψηφία των επιμέρους απαντήσεων υπολογίζεται το τελικό αποτέλεσμα.

Gradient Boosting

Το Gradient Boosting βασίζεται στην δημιουργία ενός ισχυρού συνόλου (ensemble) από ασθενή μοντέλα, συνήθως δέντρα απόφασης, τα οποία εκπαιδεύονται διαδοχικά ώστε να διορθώνουν τα σφάλματα των προηγούμενων μοντέλων [50].

Ο αλγόριθμος του Gradient Boosting αποτελείται από τρία βασικά στοιχεία:

1. **Μοντέλο Βάσης (Base Model)**: Συνήθως χρησιμοποιείται ένα δέντρο απόφασης ως το αρχικό μοντέλο.

2. **Συνάρτηση Απώλειας (Loss Function)**: Μια συνάρτηση που μετράει πόσο καλά το μοντέλο προβλέπει τα δεδομένα εκπαίδευσης.

3. **Αλγόριθμος Βελτιστοποίησης (Optimization Algorithm)**: Συνήθως χρησιμοποιείται ο αλγόριθμος gradient descent για να ελαχιστοποιηθεί η τιμή της συνάρτησης απώλειας [51].

Επιπλέον, περιλαμβάνει τα εξής βήματα:

1. Ξεκινά με ένα μοντέλο $F_0(x)$, το οποίο μπορεί να είναι απλά η μέση τιμή των στόχων.

2. Για κάθε επανάληψη m από 1 μέχρι M:

   - Υπολογίζεται το υπολειπόμενο σφάλμα $\hat{r}_i^{(m)}$ για κάθε δείγμα δεδομένων i:

   $$\hat{r}_i^{(m)} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$$

   - Εκπαιδεύεται ένα νέο δέντρο απόφασης $h_m(x)$ για να προβλέψει τα υπολειπόμενα σφάλματα.

- Το νέο μοντέλο ενημερώνεται προσθέτοντας μια αναλογική τιμή του νέου δέντρου:

$$F_m(x) = F_{m-1}(x) + \upsilon h_m(x)$$

όπου υ είναι ο ρυθμός μάθησης (learning rate), μια παράμετρος που καθορίζει το μέγεθος του βήματος στη διόρθωση.

3. Μετά από M επαναλήψεις, το τελικό μοντέλο $F_M(x)$ είναι το άθροισμα όλων των επιμέρους μοντέλων.

Για να εμβαθύνουμε στις μαθηματικές λεπτομέρειες, εξετάζουμε την απώλεια για προβλήματα, όπως η μέση τετραγωνική απόκλιση:

$$L(y, F(x)) = \frac{1}{2}(y - F(x))^2$$

Ο υπολογισμός του υπολειπόμενου σφάλματος γίνεται ως εξής:

$$\hat{r}_i^{(m)} = y_i - F_{m-1}(x_i)$$

Το νέο δέντρο $h_m(x)$ εκπαιδεύεται ώστε να προβλέπει αυτά τα υπολειπόμενα σφάλματα και το μοντέλο ενημερώνεται με:

$$F_m(x) = F_{m-1}(x) + \upsilon h_m(x)$$

Για προβλήματα ταξινόμησης με 2 κλάσεις, χρησιμοποιείται η λογιστική απώλεια:

$$L(y, F(x)) = \log(1 + \exp(-2yF(x)))$$

Παρά τα μειονεκτήματά που παρουσιάζει αυτός ο αλγόριθμος σχετικά με την ερμηνευσιμότητα, οι υψηλές επιδόσεις του σε προβλέψεις τον καθιστούν δημοφιλή αλγόριθμο.

Support Vector Machines

Η κεντρική ιδέα πίσω από τις Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines) μπορεί να συνοψιστεί στο εξής:

Δοθέντος ενός δείγματος εκπαίδευσης, το SVM κατασκευάζει ένα υπερεπίπεδο (hyperplane) ως επιφάνεια απόφασης με τρόπο τέτοιο ώστε το περιθώριο διαχωρισμού μεταξύ θετικών και αρνητικών παραδειγμάτων να μεγιστοποιείται [52].

Άρα η σχεδίαση της μηχανής βασίζεται στην εξαγωγή ενός υποσυνόλου των δεδομένων εκπαίδευσης τα οποία λειτουργούν ως διανύσματα υποστήριξης και κατά συνέπεια αντιπροσωπεύουν ένα σταθερό χαρακτηριστικό των δεδομένων.

Θεμελιώδης έννοια για την ανάπτυξη του αλγόριθμου μηχανικής μάθησης του SVM είναι ο πυρήνας εσωτερικού γινομένου μεταξύ ενός διανύσματος υποστήριξης x_i, το οποίο αποτελείται από ένα υποσύνολο των σημείων δεδομένων εκπαίδευσης, και ενός διανύσματος x, το οποίο αντλείται από τον χώρο δεδομένων εισόδου.

Για την καλύτερη κατανόηση του τρόπου λειτουργίας των SVMs, ακολουθεί ένα παράδειγμα. Έστω δείγμα εκπαίδευσης $\{(x_i, d_i)\}_{i=1}^N$, όπου x_i είναι το πρότυπο εισόδου για το i-οστό

παράδειγμα και d_i η αντίστοιχη επιθυμητή έξοδος. Έστω επίσης ότι υπάρχουν 2 κλάσεις, εκ των οποίων η μια αναπαριστάται από το υποσύνολο $d_i = -1$ και η άλλη από το υποσύνολο $d_i = 1$ και οι δύο κλάσεις είναι γραμμικά διαχωρίσιμες.

Η εξίσωση μιας επιφάνειας απόφασης με την μορφή ενός υπερεπιπέδου που εκτελεί διαχωρισμό είναι:

$$w^T x + b = 0$$

όπου x είναι ένα διάνυσμα εισόδου, w ένα προσαρμόσιμο διάνυσμα βαρέων και b είναι μια πόλωση. Άρα, ισχύει:

$$w^T x + b \geq 0, d_i = +1$$

$$w^T x + b < 0, d_i = -1$$

Για ένα συγκεκριμένο διάνυσμα βαρέων w και πόλωση b, δημιουργείται ένας χώρος ανάμεσα στο υπερεπίπεδο που ορίζεται από την εξίσωση $w^T x + b = 0$ και του πλησιέστερου σημείου δεδομένων. Ο διαχωρισμός αυτός ονομάζεται περιθώριο διαχωρισμού και συμβολίζεται με ρ. Στόχος των SVMs είναι η εύρεση των τιμών εκείνων που μεγιστοποιούν το ρ.

Logistic Regression

Η λογιστική παλινδρόμηση (Logistic Regression) είναι μια στατιστική μέθοδος για την ανάλυση συνόλων δεδομένων όπου η εξαρτημένη μεταβλητή είναι δυαδική [53]. Η βασική ιδέα πίσω από τη λογιστική παλινδρόμηση είναι η εξής: Δεδομένου ενός συνόλου δεδομένων, ο στόχος είναι να προβλεφθεί η πιθανότητα ενός δείγματος να ανήκει σε μία από δύο κατηγορίες. Αυτό επιτυγχάνεται μέσω της εφαρμογής της λογιστικής συνάρτησης σε μια γραμμική συνάρτηση των χαρακτηριστικών εισόδου.

Η λογιστική συνάρτηση, ή σιγμοειδής συνάρτηση, ορίζεται ως:

$$\sigma = \frac{1}{1 + e^{-z}}$$

όπου $z = w^T x + b$, με x να είναι το διάνυσμα χαρακτηριστικών εισόδου, w το διάνυσμα συντελεστών και b η σταθερά πόλωσης. Η συνάρτηση αυτή επιστρέφει τιμές μεταξύ 0 και 1, οι οποίες ερμηνεύονται ως πιθανότητες.

Για να κατασκευαστεί το μοντέλο λογιστικής παλινδρόμησης, χρησιμοποιείται η μέθοδος της μέγιστης πιθανοφάνειας (maximum likelihood estimation). Η συνάρτηση κόστους που χρησιμοποιείται είναι η εξής:

$$L(w, b) = \frac{1}{N} \sum_{i=1}^{N} [y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)]$$

όπου N είναι ο αριθμός των παραδειγμάτων εκπαίδευσης, y_i η πραγματική ετικέτα του i-οστού παραδείγματος και $\hat{y}_i$ η προβλεπόμενη πιθανότητα να ισχύει $y_i = 1$.

Ο στόχος είναι η ελαχιστοποίηση της συνάρτησης κόστους, γεγονός που επιτυγχάνεται συνήθως μέσω της μεθόδου (gradient descent). Οι ενημερώσεις των παραμέτρων w και b γίνονται ως εξής:

$$w := w - \eta \frac{\partial L}{\partial w}$$

$$b := b - \eta \frac{\partial L}{\partial b}$$

όπου η είναι ο ρυθμός μάθησης.

Η LG προσφέρει σημαντικά πλεονεκτήματα, όπως η απλότητα και η ερμηνευσιμότητα των αποτελεσμάτων. Επίσης, είναι αποτελεσματική όταν οι κατηγορίες είναι γραμμικά διαχωρίσιμες και μπορεί να επεκταθεί για την ανάλυση πολυκατηγορικών δεδομένων μέσω τεχνικών όπως η One-vs-Rest ή η Multinomial Logistic Regression.

Παρά τα πλεονεκτήματά της, η LG έχει και ορισμένους περιορισμούς. Για παράδειγμα, δεν αποδίδει καλά όταν υπάρχει ισχυρή μη γραμμικότητα στις σχέσεις μεταξύ των χαρακτηριστικών και των κλάσεων, καθώς και όταν τα δεδομένα είναι υπερβολικά πολυδιάστατα ή όταν υπάρχουν πολλές συσχετισμένες μεταβλητές.

3.1.3 Αξιολόγηση απόδοσης

Στον τομέα της μηχανικής μάθησης, η αξιολόγηση των μοντέλων είναι ένα κρίσιμο βήμα για την εξασφάλιση της αποτελεσματικότητάς τους. Οι μετρικές αξιολόγησης είναι τα εργαλεία που χρησιμοποιούνται για να ποσοτικοποιήσουν την απόδοση των μοντέλων. Αυτές οι μετρικές επιτρέπουν την καλύτερη κατανόηση αναφορικά με την διαδικασία εκμάθησης του μοντέλου από τα δεδομένα εκπαίδευσης και το πόσο καλά είναι σε θέση να προβλέψει τα νέα, άγνωστα δεδομένα. Χωρίς την ύπαρξη αυτών των μετρικών, δεν θα ήταν δυνατό να γίνονται αντικειμενικές συγκρίσεις μεταξύ διαφορετικών μοντέλων ή να εντοπίζονται πιθανές περιοχές που χρίζονται βελτίωσης.

Ένα από τα βασικά εργαλεία για την αξιολόγηση μοντέλων ταξινόμησης είναι ο πίνακας σύγχυσης confusion matrix. Ο πίνακας αυτός παρέχει μια συνοπτική εικόνα της απόδοσης του μοντέλου, επιτρέποντας την ανάλυση των προβλέψεων έναντι των πραγματικών τιμών. Επιπλέον, οι τιμές που συμπεριλαμβάνει θέτουν τις βάσεις για τις περισσότερες από τις μετρικές που θα παρατεθούν στην συνέχεια. Στον πίνακα συμπεριλαμβάνονται οι εξής τιμές:

- **True Positive (TP)**: Οι περιπτώσεις όπου το μοντέλο προβλέπει σωστά την θετική κατηγορία.
- **False Positive (FP)**: Οι περιπτώσεις όπου το μοντέλο προβλέπει λανθασμένα την θετική κατηγορία ενώ η πραγματική κατηγορία είναι αρνητική.
- **True Negative (TN)**: Οι περιπτώσεις όπου το μοντέλο προβλέπει σωστά την αρνητική κατηγορία.
- **False Negative (FN)**: Οι περιπτώσεις όπου το μοντέλο προβλέπει λανθασμένα την αρνητική κατηγορία ενώ η πραγματική κατηγορία είναι θετική.

Για την αξιολόγηση της απόδοσης των μοντέλων μηχανικής μάθησης χρησιμοποιούνται οι παρακάτω μετρικές, εστιάζοντας σε διαφορετική μετρική ανάλογα με το είδος του προβλήματος, δηλαδή δεν είναι απαραίτητο ότι κάθε φορά όλες οι μετρικές θα είναι κατάλληλες και απαραίτητες για την αξιολόγηση ενός μοντέλου.

Ορθότητα (Accuracy)

Η ορθότητα εκφράζει το ποσοστό των σωστών προβλέψεων (τόσο θετικών όσο και αρνητικών) σε σχέση με το συνολικό αριθμό προβλέψεων. Είναι μία από τις πιο βασικές μετρικές, ωστόσο σε ορισμένες περιπτώσεις μπορεί να είναι και παραπλανητική, ειδικά αν το πλήθος μεταξύ των τάξεων ταξινόμησης δεν είναι ίσο. Υπολογίζεται από:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Number\ of\ Instances}$$

Ακρίβεια (Precision)

Η ακρίβεια εκφράζει το ποσοστό των πραγματικά θετικών προβλέψεων από τις συνολικές προβλέψεις θετικών. Είναι σημαντική σε περιπτώσεις όπου το κόστος των ψευδώς θετικών είναι υψηλό. Υπολογίζεται από:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Ανάκληση (Recall)

Η ανάκληση εκφράζει το ποσοστό των πραγματικά θετικών περιπτώσεων που ανιχνεύθηκαν από το μοντέλο. Είναι σημαντική σε περιπτώσεις όπου το κόστος των ψευδώς αρνητικών είναι υψηλό. Υπολογίζεται από:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Εξειδίκευση (Specificity)

Η εξειδίκευση εκφράζει το ποσοστό των πραγματικά αρνητικών περιπτώσεων που ανιχνεύθηκαν από το μοντέλο ανάμεσα στις πραγματικές αρνητικές περιπτώσεις. Συμπληρώνει την μετρική της ανάκλησης και είναι σημαντική σε περιπτώσεις όπου θέλουμε να ελαχιστοποιήσουμε τα ψευδώς θετικά. Υπολογίζεται από:

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives}$$

AUROC (Area Under the Receiver Operating Characteristics)

Η ROC καμπύλη είναι ένα γράφημα που δείχνει την απόδοση ενός ταξινομητή για διάφορα κατώφλια πρόβλεψης. Η x συνιστώσα του διαγράμματος αναπαριστά τον False Positive ρυθμό, και η y συνιστώσα τον True Positive ρυθμό.

Το AUC είναι η περιοχή κάτω από την καμπύλη ROC και εκφράζει την ικανότητα του μοντέλου να διαχωρίζει μεταξύ των κατηγοριών/κλάσεων. Η τιμές του AUROC κυμαίνονται από 0 έως 1, με τις υψηλότερες τιμές να υποδηλώνουν ολοένα και καλύτερη προγνωστική απόδοση. Μια τιμή AUROC 0,5 υποδεικνύει ένα μοντέλο που κάνει τυχαίες διακρίσεις, ενώ μια τιμή ίση με 1 αντιπροσωπεύει έναν τέλειο ταξινομητή.

F-Score (F1-Score))

Το F1-Score είναι η σταθμισμένη μέση τιμή της ακρίβειας και της ανάκλησης. Είναι ιδιαίτερα χρήσιμο όταν έχουμε ακατάλληλη, δηλαδή μη-ισορροπημένη, κατανομή της κατηγοριών, διότι δίνει μια ισορροπία μεταξύ των δύο μετρικών, ακρίβειας και ανάκλησης. Υπολογίζεται από:

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

Η κατανόηση και η χρήση αυτών των μετρικών είναι ουσιαστικής σημασίας για την αξιολόγηση και τη βελτίωση των μοντέλων μηχανικής μάθησης. Με τη βοήθεια αυτών των εργαλείων, μπορούμε να διασφαλίσουμε ότι τα μοντέλα μας είναι όσο το δυνατόν πιο ακριβή και αξιόπιστα.

Η χρήση των παραπάνω μετρικών είναι ξεκάθαρη στην περίπτωση δυαδικών προβλημάτων ταξινόμησης. Στην πολυκατηγορική ταξινόμηση (multi-class classification), ωστόσο, πρέπει να λαμβάνουμε υπόψη τις μετρικές με διαφορετικούς τρόπους για να έχουμε μια σαφή και ολοκληρωμένη εικόνα των αποτελεσμάτων. Σε αυτή την περίπτωση, η αξιολόγηση των μετρικών γίνεται μέσω του υπολογισμού τους για κάθε κατηγορία ξεχωριστά και στη συνέχεια λαμβάνεται ένας μέσος όρος των τιμών. Υπάρχουν τρεις διαφορετικές στρατηγικές για τον υπολογισμό αυτού του μέσου όρου: micro-averaged, macro-averaged και weighted-averaged.

Οι παραπάνω μετρικές είναι απαραίτητες για την αξιολόγηση των πολυκατηγορικών ταξινομητών, επειδή παρέχουν διαφορετικές προοπτικές στην απόδοση του μοντέλου:

- **Micro-averaged**: Είναι χρήσιμη όταν θέλουμε να δώσουμε έμφαση στη συνολική απόδοση του μοντέλου, ιδίως όταν οι κατηγορίες έχουν παρόμοια μεγέθη. Λαμβάνουν υπόψη το σύνολο των πραγματικών θετικών (True Positives, TP), ψευδών θετικών (False Positives, FP) και ψευδών αρνητικών (False Negatives, FN) από όλες τις κατηγορίες για τον υπολογισμό μιας συνολικής μετρικής.

  $$\text{Precision (micro)} = \frac{\sum_i TP_i}{\sum_i (TP_i + FP_i)}$$

  Η στρατηγική micro-averaged δίνει περισσότερη έμφαση στις κατηγορίες που έχουν μεγαλύτερη κατανομή, καθώς αθροίζουν τις τιμές όλων των παρατηρήσεων.

- **Macro-averaged**: Χρησιμοποιείται για την εκτίμηση της απόδοσης του μοντέλου σε κάθε κατηγορία ανεξάρτητα από το μέγεθός της, αποκαλύπτοντας προβλήματα σε κατηγορίες με λίγες παρατηρήσεις.

  $$\text{Precision (macro)} = \frac{1}{N} \sum_i^N Precision_i$$

  όπου N είναι ο αριθμός των κατηγοριών. Ο συγκεκριμένος υπολογισμός μέσων τιμών είναι σημαντικός ιδιαίτερα σε προβλήματα μη ισορροπημένων δεδομένων, διότι δίνεται ίσο βάρος σε όλες τις κατηγορίες, ανεξάρτητα από την συχνότητα τους.

- **Weighted-averaged**: Παρέχεται μια ισορροπημένη εκτίμηση της απόδοσης λαμβάνο-

ντας υπόψη την κατανομή των κατηγοριών για τον υπολογισμό του μέσου όρου.

$$\text{Precision (weighted)} = \sum_{i}^{N} \frac{N_i}{N_{total}} Precision_i$$

όπου N_i είναι ο αριθμός των δειγμάτων στην κατηγορία i και N_{total} είναι το συνολικό πλήθος δειγμάτων.

Οι μετρικές αξιολόγησης είναι ένας από τους τρόπους με τους οποίους μπορούμε να εξασφαλίσουμε την αποτελεσματικότητα των μοντέλων μηχανικής μάθησης που κατασκευάζουμε. Εξίσου θεμελιώδες και σημαντικό είναι να χρησιμοποιήσουμε τεχνικές, οι οποίες επιβεβαιώνουν και εκτιμούν την ικανότητα του μοντέλου να γενικεύει, που είναι και το ουσιαστικό "στοίχημα" κάθε μοντέλου, δηλαδή να μπορέσει να αποδώσει καλά και σε νέα, άγνωστα δεδομένα.

Ένας τρόπος για αυτήν την εξασφάλιση είναι η χρήση της τεχνικής της διασταυρωμένης επικύρωσης (cross-validation). Χρησιμοποιώντας διασταυρωμένη επικύρωση, μπορούμε να μειώσουμε την πιθανότητα υπερπροσαρμογής (overfitting) του μοντέλου, να εξασφαλίσουμε δηλαδή πως το μοντέλο δεν είναι υπερβολικά προσαρμοσμένο στα δεδομένα εκπαίδευσης.

Υπάρχουν διάφοροι τύποι διασταυρωμένης επικύρωσης που χρησιμοποιούνται για την καλύτερη εκτίμηση της απόδοσης των μοντέλων. Παρακάτω αναλύονται οι πιο συνήθεις μέθοδοι.

K-fold Διασταυρωμένη Επικύρωση

Σε αυτήν τη μέθοδο, το σύνολο δεδομένων χωρίζεται σε K ισομεγέθη τμήματα ή "folds". Το μοντέλο εκπαιδεύεται K φορές, κάθε φορά χρησιμοποιώντας $K-1$ από τα τμήματα για εκπαίδευση και το υπόλοιπο 1 για επικύρωση. Στη συνέχεια, ο μέσος όρος της απόδοσης σε κάθε ένα από τα K τμήματα υπολογίζεται για να εκτιμηθεί η συνολική απόδοση του μοντέλου.

Η τεχνική αυτή μπορεί να χρησιμοποιηθεί τόσο για τον υπολογισμό των βέλτιστων υπερπαραμέτρων όσο και για την επιλογή του αποδοτικότερου μοντέλου.

Stratified K-fold Διασταυρωμένη Επικύρωση

Η Stratified K-fold διασταυρωμένη επικύρωση είναι μια παραλλαγή της K-fold που χρησιμοποιείται όταν τα δεδομένα μας είναι ανισοκατανεμημένα μεταξύ των κλάσεων. Σε αυτήν τη μέθοδο, η κατανομή των δεδομένων στις κατηγορίες διατηρείται σε κάθε fold, εξασφαλίζοντας ότι κάθε fold είναι αντιπροσωπευτικό του συνολικού συνόλου δεδομένων.

Nested Cross Validation

Η Nested Cross Validation χρησιμοποιείται κυρίως για την επιλογή μοντέλου και την εκτίμηση της απόδοσης με βέλτιστες υπερπαραμέτρους. Περιλαμβάνει δύο επίπεδα διασταυρωμένης επικύρωσης: ένα εξωτερικό και ένα εσωτερικό. Το εξωτερικό επίπεδο χρησιμοποιείται για την εκτίμηση της απόδοσης του μοντέλου, ενώ το εσωτερικό επίπεδο χρησιμοποιείται για την επιλογή των υπερπαραμέτρων.

Leave-One-Out Cross Validation (LOO-CV)

Η Leave-One-Out Cross Validation (LOO-CV) είναι μια ακραία περίπτωση της K-fold διασταυρωμένης επικύρωσης όπου το K είναι ίσο με το μέγεθος του συνόλου δεδομένων. Σε κάθε επανάληψη, ένα μόνο δείγμα χρησιμοποιείται για επικύρωση και τα υπόλοιπα για εκπαίδευση. Η διαδικασία επαναλαμβάνεται για κάθε δείγμα στο σύνολο δεδομένων.

3.1.4 Μη-ισορροπημένα Δεδομένα (Imbalanced Dataset)

Όταν οι κατηγορίες σε ένα πρόβλημα ταξινόμησης δεν είναι ίσα κατανεμημένες, δηλαδή το πλήθος των δειγμάτων για κάθε κλάση στο σύνολο δεδομένων δεν είναι ίσο, τότε γίνεται λόγος για μη ισορροπημένα δεδομένα. Αυτό σημαίνει πως μία ή περισσότερες κλάσεις έχουν σημαντικά λιγότερα παραδείγματα σε σύγκριση με άλλες, καθιστώντας δύσκολη την αποτελεσματική μάθηση των μοντέλων.

Το πρόβλημα που προκύπτει είναι ότι μη ισορροπημένα σύνολα δεδομένων τείνουν να παράγουν προβλέψεις με υψηλή ακρίβεια στην κλάση με τα στιγμιότυπα που πλειοψηφούν. Αντίστοιχα, παρουσιάζουν χαμηλή ακρίβεια για την κλάση με τα λιγότερα στιγμιότυπα. Αυτό συμβαίνει, διότι οι αλγόριθμοι ταξινόμησης που χρησιμοποιούνται θεωρούν ότι οι κλάσεις είναι ομοιόμορφα κατανεμημένες στο σύνολο των δεδομένων. Έτσι, έχουν ως στόχο την ελαχιστοποίηση του συνολικού λάθους της ταξινόμησης, στο οποίο η κλάση με τα λιγότερα στιγμιότυπα συνεισφέρει ελάχιστα.

Τα προβλήματα που προκύπτουν από μη ισορροπημένα δεδομένα μπορούν να περιγραφούν χρησιμοποιώντας αναλογίες. Για παράδειγμα, ένα πρόβλημα δυαδικής ταξινόμησης με αναλογία 1:100 σημαίνει ότι μία κατηγορία έχει 100 φορές περισσότερα παραδείγματα από την άλλη. Εναλλακτικά, οι κατανομές κατηγοριών μπορούν να εκφραστούν ως ποσοστά, όπως 80% σε μία κατηγορία, 18% σε άλλη και 2% σε μια τρίτη.

Μη ισορροπημένα δεδομένα μπορούν να προκύψουν από δύο κύριους παράγοντες: τη δειγματοληψία δεδομένων και τις ιδιότητες του πεδίου. Από την μία πλευρά υπάρχει μεροληπτική δειγματοληψία, όταν γίνεται συλλογή δεδομένων από μια συγκεκριμένη περιοχή ή χρονική περίοδο που δεν αντιπροσωπεύει τον γενικό πληθυσμό. Για παράδειγμα, για την πρόβλεψη εξέλιξη μιας νόσου, λόγω της ανάγκης για άμεση θεραπεία κατά τη διάγνωση, είναι ηθικά αδύνατη η συλλογή δεδομένων χρονοσειρών για τη μελέτη της εξέλιξης της [54]. Από την άλλη ορισμένες κατηγορίες εμφανίζονται φυσικά λιγότερο συχνά. Για παράδειγμα, η ανίχνευση απάτης συνήθως περιλαμβάνει έναν μικρό αριθμό δόλιων συναλλαγών σε σύγκριση με τις νόμιμες.

Η σοβαρότητα των μη ισορροπημένων δεδομένων κατηγοριών ποικίλλει, από ελαφριά (π.χ. αναλογία 4:6) έως ακραία (π.χ. 1:100 ή περισσότερο). Οι ελαφριές ανισοκατανομές μπορούν συχνά να αντιμετωπιστούν με τις συνήθεις τεχνικές ταξινόμησης, αλλά οι σοβαρές απαιτούν εξειδικευμένες μεθόδους.

Σε μη ίσα κατανεμημένα σύνολα δεδομένων, η κλάση με τα περισσότερα δείγματα αποτελεί και την πλειοψηφία, ενώ η κλάση με τα λιγότερα δείγματα είναι μειοψηφική. Η μειοψηφική κλάση συνήθως έχει μεγαλύτερο ενδιαφέρον επειδή οι προβλέψεις της είναι πιο πολύτιμες. Ωστόσο, οι συνήθεις αλγόριθμοι μηχανικής μάθησης τείνουν να επικεντρώνονται στην πλειοψηφική κλάση, οδηγώντας σε χαμηλή απόδοση στην μειοψηφική. Αυτή η μερο-

ληψία καθιστά δύσκολη την εκμάθηση των χαρακτηριστικών της μειοψηφικής κλάσης από τα μοντέλα [55].

Κάποιες από τις μεθόδους για την διαχείριση των μη ισορροπημένων δεδομένων, όπως παρουσιάζονται και από τον Jason Brownlee στο Βιβλίο του [55], είναι:

- **Υπερδειγματοληψία (Oversampling)**:

  Η υπερδειγματοληψία είναι μια τεχνική που χρησιμοποιείται για την αντιμετώπιση του προβλήματος των μη ισορροπημένων δεδομένων, αυξάνοντας τον αριθμό των παραδειγμάτων της μειοψηφικής κλάσης. Αυτό επιτυγχάνεται είτε με την επαναλαμβανόμενη χρήση των υπαρχόντων δειγμάτων της μειοψηφικής κλάσης είτε με τη δημιουργία νέων συνθετικών δειγμάτων. Η υπερδειγματοληψία μπορεί να βελτιώσει την απόδοση των μοντέλων ταξινόμησης, καθώς βοηθά στην εξισορρόπηση της κατανομής των κλάσεων και στην αποτροπή της μεροληψίας προς την πλειοψηφική κλάση.

  Οι πιο γνωστές και χρησιμοποιούμενες τεχνικές δημιουργίας συνθετικών δεδομένων είναι:

  - Synthetic Minority Over-sampling Technique (SMOTE), που δημιουργεί νέα συνθετικά δείγματα συνδυάζοντας χαρακτηριστικά από υπάρχοντα δείγματα της μειοψηφικής κλάσης.
  - Adaptive Synthetic Sampling Approach (ADASYN), που δημιουργεί συνθετικά δείγματα με έμφαση στα δυσκολότερα προς ταξινόμηση δείγματα, ενισχύοντας την παρουσία της μειοψηφικής κλάσης στις περιοχές του χώρου χαρακτηριστικών όπου η απόδοση του ταξινομητή είναι χαμηλότερη.

  Αναλυτικότερα:

  Η τεχνική **SMOTE** είναι μία από τις πιο διαδεδομένες μεθόδους δημιουργίας συνθετικών δεδομένων. Αναπτύχθηκε για να αντιμετωπίσει την ανισοκατανομή των κλάσεων προσθέτοντας συνθετικά δείγματα στα δεδομένα της υποεκπροσωπούμενης κλάσης. Συγκεκριμένα, η SMOTE λειτουργεί ως εξής:

  - **Επιλογή Δειγμάτων**: Για κάθε δείγμα της υποεκπροσωπούμενης κλάσης, επιλέγονται οι k πλησιέστεροι γείτονες (συνήθως k=5) χρησιμοποιώντας μεθόδους όπως η Ευκλείδεια απόσταση.
  - **Δημιουργία Συνθετικών Δεδομένων**: Για κάθε δείγμα, δημιουργούνται συνθετικά δείγματα λαμβάνοντας υπόψη τη διαφορά μεταξύ του δείγματος και των επιλεγμένων γειτόνων του. Ένα τυχαίο σημείο στη γραμμή που συνδέει το αρχικό δείγμα με τον γείτονα επιλέγεται για να δημιουργηθεί το συνθετικό δείγμα.

  Η SMOTE συμβάλλει στην εξισορρόπηση των δεδομένων δημιουργώντας νέα σημεία στο χώρο χαρακτηριστικών που αντιπροσωπεύουν τις υποεκπροσωπούμενες κλάσεις, ενισχύοντας έτσι την ικανότητα των αλγορίθμων να γενικεύουν καλύτερα.

  Η ADASYN είναι μια προσαρμοστική προσέγγιση της SMOTE, η οποία επιδιώκει να βελτιώσει περαιτέρω τη διαδικασία δημιουργίας συνθετικών δεδομένων. Η κύρια διαφορά της ADASYN από τη SMOTE είναι ότι δίνει μεγαλύτερη βαρύτητα στις περιοχές

του χώρου χαρακτηριστικών που είναι δύσκολο να ταξινομηθούν σωστά, αντί να υπερδειγματοληπτεί την μειονοτική κλάση κατευθείαν. Συγκεκριμένα, η ADASYN λειτουργεί ως εξής:

- **Υπολογισμός Δυσκολίας Δειγμάτων**: Η ADASYN υπολογίζει έναν δείκτη δυσκολίας *DR* για κάθε δείγμα της υποεκπροσωπούμενης κλάσης, βασιζόμενη στον αριθμό των γειτόνων που ανήκουν στην πλειοψηφική κλάση. Ο δείκτης δυσκολίας υπολογίζεται ως:

$$\text{Difficulty Ratio (DR)} = \frac{\text{Number of Majority Class Neighbors}}{\text{Number of Minority Class Neighbors}}$$

- **Δημιουργία Συνθετικών Δεδομένων**: Στην συνέχεια δημιουργεί δείγματα με προσαρμοστικό τρόπο, παράγοντας περισσότερα συνθετικά δείγματα για σημεία, όπου έχουν υψηλότερο δείκτη δυσκολίας.

Η ADASYN, επομένως, όχι μόνο αυξάνει τον αριθμό των δειγμάτων της υποεκπροσωπούμενης κλάσης, αλλά εστιάζει στη βελτίωση της απόδοσης του ταξινομητή στις πιο δύσκολες περιοχές, οδηγώντας σε ένα πιο ισορροπημένο και αποδοτικό μοντέλο.

- **Υποδειγματοληψία (Undersampling)**:

Η υποδειγματοληψία είναι μια άλλη τεχνική για την αντιμετώπιση των μη ισορροπημένων δεδομένων, η οποία μειώνει τον αριθμό των παραδειγμάτων της πλειοψηφικής κλάσης. Αυτό μπορεί να γίνει είτε με τυχαία διαγραφή δειγμάτων της πλειοψηφικής κλάσης είτε με την εφαρμογή πιο σύνθετων μεθόδων, όπως το Cluster Centroids, όπου τα δείγματα της πλειοψηφικής κλάσης αντικαθίστανται από τα κεντροειδή των κλάσεων. Η υποδειγματοληψία μπορεί να βοηθήσει στην εξισορρόπηση της κατανομής των κλάσεων, αλλά μπορεί επίσης να οδηγήσει σε απώλεια πληροφοριών, καθώς διαγράφονται χρήσιμα δεδομένα της πλειοψηφικής κλάσης.

- **Συσσωμάτωση (Bagging)**:

Η συσσωμάτωση (Bagging, από το Bootstrap Aggregating) είναι μια τεχνική που χρησιμοποιείται για τη βελτίωση της απόδοσης των ταξινομητών σε μη ισορροπημένα δεδομένα. Η μέθοδος αυτή δημιουργεί πολλαπλά σύνολα δεδομένων εκπαίδευσης με αντικατάσταση (bootstrap samples) από το αρχικό σύνολο δεδομένων και εκπαιδεύει ένα μοντέλο για κάθε σύνολο δεδομένων. Οι προβλέψεις των επιμέρους μοντέλων συνδυάζονται στη συνέχεια μέσω πλειοψηφικής ψήφου ή άλλων τεχνικών συνδυασμού. Η συσσωμάτωση μπορεί να βοηθήσει στη μείωση της μεροληψίας και της διακύμανσης των μοντέλων, βελτιώνοντας έτσι την ακρίβεια πρόβλεψης, ειδικά στις μειοψηφικές κλάσεις.

- **Ενίσχυση (Boosting)**:

Η ενίσχυση (Boosting) είναι μια τεχνική συνδυασμού μοντέλων που βελτιώνει την απόδοση των ταξινομητών με τη διαδοχική δημιουργία ενός συνόλου μοντέλων, όπου κάθε νέο μοντέλο εκπαιδεύεται για να διορθώσει τα σφάλματα των προηγούμενων. Οι

πιο γνωστές μέθοδοι ενίσχυσης είναι το AdaBoost και το Gradient Boosting. Στην περίπτωση των μη ισορροπημένων δεδομένων, η ενίσχυση μπορεί να τροποποιηθεί για να δώσει μεγαλύτερη βαρύτητα στις παρατηρήσεις της μειοψηφικής κλάσης. Η ενίσχυση μπορεί να οδηγήσει σε σημαντική βελτίωση της απόδοσης των μοντέλων, ιδίως όταν τα δεδομένα είναι ιδιαίτερα ανισοκατανεμημένα.

Παρά τις προόδους στη μηχανική μάθηση, η ανισοβαρής κατανομή παραμένει μια σημαντική πρόκληση. Απαιτούνται εξατομικευμένες προσεγγίσεις για κάθε σύνολο δεδομένων, ακόμα και με μεγάλα σύνολα δεδομένων, προηγμένα νευρωνικά δίκτυα και σύγχρονα μοντέλα. Συνεχής έρευνα και ανάπτυξη είναι απαραίτητες για την αντιμετώπιση των αδυναμιών των υπαρχόντων μεθόδων και τη βελτίωση της διαχείρισης ανισοβαρών δεδομένων σε πραγματικές εφαρμογές.

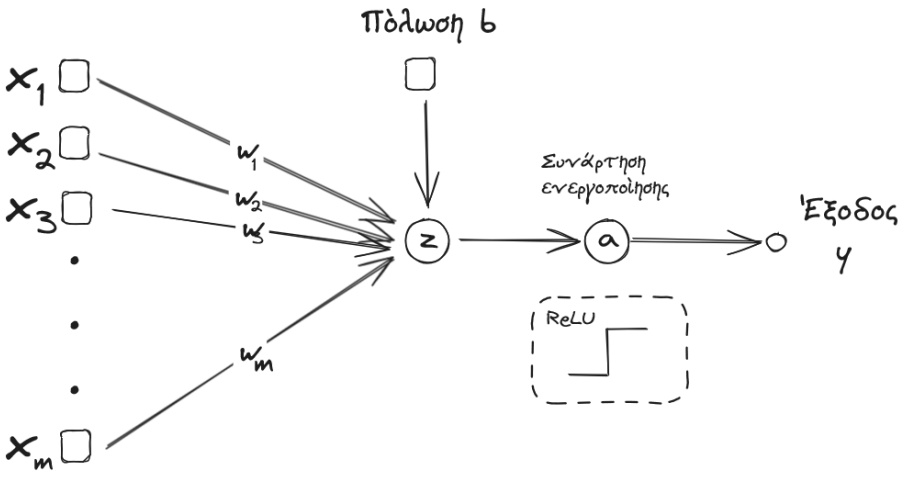
3.2 Βαθιά Μάθηση

Η βαθιά μάθηση αποτελεί υποκλάδο της μηχανικής μάθησης και εστιάζει στην εκπαίδευση πολυεπίπεδων νευρωνικών δικτύων για την επίλυση σύνθετων προβλημάτων. Η ιδέα του DL βασίζεται στη μίμηση της λειτουργίας του ανθρώπινου εγκεφάλου, αξιοποιώντας νευρωνικά δίκτυα για την αναγνώριση μοτίβων και τη λήψη αποφάσεων με βάση μεγάλα σύνολα δεδομένων [56].

Τα νευρωνικά δίκτυα αποτελούν τις βασικές δομές στην βαθιά μάθηση. Η αρχή λειτουργίας τους βασίζεται στην απομίμηση της λειτουργίας του ανθρώπινου εγκεφάλου, αξιοποιώντας δομές που προσομοιώνουν τους βιολογικούς νευρώνες και τις συνάψεις τους. Η θεμελιώδης αρχή των νευρωνικών δικτύων είναι η ικανότητά τους να μαθαίνουν από δεδομένα. Η διαδικασία μάθησης περιλαμβάνει την προσαρμογή των βαρών των συνδέσεων μεταξύ των νευρώνων με στόχο την ελαχιστοποίηση του σφάλματος πρόβλεψης. Αυτό επιτυγχάνεται μέσω αλγορίθμων βελτιστοποίησης, όπως η οπίσθια διάδοση σφάλματος (backpropagation), που επιτρέπει την ανανέωση των βαρών με βάση το σφάλμα πρόβλεψης. Στην παρούσα ενότητα θα αναλυθεί η λειτουργία των νευρωνικών δικτύων και ο μηχανισμός εκπαίδευσης τους, και διάφορες αρχιτεκτονικές τους [57].

3.2.1 Εισαγωγή στα Νευρωνικά Δίκτυα

Ένας τεχνητός νευρώνας, επίσης γνωστός ως perceptron [58], είναι μια μονάδα επεξεργασίας πληροφορίας, η οποία αποτελεί βασικό δομικό στοιχείο για την λειτουργία νευρωνικού δικτύου. Στο παρακάτω σχηματικό διάγραμμα βλέπουμε το μοντέλο ενός νευρώνα που αποτελεί την βάση για την σχεδίαση μιας μεγάλης οικογένειας νευρωνικών δικτύων που θα μελετήσουμε και παρακάτω.



Σχήμα 3.1: *Γράφημα ροής νευρώνα με συνάρτηση ενεργοποίησης*

Τα βασικά στοιχεία του νευρώνα, όπως φαίνονται και παραπάνω είναι:

- Ένα σύνολο διασυνδέσεων (συνάψεων), κάθε μια εκ των οποίων χαρακτηρίζεται από το δικό της βάρος. Συγκεκριμένα, ένα σήμα x_j στην είσοδο της σύναψης j συνδέεται με τον νευρώνα k και πολλαπλασιάζεται επί το συναπτικό βάρος w_{kj}. Επιπλέον, το βάρος ενός τεχνητού νευρώνα μπορεί να λαμβάνει σε αντίθεση με τον βιολογικό και αρνητικές τιμές.

- Έναν αθροιστή (adder) για την άθροιση των σημάτων εισόδου, σταθμισμένων από τα αντίστοιχα συναπτικά βάρη του νευρώνα.

- Μια συνάρτηση ενεργοποίησης (activation function) για τον περιορισμό του πλάτους του σήματος εξόδου ενός νευρώνα.

- Μια εξωτερικά εφαρμοζόμενη πόλωση ή προδιάθεση (bias) b_k, η οποία έχει ως αποτέλεσμα την αύξηση ή μείωση της δικτυακής διέγερσης της συνάρτησης ενεργοποίησης, ανάλογα με το αν είναι θετική ή αρνητική.

Άρα, ένας τεχνητός νευρώνας λαμβάνει πολλές εισόδους, κάθε μία από τις οποίες συνοδεύεται από ένα βάρος. Οι είσοδοι πολλαπλασιάζονται με τα αντίστοιχα βάρη και αθροίζονται για να δώσουν το συνολικό εισερχόμενο σήμα. Η διαδικασία αυτή περιγράφεται μαθηματικά ως ένας γραμμικός συνδυασμός των εισόδων και των βαρών:

$$z = \sum_{i=1}^{n} w_i * x_i + b$$

όπου z είναι το συνολικό εισερχόμενο σήμα, x_i είναι οι είσοδοι, w_i είναι τα βάρη και b είναι η προκατάληψη (bias). Αυτό το σήμα στη συνέχεια περνά από μια συνάρτηση ενεργοποίησης (activation function), η οποία καθορίζει την έξοδο του νευρώνα. Οι συνήθεις συναρτήσεις

ενεργοποίησης περιλαμβάνουν τις sigmoid, tanh και ReLU (Rectified Linear Unit).

$$a = \sigma(z)$$

όπου a είναι η έξοδος του νευρώνα και σ είναι η συνάρτηση ενεργοποίησης.

Τα βαθιά νευρωνικά δίκτυα (Deep Neural Networks) αποτελούν μια γενίκευση του Perceptron που μπορεί να προσεγγίσει πιο περίπλοκες μη γραμμικές συναρτήσεις και κατά συνέπεια να λύσει περισσότερα προβλήματα. Σε ένα νευρωνικό δίκτυο, οι νευρώνες οργανώνονται σε μορφή επιπέδων. Η απλούστερη δυνατή μορφή ενός τέτοιου δικτύου είναι ένα επίπεδο πρόσθιας τροφοδότησης (feedforward), όπου υπάρχει ένα επίπεδο εισόδου, το οποίο συνδέεται απευθείας με ένα επίπεδο νευρώνων εξόδου. Σε αυτό το δίκτυο αποδίδεται ο χαρακτηρισμός ενός επιπέδου, μιας και το επίπεδο εισόδου, αφού εκεί δεν εκτελείται κανένας υπολογισμός, δεν προσμετράται. Η εξελιγμένη μορφή αυτής της πρώτης εκδοχής νευρωνικών δικτύων είναι τα πολυεπίπεδα δίκτυα πρόσθιας τροφοδότησης. Σε αντίθεση με τα "ενός επιπέδου" χαρακτηρίζονται από την παρουσία ενός ή και περισσότερων κρυφών επιπέδων. Η λειτουργία των κρυφών νευρώνων είναι να παρεμβαίνουν μεταξύ της εξωτερικά προερχόμενης εισόδου και της εξόδου του δικτύου με κάποιο χρήσιμο τρόπο.

3.2.2 Συνάρτηση ενεργοποίησης

Η δομή και η λειτουργία ενός νευρωνικού δικτύου επηρεάζονται καθοριστικά από τις συναρτήσεις ενεργοποίησης που χρησιμοποιούνται στους νευρώνες. Οι συναρτήσεις ενεργοποίησης προσδίδουν μη γραμμικότητα στο δίκτυο, επιτρέποντας στους νευρώνες να μάθουν και να αναπαραστήσουν πιο σύνθετες σχέσεις και μοτίβα στα δεδομένα. Χωρίς τη μη γραμμικότητα που εισάγουν αυτές οι συναρτήσεις, ένα νευρωνικό δίκτυο δεν θα ήταν τίποτα περισσότερο από έναν γραμμικό πολλαπλασιαστή, περιορίζοντας έτσι την ικανότητά του να επιλύει πολύπλοκα προβλήματα. Οι πιο κοινές συναρτήσεις ενεργοποίησης περιλαμβάνουν τις sigmoid, tanh και ReLU, κάθε μία από τις οποίες έχει διαφορετικά χαρακτηριστικά και χρήσεις.

Σιγμοειδής Συνάρτηση

Η σιγμοειδής συνάρτηση, της οποίας η γραφική παράσταση έχει σχήμα "S", είναι η πλέον κοινή μορφή συνάρτησης ενεργοποίησης που χρησιμοποιείται ευρέως στα νευρωνικά δίκτυα. Η μαθηματική της έκφραση είναι η εξής:

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

όπου το z είναι το εισερχόμενο σήμα στον νευρώνα. Η συνάρτηση sigmoid έχει πεδίο τιμών μεταξύ 0 και 1. Είναι ιδιαίτερα χρήσιμη σε προβλήματα δυαδικής ταξινόμησης, καθώς η έξοδός της μπορεί να ερμηνευτεί ως πιθανότητα να ανήκει σε μία από τις δύο κατηγορίες. Ωστόσο, η συνάρτηση sigmoid υποφέρει από το πρόβλημα του vanishing gradient, όπου τα παράγωγά της γίνονται εξαιρετικά μικρά για μεγάλες ή μικρές τιμές του z, περιορίζοντας την αποτελεσματική εκπαίδευση βαθιών δικτύων.

Συνάρτηση ReLU (Rectified Linear Unit)

Η συνάρτηση ReLU είναι μία από τις πιο δημοφιλείς και ευρέως χρησιμοποιούμενες συναρτήσεις ενεργοποίησης στις σύγχρονες εφαρμογές βαθιάς μάθησης. Η μαθηματική της έκφραση είναι η εξής:

$$ReLU(z) = max(0, z)$$

όπου z είναι το εισερχόμενο σήμα στον νευρώνα. Η ReLU καθορίζει το κατώφλι της εισόδου στο μηδέν, επιστρέφοντας 0 για αρνητικές τιμές και την ίδια την είσοδο z για θετικές τιμές. Για εισόδους μεγαλύτερες του 0, η ReLU συμπεριφέρεται σαν γραμμική συνάρτηση με παράγοντα το 1. Άρα δεν μεταβάλλει την κλίμακα των θετικών εισόδων και επιτρέπει στην κλίση να περάσει αμετάβλητη κατά τη διάρκεια της οπισθοδιάδοσης. Αυτή η ιδιότητα ελαχιστοποιεί το πρόβλημα ανισηινγ γραδιεντ και επιτρέπει έτσι την εκπαίδευση βαθύτερων δικτύων.

Παρόλο που η ReLU είναι γραμμική για το μισό του χώρου εισόδου της, είναι τεχνικά μια μη γραμμική συνάρτηση επειδή έχει ένα μη διαφοροποιήσιμο σημείο στο $z = 0$, όπου αλλάζει απότομα από το z. Αυτή η μη γραμμικότητα επιτρέπει στα νευρωνικά δίκτυα να μαθαίνουν πολύπλοκα μοτίβα. Ταυτόχρονα, όμως δημιουργεί και το πρόβλημα dying ReLU, όπου οι νευρώνες μπορεί να "πεθάνουν" κατά την εκπαίδευση αν λαμβάνουν συνεχώς αρνητικές τιμές, κάτι που καθιστά τις εξόδους τους μηδενικές. Η απλή ωστόσο μορφή της, καθιστά την συνάρτηση ReLU υπολογιστικά ανέξοδη και αυτό επιτρέπει στα δίκτυα να κλιμακώνονται σε πολλά επίπεδα χωρίς σημαντική αύξηση του υπολογιστικού φόρτου, σε σύγκριση με πιο σύνθετες συναρτήσεις όπως η tanh ή η sigmoid.

Συνοψίζοντας, οι συναρτήσεις ενεργοποίησης διαδραματίζουν κρίσιμο ρόλο στην αρχιτεκτονική των νευρωνικών δικτύων, προσδίδοντας την απαραίτητη μη γραμμικότητα για την επίλυση σύνθετων προβλημάτων και την μάθηση πολύπλοκων σχέσεων από τα δεδομένα.

3.2.3 Εκπαίδευση ενός νευρωνικού δικτύου

Η εκπαίδευση των νευρωνικών δικτύων πραγματοποιείται μέσω του αλγορίθμου Gradient Descent. Για να λειτουργήσει ο αλγόριθμος αυτός, απαιτείται σε κάθε επανάληψη ο υπολογισμός της μερικής παραγώγου της συνάρτησης σφάλματος ως προς όλες τις παραμέτρους όλων των επιπέδων του δικτύου. Αυτό επιτυγχάνεται αποτελεσματικά με τη χρήση του αλγορίθμου Backpropagation.

Αρχικά, υπολογίζεται η παράγωγος της συνάρτησης σφάλματος ως προς τις παραμέτρους του τελευταίου επιπέδου με άμεσο τρόπο. Στη συνέχεια, χρησιμοποιώντας τον κανόνα της αλυσίδας $\frac{df}{dx} = \frac{df}{dy}\frac{dy}{dx}$, μπορούμε να βρούμε τις παραγώγους των παραμέτρων του αμέσως προηγούμενου επιπέδου. Αυτή η διαδικασία περιλαμβάνει την επαναχρησιμοποίηση των μερικών παραγώγων που έχουν ήδη υπολογιστεί και των γνωστών παραγώγων των συναρτήσεων ενεργοποίησης και των αφινικών μετασχηματισμών.

Η διαδικασία συνεχίζεται μέχρι το πρώτο επίπεδο του δικτύου. Τέλος, τα βάρη ανανεώνονται με κατάλληλο τρόπο ώστε να μειωθεί το σφάλμα. Η όλη διαδικασία επαναλαμβάνεται

μέχρι να επιτευχθεί το ζητούμενο τοπικό ελάχιστο, ολοκληρώνοντας την εκπαίδευση του δικτύου.

Ο αλγόριθμος Gradient Descent και η τεχνική Backpropagation αποτελούν τη βάση για την εκπαίδευση των νευρωνικών δικτύων, επιτρέποντας την προσαρμογή των βαρών μέσω επαναλαμβανόμενων βελτιώσεων, εξασφαλίζοντας έτσι τη βελτίωση της απόδοσης του δικτύου στον επιθυμητό στόχο.

3.2.4 Τύποι Νευρωνικών Δικτύων

Πολλαπλών Επιπέδων Perceptron

Τα Πολλαπλών Επιπέδων Perceptron (MLPs) είναι ένας τύπος τεχνητού νευρωνικού δικτύου που χρησιμοποιείται ευρέως σε προβλήματα ταξινόμησης και παλινδρόμησης. Τα MLPs αποτελούνται από πολλαπλά στρώματα νευρώνων, όπου κάθε νευρώνας συνδέεται με όλους τους νευρώνες του προηγούμενου και του επόμενου στρώματος. Οι νευρώνες αυτοί χρησιμοποιούν μη γραμμικές συναρτήσεις ενεργοποίησης, όπως η ReLU (Rectified Linear Unit), για να επιτρέψουν στο δίκτυο να μάθει πολύπλοκες σχέσεις από τα δεδομένα.

Η βασική αρχή λειτουργίας των MLPs είναι η εξής:

- **Είσοδος Δεδομένων**: Τα δεδομένα εισέρχονται στο δίκτυο μέσω του στρώματος εισόδου.
- **Μετασχηματισμός Δεδομένων**: Κάθε επίπεδο του δικτύου μετασχηματίζει τις εισόδους του μέσω γραμμικών συνδυασμών και μη γραμμικών συναρτήσεων ενεργοποίησης, όπως αυτές που αναφέρθηκαν στην υποενότητα 3.2.2.
- **Προσαρμογή Βαρών**: Τα βάρη των συνδέσεων μεταξύ των νευρώνων προσαρμόζονται κατά τη διάρκεια της εκπαίδευσης με στόχο την ελαχιστοποίηση μιας συνάρτησης κόστους μέσω του μηχανισμού Backpropagation.

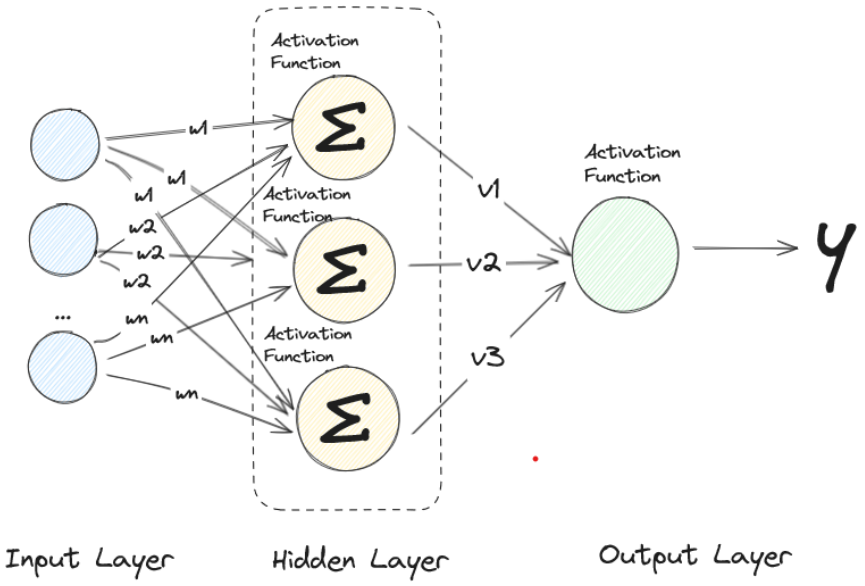
Πέρα από την αρχιτεκτονική του μοντέλου, σημαντικό ρόλο διαδραματίζουν και οι παρακάτω παράμετροι τεχνικές που μπορούν να ενσωματωθούν στο νευρωνικό δίκτυο για να το βελτιώσουν:

- **Dropout**:

  Το Dropout είναι μια τεχνική regularization που χρησιμοποιείται για την αποφυγή υπερεκπαίδευσης (overfitting) στα νευρωνικά δίκτυα. Κατά την εκπαίδευση, μέσω της τεχνικής Dropout τυχαία «απενεργοποιείται» ένα προκαθορισμένο ποσοστό νευρώνων σε κάθε στρώμα σε κάθε βήμα εκπαίδευσης. Αυτό σημαίνει πως αυτοί οι νευρώνες δεν συμμετέχουν στην επεξεργασία των δεδομένων και δεν προσαρμόζονται τα βάρη τους. Κατά τη διάρκεια της πρόβλεψης (evaluation) του μοντέλου, όλοι οι νευρώνες ωστόσο είναι ενεργοποιημένοι, όμως τα βάρη τους κλιμακώνονται κατάλληλα για να διατηρηθεί η συνοχή των εξόδων.

- **Βάρη κλάσεων στην συνάρτηση κόστους**:

  Στην περίπτωση, όπου τα δεδομένα εκπαίδευσης είναι μη-ισορροπημένα, μπορεί να είναι χρήσιμο να χρησιμοποιηθούν βάρη κλάσεων στη συνάρτηση κόστους. Αυτά τα



Σχήμα 3.2: *Τυπική μορφή ενός MultiLayer Perceptron Νευρωνικού Δικτύου*

βάρη αυξάνουν την συνεισφορά των λιγότερο αντιπροσωπευόμενων κλάσεων στη συνάρτηση κόστους, εξασφαλίζοντας ότι το μοντέλο θα μάθει να προβλέπει σωστά όλες τις κλάσεις ανεξάρτητα από την κατανομή τους.

- **Ρυθμός εκμάθησης και Ποινή L2 στην συνάρτηση βελτιστοποίησης**

  Οι συναρτήσεις βελτιστοποίησης είναι αλγόριθμοι που χρησιμοποιούνται για την προσαρμογή των βαρών του δικτύου, με στόχο την ελαχιστοποίηση της συνάρτησης κόστους. Ένας από τους πιο δημοφιλείς αλγόριθμους είναι ο Adam (Adaptive Moment Estimation).

  - **Learning Rate**: Ο ρυθμός εκμάθησης (LR) είναι μια υπερπαράμετρος που καθορίζει το βήμα που λαμβάνει ο αλγόριθμος κατά την ενημέρωση των βαρών. Μια πολύ υψηλή τιμή μπορεί να οδηγήσει σε αστάθεια, ενώ μια πολύ χαμηλή τιμή μπορεί να επιβραδύνει την εκπαίδευση.
  - **L2-Regularization Term**: Η ποινή L2 προσθέτει έναν όρο στη συνάρτηση κόστους που είναι ίσο με το τετράγωνο του μεγέθους των βαρών. Αυτό αποθαρρύνει την ανάπτυξη μεγάλων βαρών και βοηθά στην αποφυγή υπερεκπαίδευσης, καθιστώντας το μοντέλο πιο γενικεύσιμο στα δεδομένα δοκιμών.

Συνολικά, τα MLPs, σε συνδυασμό με τεχνικές όπως το Dropout, τα βάρη κλάσεων στη συνάρτηση κόστους, και προηγμένες μεθόδους βελτιστοποίησης, αποτελούν ισχυρά εργαλεία για την ανάλυση και ταξινόμηση πολύπλοκων βιολογικών δεδομένων, όπως το γονιδίωμα στον καρκίνο του μαστού.

Κεφάλαιο 4

Δεδομένα και Μέθοδοι

Στην παρούσα μελέτη χρησιμοποιούμε το σύνολο δεδομένων σε μορφή πίνακα του The Cancer Genome Atlas (TCGA) [59]. Το TCGA είναι ένα ανοιχτό (open source) έργο που στοχεύει στη δημιουργία ενός άτλαντα, ο οποίος αποτελείται από τα γονιδιωματικά προφίλ πολλών ειδών καρκίνου. Η τεράστια αυτή βάση δεδομένων είναι αποτέλεσμα της κοινής προσπάθειας μεταξύ του National Institute of Health (NIH) και του National Human Genome Research Institute που έχει ξεκινήσει από το 2006.

Μέχρι στιγμής οι ερευνητές του TCGA έχουν χαρακτηρίσει μοριακά πάνω από 20.000 πρωτογενείς καρκίνους και έχουν ταιριάξει δείγματα που καλύπτουν 33 τύπους καρκίνου. Μελέτες μεμονωμένων τύπων καρκίνου, καθώς και περιεκτικές αναλύσεις όλων των τύπων καρκίνου έχουν καταφέρει να διευρύνουν την τρέχουσα γνώση σχετικά με την ογκογένεση. Κύριος στόχος του έργου είναι η δημόσια παροχή συνόλων δεδομένων με στόχο την βελτίωση των διαγνωστικών μεθόδων, των προτύπων θεραπείας και, τέλος, την πρόληψη του καρκίνου.

4.1 Περιγραφή δεδομένων

Η διάθεση των δεδομένων από το TCGA προσφέρει την δυνατότητα για απομόνωση και ανάλυση δεδομένων σχετικά με τον καρκίνο του μαστού. Συγκεκριμένα, χρησιμοποιούνται:

- **Κλινικά μεταδεδομένα ασθενών**, τα οποία περιλαμβάνουν πληροφορίες σχετικά το ιστορικό, τη θεραπεία, την πρόγνωση και τη διάγνωση των ασθενών.

- **Δεδομένα γονιδιακής έκφρασης**, που παρέχουν λεπτομερή στοιχεία για τα επίπεδα έκφρασης των γονιδίων στους καρκινικούς ιστούς των ασθενών.

Το σύνολο των κλινικών δεδομένων που χρησιμοποιήθηκε περιλαμβάνει τις ακόλουθες μεταβλητές: ηλικία, καθαρότητα όγκου, παθολογικό στάδιο, παθολογικό T, N και M στάδιο, ιστολογικό τύπο, αριθμό λεμφαδένων, κατάσταση Estrogen Receptor (ER), κατάσταση Progesterone Receptor (PR), κατάσταση HER, φύλο, θεραπεία ραδιενέργειας, φυλή, εθνικότητα, πιθανότητα επιβίωσης.

Οι παραπάνω μεταβλητές θα φανούν χρήσιμες για την περαιτέρω συγκεκριμενοποίηση και τον περιορισμό των ασθενών για την μελέτη της γονιδιακής έκφρασης σε μικρότερα υποσύνολα ασθενών που μοιράζονται περισσότερα κοινά χαρακτηριστικά. Παρακάτω αναλύονται μερικά από τα πιο σημαντικά χαρακτηριστικά που εντοπίζονται στα κλινικά μεταδεδομένα:

- **Ηλικία**: Αναφέρεται στην ηλικία του ασθενούς τη στιγμή της διάγνωσης. Αυτή η μεταβλητή επιτρέπει τη διερεύνηση των επιδράσεων της ηλικίας στην εξέλιξη του καρκίνου του μαστού και μπορεί να παρέχει πληροφορίες σχετικά με τον αντίκτυπο της ηλικίας στην έναρξη και τη σοβαρότητα της νόσου [60].

- **Φύλο**: Ο καρκίνος του μαστού, όπως αναφέρεται και στο Κεφάλαιο 2 μπορεί να επηρεάσει τόσο το αντρικό όσο και το γυναικείο γενετικό φύλο. Ωστόσο, η παρούσα έρευνα εστιάζει στο γυναικείο φύλο [16].

- **Καθαρότητα όγκου**: Εκφράζει το ποσοστό των καρκινικών κυττάρων στον όγκο σε σχέση με τα φυσιολογικά κύτταρα. Αυτή η μεταβλητή είναι κρίσιμη για την ακρίβεια των γονιδιακών αναλύσεων, καθώς υψηλότερη καθαρότητα όγκου μπορεί να οδηγήσει σε πιο αξιόπιστα αποτελέσματα.

- **Παθολογικό στάδιο (Stage)**: Περιγράφει το στάδιο του καρκίνου του μαστού με βάση την παθολογική διάγνωση που βρίσκεται ο ασθενής. Επιπλέον, μεταβλητές παθολογικών σταδίων T, N, M περιγράφουν την έκταση του καρκίνου στον οργανισμό.

- **Ιστολογικός τύπος**: Αναφέρεται στον τύπο του ιστού από τον οποίο προέρχεται ο καρκίνος και μπορεί να επηρεάσει την πρόγνωση και την απόκριση στη θεραπεία. Εδώ θα εστιάσουμε στον διηθητικό πορογενή καρκίνο (invasive ductal carcinoma) και στον διηθητικό λοβιακό καρκίνο (invasive lobular carcinoma).

- **Φυλή, εθνικότητα**: Αυτές οι δημογραφικές μεταβλητές μπορούν να επηρεάσουν την εμφάνιση, την εξέλιξη και την ανταπόκριση στον καρκίνο του μαστού, επιτρέποντας την αναγνώριση διαφορών μεταξύ διαφορετικών πληθυσμών [61].

Οι παραπάνω μεταβλητές θα χρησιμεύσουν στην κατανόηση των διαφόρων παραγόντων που επηρεάζουν την πορεία και την έκβαση της νόσου, επιτρέποντας την καλύτερη κατηγοριοποίηση των ασθενών και την εξατομικευμένη προσέγγιση στη θεραπεία του καρκίνου του μαστού.

Το σύνολο των δεδομένων γονιδιακής έκφρασης περιλαμβάνει τη γονιδιακή έκφραση 20.155 γονίδιων για 1.097 ασθενείς. Πρόκειται για ένα σύνολο δεδομένων μεγάλων διαστάσεων, καθώς ο αριθμός των γονιδίων είναι ιδιαίτερα υψηλός σε σχέση με τον αριθμό των δειγμάτων.

Τα μεταδεδομένα του συνόλου δεδομένων έχουν τα χαρακτηριστικά RNAseq (Illumina HiSeq platform, Gene-level, RPKM). Οι πληροφορίες αυτές υποδηλώνουν πως τα δεδομένα αρχικά έχουν εξαχθεί μέσω RNA αλληλούχισης (RNA-seq) με τη χρήση της τεχνικής Illumina HiSeq. Ο χαρακτηρισμός Gene-level συνεπάγεται πως τα δεδομένα έχουν χαρτογραφηθεί και εκφραστεί με βάση τα γονίδια.

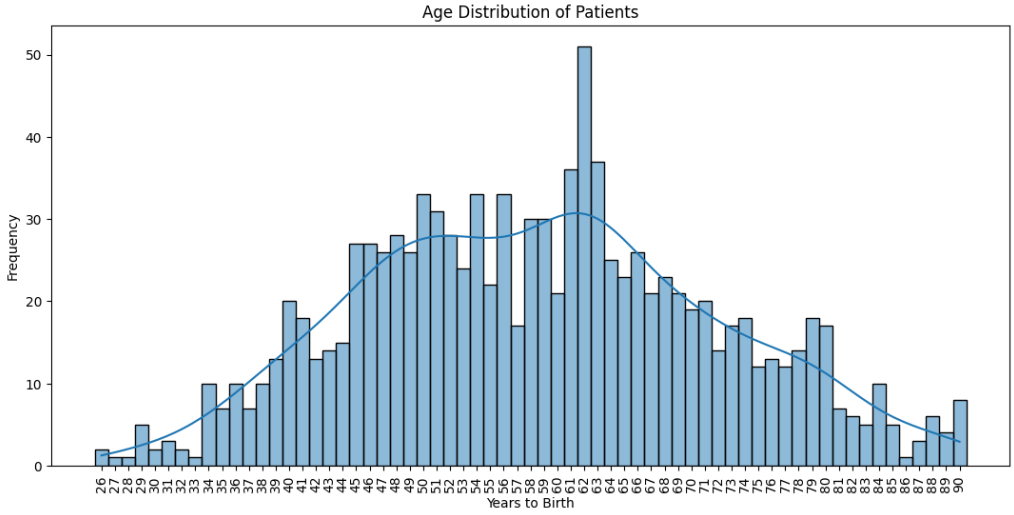
Η μέθοδος RNA-sequencing χρησιμοποιείται κυρίως για την ανάλυση των mRNA (messenger RNA). Ωστόσο μπορεί επίσης να περιλαμβάνει και άλλους τύπους RNA που αντιστοιχούν σε συγκεκριμένα γονίδια, όπως micro-RNA. Επιπλέον, αυτά τα δεδομένα είναι κανονικοποιημένα με τη μέθοδο RPKM (Reads Per Kilobase of transcript, per Million mapped reads), η οποία λαμβάνει υπόψη το βάθος της αλληλούχισης και το μήκος των γονιδίων. Η

κανονικοποίηση με RPKM επιτρέπει τη σύγκριση της έκφρασης γονιδίων μεταξύ διαφορετικών δειγμάτων.

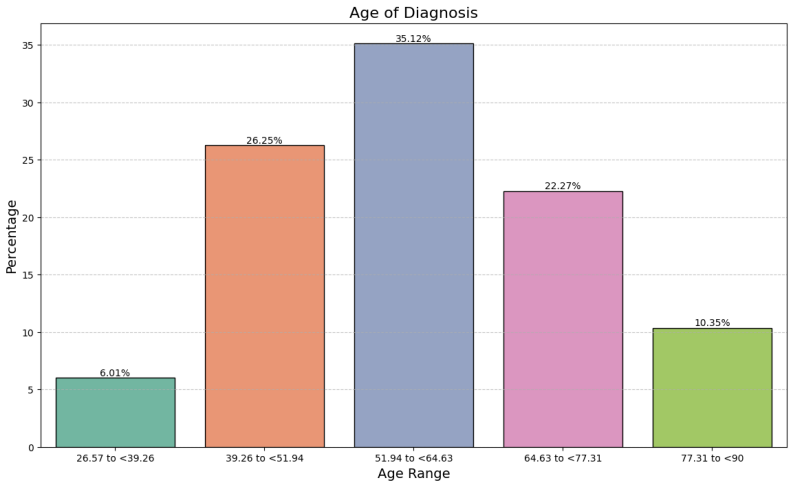
Η ύπαρξη μεγάλου αριθμού γονιδίων, εδώ 20.155, προσφέρει την ευκαιρία για ευρεία και λεπτομερή διερεύνηση της γονιδιακής δραστηριότητας, αλλά ταυτόχρονα απαιτεί προσεκτικό χειρισμό και έλεγχο. Συνεπώς, απαραίτητα βήματα για την επιτυχή αξιοποίηση αυτού του συνόλου δεδομένων είναι η σωστή επεξεργασία των δεδομένων, η κανονικοποίηση και η ανάλυση με κατάλληλα υπολογιστικά εργαλεία.

Για την καλύτερη οπτικοποίηση και αντίληψη τόσο των κλινικών μεταδεδομένων, όσο και των γονιδιακών, είναι σημαντική η καλύτερη μελέτη τους.

Στο Σχήμα 4.1 παρουσιάζεται η κατανομή των ασθενών με βάση την ηλικία τους.



Σχήμα 4.1: *Κατανομή Ασθενών με βάση την ηλικία*



Σχήμα 4.2: *Κατανομή ασθενών με βάση την ηλικιακή ομάδα διάγνωσης*

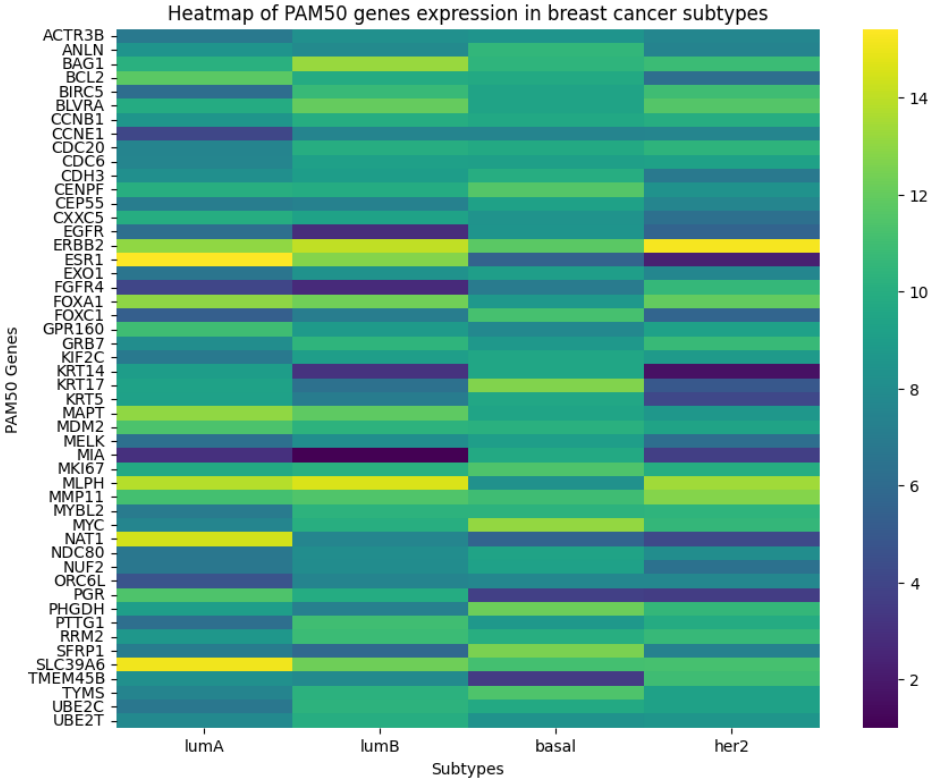
Είναι γνωστό από τη βιβλιογραφία ότι η αντιμετώπιση και η διάγνωση του καρκίνου του μαστού επηρεάζονται σημαντικά από αυτόν τον παράγοντα [60]. Παρατηρούμε ότι στο σύνολο δεδομένων που διαθέτουμε ισχύουν τα εξής :

- Οι ηλικίες των ασθενών κατανέμονται σε ένα ευρύ φάσμα, με συγκεκριμένες ηλικιακές ομάδες να παρουσιάζουν υψηλότερη συχνότητα. Οι ηλικίες κυμαίνονται από 26 έως 90 έτη, με την πλειοψηφία των ασθενών να βρίσκεται μεταξύ 40 και 75 ετών.

- Οι περισσότερες περιπτώσεις συγκεντρώνονται στις ηλικίες μεταξύ 50 και 65 ετών. Συγκεκριμένα, παρατηρείται μία κορύφωση γύρω στην ηλικία των 60 ετών, που είναι γνωστή ως η ηλικιακή ομάδα με τη μεγαλύτερη συχνότητα εμφάνισης καρκίνου του μαστού. Αυτή η παρατήρηση συμβαδίζει με τα ευρήματα της υπάρχουσας βιβλιογραφίας, η οποία υποδεικνύει αυξημένο κίνδυνο για καρκίνο του μαστού στις μετεμμηνοπαυσιακές γυναίκες [60].

Η παραπάνω κατανομή διασφαλίζει πως το σύνολο των δεδομένων αποτελεί ένα αντιπροσωπευτικό δείγμα ηλικιακού εύρους ασθενών με βάση τα όσα είναι γνωστά από τις αντίστοιχες βιβλιογραφικές αναφορές.

Το σύνολο των δεδομένων περιέχει τη γονιδιακή έκφραση χιλιάδων γονιδίων. Το πλήθος αυτό, καθιστά απαγορευτική, αλλά και μη ουσιώδη την μελέτη όλων των γονιδίων - ειδικά πριν την επεξεργασία τους. Ωστόσο, στο πλαίσιο της περιγραφής των δεδομένων θα παρατηρήσουμε πιο προσεκτικά το υποσύνολο γονιδίων PAM50. Όπως αναφέρεται και στην Ενότητα 2.2.2, πρόκειται για ένα πάνελ 50 γονιδίων που είναι ερευνητικά αποδεδειγμένο πως συσχετίζεται με τον καρκίνο του μαστού και συγκεκριμένα με την ταξινόμηση υποτύπου της νόσου [62].

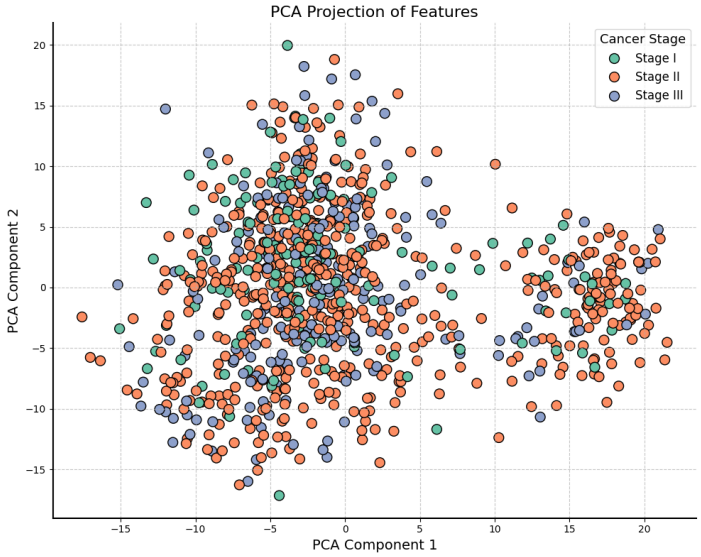


Σχήμα 4.3: *Διάγραμμα θερμότητας της έκφρασης γονιδίων PAM50 για 4 διαφορετικούς υποτύπους της νόσου*

Στο Σχήμα 4.3 φαίνεται το διάγραμμα θερμότητας των 50 γονιδίων ως προς 4 ασθενείς

που έχουν διαγνωστεί με διαφορετικούς υποτύπους του IDC και που έχουν υψηλή καθαρότητα όγκου. Παρατηρούνται τα διαφορετικά χαρακτηριστικά του κάθε υποτύπου που αντικατοπτρίζονται και στην γονιδιακή έκφραση των γονιδίων.

Επιπλέον, για την καλύτερη οπτικοποίηση των δεδομένων πραγματοποιούμε ανάλυση κύριων συνιστωσών PCA, η οποία απεικονίζει τα δεδομένα χαρακτηριστικών σε δύο κύριες συνιστώσες, όπως φαίνεται στο Σχήμα 4.4. Κάθε χρώμα αντιστοιχεί σε μία κατηγορία (Stage *I* - πράσινο, Stage *II* - πορτοκαλί, Stage *III* - μπλε). Παρατηρούμε πως τα δεδομένα είναι διασκορπισμένα και στις δύο συνιστώσες, με αρκετή επικάλυψη μεταξύ των κατηγοριών. Δεν υπάρχει σαφής διαχωρισμός μεταξύ των τριών κατηγοριών στον χώρο των δύο πρώτων συνιστωσών, υποδεικνύοντας ότι οι δύο πρώτες συνιστώσες δεν επαρκούν για να διαχωρίσουν πλήρως τα δεδομένα βάσει των κατηγοριών τους. Αυτό μπορεί να σημαίνει ότι οι διαφορές μεταξύ των κατηγοριών είναι πιο εμφανείς σε υψηλότερες διαστάσεις, ή ότι τα χαρακτηριστικά που χρησιμοποιήθηκαν δεν είναι τα πλέον κατάλληλα για την διάκριση των κατηγοριών, ωστόσο σε κάθε περίπτωση μας προσδίδει μια καλύτερη εικόνα για τα δεδομένα που έχουμε στην διάθεση μας και την ευκολία ταξινόμησης τους.



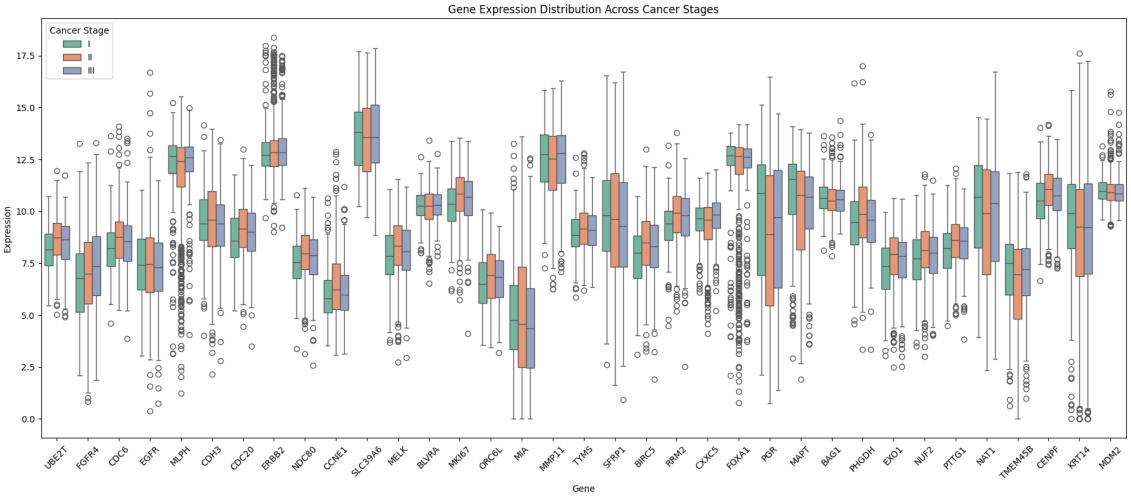
Σχήμα 4.4: *Ανάλυση κυρίων συνιστωσών των δεδομένων γονιδιακής έκφρασης σε 2 συνιστώσες*

Στο Σχήμα 4.5 απεικονίζεται η κατανομή των επιπέδων έκφρασης γονιδίων που ανήκουν στο PAM50 σε διαφορετικά στάδια καρκίνου (*I*, *II* και *III*).

Τα επίπεδα έκφρασης ποικίλλουν σημαντικά μεταξύ διαφορετικών γονιδίων. Ορισμένα γονίδια παρουσιάζουν συνολικά υψηλότερα επίπεδα έκφρασης, ενώ άλλα έχουν χαμηλότερα επίπεδα έκφρασης. Για παράδειγμα, τα γονίδια UBE2T, CDC6, ERBB2, MKI67 παρουσιάζουν υψηλότερη έκφραση σε μεταγενέστερα στάδια (*II* και *III*) σε σύγκριση με το στάδιο *I*. Αντίθετα, τα γονίδια PGR, FOXA1, MAPT τείνουν να εμφανίζουν υψηλότερη έκφραση στα στάδια (*I* και *II*) σε σύγκριση με το στάδιο *III*.

Ωστόσο, το μεγαλύτερο μέρος των γονιδίων εμφανίζει παρόμοια επίπεδα έκφρασης σε όλα τα στάδια, γεγονός που υποδηλώνει ότι η έκφρασή των γονιδίων PAM50 μπορεί να μην επηρεάζεται σημαντικά από την εξέλιξη των σταδίων καρκίνου. Επιπλέον, από τον μεγάλο

αριθμό ακραίων τιμών που υπάρχουν στα δεδομένα έκφρασης (outliers), υποδεικνύεται μεγάλη μεταβλητότητα στα επίπεδα έκφρασης μεταξύ των δειγμάτων, υπογραμμίζοντας την ετερογένεια στη γονιδιακή έκφραση μεταξύ διαφορετικών ασθενών.



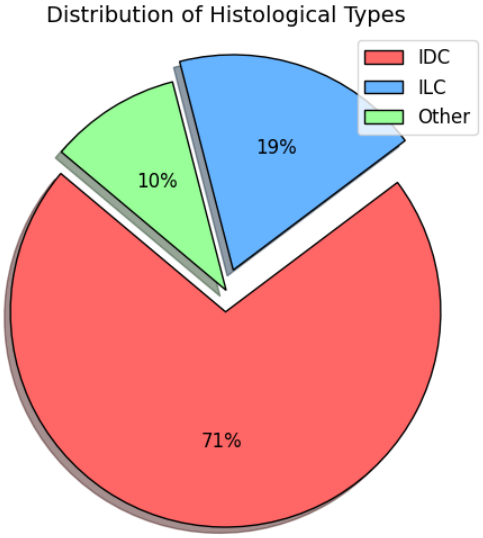
Σχήμα 4.5: *Κατανομή των επιπέδων έκφρασης γονιδίων PAM50 σε διαφορετικά στάδια καρκίνου*

4.2 Προεπεξεργασία δεδομένων

Η επεξεργασία των δεδομένων ξεκινά από τα κλινικά μεταδεδομένα που έχουμε στην διάθεση μας.

Το σύνολο δεδομένων που χρησιμοποιήθηκε σε αυτή τη μελέτη αποτελούνταν αρχικά από δεδομένα 1.097 διαφορετικών ασθενών διαγνωσμένοι με καρκίνο του μαστού. Κάθε ασθενής χαρακτηρίζεται από ένα μοναδικό αναγνωριστικό ασθενούς. Η κατανομή των ασθενών στις διάφορες διαγνωστικές κατηγορίες εντός του αρχικού συνόλου δεδομένων παρουσιάζεται παρακάτω:

Για να διασφαλιστεί η ομοιογένεια ανάμεσα στα δείγματα εστιάζουμε στον διηθητικό πορογενή καρκίνο (IDC) και στον διηθητικό λοβιακό καρκίνο (ILC) αφαιρώντας τους ασθενείς από τις υπόλοιπες κατηγορίες. Παρατηρούμε και από το Σχήμα 4.6 πως οι αυτές είναι οι ιστολογικές κατηγορίες που εκπροσωπούνται περισσότερο στο σύνολο δεδομένων. Αυτή η επιλογή επιτρέπει την εστίαση στους δύο πιο κοινούς τύπους καρκίνου του μαστού, μειώνοντας την ετερογένεια και αυξάνοντας την αξιοπιστία των αποτελεσμάτων.



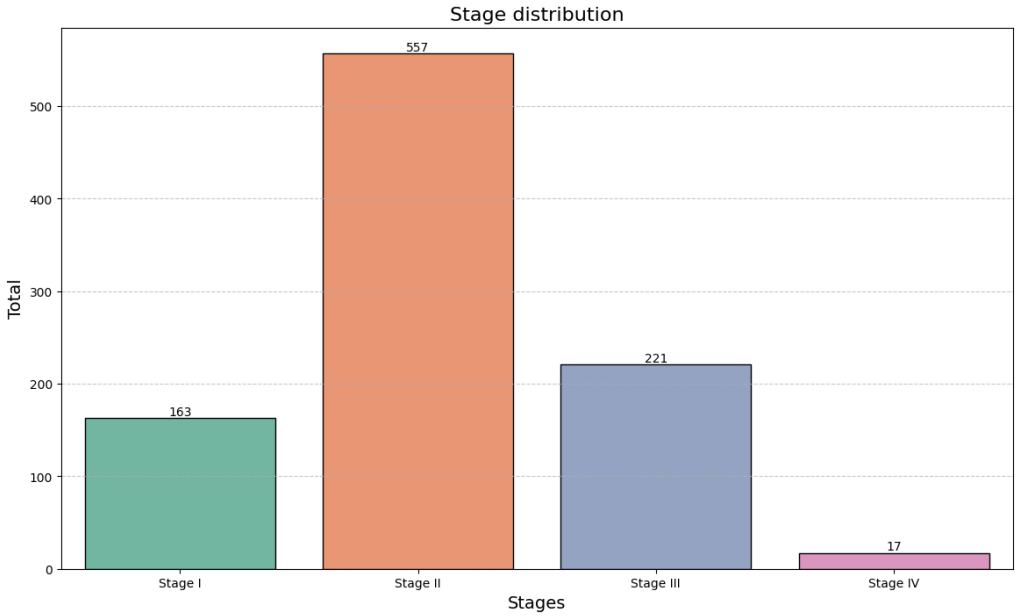
Σχήμα 4.6: *Κατανομή Ιστολογικών Τύπων*

Για τον ίδιο λόγο, αφαιρούμε τις περιπτώσεις και τα δείγματα αρσενικού φύλου από τα δεδομένα. Αφενός, γιατί αυτά απαρτίζουν μόνο ένα μικρό ποσοστό των συνολικών δεδομένων, περίπου 0,01%, και αφετέρου, γιατί η συμπερίληψή τους θα επηρέαζε την ομοιογένεια των δειγμάτων. Η αφαίρεση αυτών των περιπτώσεων επιτρέπει να διατηρήσουμε μια συνεπή και συγκρίσιμη ομάδα δειγμάτων, επικεντρωμένη αποκλειστικά στις γυναίκες ασθενείς με τους συγκεκριμένους τύπους καρκίνου του μαστού.

Η σημαντικότερη ίσως μεταβλητή που υπάρχει στο σύνολο δεδομένων είναι το παθολογικό στάδιο στο οποίο βρίσκεται ο καρκίνος του μαστού της ασθενούς. Αυτή η μεταβλητή αποτελεί την πεμπτουσία του μοντέλου ταξινόμησης.

Αναλύοντας την κατανομή των ασθενών ανά στάδιο, γίνεται αντιληπτό πως υπάρχουν πολύ περισσότερα δείγματα για ασθενείς που βρίσκονται στάδιο *II*, με τα δείγματα αυτής της κατηγορίας να είναι τουλάχιστον τριπλάσια σε σύγκριση με τα δείγματα του σταδίου *I*. Η μεγάλη αυτή συγκέντρωση δειγμάτων στο στάδιο *II* υποδηλώνει ότι ένας σημαντικός αριθμός ασθενών διαγιγνώσκεται σε αυτό το μεσαίο στάδιο της νόσου. Παρατηρείται επίσης πως είναι διαθέσιμα μόνο ελάχιστα δείγματα για το στάδιο *IV*. Λόγω του μικρού αυτού αριθμού, καθίσταται στατιστικά δύσκολη η συμπερίληψη αυτών των ασθενών στις αναλύσεις. Συνεπώς, για την διασφάλιση της αξιοπιστίας των αποτελεσμάτων και την ομοιογένεια του συνόλου δεδομένων, θα αφαιρεθούν οι συγκεκριμένες ασθενείς από την ανάλυσή και η εστίαση του προγνωστικού μοντέλου θα γίνει με βάση τις υπόλοιπες κατηγορίες, δηλαδή τα στάδια *I, II* και *III*.

Γίνεται φυσικά αντιληπτό και από το Σχήμα 4.7 πως το συγκεκριμένο σύνολο δεδομένων δεν είναι ισορροπημένο ως προς το παθολογικό στάδιο της νόσου.



Σχήμα 4.7: *Κατανομή Παθολογικών Σταδίων*

Όσον αφορά τη διαδικασία προεπεξεργασίας των δεδομένων γονιδιακής έκφρασης για την ταξινόμηση σταδίων καρκίνου του μαστού, αυτή περιλαμβάνει διάφορα βήματα που αφορούν την αφαίρεση χαμηλά εκφρασμένων γονιδίων, την επιλογή γονιδίων με υψηλή μέση τιμή και διακύμανση, και την ανίχνευση και απομάκρυνση ανωμαλιών (outliers). Παρακάτω περιγράφονται αναλυτικά τα βήματα που ακολουθήθηκαν.

1. **Αφαίρεση Χαμηλά Εκφρασμένων Γονιδίων**: Το πρώτο βήμα της προεπεξεργασίας περιλαμβάνει την απομάκρυνση των γονιδίων που παρουσιάζουν χαμηλή έκφραση σε μεγάλο ποσοστό των δειγμάτων. Συγκεκριμένα, αφαιρούμε τα γονίδια τα οποία εκφράζονται στο μηδενικό επίπεδο σε ποσοστό μεγαλύτερο από 20% των δειγμάτων. Αυτό το βήμα είναι σημαντικό, καθώς τα γονίδια που δεν εκφράζονται σε αρκετά δείγματα δεν παρέχουν χρήσιμη πληροφορία για την ανάλυση και μπορεί να προσθέσουν θόρυβο στα δεδομένα.

2. **Επιλογή Γονιδίων με Υψηλή Μέση Τιμή και Διακύμανση**: Αφού αφαιρέσουμε τα χαμηλά εκφρασμένα γονίδια, επιλέγουμε τα γονίδια των οποίων η μέση τιμή και η διακύμανση είναι μεγαλύτερες από 0.5. Η μέση τιμή (mean) κάθε γονιδίου υπολογίζεται ως ο μέσος όρος των τιμών έκφρασης του γονιδίου σε όλα τα δείγματα, ενώ η διακύμανση (variance) υπολογίζεται ως η μέση τετραγωνική απόκλιση των τιμών από τη μέση τιμή. Η επιλογή των γονιδίων με υψηλή μέση τιμή και διακύμανση εξασφαλίζει ότι τα γονίδια που παραμένουν στο σύνολο δεδομένων έχουν σημαντική έκφραση και ποικιλία, γεγονός που μπορεί να συμβάλλει στη βελτίωση της ακρίβειας της ταξινόμησης.

3. **Ανίχνευση και Απομάκρυνση Ανωμαλιών (Outliers) με Χρήση της Mahalanobis Απόστασης**: Η Mahalanobis απόσταση είναι μια μετρική που χρησιμοποιείται για την ανίχνευση ακραίων τιμών (outlier detection) σε δεδομένα πολλαπλών διαστάσεων.

Αυτή η απόσταση λαμβάνει υπόψη τόσο τη διασπορά όσο και τις συσχετίσεις μεταξύ των μεταβλητών, γεγονός που την καθιστά ιδιαίτερα χρήσιμη για τη μείωση διαστασιμότητας σε γονιδιακά δεδομένα.

Για την εκτίμηση της συνδιακύμανσης των δεδομένων χρησιμοποιείται η μέθοδος της ελάχιστης εκτιμήτριας συνδιακύμανσης (Minimum Covariance Determinant, MCD), η οποία παρέχει μια ανθεκτική εκτίμηση που δεν επηρεάζεται από τις ακραίες τιμές. Αφού εκτιμηθεί η συνδιακύμανση, υπολογίζεται η Mahalanobis απόσταση για κάθε δείγμα.

Για τον καθορισμό των ακραίων τιμών, χρησιμοποιείται ένα επίπεδο σημαντικότητας 5% ($a = 0.05$). Τα δείγματα που έχουν Mahalanobis απόσταση μεγαλύτερη από το καθορισμένο όριο, το οποίο υπολογίζεται βάσει της κατανομής x^2, θεωρούνται ακραίες τιμές (outliers). Τα δεδομένα αυτά αφαιρούνται από το σύνολο δεδομένων, βελτιώνοντας έτσι την ποιότητα των δεδομένων και την ακρίβεια των μελλοντικών αναλύσεων.

Με την ολοκλήρωση αυτών των βημάτων, επιτυγχάνεται η δημιουργία ενός πιο καθαρού συνόλου δεδομένων γονιδιακής έκφρασης, έχοντας αφαιρέσει όλη εκείνη την επιπρόσθετη πληροφορία, η οποία είναι περιττή για την εκπαίδευση του μοντέλου.

Ταυτόχρονα, εξίσου σημαντική είναι και η κανονικοποίηση των δεδομένων, καθώς τα δεδομένα γονιδιακής έκφρασης μπορεί να έχουν διαφορετικές κλίμακες. Η κανονικοποίηση διασφαλίζει ότι όλα τα δεδομένα βρίσκονται σε ένα κοινό εύρος τιμών, διευκολύνοντας την ανάλυση από τα μοντέλα μηχανικής μάθησης. Οι τεχνικές κανονικοποίησης που χρησιμοποιήσαμε περιλαμβάνουν:

- **Min-Max Scaling**: Μετασχηματίζει τα δεδομένα ώστε να βρίσκονται σε ένα προκαθορισμένο εύρος, συνήθως από 0 έως 1. Κάθε τιμή x μετατρέπεται ως εξής:

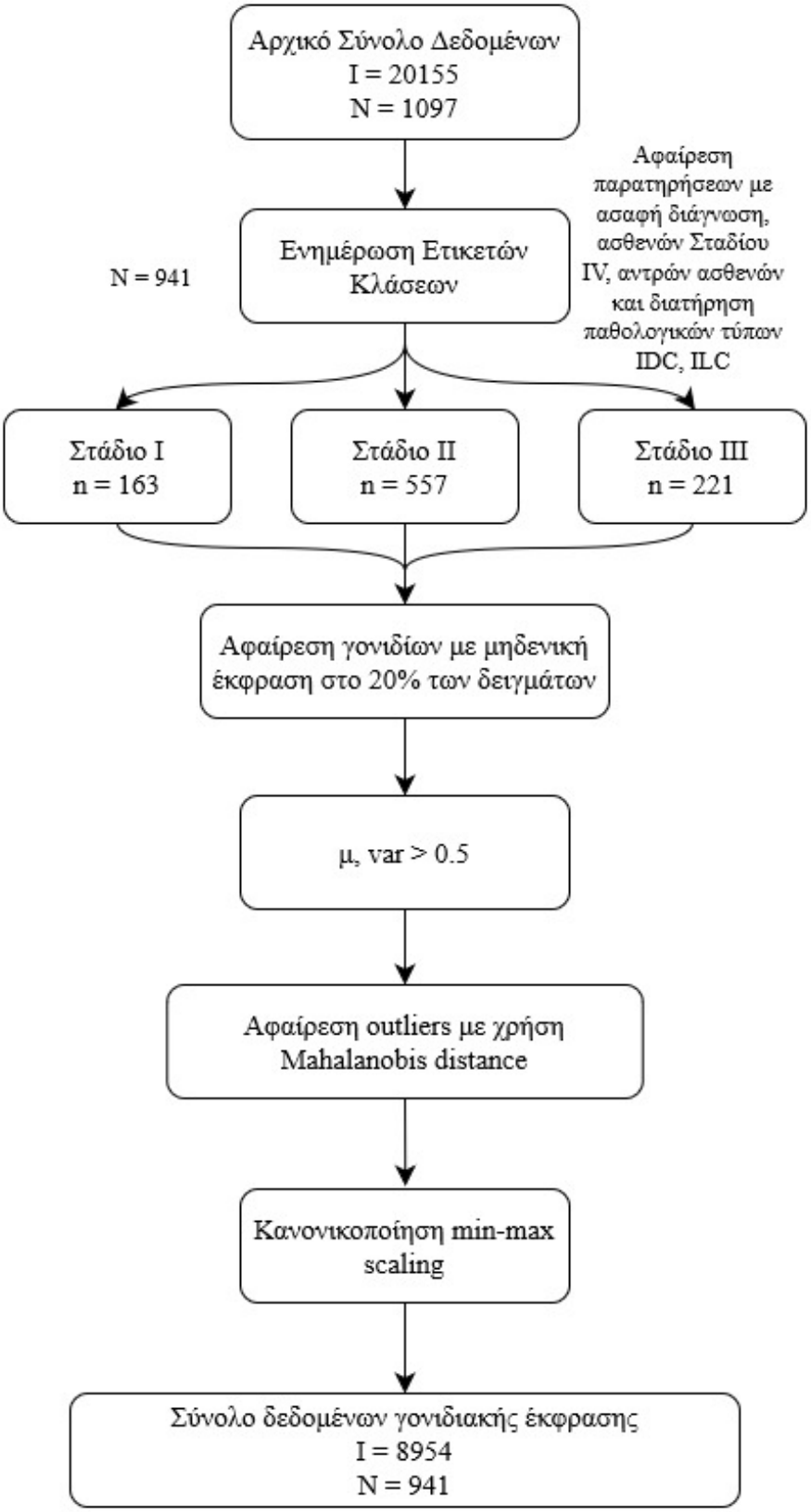
$$x' = \frac{x - min(x)}{max(x) - min(x)}$$

- **Z-score Normalization**: Μετασχηματίζει τα δεδομένα έτσι ώστε να έχουν μέση τιμή 0 και τυπική απόκλιση 1. Κάθε τιμή x μετατρέπεται ως εξής:

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

όπου μ είναι η μέση τιμή και σ η τυπική απόκλιση των δεδομένων.

Η κανονικοποίηση είναι απαραίτητη για την αποφυγή υπεροχής ορισμένων γονιδίων στις αναλύσεις, εξασφαλίζοντας ότι κάθε γονίδιο συνεισφέρει ισότιμα στην ανάλυση. Με την εφαρμογή των παραπάνω τεχνικών, έχει μειωθεί ο αριθμός των γονιδιακών εκφράσεων από 20.155 σε 8.954, περιορίζοντας σημαντικά τη διαστασιμότητα του προβλήματος. Αυτό πέραν του ότι απλοποιεί την ανάλυση, υπόσχεται και την απλοποίηση επίσης την ακρίβεια και την αποδοτικότητα των μοντέλων ταξινόμησης, αφού απομακρύνουν τα δεδομένα, τα οποία ενδεχομένως να αποτελούν θόρυβο. Στη βιβλιογραφία, η χρήση τέτοιων τεχνικών έχει αποδειχθεί ιδιαίτερα αποτελεσματική στη βελτίωση της απόδοσης των αναλυτικών μεθόδων σε δεδομένα υψηλής διαστασιμότητας [35, 5].



Σχήμα 4.8: *Προεπεξεργασία του αρχικού συνόλου δεδομένων. Ο εκάστοτε συνολικός αριθμός χαρακτηριστικών συμβολίζεται με I και ο συνολικός αριθμός ασθενών με N.*

Ωστόσο, το πρόβλημα της υψηλής διαστασιμότητας και ο συγκριτικά μικρός αριθμός δειγμάτων αποτελούν σημαντικές προκλήσεις που πρέπει να αντιμετωπιστούν. Σε αυτήν την περίπτωση, η χρήση μεθόδων feature engineering κρίνεται απαραίτητη για την αποτελεσματική ανάλυση και επεξεργασία των δεδομένων. Feature engineering είναι η διαδικασία της μετατροπής και αναδιαμόρφωσης των ανεπεξέργαστων δεδομένων σε χαρακτηριστικά που αναδεικνύουν τη σημασία και τη δύναμη της πληροφορίας που περιέχουν. Αυτή η διαδικασία βελτιώνει τις δυνατότητες ανάλυσης των δεδομένων από τα μοντέλα μηχανικής μάθησης.

Θα χρειαστεί για αυτόν τον λόγο να εξετάσουμε και άλλες τεχνικές που λειτουργούν είτε ανεξάρτητα είτε σε συνδυασμό με τις παραπάνω μεθόδους με στόχο τη μείωση της διαστασιμότητας και την αποτελεσματική ανάλυση των δεδομένων γονιδιακής έκφρασης. Παρακάτω αναφέρουμε επιγραμματικά τις πιο σημαντικές από αυτές τις τεχνικές, οι οποίες θα αναλυθούν εκτενέστερα στην αντίστοιχη ενότητα της μεθοδολογίας:

- **Pathway Analysis**: Αυτή η τεχνική χρησιμοποιείται ευρέως για την εξαγωγή βιολογικού νοήματος από δεδομένα υψηλής διαμέτρου. Εστιάζει στον προσδιορισμό των μονοπατιών που έχουν επηρεαστεί λόγω των διαφορικών μοτίβων γονιδιακής έκφρασης.
- **Χρήση των PAM50 Genes**: Η χρήση των PAM50 γονιδίων μας επιτρέπει να μειώσουμε τη διαστασιμότητα επιλέγοντας ένα υποσύνολο γονιδίων που έχει αποδειχθεί ότι είναι κατάλληλο για τη συγκεκριμένη ανάλυση.

Η εφαρμογή αυτών των τεχνικών στα δεδομένα στοχεύει στην περιορισμό των δεδομένων με στόχο την μείωση των μη σχετικών γονιδίων και την συγκρότηση πιο στοχευμένης πληροφορίας σε μικρότερο πλήθος δεδομένων. Η εφαρμογή των παραπάνω θα παρουσιαστεί αναλυτικά στο Κεφάλαιο 5.

4.3 Μέθοδοι

Στόχος είναι η εύρεση ενός μοντέλου που καταφέρνει να ταξινομήσει τα διάφορα στάδια καρκίνου του μαστού ή να εντοπίσει εκείνα τα χαρακτηριστικά (biomarkers) που διακρίνουν κάθε στάδιο. Προς επίτευξη του στόχου αυτού παρακάτω παρουσιάζονται οι μέθοδοι, αλλά και τα μοντέλα, τα οποία θα χρησιμοποιηθούν σε αυτήν την διερεύνηση.

4.3.1 Κλασσικοί Αλγόριθμοι Ταξινόμησης Μηχανικής Μάθησης

Για τη διαδικασία ταξινόμησης, χρησιμοποιήθηκαν κλασσικοί αλγόριθμοι μηχανικής μάθησης λόγω της αποτελεσματικότητας τους στη επίλυση προβλημάτων ταξινόμησης, οι οποίοι είναι συμβατοί με το σύνολο δεδομένων. Στην παρούσα έρευνα χρησιμοποιούνται οι ακόλουθοι αλγόριθμοι, οι οποίοι παρουσιάστηκαν αναλυτικά στο Κεφάλαιο 3:

- **Δέντρο Απόφασης (Decision Tree)**: Το Δέντρο Απόφασης είναι μια ιεραρχική δομή που μοιάζει με δέντρο και διαχωρίζει τα δεδομένα βάσει διαφορετικών χαρακτηριστικών για να κάνει προβλέψεις. Κάθε κόμβος του δέντρου αντιπροσωπεύει μια απόφαση ή διαχωρισμό με βάση ένα συγκεκριμένο χαρακτηριστικό, ενώ τα φύλλα του δέντρου αντιπροσωπεύουν τις τελικές κατηγορίες.

- **Τυχαία Δάση (Random Forest)**: Τα Τυχαία Δάση είναι μια μέθοδος συνόλου που συνδυάζει πολλαπλά Δέντρα Απόφασης για να κάνει προβλέψεις.

- **Gradient Boosting**: Το Gradient Boosting είναι επίσης μια μέθοδος συνόλου, που εκπαιδεύει το μοντέλο διαδοχικά, ενώ κάθε νέο μοντέλο προσπαθεί να βελτιώσει το προηγούμενο. Σε κάθε επανάληψη, ο αλγόριθμος υπολογίζει την κλίση της συνάρτησης απώλειας σε σχέση με τις προβλέψεις του τρέχοντος μοντέλου και στη συνέχεια εκπαιδεύει ένα νέο μοντέλο για να ελαχιστοποιήσει αυτή την κλίση.

- **Μηχανές Διανυσμάτων Υποστήριξης (SVM)**: Οι Μηχανές Διανυσμάτων Υποστήριξης στοχεύουν στην εύρεση ενός βέλτιστου υπερεπιπέδου που διαχωρίζει τις διαφορετικές κατηγορίες στο σύνολο δεδομένων.

- **Λογιστική Παλινδρόμηση (Logistic Regression)**: Η Λογιστική Παλινδρόμηση είναι μία μέθοδος πολυπαραγοντικής στατιστικής ανάλυσης που χρησιμοποιείται για την πρόβλεψη της πιθανότητας εμφάνισης μιας δυαδικής εξαρτημένης μεταβλητής.

4.3.2 Αρχιτεκτονικές Μοντέλων Νευρωνικών Δικτύων

Multilayer Perceptron

Στην παρούσα μελέτη, διερευνήθηκε η ανάπτυξη και βελτιστοποίηση ενός πολυεπίπεδου νευρωνικού δικτύου τύπου Perceptron (MLP) με στόχο την εύρεση της βέλτιστης αρχιτεκτονικής για την συγκεκριμένη εφαρμογή. Η διαδικασία επιλογής της αρχιτεκτονικής και των υπερπαραμέτρων του μοντέλου έγινε πειραματικά, μέσω σύγκρισης των αποτελεσμάτων που προέκυψαν από διάφορες παραλλαγές του δικτύου. Παρακάτω παρουσιάζεται αναλυτικά η αρχιτεκτονική του μοντέλου που απέφερε τις καλύτερες επιδόσεις.

Για την εύρεση του βέλτιστου μοντέλου χρησιμοποιήσαμε διάφορες τεχνικές, οι οποίες στις περισσότερες περιπτώσεις λειτουργούν συνδυαστικά μεταξύ τους. Αυτές περιλαμβάνουν:

1. Προσαρμογή των Υπερπαραμέτρων:

    - Learning Rate: Ο ρυθμός μάθησης είναι μια από τις πιο κρίσιμες υπερπαραμέτρους, καθώς καθορίζει το βήμα με το οποίο το μοντέλο προσαρμόζει τις παραμέτρους του κατά την εκπαίδευση.
    - L2 Regularization Term: Η παράμετρος κανονικοποίησης L2 βοηθά στην αποφυγή υπερπροσαρμογής προσθέτοντας ένα ποινικό όρο στην απώλεια που εξαρτάται από το μέγεθος των βαρών.
    - Class Weights: Οι συντελεστές βαρών για κάθε κλάση μπορούν να χρησιμοποιηθούν για να αντιμετωπιστεί η μη-ισορροπημένη κατανομή των δεδομένων μεταξύ των διαφόρων κλάσεων.

2. Υπολογισμός Loss μέσω Cross Entropy:

    - Η συνάρτηση απώλειας cross-entropy χρησιμοποιείται για την εκτίμηση της απόδοσης του μοντέλου, καθώς μετρά τη διαφορά μεταξύ των προβλεπόμενων τιμών και των πραγματικών ετικετών.

Το μοντέλο MLP που χρησιμοποιήθηκε διαμορφώθηκε με την παρακάτω αρχιτεκτονική:

1. Ένα στρώμα Dropout: Χρησιμοποιείται στην είσοδο για να μειώσει την υπερπροσαρμογή διαγράφοντας τυχαία ένα ποσοστό των εισόδων κατά την εκπαίδευση.

2. Πρώτο Πυκνό Στρώμα (Dense Layer): Αποτελείται από 8954 νευρώνες στην είσοδο και 256 νευρώνες στην έξοδο. Χρησιμοποιεί τη συνάρτηση ενεργοποίησης ReLU (Rectified Linear Unit), η οποία προσθέτει μη γραμμικότητα στο μοντέλο.

3. Δεύτερο Στρώμα Dropout: Εφαρμόζεται μετά το πρώτο πυκνό στρώμα για να συνεχίσει την προσπάθεια μείωσης της υπερπροσαρμογής.

4. Δεύτερο Dense Layer: Έχει 256 εισόδους και 128 εξόδους. Χρησιμοποιεί επίσης τη συνάρτηση ενεργοποίησης ReLU.

5. Τρίτο Στρώμα Dropout: Προστίθεται μετά το δεύτερο πυκνό στρώμα για την περαιτέρω βελτίωση της γενίκευσης του μοντέλου.

6. Τελικό Projection Layer: Είναι ένα πυκνό στρώμα με 128 εισόδους και έναν αριθμό εξόδων ίσο με τον αριθμό των κατηγοριών που πρόκειται να προβλεφθούν, ο οποίος στην προκειμένη περίπτωση είναι 3 (Στάδιο *I, II* και *III*). Η έξοδος αυτού του στρώματος παρέχει τις τελικές προβλέψεις του μοντέλου.

Με την παραπάνω αρχιτεκτονική, το MLP μοντέλο πετυχαίνει τις καλύτερες επιδόσεις συγκριτικά με τα υπόλοιπα μοντέλα που δοκιμάστηκαν κατά την διάρκεια της διερεύνησης. Η χρήση των στρωμάτων Dropout, η επιλογή της συνάρτησης ενεργοποίησης, και η προσεκτική προσαρμογή των υπερπαραμέτρων συμβάλλουν στην βελτίωση της απόδοσης του μοντέλου. Η συγκεκριμένη προσέγγιση επιλέχθηκε επιπλέον, επειδή εξασφαλίζει στο μοντέλο την ικανότητα να γενικεύει σε νέα δεδομένα (μέσω της χρήσης των Dropout επιπέδων) αποφεύγοντας την υπερπροσαρμογη.

4.3.3 Μετρικές Αξιολόγησης

Από την περιγραφή που προηγήθηκε στην αρχή του Κεφαλαίου, στην υποενότητα 4.1 του συνόλου δεδομένων γίνεται κατανοητό πως το σύνολο δεδομένων δεν είναι σε καμία περίπτωση ίσα κατανεμημένο ανάμεσα στις διαφορετικές κατηγορίες. Αυτή η μη ισορροπημένη κατανομή των δεδομένων επηρεάζει σημαντικά την ανάλυση και καθορίζει τη βαρύτητα και τη σημασία που αποδίδεται στις μετρικές για την αξιολόγηση της επίδοσης των μοντέλων 3.1.4.

Για αυτόν τον λόγο η βασική μετρική που αξιοποιείται και στην οποία εστιάζει η αξιολόγηση των μοντέλων είναι το F1-score, το οποίο συνδυάζει τις μετρικές Recall και Precision σε ένα ενιαίο μέτρο απόδοσης. Το F1-score είναι ιδιαίτερα χρήσιμο στις περιπτώσεις μη ισορροπημένων δεδομένων, καθώς λαμβάνει υπόψη τόσο τις ψευδώς θετικές όσο και τις ψευδώς αρνητικές προβλέψεις, προσφέροντας μια ολοκληρωμένη εικόνα της απόδοσης του μοντέλου.

Τέλος, από τις μετρικές που παράγει η κάθε κατηγορία υπολογίζεται ο μέσος όρος των τιμών αυτών με την τεχνική macro-average, όπως αυτή περιγράφηκε στην υποενότητα 3.1.3.

4.3.4 Εγκυρότητα Αποτελεσμάτων

Για να διασφαλιστεί η αξιόπιστη εκτίμηση των επιδόσεων των μοντέλων ταξινόμησης, χρησιμοποιείται η τεχνική διασταυρωμένης επικύρωσης Stratified K-Fold cross-validation. Αυτή η τεχνική αποτελεί μια παραλλαγή της κλασικής μεθόδου K-Fold cross-validation, με κύριο στόχο τη διασφάλιση ότι κάθε fold παρουσιάζει ανάλογη κατανομή κλάσεων με αυτή του αρχικού συνόλου δεδομένων.

Συγκεκριμένα, χρησιμοποιήθηκε εδώ διασταυρωμένη επικύρωση 5 πτυχών, η οποία χώρισε το σύνολο δεδομένων σε πέντε ισόποσες αναδιπλώσεις και η αξιολόγηση εκτελέστηκε πέντε φορές. Ο προσδιορισμός των αποτελεσμάτων των 5 επαναλήψεων υπολογίζεται ως ο μέσος όρος των μετρικών (F1-score, Recall, Precision) από κάθε αναδίπλωση.

Με την χρήση της μεθόδου επικύρωσης Stratified K-Fold μειώνεται η πιθανότητα υπερπροσαρμογής (overfitting) και εξασφαλίζεται πως η απόδοση του μοντέλου είναι αντιπροσωπευτική και γενικεύσιμη σε διαφορετικά υποσύνολα δεδομένων. Με αυτόν τον τρόπο, επιτυγχάνεται μια πιο δίκαιη και ολοκληρωμένη αξιολόγηση των μοντέλων, εξασφαλίζοντας ότι η αξιολόγησή μας είναι ανθεκτική στην τυχαιότητα και τη μεταβλητότητα των δεδομένων.

Η αξιοπιστία των αποτελεσμάτων εξασφαλίζεται επιπλέον, αφού καθε πείραμα/μοντέλο που συμπεριλαμβάνει και την παραπάνω αναφερόμενη μέθοδο διασταυρωμένης επικύρωσης, εκτελείται 30 φορές. Κάθε ένα από αυτά τα πειράματα αρχικοποιείται με διαφορετικούς random seeds. Τα random seeds είναι αριθμοί που χρησιμοποιούνται για να καθορίσουν την αρχική κατάσταση μιας γεννήτριας τυχαίων αριθμών και χρησιμοποιούνται αρκετά στο εσωτερικό των αλγορίθμων μηχανικής μάθησης. Με αυτόν τον τρόπο εξασφαλίζουμε αφενός την αναπαραγωγή των αποτελεσμάτων και αφετέρου μπορούμε να εγγυηθούμε ότι τα αποτελέσματα δεν επηρεάζονται από την τυχαία επιλογή των αρχικών συνθηκών, παρέχοντας μια πιο ακριβή εκτίμηση της απόδοσης του μοντέλου.

4.3.5 Προσδιορισμός Βέλτιστων Υπερπαραμέτρων

Αναπτύσσοντας μοντέλα τόσο με τη χρήση αλγορίθμων μηχανικής μάθησης όσο και με τη χρήση νευρωνικών δικτύων, γίνεται φανερό πως τα μοντέλα αυτά μπορούν να βελτιστοποιηθούν με πάνω από έναν τρόπο. Για παράδειγμα, στο Decision Tree μπορούμε να προσθέσουμε διαφορετικό βάθος (depth), ενώ στα νευρωνικά δίκτυα μπορούμε να τροποποιήσουμε το πλήθος των επαναλήψεων (iterations) και εποχών (epochs) ή να βρούμε την κατάλληλη ποσότητα για dropout. Είναι γεγονός πως η χρήση διαφορετικών υπερπαραμέτρων μπορεί να βελτιώσει κατά πολύ την ικανότητα των μοντέλων για ταξινόμηση.

Ένας τρόπος να γίνει αυτό είναι με τη χρήση περαιτέρω αλγορίθμων, όπως ο GridSearchCV. Με την αναζήτηση πλέγματος διερευνούμε διαφορετικούς συνδυασμούς υπερπαραμέτρων για κάθε αλγόριθμο ταξινόμησης και για τη μέθοδο επιλογής χαρακτηριστικών. Η διαδικασία αναζήτησης πλέγματος περιλαμβάνει τη συστηματική δοκιμή διαφορετικών τιμών παραμέτρων εντός προκαθορισμένων ευρών και την αξιολόγηση της απόδοσης του μοντέλου χρησιμοποιώντας τις μετρικές αξιολόγησης. Οι παράμετροι εκείνοι που οδηγούν στην καλύτερη απόδοση επιλέγονται και ως βέλτιστες τιμές του συγκεκριμένου μοντέλου.

Κεφάλαιο 5

Διερεύνηση Μεθοδολογίας και Ανάπτυξη Μοντέλων

Σε αυτό το κεφάλαιο παρουσιάζεται η διαδικασία διερεύνησης και υλοποίησης υποψήφιων μοντέλων πρόβλεψης με βάση τις μεθόδους που αναπτύχθηκαν στο προηγούμενο κεφάλαιο. Στόχος των ταξινομητών που παρουσιάζονται είναι να διακρίνουν τα διαφορετικά στάδια του καρκίνου του μαστού, αξιοποιώντας γονιδιακά δεδομένα και παράγωγά τους.

Αρχικά, εξετάζεται η ταξινόμηση με χρήση των πρωτότυπων γονιδιακών δεδομένων, χρησιμοποιώντας τόσο κλασσικούς αλγόριθμους μηχανικής μάθησης όσο και αρχιτεκτονικές νευρωνικών δικτύων τύπου MLP. Στη συνέχεια, αναλύεται πως η τροποποίηση των αρχικών - γονιδιακών - δεδομένων και η εξαγωγή βιολογικών χαρακτηριστικών μέσω ανάλυσης GSEA σε συνδυασμό με αλγορίθμους μηχανικής μάθησης, μπορεί να βελτιώσει την διαδικασία ταξινόμησης. Λαμβάνοντας υπόψιν τα κλινικά μεταδεδομένα, θα μελετηθεί υπό νέο πρίσμα η κατηγοριοποίηση της νόσου σε στάδια. Ύστερα, θα εξεταστεί και η χρήση συνθετικών δεδομένων. Τέλος, ένα ακόμη σημαντικό μέρος της παρούσας διερεύνησης θα αποτελέσει η ταξινόμηση των μεταβάσεων μεταξύ των σταδίων του καρκίνου. Η προσέγγιση αυτή, αν και πιο σύνθετη, δύναται να προσφέρει πληροφορίες για την εξέλιξη της νόσου.

Συνολικά, το κεφάλαιο αυτό έχει ως στόχο να εμβαθύνει στις διάφορες μεθοδολογίες που δοκιμάστηκαν κατά την διάρκεια επίλυσης του προβλήματος ταξινόμησης. Στο επόμενο κεφάλαιο θα παρουσιαστούν τα αποτελέσματα αυτών των μεθόδων, παρέχοντας μια ολοκληρωμένη εικόνα της αποδοτικότητας των παρακάτω προσεγγίσεων.

5.1 Ταξινόμηση Σταδίων Καρκίνου του Μαστού με δεδομένα γονιδιακής έκφρασης

Σε αυτήν την ενότητα παρουσιάζεται η ανάπτυξη μοντέλων ταξινόμησης των σταδίων του καρκίνου του μαστού, βασισμένων σε δεδομένα γονιδιακής έκφρασης. Αυτό αποτελεί ουσιαστικά την πρώτη προσπάθεια δημιουργίας ενός μοντέλου ταξινόμησης. Τα δεδομένα αυτά προέρχονται από 941 ασθενείς και περιλαμβάνουν τις εκφράσεις 8954 γονιδίων, οργανωμένα σε πίνακα διαστάσεων 941×8954. Η προεπεξεργασία των δεδομένων έχει ήδη παρουσιαστεί στην Ενότητα 4.2, και δεν πραγματοποιούνται περαιτέρω τροποποιήσεις σε αυτό το στάδιο.

Αξιοποιούνται διάφορες μέθοδοι που αναφέρθηκαν στην Ενότητα 4.3 για την ανάπτυξη και εκπαίδευση των μοντέλων. Συγκεκριμένα, οι μέθοδοι που δοκιμάζονται είναι οι ακόλου-

θες:

- **Κλασσικοί Αλγόριθμοι Μηχανικής Μάθησης**: Εκπαιδεύουμε το μοντέλο μας χρησιμοποιώντας τους εξής αλγόριθμους:
  - Random Forest Classifier
  - Decision Tree Classifier
  - Gradient Boosting (xgboost, catboost)
  - Support Vector Machine (SVM)
  - Logistic Regression

  Η εκπαίδευση πραγματοποιείται χρησιμοποιώντας τη μέθοδο StratifiedKFold με 5 αναδιπλώσεις και εκτελούμε 30 πειράματα με διαφορετικές αρχικές συνθήκες (random seeds) για να εξάγουμε τον μέσο όρο των μετρικών. Οι υπερπαράμετροι των αλγορίθμων ρυθμίζονται μέσω της διαδικασίας υπερπαραμετροποίησης.

- **Πολλαπλών επιπέδων Perceptron (MLP)**: Τα MLPs είναι ο de facto αλγόριθμος για την εύρεση μη-γραμμικού συσχέτισης μεταξύ των δεδομένων εισόδου. Για αυτόν τον λόγο, η επόμενη αρχιτεκτονική που εξετάζεται είναι η ταξινόμηση μέσω MLPs.

  Κατά την διάρκεια των πειραμάτων δοκιμάστηκαν πολλαπλές αρχιτεκτονικές MLPs. Ωστόσο, επιλέχθηκε και παρουσιάζεται αυτό που εμφάνισε συγκριτικά την καλύτερη επίδοση. Στο μοντέλο που επιλέχθηκε, εφαρμόστηκαν τεχνικές όπως το Dropout, Learning Rate και L2 Regularization για τη βελτίωση της απόδοσης και την αποφυγή υπερεκπαίδευσης.

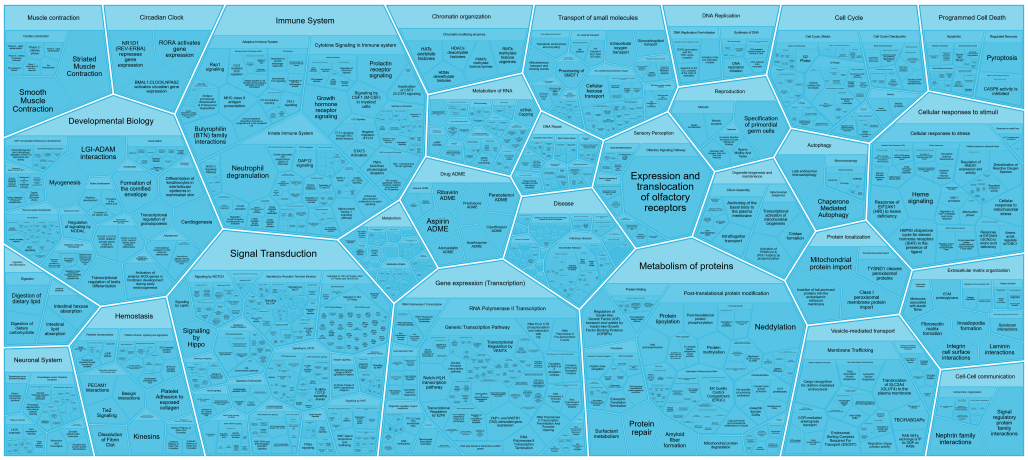
5.2 Ταξινόμηση Σταδίων Καρκίνου του Μαστού με χρήση Pathway Analysis

Η ταξινόμηση των σταδίων του καρκίνου του μαστού αποτελεί ένα πολύπλοκο πρόβλημα, απαιτώντας προηγμένες μεθόδους ανάλυσης γονιδιακών δεδομένων. Μια από τις ευρέως χρησιμοποιούμενες προσεγγίσεις είναι η ανάλυση βιολογικών μονοπατιών (Pathway Analysis), η οποία εστιάζει στη μελέτη συγκεκριμένων βιολογικών μονοπατιών και τη συσχέτισή τους με φαινοτυπικές εκδηλώσεις της νόσου. Μέσω αυτής της μεθόδου, εξάγεται βιολογική σημασία από τα γονιδιακά δεδομένα, εστιάζοντας στον προσδιορισμό των μονοπατιών που μπορεί να έχουν διαταραχθεί λόγω διαφορετικών μοτίβων γονιδιακής έκφρασης. Οι υπάρχουσες μέθοδοι αξιοποιούν τα επίπεδα γονιδιακής έκφρασης και την υφιστάμενη γνώση για τον οργανισμό, με σκοπό τον εντοπισμό των υποκείμενων βιολογικών διαδικασιών και μηχανισμών.

Τα βιολογικά μονοπάτια έχουν χρησιμοποιηθεί στο παρελθόν για την αποτελεσματική διάκριση διαφορετικών τύπων καρκίνου. Για παράδειγμα, το μονοπάτι της αδιποκυτοκίνης έχει αποδειχθεί ότι διαχωρίζει αποτελεσματικά τον καρκίνο του μαστού από τους όγκους του παχέος εντέρου και του στομάχου [63].

Σύμφωνα με τον Baverstock [64], τα γονίδια διαδραματίζουν έναν παθητικό ρόλο αποθήκευσης κρίσιμων πληροφοριών, ενώ ο φαινότυπος, δηλαδή η έκφραση αυτών των πληροφοριών σε συγκεκριμένες βιολογικές λειτουργίες και χαρακτηριστικά, έχει τον ενεργό ρόλο στην κληρονομικότητα, την ανάπτυξη και τη μορφογένεση. Τα δεδομένα γονιδιακής έκφρασης που διαθέτουμε δεν αντιπροσωπεύουν άμεσα ούτε τα γονίδια ούτε τον φαινότυπο. Συνεπώς, η χρήση φαινοτύπων αντί γονιδίων ίσως να επιφυλάσσει ποιοτικότερη και πιο αξιόπιστη πληροφορία.

Το πρώτο βήμα αυτής της διαδικασίας είναι η εφαρμογή της ανάλυσης Gene Set Enrichment (GSE), η οποία θα επιτρέψει την απομόνωση και εξέταση των διαφοροποιημένων εκφρασμένων γονιδίων (differentially expressed genes) και τη σύνδεση τους με συγκεκριμένα άλλα σύνολα γονιδίων και τους αντίστοιχους φαινότυπούς τους. Μέσω αυτής της ανάλυσης εντοπίζονται 29 διαφορετικοί φαινότυποι ή αλλιώς βιολογικά μονοπάτια (biological pathways). Ωστόσο, επειδή σε αρκετά από αυτά υπάρχουν μηδενικές τιμές, προκειμένου να διατηρηθεί η ακεραιότητα των αποτελεσμάτων, αφαιρούνται όλα τα μονοπάτια όπου πάνω από το 60% των τιμών είναι μηδενικά. Μετά από αυτή τη διαδικασία, τα βιολογικά μονοπάτια που παραμένουν είναι συνολικά 22. Αυτά τα μονοπάτια θα χρησιμοποιηθούν ως χαρακτηριστικά features για τα μοντέλα εκπαίδευσης, δεδομένου ότι μπορούν να παρέχουν πολύτιμες πληροφορίες για την ταξινόμηση των σταδίων του καρκίνου του μαστού.



Σχήμα 5.1: *Απεικόνιση των 29 βιολογικών μονοπατιών που χρησιμοποιήθηκαν*

Στην ανάλυση της προηγούμενης υποενότητας 5.1, χρησιμοποιήθηκαν πέντε διαφορετικοί αλγόριθμοι ταξινόμησης: Random Forest, Decision Tree, Gradient Boosting (xgboost, catboost), Support Vector Machine, Logistic Regression. Ωστόσο, στην παρούσα προσέγγιση, τα δεδομένα που είναι πλέον τα βιολογικά μονοπάτια, παρουσιάζουν μη αριθμητικές τιμές για κάποιους ασθενείς. Η χρήση διαφόρων τεχνικών για την αντιμετώπιση των NaN τιμών, όπως η αντικατάστασή τους με τον γενικό μέσο όρο του συγκεκριμένου μονοπατιού, δεν είναι πολύ αποδοτική και μπορεί να εισάγει σφάλματα στα αποτελέσματά μας. Επομένως, χρησιμοποιούνται μόνο ταξινομητές που μπορούν να διαχειριστούν μη αριθμητικές τιμές στα δεδομένα.

Οι αλγόριθμοι ταξινόμησης που εφαρμόζονται σε αυτή την περίπτωση είναι:

- RandomForestClassifier

- DecisionTreeClassifier
- Gradient Boosting μέσω της εφαρμογής (xgboost, catboost)

Αυτοί οι ταξινομητές έχουν την ικανότητα να διαχειρίζονται τα δεδομένα με μη αριθμητικές τιμές χωρίς να απαιτούν προηγούμενη επεξεργασία για την αντιμετώπιση των κενών, διατηρώντας έτσι την ακεραιότητα των δεδομένων και ελαχιστοποιώντας την εισαγωγή σφαλμάτων.

Η εκπαίδευση πραγματοποιείται χρησιμοποιώντας τη μέθοδο StratifiedKFold με 5 αναδιπλώσεις και αφού εκτελεστούν 30 πειράματα εξάγεται ο μέσος όρος των μετρικών απόδοσης με την μέθοδο macro-weighted. Οι μετρικές αξιολόγησης περιλαμβάνουν το (F1-score), την ακρίβεια (Precision) και την ανάκληση (Recall), προκειμένου να εξασφαλιστεί μια ολοκληρωμένη εικόνα της απόδοσης κάθε μοντέλου.

Αυτήν η προσέγγιση αποσκοπεί στην ακριβή ταξινόμηση των σταδίων του καρκίνου του μαστού, αξιοποιώντας τις δυνατότητες της ανάλυσης των μονοπατιών (Pathway Analysis) και των αλγορίθμων μηχανικής μάθησης.

5.3 Ταξινόμηση Σταδίων με Κλινικά Μεταδεδομένα

Σε αυτό το σενάριο πραγματοποιείται ταξινόμηση σταδίων με δεδομένα γονιδιακής έκφρασης, στα οποία όμως συμπεριλαμβάνονται και τα κλινικά μεταδομένα. Οι κλινικές πληροφορίες των ασθενών χρησιμοποιούνται για να περιορίζουν το σύνολο δεδομένων. Είναι θεμιτό να δημιουργηθούν υποσύνολα ασθενών, τα οποία θα μοιράζονται περισσότερα κοινά χαρακτηριστικά - και άρα πιο παρόμοιο γονιδίωμα - ανεξάρτητα με την ταυτοποίηση της νόσου.

Ένα πιθανό υποσύνολο είναι αυτό που περιέχει όλες τις γυναίκες ασθενείς εντός του ηλικιακού εύρους 25-45. Ένα άλλο πιθανό υποσύνολο είναι αυτό που περιέχει μόνο λευκές γυναίκες ασθενείς. Επομένως, μέσω της συγκεκριμένης μεθοδολογίας αφενός περιορίζεται το σύνολο των δεδομένων, αλλά ταυτόχρονα γίνεται πιο στοχευμένο σε μια συγκεκριμένη κατηγορία ασθενών, η οποία μοιράζεται κοινά χαρακτηριστικά.

Συγκεκριμένα, τα υποσύνολα που εξετάζονται είναι διαδοχικά πιο συμπεριληπτικά και αποτελούνται από συνδυασμούς των παρακάτω χαρακτηριστικών:

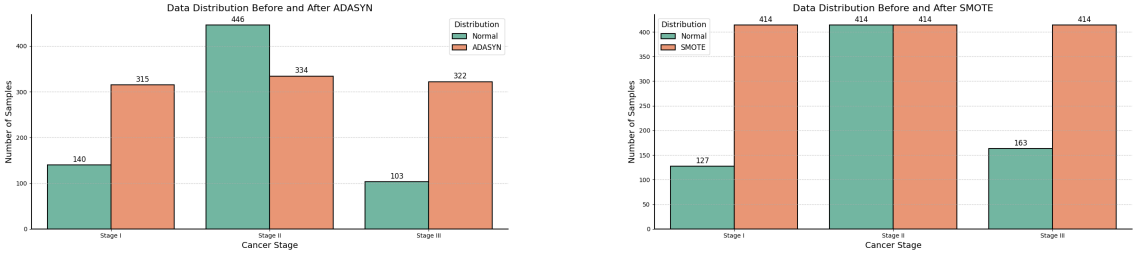
- Ηλικιακές ομάδες: 25-54, 55-94
- Φυλή: Λευκή
- Εθνικότητα: Μη Ισπανική, μη Λατινική
- Ιστολογικός Τύπος: Διηθητικό πορογενές καρκίνωμα, Διηθητικό λοβιακό καρκίνωμα

Αυτά τα υποσύνολα επιτρέπουν τη λεπτομερή ανάλυση και σύγκριση διαφορετικών δημογραφικών και κλινικών παραμέτρων, διευκολύνοντας στην κατανόηση των επιπτώσεων τους στην πρόγνωση και την αντιμετώπιση του καρκίνου του μαστού. Στόχος μέσω της συγκεκριμένης μεθόδου είναι να δούμε αν η ανάλυση γονιδιακών δεδομένων μιας πιο συγκεκριμένης υποκατηγορίας ασθενών, όπως της ηλικιακής ομάδας 25-45, μπορεί να βελτιώσει την απόδοση του μοντέλου.

5.4 Ταξινόμηση με χρήση συνθετικών δεδομένων

Στην παρούσα έρευνα, λαμβάνοντας υπόψιν την ύπαρξη μη ισορροπημένων δεδομένων, η εφαρμογή των τεχνικών SMOTE και ADASYN φαίνεται υποσχόμενη για την αντιμετώπιση αυτού του προβλήματος. Το σύνολο δεδομένων παρουσιάζει σημαντική ανισορροπία στις κλάσεις πρόβλεψης, με τους ασθενείς σταδίου *II* να είναι υπερεκπροσωπούμενοι σε σύγκριση με τους ασθενείς σταδίων *I* και *III*.

Στο Σχήμα 5.2α΄ παρουσιάζεται η αρχική κατανομή των σταδίων των ασθενών και η κατανομή μετά την εφαρμογή της τεχνικής ADASYN, ενώ στο Σχήμα 5.2β΄ αντιπαραβάλλεται η αρχική κατανομή μαζί με την κατανομή ύστερα από εφαρμογή της τεχνικής SMOTE. Η νέα κατανομή, που αφορά μόνο το σύνολο εκπαίδευσης, στην πρώτη περίπτωση περιλαμβάνει 971 δείγματα, με τις κλάσεις να είναι σχεδόν ισοκατανεμημένες, ενώ στην δεύτερη περίπτωση περιλαμβάνει 1242 δεδομένα και οι κλάσεις πλέον είναι ομοιόμορφα κατανεμημένες.



(α΄) *ADASYN* (β΄) *SMOTE*

Σχήμα 5.2: *Σύγκριση κατανομών δεδομένων πριν και μετά από την εφαρμογή τεχνικών υπερδειγματοληψίας (SMOTE, ADASYN) στο σύνολο δεδομένων εκπαίδευσης*

Από τις κατανομές των δεδομένων διακρίνονται οι κύριες διαφορές μεταξύ των δύο τεχνικών. Συγκεκριμένα, η SMOTE οδηγεί σε ομοιόμορφη κατανομή των δεδομένων σύμφωνα με την κατανομή της πλειοψηφικής τάξης. Αντίθετα, η ADASYN δίνει μεγαλύτερη έμφαση στη δημιουργία νέων δειγμάτων σε περιοχές όπου η ταξινόμηση είναι δύσκολη, με αποτέλεσμα η κατανομή μετά την εφαρμογή της να μην είναι απόλυτα ομοιόμορφη.

Η χρήση των τεχνικών SMOTE και ADASYN στοχεύει στην αύξηση της αντιπροσώπευσης των υποεκπροσωπούμενων κλάσεων, δημιουργώντας συνθετικά δεδομένα που ενισχύουν την ακρίβεια και τη γενικευσιμότητα του ταξινομητή. Με αυτόν τον τρόπο, επιτυγχάνεται μια πιο ισορροπημένη κατανομή των δεδομένων, που αποσκοπεί στην βελτίωση της ικανότητας του μοντέλου να προβλέπει σωστά τα στάδια του καρκίνου του μαστού και μειώνοντας τη μεροληψία προς τις υπερεκπροσωπούμενες κλάσεις.

5.5 Ταξινόμηση Μεταβάσεων Μεταξύ Σταδίων

Η συγκέντρωση δεδομένων σε μορφή χρονοσειρών μέσω της επαναλαμβανόμενης συλλογής δειγματοληψιών κατά την διάρκεια εξέλιξης της νόσου θα παρείχε ουσιαστικές πληροφορίες. Με αυτό τον τρόπο καθίσταται εφικτή η πλήρης εκμετάλλευση των τεχνολογιών που έχουν αναπτυχθεί σήμερα και που επιτρέπουν την μελέτη των γονιδιωμάτων του καρκίνου. Ωστόσο, λόγω της ανάγκης για άμεση θεραπεία μετά τη διάγνωση, είναι ηθικά ανέφικτη η

συλλογή δεδομένων χρονοσειρών για τη μελέτη της εξέλιξης του καρκίνου του μαστού [54].

Η σκιαγράφηση αυτής της δυναμικής διαδικασίας και ο προσδιορισμός των κομβικών μοριακών γεγονότων που οδηγούν στη σταδιακή εξέλιξη προς την ανάπτυξη καρκίνου θα αποτελέσουν ένα κρίσιμο θεμέλιο και οδηγό για την ανάπτυξη διαγνωστικών, προγνωστικών και στοχευμένων θεραπευτικών μεθόδων για τον καρκίνο.

Από τα παραπάνω, καθίσταται σαφές ότι η δημιουργία ενός πλήρους πορτραίτου καθ' όλη την διάρκεια της νόσησης για κάθε ασθενή θα είχε σημαντικά οφέλη. Ωστόσο, επειδή αυτό είναι αδύνατο, η επόμενη καλύτερη προσέγγιση είναι η μελέτη της εξέλιξης της νόσου μέσω της ομαδοποίησης ασθενών που μοιράζονται κοινά χαρακτηριστικά, τόσα πολλά, ώστε η μόνη κύρια διαφορά τους να είναι το στάδιο καρκίνου του μαστού στο οποίο βρίσκονται.

Σε αυτήν την περίπτωση, η μορφή των δεδομένων τροποποιείται για να ταιριάξει στα παραπάνω. Όμοιοι ασθενείς ομαδοποιούνται με σκοπό την δημιουργία συνθετικών τροχιών (trajectories) μέσω Ευκλείδιας Απόστασης. Η προσέγγιση αυτή στοχεύει στη διευκόλυνση του προβλήματος, καθώς κάθε δείγμα πλέον περιέχει περισσότερη πληροφορία.

Με τις τεχνητές τροχιές των ασθενών, μπορούν να σχηματιστούν ζεύγη ασθενών που αναπαριστούν τη μετάβαση από το Στάδιο 1 στο Στάδιο 2 και από το Στάδιο 2 στο Στάδιο 3. Έστω ότι η γονιδιακή έκφραση κάθε ασθενούς μπορεί να αναπαρασταθεί από ένα διάνυσμα x. Η δημιουργία των ζευγών γίνεται με δύο τρόπους:

- Υπολογισμός της διαφοράς των διανυσμάτων δύο ασθενών (Patient Difference Method). Έστω x_1 το διάνυσμα ενός ασθενούς στο Στάδιο 1 και x_2 το διάνυσμα ενός ασθενούς που βρίσκεται στην ίδια συνθετική τροχιά με τον ασθενή x_1. Το νέο διάνυσμα μετάβασης είναι $x_{1-2} = x_1 - x_2$. Σε αυτή την περίπτωση, το νέο διάνυσμα έχει το ίδιο μέγεθος με το αρχικό.

- Συνένωση των δύο διανυσμάτων (Patient Concatination Method). Έστω x_1 το διάνυσμα ενός ασθενούς στο Στάδιο 1 και x_2 το διάνυσμα ενός ασθενούς που βρίσκεται στην ίδια συνθετική τροχιά με τον ασθενή x_1. Το νέο διάνυσμα μετάβασης είναι $x_{1-2} = [x_1, x_2]$. Σε αυτή την περίπτωση, το νέο διάνυσμα έχει το διπλάσιο μέγεθος από το αρχικό.

Με την εφαρμογή των παραπάνω μεθόδων προκύπτουν δύο διαφορετικά σύνολα δεδομένων. Με την χρήση αυτών των δεδομένων είναι δυνατό να αναπτυχθούν μοντέλα τόσο βασισμένα σε αλγορίθμους μηχανικής μάθησης, σύμφωνα με τις μεθόδους και την μεθοδολογία που αναπτύξαμε παραπάνω.

Λόγω του τρόπου με τον οποίο έχουν δημιουργηθεί τα δεδομένα των ζευγών ασθενών, είναι δυνατό ένας ασθενείς και ειδικά οι ασθενείς Σταδίου *I, III* που μειοψηφούν, να εμφανίζεται σε περισσότερες από μία τροχιές. Αυτό έχει ως αποτέλεσμα μέσω της εφαρμογής της μεθόδου "Συνένωσης δύο διανυσμάτων ασθενών" να υπάρχει μερική διαρροή δεδομένων κατά τον διαχωρισμό του συνόλου δεδομένων σε δεδομένα εκπαίδευσης και ελέγχου. Για να αντιμετωπίσουμε αυτό το πρόβλημα δημιουργούμε εκ των προτέρων σύνολα ασθενών, τα οποία δεν έχουν επικαλυπτόμενους ασθενείς στο σύνολο δεδομένων εκπαίδευσης και ελέγχου. Με αυτόν τον τρόπο δημιουργούμε σύνολα δεδομένων εκπαίδευσης χωρίς επικάλυψη, αποφεύγοντας το πρόβλημα διαρροής δεδομένων.

Η παρούσα ανάλυση μπορεί να επεκταθεί περαιτέρω συνδυάζοντας πληροφορίες από τις προαναφερθείσες μεθόδους, ειδικότερα όσες αναφέρονται στην υποενότητα για τα βιολογικά μονοπάτια 5.2. Εφόσον, πλέον, η πληροφορία για τις συνθετικές τροχιές των ασθενών είναι διαθέσιμη, είναι εφικτή η ομαδοποίηση των ασθενών και με το σύνολο δεδομένων που περιέχει πληροφορίες για τα βιολογικά μονοπάτια, πέρα από τα γονιδιακά δεδομένα.

Τέλος, αξίζει να τονιστεί πως στην παρούσα έρευνα η έννοια της ταξινόμησης σταδίων διευρύνεται αρκετά. Με απώτερο σκοπό την εύρεση συσχετίσεων και την εκπαίδευση του μοντέλου, πολλές φορές η πρόβλεψη θα ξεφύγει από την 3-κλάσεων ταξινόμηση σταδίων (Στάδιο *I*, Στάδιο *II*, Στάδιο *III*) και θα προχωρήσει στην πρόβλεψη και άλλων ειδών ταξινόμησης με βάση τα στάδια. Σε αυτήν την προσπάθεια μετουσιώνουμε το πολλαπλών κλάσεων πρόβλημα σε δυαδικό πρόβλημα ταξινόμησης μελετώντας την ικανότητα των μοντέλων να διακρίνουν τα στάδια μεταξύ τους. Αυτό συμβαίνει σίγουρα στην μεθοδολογία Ταξινόμησης Μεταβάσεων (5.5), αλλά στην επόμενη ενότητα αποτελεσμάτων αρκετές από τις παραπάνω μεθοδολογίες θα επεκταθούν σε διαφορετικές προβλέψεις, όπως μεταξύ πρώιμων και όψιμων (Early and Late) σταδίων καρκίνων του μαστού.

Κεφάλαιο 6

Αποτελέσματα

Σε αυτό το κεφάλαιο παρουσιάζονται τα αποτελέσματα της προτεινόμενης μεθοδολογίας, όπως αυτή περιγράφεται στο Κεφάλαιο 5.

Η ανάλυση ακολουθεί την ίδια σειρά μελέτης με το προηγούμενο κεφάλαιο. Αρχικά, συγκρίνονται τα αποτελέσματα και η απόδοση των ταξινομητών, κλασσικών αλγορίθμων και νευρωνικών δικτύων, που έχουν εκπαιδευτεί με γονιδιακά δεδομένα. Στη συνέχεια, εξετάζεται η απόδοση των μοντέλων στην εκπαίδευση τους με biological pathways έναντι γονιδιακών δεδομένων.

Λαμβάνοντας υπόψη τα διαθέσιμα κλινικά μεταδεδομένα αναλύεται η επίδραση τους στα αποτελέσματα, όταν το σύνολο των δειγμάτων εκπαίδευσης περιοριστεί με βάση αυτά. Ύστερα μελετάται η επίδραση στην αποτελεσματικότητα των μοντέλων, όταν τα δεδομένα εισόδου, σε αυτήν την περίπτωση τα Biological Pathways ενισχυθούν με συνθετικά δεδομένων. Τέλος, διερευνάται η απόδοση των μοντέλων όταν το ζητούμενο πρόβλημα πολλαπλών κατηγοριών ταξινόμησης σταδίων μετασχηματιστεί σε πρόβλημα ταξινόμησης μεταβάσεων σταδίων.

Μέσω των παραπάνω προσεγγίσεων, επιδιώκεται η εξαγωγή ολοκληρωμένων και αξιόπιστων συμπερασμάτων σχετικά με την απόδοση και τη βελτιστοποίηση των μοντέλων ταξινόμησης σε διάφορα σενάρια για τα στάδια του καρκίνου του μαστού.

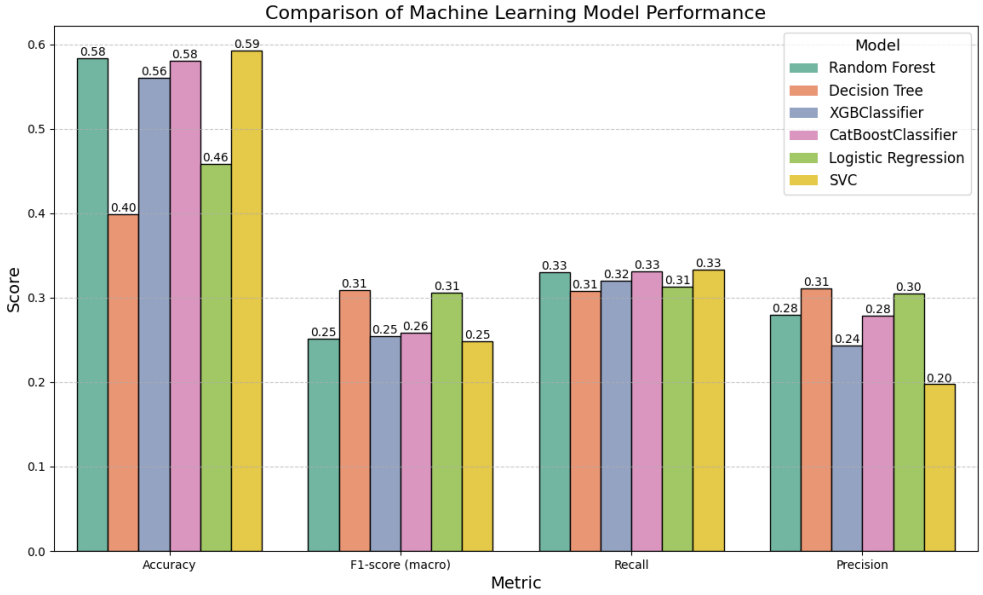
6.1 Ταξινόμηση Σταδίων Καρκίνου του Μαστού με δεδομένα γονιδιακής έκφρασης

Η πρώτη προσπάθεια ανάπτυξης μοντέλου ταξινόμησης των σταδίων του καρκίνου του μαστού χρησιμοποιεί ως είσοδο τα δεδομένα γονιδιακής έκφρασης και χρησιμοποιεί και κλασσικούς αλγόριθμους και νευρωνικά δίκτυα MLPs για τον σκοπό αυτό.

Αλγόριθμοι Μηχανικής Μάθησης

Στο παρακάτω Σχήμα 6.1 παρατηρούμε την απόδοση του μοντέλου σε 6 διαφορετικές υλοποιήσεις αλγορίθμων μηχανικής μάθησης. Στόχος των παρακάτω μοντέλων είναι η ταξινόμηση των γονιδιακών δεδομένων σε 3 κλάσεις (Στάδιο *I*, Στάδιο *II*, Στάδιο *III*).

Προκειμένου να γίνει ολοκληρωμένη αξιολόγηση των αποτελεσμάτων, πρέπει να ληφθεί υπόψη η μη-ισορροπημένη κατανομή των δεδομένων. Αυτό είναι σημαντικό, καθώς όπως φαίνεται στο Σχήμα 6.1 η μετρική *accuracy* καταφέρνει να έχει απόδοση κοντά στο 60%,



Σχήμα 6.1: *Σύγκριση μετρικών απόδοσης μοντέλων μηχανικής μάθησης*

ωστόσο, λόγω της ανισοκατανομής των κλάσεων, η απόδοση αυτή είναι παραπλανητική για την συνολική ικανότητα του μοντέλου.

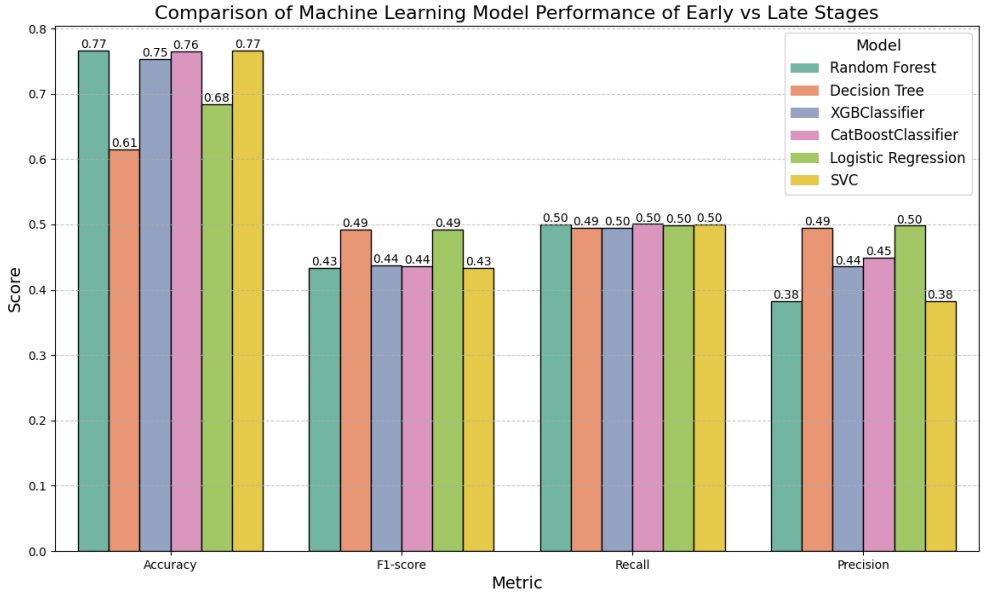
Η μετρική *Accuracy* δεν λαμβάνει υπόψη την ανισορροπία μεταξύ των κλάσεων. Σε ένα σύνολο δεδομένων, όπως αυτό που μελετάται στην παρούσα εργασία, όπου μία κλάση είναι πολύ πιο συχνή από τις άλλες ένας ταξινομητής μπορεί να επιτύχει υψηλή ακρίβεια απλά προβλέποντας πάντα τη συχνότερη κλάση. Πιο αντιπροσωπευτικά αποτελέσματα απόδοσης γίνονται αντιληπτά μέσω των υπόλοιπων μετρικών. Για αυτόν τον λόγο, θα δίνεται περισσότερη έμφαση στις υπόλοιπες μετρικές.

Στο συγκεκριμένο πρόβλημα ταξινόμησης, παρατηρούμε πως το F1-score είναι σχετικά χαμηλό για όλα τα μοντέλα, υποδεικνύοντας δυσκολία στη διατήρηση καλής ισορροπίας μεταξύ ανάκλησης και ακρίβειας. Η χαμηλή αυτή τιμή προκύπτει από τις εξίσου χαμηλές τιμές που λαμβάνουν οι μετρικές της ανάκλησης και της ακρίβειας. Οι τιμές των διαφορετικών μοντέλων για την ανάκληση είναι σχεδόν ίδιες, κυμαινόμενες ανάμεσα σε 31 – 33%. Η ακρίβεια ποικίλλει περισσότερο μεταξύ των μοντέλων, με το SVC να έχει τη χαμηλότερη ακρίβεια, 20% και το XGBoost την υψηλότερη, 31%.

Συνολικά, η γενικευμένη χαμηλή απόδοση σε όλες τις μετρικές υποδεικνύει ότι υπάρχει περιθώριο βελτίωσης στην ικανότητα των μοντέλων να ανιχνεύουν και να προβλέπουν σωστά τις διαφορετικές κλάσεις. Η ανάλυση αυτή δείχνει ότι ενώ κάποια μοντέλα έχουν καλύτερη συνολική απόδοση, κανένα από τα εξεταζόμενα μοντέλα δεν έχει καλές επιδόσεις σε όλες τις μετρικές.

Στο Σχήμα 6.2 φαίνονται τα αποτελέσματα από το πρόβλημα ταξινόμησης πρώιμων και όψιμων δειγμάτων, το οποίο είναι ένα δυαδικό πρόβλημα ταξινόμησης.

Η μετρική Accuracy διατηρείται υψηλή (75-77%) σχεδόν σε όλα τα μοντέλα πέρα από Decision Trees και Logistic Regression. Αυτό δικαιολογείται, καθώς και εδώ συνεχίζει να υ-



Σχήμα 6.2: *Σύγκριση μετρικών απόδοσης μοντέλων μηχανικής μάθησης για ταξινόμηση πρώιμων και όψιμων (Early vs Late) κλάσεων*

πάρχει το πρόβλημα των μη-ισορροπημένων δεδομένων με την κλάση των πρώιμων δειγμάτων να είναι η πλειοψηφούσα.

Οι τιμές κοντά στο 50%, για τις υπόλοιπες μετρικές, υποδηλώνουν τυχαιότητα και αδυναμία του μοντέλου να ταξινομήσει τα δεδομένα του συνόλου ελέγχου. Όλα τα μοντέλα έχουν ανάκληση 50%, δηλαδή αναγνωρίζουν σωστά το 50% των πραγματικών θετικών περιπτώσεων. Αυτό δείχνει ότι τα μοντέλα είναι εξίσου καλά (ή κακά) στον εντοπισμό πρώιμων και όψιμων σταδίων χωρίς προτίμηση, αφού χάνουν τις μισές από τις πραγματικές θετικές περιπτώσεις.

Πολλαπλών Στρωμάτων Perceptron - MLPs

Στην συνέχεια εξετάζεται η απόδοση του μοντέλου με χρήση MLPs. Παρακάτω παρατίθεται η απόδοση της επικρατέστερης αρχιτεκτονικής MLP μαζί με διαφορετικές υπερπαραμέτρους που χρησιμοποιήθηκαν για την εκπαίδευση του μοντέλου.

Από τα δεδομένα που παρουσιάζονται στον Πίνακα 6.1, μπορούμε να εξάγουμε διάφορα συμπεράσματα για την απόδοση του μοντέλου με διαφορετικές υπερπαραμέτρους. Παρακάτω αναλύονται οι βασικοί παράγοντες και οι αντίστοιχες επιδόσεις τους:

Οι τιμές του Dropout που εξετάστηκαν είναι 0.1, 0.2 και 0.3. Γενικά, η τιμή 0.1 φαίνεται να δίνει καλύτερα αποτελέσματα στις περισσότερες μετρικές, όπως το F1-score, το Recall και το Precision. Οι τιμές του Learning Rate (LR) που εξετάστηκαν είναι 5e-05, 1e-05, 0.001 και 0.0001. Το LR 1e-05 φαίνεται να δίνει τις καλύτερες επιδόσεις, όπως φαίνεται από τις υψηλότερες τιμές F1-score, Recall και Precision σε συνδυασμό με Dropout 0.1.

Υπάρχουν τρεις διαφορετικοί συνδυασμοί βαρών (2.5, 1, 2.5), (1, 0.25, 1), και (3, 1.5, 3), όπου ο καλύτερος συνδυασμός βαρών φαίνεται να είναι (2.5, 1, 2.5). Το σύνολο αυτό βαρών

σχηματίστηκε λαμβάνοντας υπόψη τις κατανομές μεταξύ των κατηγοριών. Οι τιμές της L2 κανονικοποίησης που εξετάστηκαν είναι 0.1 και 0.2. Γενικά, η τιμή 0.1 δίνει καλύτερα αποτελέσματα σε σύγκριση με την τιμή 0.2, ειδικά όταν συνδυάζεται με Dropout 0.1 και LR 1e-05.

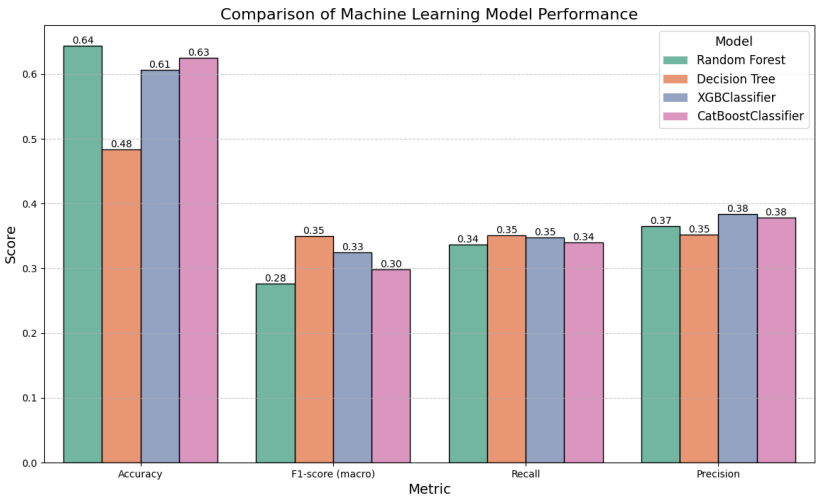
Η καλύτερη απόδοση παρατηρείται με τις εξής υπερπαραμέτρους: Dropout: 0.1, LR: 1e-05, Weights: (2.5, 1, 2.5), L2-Reg: 0.1. Αυτή η ρύθμιση δίνει Accuracy 51%, F1-score, Recall, και Precision στο 41%.

Ωστόσο, παρά το γεγονός ότι ορισμένοι συνδυασμοί υπερπαραμέτρων δίνουν σχετικά καλύτερες επιδόσεις, η γενική εικόνα της απόδοσης του μοντέλου δεν είναι ικανοποιητική. Οι τιμές των μετρικών, όπως η ακρίβεια (Accuracy), η F1-score, η ανάκληση (Recall) και η ακρίβεια (Precision), είναι χαμηλές σε απόλυτες τιμές. Η καλύτερη F1-score που επιτυγχάνεται είναι μόλις 41%. Αυτές οι τιμές υποδεικνύουν ότι το μοντέλο δεν έχει καλή ικανότητα γενίκευσης και πιθανότατα δεν καταφέρνει να αναγνωρίσει επαρκώς τα πρότυπα στα δεδομένα.

Υπερπαράμετροι				**Μετρικές**			
Dropout	LR	Weights	L2-Reg	Accuracy	F1-score	Recall	Precision
0.1	5e-05	(2.5, 1, 2.5)	0.1	0.4431	0.3091	0.3748	0.3501
0.1	5e-05	(2.5, 1, 2.5)	0.2	0.5344	0.2284	0.3331	0.1782
0.2	5e-05	(2.5, 1, 2.5)	0.1	0.4901	0.3021	0.3692	0.3593
0.2	5e-05	(2.5, 1, 2.5)	0.2	0.5452	0.2898	0.3466	0.2812
0.1	**1e-05**	**(2.5, 1, 2.5)**	**0.1**	**0.5182**	**0.4104**	**0.4161**	**0.4190**
0.1	1e-05	(2.5, 1, 2.5)	0.2	0.5565	0.3130	0.3551	0.3098
0.2	1e-05	(2.5, 1, 2.5)	0.1	0.5146	0.4009	0.4098	0.4128
0.2	1e-05	(2.5, 1, 2.5)	0.2	0.5540	0.3232	0.3622	0.3129
0.1	1e-05	(1, 0.25, 1)	0.1	0.3321	0.3280	0.4401	0.4351
0.1	1e-05	(1, 0.25, 1)	0.2	0.2364	0.1394	0.3382	0.1245
0.2	1e-05	(1, 0.25, 1)	0.1	0.3203	0.3124	0.4292	0.4427
0.2	1e-05	(1, 0.25, 1)	0.2	0.2345	0.1328	0.3343	0.1253
0.1	0.001	(2.5, 1, 2.5)	0.2	0.5421	0.2354	0.3347	0.1872
0.1	0.001	(2.5, 1, 2.5)	0.1	0.5921	0.2479	0.3333	0.1974
0.2	0.001	(2.5, 1, 2.5)	0.2	0.5495	0.2419	0.3340	0.2033
0.3	0.001	(2.5, 1, 2.5)	0.2	0.5747	0.2440	0.3326	0.1948
0.3	0.0001	(2.5, 1, 2.5)	0.2	0.5604	0.2544	0.3397	0.2548
0.1	1e-05	(3, 1.5, 3)	0.1	0.5772	0.3265	0.3605	0.4352
0.1	1e-05	(3, 1.5, 3)	0.2	0.5921	0.2479	0.3333	0.1974
0.2	1e-05	(3, 1.5, 3)	0.1	0.5787	0.3184	0.3571	0.4267
0.2	1e-05	(3, 1.5, 3)	0.2	0.5921	0.2479	0.3333	0.1974
0.1	5e-05	(3, 1.5, 3)	0.1	0.5685	0.2736	0.3431	0.2456
0.1	5e-05	(3, 1.5, 3)	0.2	0.5921	0.2479	0.3333	0.1974
0.2	5e-05	(3, 1.5, 3)	0.1	0.5862	0.2582	0.3382	0.2355
0.2	5e-05	(3, 1.5, 3)	0.2	0.5921	0.2479	0.3333	0.1974

Πίνακας 6.1: *Διαφορετικές ρυθμίσεις υπερπαραμέτρων για το μοντέλο MLP*

6.2 Ταξινόμηση Σταδίων Καρκίνου του Μαστού με χρήση Pathway Analysis



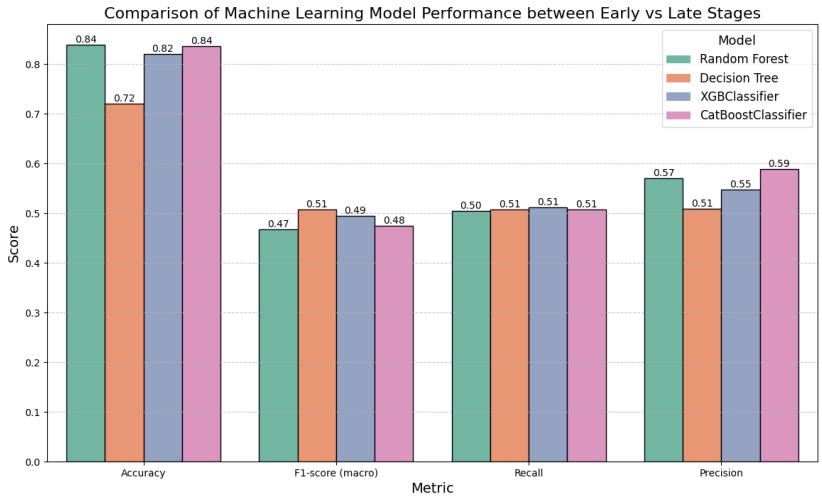
Σχήμα 6.3: *Σύγκριση αποτελεσμάτων ταξινόμησης 3-κλάσεων με κλασσικούς αλγόριθμους μηχανικής μάθησης και είσοδο τα 22 βιολογικά μονοπάτια*

Το Σχήμα 6.3 απεικονίζει τα αποτελέσματα της ταξινόμησης σταδίων με τροποποιημένα δεδομένα εισόδου ύστερα από ανάλυση μονοπατιών.

Παρατηρούμε πως οι τιμές των μετρικών Precision, Recall, F1-score είναι σχετικά σταθερές τόσο μεταξύ τους όσο και ανάμεσα στα διαφορετικά μοντέλα, και κυμαίνονται ανάμεσα στο $34 - 38\%$. Οι χαμηλές τιμές των μετρικών Precision και Recall υποδηλώνουν πως τα μοντέλα δεν είναι ιδιαίτερα καλά στον εντοπισμό πραγματικών θετικών αποτελεσμάτων ούτε στην ελαχιστοποίηση των ψευδώς θετικών αποτελεσμάτων.

Στο Σχήμα 6.4 έχει υπολογιστεί η απόδοση του ταξινομητή στο δυαδικό πρόβλημα ταξινόμησης μεταξύ πρώιμων (Στάδιο *I, II*) και όψιμων κατηγοριών (Στάδιο *III*).

Τα αποτελέσματά δείχνουν πως ενώ τα μοντέλα δείχνουν κάποια ικανότητα διάκρισης μεταξύ πρώιμων και όψιμων σταδίων καρκίνου, η απόδοσή τους είναι οριακά τυχαία. Η χρήση της ακρίβειας, της ανάκλησης και του F1-score συμβάλλει στην ανάδειξη αυτού του ζητήματος, καταδεικνύοντας ότι ακόμη και με υψηλό Accuracy, τα μοντέλα δεν είναι αξιόπιστα λόγω της μη ιδανικής ισορροπίας μεταξύ αληθώς θετικών και ψευδώς θετικών αποτελεσμάτων.



Σχήμα 6.4: *Σύγκριση μετρικών απόδοσης μοντέλων μηχανικής μάθησης για ταξινόμηση πρώιμων και όψιμων (Early vs Late) κλάσεων με είσοδο τα 22 βιολογικά μονοπάτια*

6.3 Ταξινόμηση Σταδίων με Κλινικά Μεταδεδομένα

Στόχος υλοποίησης της συγκεκριμένης ταξινόμησης είναι να εξεταστεί αν η ετερογένεια που παρατηρείται μεταξύ των ασθενών επηρεάζει αρνητικά τα αποτελέσματα. Η ετερογένεια μεταξύ των ασθενών αναφέρεται στην ποικιλομορφία των ασθενών που συμπεριλαμβάνονται στα δεδομένα, όπως ασθενείς διαφορετικών ηλικιών, εθνικοτήτων, φυλών, αλλά και διαφορετικών ιστολογικών τύπων, πέρα φυσικά από την ετερογένεια που φέρει η διαφοροποίηση των ασθενών σε στάδια.

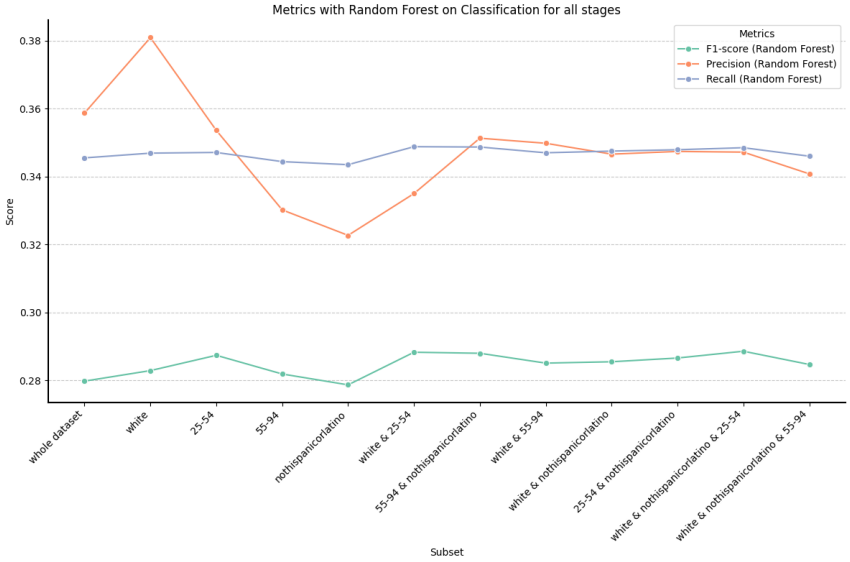
Για να μελετηθεί η επιρροή της ετερογένειας στην απόδοση των μοντέλων εφαρμόζονται περιορισμοί στα δεδομένα. Αυτό συνεπάγεται πως δημιουργούνται υποσύνολα ασθενών, όπου όλοι οι ασθενείς μοιράζονται π.χ έναν συγκεκριμένο ιστολογικό τύπο και βρίσκονται στην ίδια ηλικιακή ομάδα. Παρακάτω παρουσιάζεται η επίδοση δύο εκ των έξι ταξινομητών, οι οποίοι απέδωσαν καλύτερα στην ταξινόμηση που παρουσιάστηκε στην υποενότητα 6.1. Συγκεκριμένα, επιλέχθηκαν οι αλγόριθμοι Random Forest και Gradient Boosting - ο τελευταίος μέσω της υλοποίησης του CatboostClassifier.

Οι γραφικές παραστάσεις παρουσιάζουν τις επιδόσεις των δύο ταξινομητών (CatBoostClassifier και Random Forest), μετρώντας τις τιμές των F1-score, Precision και Recall για διάφορα υποσύνολα δεδομένων. Τα αποτελέσματα για την ταξινόμηση όλων των σταδίων (multiclass classification) είναι τα παρακάτω.

Η αρχική τιμή στο διάγραμμα γραμμών είναι θεωρητικά ταυτόσημη με αυτές που παρατηρήθηκαν στα μοντέλα της πρώτης υποενότητας αυτού του Κεφαλαίου. Οποιαδήποτε απόκλιση μπορεί να αποδοθεί σε θέματα αρχικοποίησης μέσω random seed. Οι τιμές που ακολουθούν παρουσιάζουν την απόδοση του μοντέλου για τις υπόλοιπες μετρικές.

Η συνολική εικόνα υποδεικνύει ότι τα αποτελέσματα δεν εμφανίζουν αισθητή βελτίωση μέσω της εφαρμογής αυτής της μεθόδου. Αυτό υποδηλώνει ότι η μείωση της ετερογένειας των δειγμάτων δεν οδηγεί απαραίτητα σε βελτίωση της απόδοσης. Επιπλέον, δεν προχωράμε

σε περαιτέρω περιορισμούς με χρήση περισσότερων κλινικών περιορισμών, καθώς το σύνολο των δεδομένων γίνεται σημαντικά μικρότερο, καθιστώντας την εκπαίδευση του μοντέλου πιο δύσκολη.

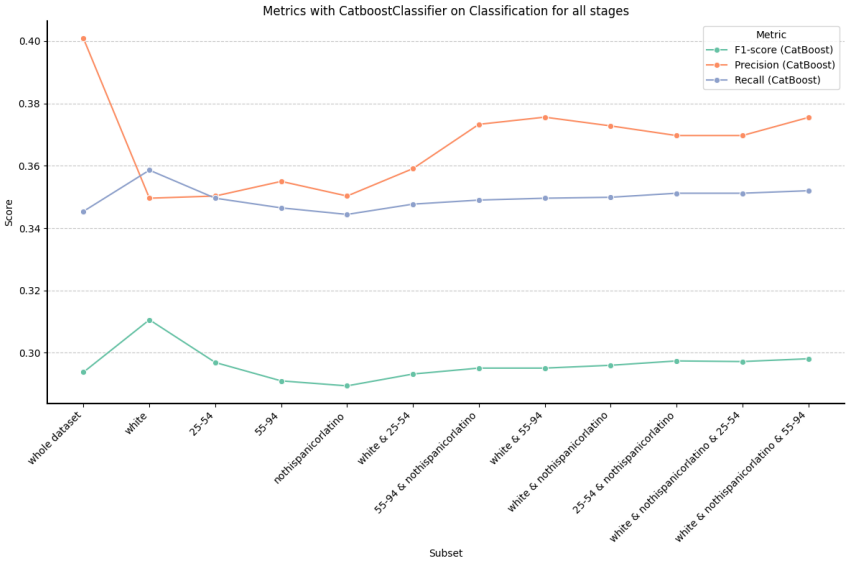


Σχήμα 6.5: *Σύγκριση αποτελεσμάτων μετρικών με διαφορετικά υποσύνολα των δεδομένων με Random Forest Classifier*

Στην περίπτωση του Random Forest, το F1-score βελτιώνεται ελάχιστα (+2%) κατά τη διάρκεια αυτής της διαδικασίας, με τη μεγαλύτερη τιμή να παρατηρείται στο υποσύνολο δειγμάτων που περιλαμβάνει λευκές γυναίκες ηλικίας 25-54 ετών. Ωστόσο, οι υπόλοιπες μετρικές (Precision, Recall) παραμένουν χαμηλές.

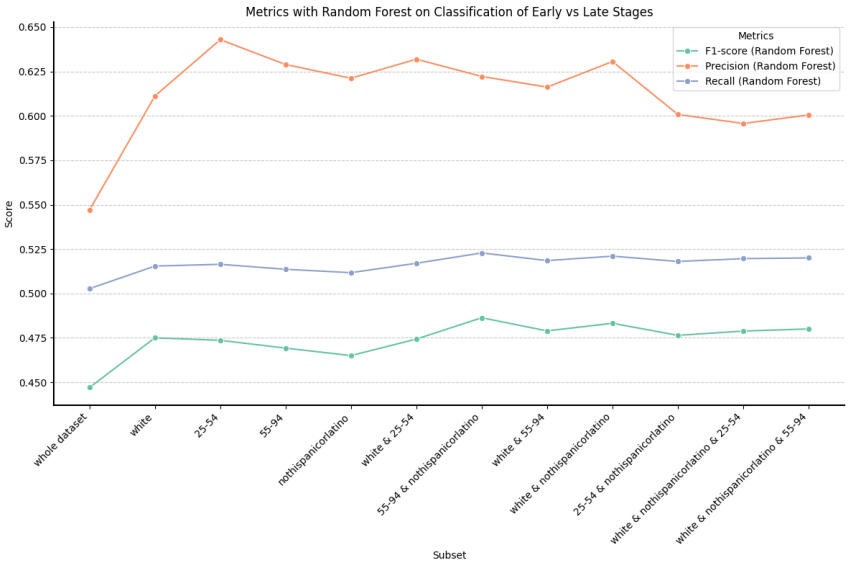
Αντίστοιχα, στην περίπτωση του CatBoostClassifier, η μέθοδος δεν δείχνει σημαντική βελτίωση σε σύγκριση με τα αρχικά και πλήρη δεδομένα. Η υψηλότερη τιμή F1-score παρατηρείται στο υποσύνολο των λευκών γυναικών (περίπου 31%), ενώ η τιμή της ανάκλησης μεγιστοποιείται σε σχέση με όλα τα υπόλοιπα υποσύνολα (36%). Γενικά, το F1-score παραμένει σχετικά σταθερό, με μικρές διακυμάνσεις ανάμεσα στα υποσύνολα.

Τα αποτελέσματα δείχνουν ότι η μείωση της ετερογένειας στα γονιδιακά δεδομένα δεν διευκολύνει την ταξινόμηση σε τρία στάδια, εκτός από ελάχιστες περιπτώσεις και αυτό σε πολύ μικρό βαθμό. Αυτή η προσπάθεια δεν παρέχει επαρκή επιβεβαίωση για την ανάπτυξη ενός αξιόπιστου μοντέλου.



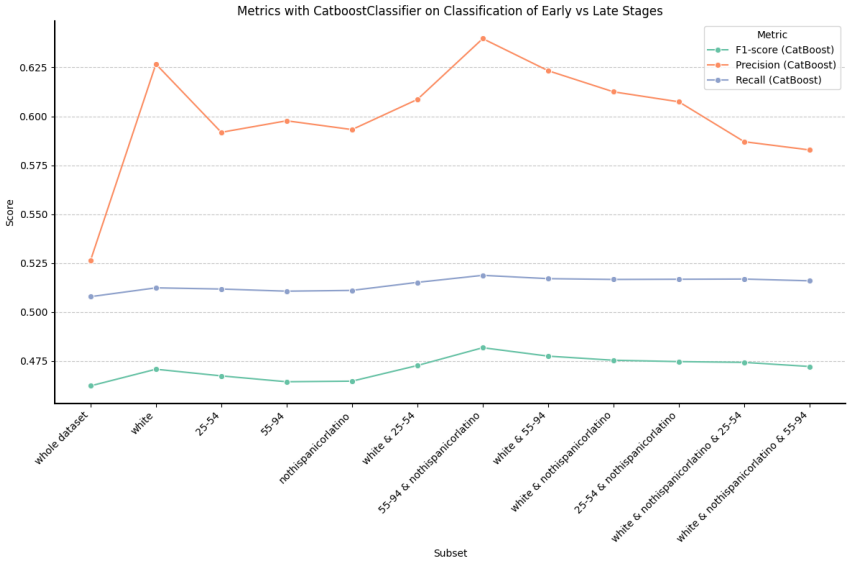
Σχήμα 6.6: *Σύγκριση αποτελεσμάτων μετρικών με διαφορετικά υποσύνολα των δεδομένων με Catboost Classifier*

Όπως και στις προηγούμενες προσεγγίσεις, πραγματοποιείται και εδώ ταξινόμηση μεταξύ πρώιμων και όψιμων σταδίων.



Σχήμα 6.7: *Σύγκριση αποτελεσμάτων μετρικών με διαφορετικά υποσύνολα των δεδομένων με Random Forest Classifier για ταξινόμηση μεταξύ πρώιμων και όψιμων σταδίων*

Η γενική εικόνα δείχνει ότι το δυαδικό πρόβλημα ταξινόμησης συνεχίζει να παρουσιάζει τιμές στις μετρικές Recall και F1-score κοντά στο 50%. Αν και το F1-score αυξάνεται σε σύγκριση με τα πρωτότυπα δείγματα, αυτή η αύξηση πλησιάζει όλο και περισσότερο το 50%, γεγονός που υποδηλώνει μια πιο τυχαία απόδοση του μοντέλου. Αυτό συνεπάγεται ότι το πρόβλημα δυαδικής ταξινόμησης γίνεται πιο δύσκολο. Επιπλέον, η μείωση του πλήθους των δειγμάτων φαίνεται να παίζει καθοριστικό ρόλο στην απόδοση του μοντέλου, καθώς καθιστά δυσκολότερη την εύρεση των βέλτιστων παραμέτρων για την ταξινόμηση.



Σχήμα 6.8: *Σύγκριση αποτελεσμάτων μετρικών με διαφορετικά υποσύνολα των δεδομένων με Catboost Classifier για ταξινόμηση μεταξύ πρώιμων και όψιμων σταδίων*

Παράλληλα, παρατηρείται αύξηση στο Precision κατά περίπου 10% και στους δύο ταξινομητές. Αυτό υποδηλώνει ότι το μοντέλο γίνεται πιο αποτελεσματικό στον εντοπισμό των σωστών κλάσεων, με λιγότερες ψευδώς θετικές προβλέψεις. Ταυτόχρονα, ενώ η ακρίβεια βελτιώνεται, η ανάκληση παραμένει σταθερή.

Συνολικά, τα αποτελέσματα δείχνουν ότι η προσπάθεια βελτίωσης της ταξινόμησης μέσω της μείωσης της ετερογένειας των δειγμάτων οδηγεί σε μικτά αποτελέσματα. Ενώ υπάρχει μια βελτίωση στην ακρίβεια, η γενική απόδοση του μοντέλου δεν βελτιώνεται ουσιαστικά, λόγω της αυξημένης δυσκολίας στην ταξινόμηση και της μειωμένης επάρκειας δεδομένων.

6.4 Ταξινόμηση με Συνθετικά Δεδομένα

Στην παρούσα υποενότητα θα παρουσιαστεί η απόδοση των ταξινομητών με τη χρήση συνθετικών δεδομένων, τα οποία έχουν παραχθεί μέσω των τεχνικών ADASYN και SMOTE, για δύο είδη ταξινομήσεων. Η πρώτη ταξινόμηση αφορά τις τρεις κλάσεις, Στάδιο *I*, *II*, και *III*, ενώ η δεύτερη ταξινόμηση αφορά τις δύο κλάσεις, πρώιμων και όψιμων δειγμάτων.

Από την ανάλυση των συνθετικών δεδομένων προκύπτουν τα ακόλουθα:

Τα αποτελέσματα της ταξινόμησης μεταξύ των τεχνικών SMOTE και ADASYN είναι πολύ παρόμοια. Οι ομοιότητα αυτή στα αποτελέσματα μπορεί να οφείλεται σε διάφορες αιτίες.

Όταν υπάρχει σημαντική επικάλυψη μεταξύ των κλάσεων, τόσο η SMOTE όσο και η ADASYN ενδέχεται να παράγουν παρόμοια συνθετικά δείγματα, οδηγώντας σε παρόμοιες μετρικές απόδοσης. Επειδή και οι δύο τεχνικές παράγουν νέα δείγματα με βάση τα υπάρχοντα και σε περιπτώσεις, όπου τα όρια των κλάσεων δεν είναι σαφή, τα νέα δείγματα ενδέχεται να μην παρέχουν πρόσθετες χρήσιμες πληροφορίες. Αυτό έγινε φανερό και από την ανάλυση κύριων συνιστωσών σε έναν βαθμό, όπως φανερώθηκε στο Κεφάλαιο 4 στο Σχήμα 4.4.

Ταξινομητές	ADASYN			
	Accuracy	F1-score	Recall	Precision
Random Forest	0.5557	0.3842	0.3865	0.4187
Decision Trees	0.4094	0.3461	0.3586	0.3502
XGBClassifier	0.5515	0.3885	0.3877	0.4241
SVC	0.4587	**0.4241**	**0.4547**	**0.4306**
MLP	0.2113	0.1766	0.3535	0.1229

Πίνακας 6.2: *Αποτελέσματα ταξινομητών μεταξύ όλων των σταδίων με ADASYN*

Σχετικά με το πρώτο είδος ταξινόμησης μεταξύ των 3 κλάσεων τα ποσοστά συνεχίζουν να είναι αρκετά χαμηλά. Σε σύγκριση ωστόσο με τα αποτελέσματα από την ταξινόμηση των πρωτότυπων δεδομένων γονιδιακής έκφρασης χωρίς προσθήκη συνθετικών δεδομένων, όπως φαίνονται στο Σχήμα 6.1, παρατηρούμε μια αισθητή βελτίωση της τάξης του 10% για τους κλασσικούς αλγόρθμους. Επομένως, εκεί τα συνθετικά δεδομένα φαίνεται σε έναν βαθμό να βοήθησαν στην διάκριση των κλάσεων. Αντίθετα, τα MLPs εμφάνισαν την χειρότερη απόδοση συγκριτικά με τους υπόλοιπους αλγόριθμους και σημαντική μείωση απόδοσης σε σύγκριση με την χρήση μόνο των πρωτότυπων δεδομένων.

Ταξινομητές	SMOTE			
	Accuracy	F1-score	Recall	Precision
Random Forest	0.5574	0.3889	0.3929	0.4268
Decision Trees	0.4145	0.3343	0.3398	0.3385
XGBClassifier	0.5480	0.3654	0.3712	0.4036
SVC	0.4774	**0.4428**	**0.4760**	**0.4447**
MLP	0.2343	0.1265	0.3333	0.0781

Πίνακας 6.3: *Αποτελέσματα ταξινομητών μεταξύ όλων των σταδίων με SMOTE*

Στην συνέχεια πραγματοποιείται δυαδική ταξινόμηση μεταξύ Πρώιμων και Όψιμων σταδίων καρκίνου του μαστού. Εδώ ο ταξινομητής SVC και με τις δύο τεχνικές SMOTE και ADASYN επιτυγχάνει την βέλτιστη απόδοση ανάμεσα στους υπόλοιπους αλγορίθμους ταξινόμησης. Η τιμή Recall, 59%, υποδηλώνει καλή απόδοση στην ανίχνευση θετικών δειγμάτων, και η τιμή Precision 57% είναι επίσης αρκετά καλή. Παρόλο που το Accuracy του SVC είναι χαμηλότερο από τον Random Forest και τον XGBClassifier, η συνολική ισορροπία που παρουσιάζει ο SVC τον καθιστά αξιόπιστο.

Ταξινομητές	ADASYN			
	Accuracy	F1-score	Recall	Precision
Random Forest	0.7506	0.4916	0.5217	0.5914
Decision Trees	0.5983	0.4961	0.4987	0.4990
XGBClassifier	0.7370	0.54423	0.5473	0.5913
SVC	0.6468	**0.5724**	**0.5880**	0.5720

Πίνακας 6.4: *Αποτελέσματα ταξινομητών μεταξύ Early vs Late με δημιουργία συνθετικών δεδομένων μέσω τεχνικής ADASYN*

Αξίζει επιπλέον να αναφερθεί ότι, συγκριτικά με την επίδοση των μοντέλων που χρησι-

μοποιούν τα πρωτότυπα δεδομένα για ταξινόμηση, όπως αναλύθηκε στο πρώτο μέρος του παρόντος κεφαλαίου και απεικονίζεται στο σχήμα 6.2, η απόδοση των μοντέλων με τη χρήση συνθετικών δεδομένων είναι αισθητά βελτιωμένη. Αυτή η βελτίωση αποδεικνύει την αποτελεσματικότητα των τεχνικών ADASYN και SMOTE στην εξισορρόπηση του συνόλου δεδομένων και στην αύξηση της ακρίβειας των ταξινομητών, επιτρέποντας τους να αποδώσουν καλύτερα στην ταξινόμηση των σταδίων καρκίνου του μαστού.

Ταξινομητές	SMOTE			
	Accuracy	F1-score	Recall	Precision
Random Forest	0.7617	0.4958	0.5279	0.6695
Decision Trees	0.6145	0.5190	0.5236	0.5211
XGBClassifier	0.7455	0.5444	0.5493	0.6062
SVC	0.6357	**0.5726**	**0.5962**	0.5754

Πίνακας 6.5: *Αποτελέσματα ταξινομητών μεταξύ Early vs Late με δημιουργία συνθετικών δεδομένων μέσω τεχνικής SMOTE*

Μέχρι στιγμής, η χρήση των συνθετικών δεδομένων έχει αποδειχθεί ιδιαίτερα αποτελεσματική, επιτυγχάνοντας τα υψηλότερα ποσοστά απόδοσης σε όλες τις μετρικές συνδυαστικά ανάμεσα σε όλες τις παραπάνω μεθοδολογίες.

6.5 Ταξινόμηση Μεταβάσεων Μεταξύ Σταδίων

Η τελευταία προσέγγιση που διερευνάται στο πλαίσιο αυτής της εργασίας είναι η ταξινόμηση μεταβάσεων μεταξύ σταδίων του καρκίνου του μαστού. Πρόκειται για ένα δυαδικό πρόβλημα ταξινόμησης, το οποίο αξιολογείται με δύο διαφορετικές μεθόδους: Συνένωση Διανυσμάτων και Διαφορά Διανυσμάτων Ασθενών, χρησιμοποιώντας δεδομένα από Βιολογικά Μονοπάτια (Biological Pathways).

Τα αποτελέσματα των ταξινομητών με τη μέθοδο "Συνένωσης Διανυσμάτων" παρουσιάζονται στον Πίνακα 6.6:

Ταξινομητές	Μετρικές Απόδοσης			
	Accuracy	F1-score	Recall	Precision
Random Forest	0.7058	0.4848	0.4959	0.4989
Decision Trees	0.5767	0.4830	0.5143	0.5112
XGBClassifier	0.6425	0.4996	0.5117	0.5146
CatBoostClassifier	0.6925	**0.5160**	0.5266	0.5327

Πίνακας 6.6: *Αποτελέσματα ταξινομητών με υλοποίηση της μεθόδου "Συνένωσης Διανυσμάτων" με χρήση Biological Pathways*

Τα αποτελέσματα ταξινόμησης για το πρόβλημα μετάβασης σταδίων μέσω της μεθόδου "Συνένωσης Διανυσμάτων" υποδεικνύουν μη βέλτιστη απόδοση στους αξιολογούμενους ταξινομητές. Η μετρική F1-score, η οποία είναι κρίσιμη για την αξιολόγηση του μοντέλου σε μη ισορροπημένα δεδομένα, είναι ιδιαίτερα χαμηλή και κυμαίνεται γύρω στο 50%. Αυτό υποδηλώνει ότι τα μοντέλα δυσκολεύονται να διακρίνουν αποτελεσματικά τις δύο μεταβατικές κατηγορίες.

Συγκριτικά, με τους υπόλοιπους ταξινομητές, ο Catboost Classifier λαμβάνει την υψηλότερη τιμή στο F1-score (51,6%) με ακρίβεια 69,25%. Ωστόσο, η απόδοση εξακολουθεί να μην είναι ικανοποιητική για την ανάπτυξη ενός μοντέλου.

Συνολικά, τα αποτελέσματα υπογραμμίζουν τις προκλήσεις στη χρήση biological pathways ως χαρακτηριστικών. Οι ταξινομητές, ενώ δείχνουν διαφορετικούς βαθμούς ακρίβειας, αποτυγχάνουν σταθερά να επιτύχουν βαθμολογία F1-score σημαντικά πάνω από 50%, υπογραμμίζοντας την περιορισμένη διακριτική τους δύναμη σε αυτό το πλαίσιο.

Τα αποτελέσματα των ταξινομητών με τη μέθοδο "Διαφοράς Διανυσμάτων" Ασθενών παρουσιάζονται στον Πίνακα 6.7:

Ταξινομητές	Μετρικές Απόδοσης			
	Accuracy	F1-score	Recall	Precision
Random Forest	0.7956	0.4876	0.5214	0.7100
Decision Trees	0.7224	**0.5840**	**0.5863**	0.5842
XGBClassifier	0.7859	0.5617	0.5604	0.6459
CatBoostClassifier	0.7980	0.5301	0.5433	**0.7142**

Πίνακας 6.7: *Αποτελέσματα ταξινομητών με υλοποίηση της μεθόδου "Διαφοράς Διανυσμάτων Ασθενών" με χρήση Biological Pathways*

Συνολικά, η μέθοδος "Διαφοράς Διανυσμάτων" δείχνει μια σαφή βελτίωση στην απόδοση ταξινόμησης. Οι βαθμολογίες F1-score είναι σταθερά υψηλότερες, ιδιαίτερα για τα Decision Trees και τον XGBClassifier, υποδηλώνοντας καλύτερη αποτελεσματικότητα του μοντέλου στη διάκριση μεταξύ των δύο μεταβάσεων σταδίου καρκίνου. Την καλύτερη απόδοση με βάση το F1-score επιδεικνύει ο ταξινομητής Decision Tree (58%) και ταυτόχρονα έχει και την μεγαλύτερη τιμή Recall (58%), υποδεικνύοντας καλύτερη ικανότητα στη σωστή ταξινόμηση των θετικών κλάσεων.

Αυτά τα αποτελέσματα δείχνουν ότι η ταξινόμηση μέσω της μεθόδου "Διαφοράς Διανυσμάτων" αποτελεί μια πιο αποτελεσματική προσέγγιση σε αυτό το πλαίσιο, πιθανώς λόγω της καλύτερης αποτύπωσης των δυναμικών αλλαγών μεταξύ των σταδίων.

Οι παραπάνω μέθοδοι χρησιμοποίησαν ως είσοδο βιολογικά μονοπάτια. Παρακάτω βλέπουμε τα αποτελέσματα της ταξινόμησης μεταβάσεων μεταξύ σταδίων χρησιμοποιώντας ως είσοδο αρχικά τα γονίδια εκείνα που ανήκουν στο PAM50 και στην συνέχεια όλο το γονιδίωμα. Πραγματοποιώντας την σύγκριση αυτή, όπως φαίνεται και στον Πίνακα 6.8, μπορούμε να παρατηρήσουμε την διαγνωστική ικανότητα των γονιδίων που ανήκουν στο πάνελ 50 γονιδίων στην σταδιοποίηση και να πραγματοποιήσουμε μια άμεση σύγκριση με την ικανότητα που παρουσιάζει όλο το γονιδίωμα στην επίλυση αυτού του προβλήματος.

Στα αποτελέσματα του Πίνακα 6.8 ότι ο XGB ταξινομητής καταφέρνει συγκριτικά με τους υπόλοιπους να αποσπάσει τις υψηλότερες τιμές σε όλες τις μετρικές αξιολόγησης. Φαίνεται επίσης πως συνολικά όλο το γονιδίωμα έχει καλύτερη ικανότητα αποτύπωσης των σταδίων σε σύγκριση μόνο με τα γονίδια του PAM50.

Ταξινομητές	Γονίδια PAM50			Όλο το γονιδίωμα (8954 Γονίδια)		
	F1-score	Recall	Precision	F1-score	Recall	Precision
Random Forest	0.4959	0.5134	0.5156	0.5320	0.5446	0.5536
Decision Trees	0.4225	0.4706	0.4792	0.4358	0.5045	0.4919
XGBClassifier	**0.5185**	**0.5492**	**0.5363**	**0.5475**	**0.5703**	**0.5698**
CatBoostClassifier	0.4808	0.5064	0.5113	0.5047	0.5270	0.5401

Πίνακας 6.8: *Σύγκριση αποτελεσμάτων ταξινομητών με χρήση της μεθόδου "Συνένωσης Διανυσμάτων Ασθενών" με είσοδο PAM50 και όλο το γονιδίωμα*

Ο πίνακας γονιδίων PAM50, αν και χρήσιμος για τον χαρακτηρισμό υποτύπου καρκίνου του μαστού, δεν λειτουργεί αποτελεσματικά στην αποτύπωση της πλήρους πολυπλοκότητας των μεταβάσεων σταδίου σε σύγκριση με ολόκληρο το γονιδίωμα. Η ενσωμάτωση ενός ευρύτερου συνόλου γονιδιωματικών χαρακτηριστικών δείχνει σταθερά καλύτερες μετρήσεις απόδοσης σε όλους τους ταξινομητές.

Παρόλα αυτά και οι δύο είσοδοι συνολικά δεν καταφέρνουν να αναπτύξουν έναν ισχυρό ταξινομητή. Περαιτέρω έρευνα θα μπορούσε να διερευνήσει υβριδικά μοντέλα που αξιοποιούν τόσο τα γονίδια PAM50 όσο και πρόσθετα γονιδιωματικά χαρακτηριστικά, εξισορροπώντας πιθανώς την ερμηνευσιμότητα του PAM50 με τη βελτιωμένη απόδοση δεδομένων ολόκληρου του γονιδιώματος.

Τα αποτελέσματα των ταξινομητών με τη μέθοδο "Διαφοράς Διανυσμάτων" Εισόδου παρουσιάζονται στον Πίνακα 6.9:

Ταξινομητές	Γονίδια PAM50			Όλο το γονιδίωμα (8954 Γονίδια)		
	F1-score	Recall	Precision	F1-score	Recall	Precision
Random Forest	0.4540	0.5058	0.5607	0.4731	0.5130	0.6331
Decision Trees	0.5674	0.5693	0.5684	0.5599	0.5621	0.5609
XGBClassifier	**0.5898**	**0.5808**	0.6731	**0.6172**	**0.6007**	0.8133
CatBoostClassifier	0.5472	0.5527	**0.6842**	0.5462	0.5552	**0.8395**

Πίνακας 6.9: *Σύγκριση αποτελεσμάτων ταξινομητών με χρήση της μεθόδου "Διαφοράς Διανυσμάτων Ασθενών" με είσοδο PAM50 και όλο το γονιδίωμα*

Στα αποτελέσματα του Πίνακα 6.9 παρατηρούμε ότι ο ταξινομητής XGB καταφέρνει να έχει την υψηλότερη τιμή F1-score και στην περίπτωση της εισόδου μόνο των PAM50 γονιδίων, όσο και στην περίπτωση που ως είσοδο δέχεται όλο το γονιδίωμα. Επιπλέον, γίνεται ξεκάθαρο και σε αυτήν την μέθοδο, πως τα γονίδια PAM50 δεν είναι ικανά να αποτυπώσουν την μετάβαση μεταξύ των σταδίων. Αντίθετα, σε αυτήν την περίπτωση, με είσοδο όλο το γονιδίωμα η μετρική F1-score βρίσκεται στο 61%. Ταυτόχρονα, τόσο το Recall, όσο και το Precision λαμβάνουν εξίσου υψηλές τιμές, υποδεικνύοντας ότι το μοντέλο συνολικά καταφέρνει να έχει καλή απόδοση.

Η παρατήρηση για την καλύτερη ικανότητα του μοντέλου να διακρίνει μεταξύ των δύο μεταβάσεων στην περίπτωση της εισόδου όλου του γονιδιώματος είναι καθολική και ισχύει και για τους τέσσερις ταξινομητές που βρίσκονται υπό αξιολόγηση. Ταυτόχρονα, η υψηλή τιμή της μετρικής Precision σε όλες τις περιπτώσεις υποδηλώνει την ικανότητα του μοντέλου να προβλέπει σωστά τις δύο κατηγορίες, υποδεικνύοντας πως ένα ευρύτερο σύνολο γονιδίων

βοηθά στην ακριβέστερη διάκριση.

Κεφάλαιο 7

Συμπεράσματα και Μελλοντικές Προεκτάσεις

7.1 Σύνοψη και Συμπεράσματα

Στην παρούσα διπλωματική εργασία διερευνήθηκε η ανάπτυξη ενός μοντέλου που βασίζεται σε κλασσικούς αλγορίθμους μηχανικής μάθησης και νευρωνικά δίκτυα τύπου MLPs για την ταξινόμηση των σταδίων καρκίνου του μαστού χρησιμοποιώντας γονιδιακά δεδομένα. Συγκεκριμένα, αυτό επιτεύχθηκε μέσω της αξιοποίησης δεδομένων γονιδιακής έκφρασης mRNA, τα οποία προέκυψαν από RNA-Sequencing, καθώς και μέσω των παραγώγων τους. Επομένως, πρόκειται για ένα πρόβλημα ταξινόμησης στα στάδια καρκίνου του μαστού (Στάδιο *I, II* και *III*).

Για την επίτευξη των στόχων της έρευνας, το πρόβλημα μετασχηματίστηκε και έλαβε διάφορες μορφές. Πέρα από την πολλαπλών κατηγοριών ταξινόμηση για όλα τα στάδια της νόσου, η επόμενη σημαντική προσέγγιση που μελετήθηκε, είναι η δυαδική ταξινόμηση μεταξύ πρώιμων (Στάδιο *I, II*) και όψιμων (Στάδιο *III*) σταδίων.

Αφού πραγματοποιήθηκε η κατάλληλη προεπεξεργασία των δεδομένων, ακολούθησαν εκτενείς διερευνήσεις με σκοπό την εύρεση κατάλληλου ταξινομητή. Η έρευνα ξεκίνησε με την ταξινόμηση των δεδομένων γονιδιακής έκφρασης mRNA μέσω απλών αλγορίθμων μηχανικής μάθησης, όπως Random Forest (RF), Decision Trees (DT), Support Vector Classifier (SVC), Gradient Boosting και Linear Regression. Τα αποτελέσματα έδειξαν ότι το συγκεκριμένο πρόβλημα είναι σύνθετο και δεν επιδέχεται εύκολη λύση, αποδεικνύοντας ότι αποτελεί μια σημαντική πρόκληση. Η δυσκολία αυτή επιβεβαιώθηκε και όταν το πρόβλημα επιχειρήθηκε να επιλυθεί με Νευρωνικά Δίκτυα, μέσω μοντέλων Convolutional Neural Networks (CNN) και Multi-Layer Perceptron (MLP).

Στη συνέχεια, η μεθοδολογία επεκτάθηκε με τη μετατροπή των δεδομένων γονιδιακής έκφρασης σε βιολογικά μονοπάτια μέσω της εφαρμογής Gene Set Enrichment Analysis (GSEA). Σε αυτή την περίπτωση, τα αποτελέσματα ήταν ελαφρώς βελτιωμένα σε σύγκριση με την αρχική μεθοδολογία, αλλά συνολικά όχι αρκετά, ώστε να θεωρηθεί ότι έχει αναπτυχθεί ένας ισχυρός ταξινομητής. Λαμβάνοντας υπόψη την έντονη ετερογένεια της νόσου και των δεδομένων, η επόμενη μεθοδολογία επικεντρώθηκε στην προσπάθεια ταξινόμησης των σταδίων καρκίνου του μαστού σε ομογενή σύνολα ασθενών. Μεταξύ των χαρακτηριστικών αυτών είναι η ηλικιακή ομάδα, ο ιστολογικός τύπος, και η εθνικότητα των ασθενών. Ωστόσο, παρατηρήθηκε πως η μείωση της ετερογένειας των δεδομένων δεν βοήθησε ουσιαστικά στη διαδικασία ταξινόμησης. Αυτό συνεπάγεται πως είτε η ομογένεια που προκλήθηκε με τους

περιορισμούς μεταξύ των ασθενών δεν είναι επαρκής για τον διαχωρισμό των κλάσεων, είτε η δυσκολία του προβλήματος της ταξινόμησης σταδίων δεν οφείλεται αποκλειστικά στην ετερογένεια των ασθενών.

Το επόμενο βήμα, δεδομένης της μη ίσης κατανομής των δεδομένων, ήταν η χρήση τεχνικών για τη δημιουργία συνθετικών δεδομένων. Η εφαρμογή αυτών των τεχνικών, οι οποίες αντιμετωπίζουν ένα από τα βασικά προβλήματα του συνόλου δεδομένων, παρουσίασαν τα καλύτερα αποτελέσματα μέχρι στιγμής. Το ίδιο ισχύει και για την απόδοση του μοντέλου στο πρόβλημα δυαδικής ταξινόμησης μεταξύ πρώιμων και όψιμων δειγμάτων, όπου τα αποτελέσματα της μετρικής F1-score έφτασαν κοντά στο 57%.

Η τελευταία προσέγγιση που δοκιμάστηκε στο πλαίσιο αυτής της διπλωματικής εργασίας ήταν η ταξινόμηση μεταβάσεων μεταξύ σταδίων. Σε αυτή την περίπτωση, χρησιμοποιώντας αλγορίθμους μηχανικής μάθησης, δοκιμάστηκαν δύο μέθοδοι ταξινόμησης σταδίων. Πριν από την έναρξη της ταξινόμησης, πραγματοποιήθηκε κατάλληλη επεξεργασία δεδομένων. Το σύνολο δεδομένων δεν παρέχει πληροφορίες ίδιων ασθενών για την εξέλιξη της νόσου, δηλαδή δεν περιέχει δεδομένα γονιδιακής έκφρασης για του ασθενή για το Στάδιο *I* και *II*. Η καλύτερη δυνατή προσομοίωση της εξέλιξης της νόσου γίνεται συνδυάζοντας και συνενώνοντας τους υπάρχοντες ασθενείς με τέτοιον τρόπο, ώστε να μοιράζονται αρκετά κοινά χαρακτηριστικά, ώστε η μόνη τους διαφορά, όσο είναι δυνατό, να είναι το διαφορετικό στάδιο της νόσου. Από τον μετασχηματισμό του συνόλου δεδομένων και της συσχέτισης των ασθενών, υλοποιήθηκαν στην συνέχεια δύο διαφορετικοί μέθοδοι συνδυασμού των ασθενών.

Η μεθοδολογία που χρησιμοποιήθηκε περιλάμβανε αρχικά την ανάλυση των βιολογικών μονοπατιών των ασθενών, και στη συνέχεια τη χρήση των δεδομένων γονιδιακής έκφρασης. Τα δεδομένα αυτά αξιοποιήθηκαν τόσο στην πλήρη τους μορφή μετά την προεπεξεργασία, όσο και εστιάζοντας μόνο στην γονιδιακή έκφραση των 50 γονιδίων που ανήκουν στο PAM50. Από αυτήν την μεθοδολογία έγινε ξεκάθαρη η αδυναμία των PAM50 γονιδίων να αποτυπώσουν την διαφορά μεταξύ των μεταβάσεων. Αντίθετα, η ολοκληρωμένη γονιδιωματική προσέγγιση ενισχύει σημαντικά την ικανότητα των μοντέλων να ταξινομούν με μεγαλύτερη ακρίβεια τις μεταβάσεις σταδίου καρκίνου. Αυτό έχει σημαντικές επιπτώσεις για κλινικές εφαρμογές, υποδηλώνοντας ότι η ενσωμάτωση ευρύτερων γονιδιωματικών δεδομένων θα μπορούσε να οδηγήσει σε πιο ακριβή και αξιόπιστα διαγνωστικά εργαλεία.

Συνολικά, παρατηρείται ότι είναι πιο εύκολο τα δεδομένα γονιδιακής έκφρασης να ομαδοποιηθούν με βάση το υποείδος του καρκίνου του μαστού παρά με το στάδιο της νόσου, όπως επιβεβαιώνεται και από άλλες έρευνες, όπως αυτή των Robert Lesurf κ.α [65]. Η ταξινόμηση των σταδίων καρκίνου του μαστού με χρήση γονιδιακών δεδομένων αποδεικνύεται αρκετά δύσκολη. Η παρούσα διπλωματική εργασία αποτελεί μια εκτενή διερεύνηση του συγκεκριμένου πεδίου και των δυνατοτήτων που προσφέρουν αυτά τα δεδομένα στην ταξινόμηση των σταδίων, υποδεικνύοντας ταυτόχρονα τις προκλήσεις και τους περιορισμούς που πρέπει να αντιμετωπιστούν για την επίτευξη βελτιωμένων αποτελεσμάτων. Ωστόσο, η έρευνα σε αυτό το πεδίο παραμένει αρκετά ανοιχτή, παρέχοντας έτσι πολλές δυνατότητες για μελλοντικές επεκτάσεις και περαιτέρω διερεύνηση. Στη συνέχεια, παρουσιάζονται προτεινόμενες κατευθύνσεις για την εξέλιξη και εμπλουτισμό της μελέτης αυτής.

7.2 Προοπτικές - Μελλοντικές Προεκτάσεις

Μελλοντικές προεκτάσεις και διερευνήσεις που μπορούν να πραγματοποιηθούν γύρω από την θεματική και έρευνα της παρούσας διπλωματικής εργασίας είναι οι εξής:

- **Διερεύνηση και συμπερίληψη δειγμάτων κατηγορίας Σταδίου *IV* και φυσιολογικών ιστολογικών γονιδιακών δειγμάτων**:

  Κατά τη διάρκεια της παρούσας διερευνητικής εργασίας, τα δεδομένα που χρησιμοποιήθηκαν περιορίστηκαν στα δεδομένα γονιδιακής έκφρασης Σταδίου *I, II,* και *III* για λόγους ομοιογένειας και επαρκούς αναπαράστασης. Μια περαιτέρω προσθήκη στη μεθοδολογία που αναπτύχθηκε σχετίζεται με τη χρήση επιπλέον δειγμάτων στη μελέτη, τόσο από φυσιολογικό ιστό όσο και από ασθενείς με προχωρημένο στάδιο καρκίνου του μαστού (Στάδιο *IV*). Η συμπερίληψη αυτών των δεδομένων μπορεί να βελτιώσει σημαντικά την ικανότητα του μοντέλου να διακρίνει μεταξύ των διαφορετικών σταδίων της νόσου και να προσφέρει πιο σαφή συμπεράσματα σχετικά με τη γονιδιωματική εξέλιξη του καρκίνου και τα μοτίβα που προκύπτουν κατά την εξέλιξή του.

- **Χρήση γονιδιακών δεδομένων micro-RNA για την ταξινόμηση σταδίων καρκίνου του μαστού**:

  Σε αυτήν την διπλωματική εργασία διερευνήθηκε η ανάπτυξη ενός προγνωστικού μοντέλου, το οποίο αξιοποιεί γονιδιακά δεδομένα, τύπου mRNA, για την σταδιοποίηση του καρκίνου του μαστού. Ωστόσο, σύμφωνα με τους Abidalkareem κ.α [66] υπάρχει ένα σημαντικό κενό στην εφαρμογή προηγμένων τεχνικών μηχανικής μάθησης με τη χρήση των micro-RNA (miRNAs) γονιδιακών δεδομένων ως βιοδεικτών για την ταξινόμηση σταδίων του καρκίνου του μαστού. Παρά τον αναγνωρισμένο ρόλο των miRNAs στην παθογένεση και την εξέλιξη του καρκίνου [67, 68], η χρήση και η αναγνώριση των miRNAs δεδομένων ως βιοδεικτών μεταξύ των σταδίων του καρκίνου του μαστού παραμένει ανεξερεύνητη. Επομένως, μια πιθανή εξέλιξη της παρούσας εργασίας θα ήταν η δοκιμή των μεθόδων που αναπτύχθηκαν στο Κεφάλαιο 5 πλέον με την χρήση miRNA δεδομένων. Τα δεδομένα αυτά παρατίθενται επίσης δημόσια στο TCGA [59].

- **Αξιοποίηση Gene Embeddings και Transformers για την Βελτίωση της Ταξινόμησης Σταδίων Καρκίνου του Μαστού**:

  Μια επιπλέον μελλοντική επέκταση της προτεινόμενης μεθοδολογίας περιλαμβάνει τη χρήση high dimensional embeddings για semantic representation για δεδομένα γονιδιακής έκφρασης σε συνδυασμό με μοντέλα transformers. Η ιδέα των embeddings προέρχεται από τον τομέα της επεξεργασίας φυσικής γλώσσας (Natural Language Processing, NLP), όπου τεχνολογίες όπως το word2vec έχουν αποδείξει την αποτελεσματικότητά τους στην αναπαράσταση λέξεων σε έναν χώρο υψηλών διαστάσεων. Η επιτυχία των word2vec embeddings στον τομέα του NLP μπορεί να μεταφερθεί στη βιοπληροφορική για την αναπαράσταση γονιδίων.

  Στην παρούσα εργασία, η τεχνική των gene embeddings μπορεί να χρησιμοποιηθεί για να αναπαραστήσει τα γονίδια σε έναν πολυδιάστατο χώρο, όπου η εγγύτητα μεταξύ των

διανυσμάτων θα αντανακλά τη λειτουργική ή βιολογική τους σχέση. Υπάρχουν έρευνες που έχουν ασχοληθεί με την κατανεμημένη αναπαράσταση των γονιδίων βάσει της συνέκφρασης (co-expression) τους, όπως αυτή των Jincheng κ.α [69]. Η χρήση αυτών δεδομένων σε συνδυασμό με μοντέλα βαθιάς μάθησης transformers θα επιτρέψει την επεξεργασία αυτών των αναπαραστάσεων γονιδίων για την εξαγωγή προγνωστικών μοντέλων.

- **Χρήση Τεχνικών Μεταφοράς Μάθησης (Transfer Learning)**:

  Η εφαρμογή transfer learning αποτελεί μια σημαντική μελλοντική προέκταση για την επίλυση προβλημάτων στον τομέα της μηχανικής μάθησης και της βιοπληροφορικής. Η ιδέα βασίζεται στην αξιοποίηση των ήδη εκπαιδευμένων μοντέλων από άλλα παρόμοια προβλήματα ή pretext tasks, ώστε να βελτιωθεί η απόδοση του μοντέλου.

  Ένα πιθανό pretext task είναι η δημιουργία ενός μοντέλου, το οποίο προσπαθεί να προβλέψει ή να συμπληρώσει απουσιάζουσες τιμές γονιδιακής έκφρασης. Αυτή η περίπτωση θα μας επέτρεπε παράλληλα να αξιοποιήσουμε ένα ευρύ φάσμα δεδομένων, καθώς το μοντέλο θα μπορούσε να εκπαιδευτεί ανεξάρτητα από τον τύπου του καρκίνου και επομένως σε όλη την βάση δεδομένων του TCGA, καλύπτει πάνω από 33 διαφορετικούς τύπους καρκίνου και περιέχει χιλιάδες δείγματα.

  Αφού το μοντέλο εκπαιδευτεί με την προαναφερόμενη μέθοδο, στη συνέχεια μπορεί να γίνει περαιτέρω βελτίωση (fine tuning) στο συγκεκριμένο πρόβλημα ταξινόμησης σταδίων καρκίνου του μαστού. Αυτή η διαδικασία επιτρέπει στο μοντέλο να επωφεληθεί από τις γνώσεις που έχει αποκτήσει σε προηγούμενες φάσεις της εκπαίδευσης, βελτιώνοντας έτσι την απόδοσή του στο τελικό και συγκεκριμένο πρόβλημα.

- **Επέκταση μεθοδολογίας σε άλλα είδη καρκίνου**:

  Όπως αναφέρθηκε στην υποενότητα 2.1.2, ο καρκίνος του μαστού είναι μια εξαιρετικά ετερογενής νόσος, παρουσιάζοντας μεγαλύτερη ποικιλομορφία συγκριτικά με άλλους τύπους καρκίνου. Μια πιθανή κατεύθυνση για μελλοντική έρευνα είναι η εφαρμογή της μεθοδολογίας που παρουσιάστηκε στην ενότητα 5 σε άλλους τύπους καρκίνου, με σκοπό την αξιολόγηση των αποτελεσμάτων και τη διερεύνηση της αποτελεσματικότητας και της γενικότητάς της.

Βιβλιογραφία

[1] National Institutes of Health (US). *NIH Curriculum Supplement Series*. Biological Sciences Curriculum Study, 2007.

[2] American Cancer Society. *Ductal Carcinoma In Situ (DCIS)*. https://www.cancer.org/cancer/types/breast-cancer/about/types-of-breast-cancer/dcis.html.

[3] L Guo, D Kong, J Liu και et al. *Breast cancer heterogeneity and its implication in personalized precision therapy*. *Experimental Hematology Oncology*, 12:3, 2023.

[4] LGMD Frederick, David L Page, Irvin D Fleming, April G Fritz, Charles M Balch, Daniel G Haller, Monica Morrow και others. *AJCC cancer staging manual*. Springer Science & Business Media, 2002.

[5] M West, C Blanchette, H Dressman, E Huang, S Ishida, R Spang, H Zuzan, J. A Olson, J. R Jr, Marks και J. R. Nevins. *Predicting the clinical status of human breast cancer by using gene expression profiles*. *Proceedings of the National Academy of Sciences of the United States of America*, 98«20»:11462–11467, 2001.

[6] Dejun Zhang, Lu Zou, Xionghui Zhou και Fazhi He. *Integrating Feature Selection and Feature Extraction Methods With Deep Learning to Predict Clinical Outcome of Breast Cancer*. *IEEE Access*, 6:28936–28944, 2018.

[7] Y Xiao, J Wu, Z Lin και X Zhao. *A deep learning-based multi-model ensemble method for cancer prediction*. *Computer Methods and Programs in Biomedicine*, 153:1–9, 2018.

[8] Murtada K. Elbashir, Mohamed Ezz, Mohanad Mohammed και Said S. Saloum. *Lightweight Convolutional Neural Network for Breast Cancer Classification Using RNA-Seq Gene Expression Data*. *IEEE Access*, 7:185338–185348, 2019.

[9] M Mostavi, YC Chiu, Y Huang και et al. *Convolutional neural network models for cancer type prediction based on gene expression*. *BMC Med Genomics*, 13(5):44, 2020.

[10] World Health Organization. *Breast cancer*. https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

[11] INC National Breast Cancer Foundation. *Breast Cancer Facts & Stats*. https://www.nationalbreastcancer.org/breast-cancer-facts/.

[12] T Naito, Y και Urasaki. *Precision medicine in breast cancer*. *Chinese clinical oncology*, 7(3):29, 2018.

[13] E Aličković και A Subasi. *Breast cancer diagnosis using GA feature selection and Rotation Forest*. *Neural Computing & Application*, 28:753–764, 2017.

[14] American Cancer Society. *Survival Rates for Breast Cancer*. https://www.cancer.org/cancer/types/breast-cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-survival-rates.html.

[15] J Holmes, L Sacchi, R Bellazzi και others. *Artificial intelligence in medicine*. *Ann R Coll Surg Engl*, 86:334–8, 2004.

[16] D Ly, D Forman, J Ferlay, L. A Brinton και M. B Cook. *An international comparison of male and female breast cancer incidence rates*. *International journal of cancer*, 132(8):1918–1926, 2013.

[17] Johns Hopkins Medicing Pathology. *Overview of the Breast*. https://pathology.jhu.edu/breast/overview.

[18] National Cancer Institute. *SEER Training Modules, Breast Cancer*. https://training.seer.cancer.gov/breast/anatomy/.

[19] J Wang, B Li, M Luo και et al. *Progression from ductal carcinoma in situ to invasive breast cancer: molecular features and clinical significance*. *Signal Transduction and Targeted Therapy*, 9:83, 2024.

[20] M. C Braasch, A. L Amin, C. R Balanoff, J. L Wagner και K. E Larson. *Prognostic Significance of Lobular Carcinoma In-Situ (LCIS) Diagnosed Alongside Invasive Breast Cancer*. *Breast cancer : basic and clinical research*, 16, 2022.

[21] C Fumagalli και M Barberis. *Breast Cancer Heterogeneity*. *Diagnostics (Basel, Switzerland)*, 11(9):1555, 2021.

[22] National Human Genome Research Institute. https://www.genome.gov/genetics-glossary/Cell.

[23] National Cancer Institute. *Model-Based Small Area Estimates of Cancer-Related Knowledge*. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/cell, 2024. Ν"Ι Διςτιοναρψ οφ ἀνςερ Τερμς, ςελλ.

[24] M Cuffe, Stein, D Wilfred, Staehelin, Andrew L, Alberts, M Bruce, Bernfield, R Merton, Slack, M.W. Jonathan, Chow, Lodish Christopher, F Harvey, Cooper, John A., Laskey και Ronald A. *cell*. *Encyclopedia Britannica*, 2024.

[25] Alberts, B. *Molecular Biology of the Cell. 6th Edition*. Garland Science, Taylor and Francis Group, New York, 2015.

[26] Pokapū Akoranga Pūtaiao. *DNA, chromosomes and gene expression* . https://www.sciencelearn.org.nz/resources/206-dna-chromosomes-and-gene-expression. Σϛιενϛε Λεαρνινγ Ηυβ.

[27] National Cancer Institute. *biomarker*. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/biomarker.

[28] MJ Duffy, N Harbeck, M Nap, R Molina, A Nicolini, E Senkus και F Cardoso. *Clinical use of biomarkers in breast cancer: Updated guidelines from the European Group on Tumor Markers (EGTM)*. *Eur J Cancer. 2017 Apr;75:284-298. doi: 10.1016/j.ejca.2017.01.017. Epub 2017 Feb 28. PMID: 28259011*, 2017.

[29] Y Miki, J Swensen, D Shattuck-Eidens, P. A Futreal, K Harshman, S Tavtigian, Q Liu, C Cochran, L. M Bennett και W. Ding. *A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1*. *Science (New York, N.Y.)*, 266(5182):66–71, 1994.

[30] S Parker, J, M Mullins, M. C Cheang, S Leung, D Voduc, T Vickery, S Davies, C Fauron, X He, Z Hu, J. F Quackenbush, I. J Stijleman, J Palazzo, J. S Marron, A. B Nobel, E Mardis, T. O Nielsen, M. J Ellis, C. M Perou και P. S. Bernard. *Supervised risk predictor of breast cancer based on intrinsic subtypes*. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology, 27(8), 1160–1167*, 2009.

[31] B Wallden, J Storhoff, T Nielsen, N Dowidar, C Schaper, S Ferree, S Liu, S Leung, G Geiss, J Snider, T Vickery, S. R Davies, E. R Mardis, M Gnant, I Sestak, M. J Ellis, C. M Perou, P. S Bernard και J. S Parker. *Development and verification of the PAM50-based Prosigna breast cancer gene signature assay*. *BMC medical genomics, 8, 54*, 2015.

[32] K. B Johnson, W. Q Wei, D Weeraratne, M. E Frisse, K Misulis, K Rhee, J Zhao και J. L Snowdon. *Precision Medicine, AI, and the Future of Personalized Health Care*. *Clinical and translational science*, 14(1):86–93, 2021.

[33] J. E Lunshof, J Bobe, J Aach, M Angrist, J. V Thakuria, D. B Vorhaus, M. R Hoehe και G. M. Church. *Personal genomes in progress: from the human genome project to the personal genome project*. *Dialogues in clinical neuroscience*, 12(1):47–60, 2010.

[34] J Xu, P Yang, S Xue και et al. *Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives*. *Hum Genet*, 138:109–124, 2019.

[35] Malihe Ram, Ali Najafi και Mohammad Taghi Shakeri. *Classification and Biomarker Genes Selection for Cancer Gene Expression Data Using Random Forest*. *Iranian Journal of Pathology*, 12(4):339–347, 2017.

[36] Y Zhang, Q Deng, W Liang και X. Zou. *An Efficient Feature Selection Strategy Based on Multiple Support Vector Machine Technology with Gene Expression Data*. *BioMed research international*, 2018.

[37] Y Wang, JGM Klijn, Y Zhang, AM Sieuwerts, MP Look, F Yang, D Talantov, M Timmermans, MEMeijer van Gelder και J Yu. *Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer*. *Lancet*, 365(9460):671–679, 2005.

[38] LJvan't Veer, H Dai, MJvan de Vijver, YD He, AAM Hart, M Mao, HL Peterse, Kvan der Kooy, MJ Marton, AT Witteveen και et al. *Gene expression profiling predicts clinical outcome of breast cancer*. *Nature Science*, 415(6871):530–536, 2002.

[39] H. Y Chuang, E Lee, Y. T Liu, D Lee και T Ideker. *Network-based classification of breast cancer metastasis*. *Molecular systems biology*, 3(140), 2007.

[40] A Subramanian, P Tamayo, V. K Mootha, S Mukherjee, B. L Ebert, M. A Gillette, A Paulovich, S. L Pomeroy, T. R Golub, E. S Lander και J. P Mesirov. *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles*. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.

[41] S Kim, M Kon και C. DeLisi. *Pathway-based classification of cancer subtypes*. *Biology direct*, 7:21, 2012.

[42] S Roy, R Kumar, V Mittal και et al. *Classification models for Invasive Ductal Carcinoma Progression, based on gene expression data-trained supervised machine learning*. *Scientific Reports, Nature*, 10:4113, 2020.

[43] F Gao, W Wang, M Tan και et al. *DeepCC: a novel deep learning-based framework for cancer molecular subtype classification*. *Oncogenesis*, 8:44, 2019.

[44] A. L. Samuel. *Some studies in machine learning using the game of checkers*. *IBM J. Res. Dev.*, 3(3):210–229, 1959.

[45] S. B Kotsiantis. *Supervised Machine Learning: A Review of Classification Techniques*. *Informatica*, 31:249–268, 2007.

[46] H. B Barlow. *Unsupervised Learning*. *Neural Computation*, 1(3):295–311, 1989.

[47] Leslie Pack Kaelbling, Michael L Littman και Andrew W Moore. *Reinforcement learning: A survey*. *Journal of artificial intelligence research*, 4:237–285, 1996.

[48] Lior Rokach και Oded Maimon. *Decision trees*. *Data mining and knowledge discovery handbook*, σελίδες 165–192, 2005.

[49] L Breiman. *Random forests*. *Machine Learning*, 45(1):5–32, 2001.

[50] J. H. Friedman. *Greedy function approximation: A gradient boosting machine*. *Annals of Statistics*, 29:5, 1999.

[51] Friedman και H Jerome. *Greedy function approximation: a gradient boosting machine*. *Annals of statistics*, σελίδες 1189–1232, 2001.

[52] S. S Haykin. *Neural networks and learning machines*. Upper Saddle River, NJ: Pearson Education, 2009.

[53] C. Y. J Peng, K. L Lee και G. M. Ingersoll. *An Introduction to Logistic Regression Analysis and Reporting*. *The Journal of Educational Research*, 96(1):3–14, 2002.

[54] Y Sun, J Yao, N.J Nowak και et al. *Cancer progression modeling using static sample data*. *Genome Biol*, σελίδα 440, 2014.

[55] J. Brownlee. *Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning*. Machine Learning Mastery, 2020.

[56] I.H Sarker. *Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions*. *SN Computer Science*, 2:420, 2021.

[57] Michael A Nielsen. *Neural networks and deep learning*, τόμος 25. Determination press San Francisco, CA, USA, 2015.

[58] Frank Rosenblatt. *The perceptron: a probabilistic model for information storage and organization in the brain*. *Psychological review*, 1958. 65(6):386.

[59] K Tomczak, P Czerwińska και M. Wiznerowicz. *The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge*. *Contemporary oncology (Poznan, Poland)*, 2015. 19(1A), A68–A77.

[60] M Kamińska, T Ciszewski, KŁopacka Szatan, P Miotła και E Starosławska. *Breast cancer risk factors*. *Menopause review*, 14(3):196–202, 2015.

[61] K.A Hirko, G Rocque, E Reasor και et al. *The impact of race and ethnicity in breast cancer—disparities and implications for precision oncology*. *BMC Med*, 20:72, 2022.

[62] B Wallden, J Storhoff, T Nielsen και et al. *Development and verification of the PAM50-based Prosigna breast cancer gene signature assay*. *BMC Med Genomics*, 8:54, 2015.

[63] M Dalamaga. *Obesity, Insulin Resistance, Adipocytokines and Breast Cancer: New Biomarkers and Attractive Therapeutic Targets*. *World journal of experimental medicine*, 3:34–42, 2013.

[64] Keith Baverstock. *The gene: An appraisal*. *Progress in Biophysics and Molecular Biology*, 164:46–62, 2021.

[65] R Lesurf, M. R Aure, H. H Mørk και V. Vitelli. *Molecular Features of Subtype-Specific Progression from Ductal Carcinoma In Situ to Invasive Breast Cancer*. *Cell Reports*, 16(4):1166–1179, 2016.

[66] A Abidalkareem, AK Ibrahim, M Abd, O Rehman και H Zhuang. *Identification of Gene Expression in Different Stages of Breast Cancer with Machine Learning*. *Cancers*, 16(10):1864, 2024.

[67] J Lu, G Getz, E Miska και et al. *MicroRNA expression profiles classify human cancers*. *Nature*, 435:834–838, 2005.

[68] T Sørlie, C.M Perou, R Tibshirani, T Aas, S Geisler, H Johnsen, T Hastie, M.B Eisen, M Van De Rijn, S.S Jeffrey και et al. *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications*. *Proc. Natl. Acad. Sci. USA*, 98:10869–10874, 2001.

[69] J Du, P Jia, Y Dai και et al. *Gene2vec: distributed representation of genes based on co-expression*. *BMC Genomics*, 20(1):82, 2019.