ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Ανάπτυξη μοντέλου υποστήριξης της ανίχνευσης του καρκίνου του μαστού από εξετάσεις αίματος, βασισμένο σε τεχνικές συλλογικής μάθησης.

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΒΑΣΙΛΕΙΟΥ ΕΛΕΥΘΕΡΙΑΔΗ

Επιβλέπουσα : Κωνσταντίνα Σ. Νικήτα
Καθηγήτρια Ε.Μ.Π.

Συνεπιβλέπουσα : Κωνσταντία Ζαρκογιάννη
Εργαστηριακό Διδακτικό Προσωπικό Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2020

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Ανάπτυξη μοντέλου υποστήριξης της ανίχνευσης του καρκίνου του μαστού από εξετάσεις αίματος, βασισμένο σε τεχνικές συλλογικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΒΑΣΙΛΕΙΟΥ ΕΛΕΥΘΕΡΙΑΔΗ

Επιβλέπουσα : Κωνσταντίνα Σ. Νικήτα
Καθηγήτρια Ε.Μ.Π.

Συνεπιβλέπουσα : Κωνσταντία Ζαρκογιάννη
Εργαστηριακό Διδακτικό Προσωπικό Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την .....

…….....	..…….....	.....……
Κωνσταντίνα Σ. Νικήτα	Ανδρέας Γεώργιος Σταφυλοπάτης	Γεώργιος Στάμου
Καθηγήτρια Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Αναπ. Καθηγητής Ε.Μ.Π

Αθήνα, Φεβρουάριος 2020

.....….

Βασίλειος Ελευθεριάδης

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών

Ευχαριστίες

Θα ήθελα να ευχαριστήσω την επιβλέπουσα καθηγήτρια της διπλωματικής μου εργασίας κα. Κωνσταντίνα Νικήτα για την ευκαιρία που μου έδωσε να εργαστώ σε ένα θέμα το οποίο με ενδιέφερε ιδιαίτερα και με οδήγησε σε συνεχή αναζήτηση, από την πρώτη μέχρι και την τελευταία στιγμή της εκπόνησης. Θα ήθελα, επίσης, να ευχαριστήσω την κα. Κωνσταντία Ζαρκογιάννη, η οποία συμμετείχε στην επίβλεψη της εργασίας, για την καθοδήγηση, τις συμβουλές και τις παρατηρήσεις της, ωθώντας με σε βελτίωση των αποτελεσμάτων και της ποιότητας της εργασίας μου μετά από κάθε μας συνάντηση.

Περίληψη

Στην παρούσα διπλωματική εργασία παρουσιάζεται ένα μοντέλο ανίχνευσης του καρκίνου του μαστού εφαρμόζοντας τεχνικές μηχανικής μάθησης σε αποτελέσματα αιματολογικών εξετάσεων και ανθρωπομετρικά δεδομένα.

Για την ανάπτυξη και την αξιολόγηση του μοντέλου, χρησιμοποιήθηκε η βάση δεδομένων Coimbra Breast Cancer Dataset, η οποία παραχωρείται μέσω της δημοσίως διαθέσιμης βιβλιοθήκης UCI Machine Learning Repository. Τα δεδομένα περιλαμβάνουν καταγραφές από 116 γυναίκες, για κάθε μία από τις οποίες πραγματοποιήθηκαν αιματολογικές εξετάσεις των επιπέδων γλυκόζης, ινσουλίνης, λεπτίνης, αντιπονεκτίνης, ρεζιστίνης και MCP-1, καθώς και σωματομετρικών δεδομένων όπως η ηλικία και ο δείκτης μάζας σώματος (BMI). Από τις 116 συμμετέχοντες, οι 64 είχαν πρόσφατα διαγνωστεί με καρκίνο του μαστού, ενώ οι 52 αποτέλεσαν την υγιή ομάδα ελέγχου.

Για την ανάπτυξη του μοντέλου διερευνήθηκε η χρήση διαφόρων αλγορίθμων μηχανικής μάθησης (SVM, Logistic Regression, Random Forest, XGBoost, Naive Bayes), ενώ πραγματοποιήθηκε εκτενής αναζήτηση βέλτιστου συνδυασμού βιοδεικτών που αντιστοιχούν σε κάθε αλγόριθμο. Για την επίτευξη υψηλής ακρίβειας διερευνήθηκε η χρήση μεθόδων συλλογικής μάθησης, συνδυάζοντας την έξοδο των αρχικών μοντέλων. Το μοντέλο με την βέλτιστη απόδοση βασίστηκε στη συνδυασμένη χρήση των μοντέλων XGBoost, Random Forest, Support Vector Machines και KNN μέσω σταθμισμένης ψηφοφορίας, επιτυγχάνοντας AUC: 0.928 , Ακρίβεια: 92.24%, Ευαισθησία: 93.75% και Ειδικότητα: 90.38%, F1-score 0.930 και Precision 0.923.

Σημαντικοί Όροι: ΚΑΡΚΙΝΟΣ ΜΑΣΤΟΥ, ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ, ΑΝΙΧΝΕΥΣΗ, ΑΙΜΑΤΟΛΟΓΙΚΕΣ ΕΞΕΤΑΣΕΙΣ, ΣΥΛΛΟΓΙΚΗ ΜΑΘΗΣΗ.

Abstract

The objective of the present thesis is the design, development and evaluation of a breast cancer detection model through applying machine learning techniques on data from blood analysis and anthropometric data.

For the development and evaluation of the model the Coimbra Breast Cancer Dataset has been used, which is publicly available through the UCI Machine Learning Repository. This dataset consists of 116 instances. Out of the 116 women participants, 64 had been diagnosed with breast cancer . For each participant the values of serum Glucose, Leptin, Adiponectin, Resistin and Chemokine Monocyte Chemoattractant Protein 1 (MCP-1), and levels of plasma Insulin were assessed by blood analysis. The Homeostasis Model Assessment (HOMA) index and the Leptin/Antiponectin ratios were calculated. Moreover, anthropometric data such as age and Body Mass Index (BMI) were recorded for each participant.

For the development of the model, the use of several machine learning models (SVM, Logistic Regression, Random Forest, XGBoost, Naive Bayes) has been investigated, while extended search of the optimal combination of biomarkers for each model has been conducted. Aiming at achieving high levels of accuracy, ensemble learning techniques have been applied in order to combine the outputs of the primary models towards producing the final decision. The model which achieved the highest accuracy was based on the combined use of the XGBoost, Random Forest, Support Vector Machines and KNN models and the application of the Dynamic Weighted Voting based on Accuracy on the outputs of the primary models. The obtained results demonstrate the effectiveness of the model (AUC : 0.928 , Accuracy : 92.24% , Sensitivity : 93.75%, Specificity 90.38%, F1-score 0.930 and Precision 0.923).

KEYWORDS : BREAST CANCER,DETECTION, MACHINE LEARNING, BLOOD ANALYSIS, ENSEMBLE LEARNING.

Πίνακας περιεχομένων

Υπότιτλοι Εικόνων

Υπέρτιτλοι Πινάκων

Κεφάλαιο 1

Καρκίνος του μαστού

1.1 Ιατρικό υπόβαθρο

1.1.1 Εισαγωγή

Ο καρκίνος χαρακτηρίζεται από τη μη φυσιολογική κυτταρική ανάπτυξη και διαίρεση και μπορεί να εμφανιστεί οπουδήποτε στο ανθρώπινο σώμα. Υπό φυσιολογικές συνθήκες τα ανθρώπινα κύτταρα αναπτύσσονται και διαιρούνται ούτως ώστε να δημιουργήσουν νέα κύτταρα όταν και όπου το σώμα τα χρειάζεται. Όταν τα κύτταρα γεράσουν ή υποστούν ζημιά πεθαίνουν και εν συνεχεία νέα κύτταρα τα αντικαθιστούν. Κατά την ανάπτυξη όμως του καρκίνου αυτή η φυσιολογική διαδικασία παύει να υφίσταται.

Λόγο μεταλλάξεων στο γενετικό τους υλικό, τα καρκινικά κύτταρα διαφέρουν από τα φυσιολογικά κύτταρα με τρόπους που τους επιτρέπουν να αναπτύσσονται ανεξέλεγκτα και να γίνονται επεμβατικά. Μία σημαντική διαφορά είναι ότι τα καρκινικά κύτταρα είναι λιγότερο εξειδικευμένα από τα κανονικά. Δηλαδή, σε αντίθεση με τα φυσιολογικά κύτταρα, τα καρκινικά δεν ωριμάζουν σε διαφορετικούς κυτταρικούς τύπους. Αυτό προκαλεί τα καρκινικά κύτταρα να αγνοούν τα σήματα του οργανισμού τα οποία σηματοδοτούν την λήξη της διαδικασίας διαίρεσης, με αποτέλεσμα να διαιρούνται ασταμάτητα, σε αντίθεση με τα φυσιολογικά τα οποία δεχόμενα το σήμα τερματισμού διαίρεσης από τον οργανισμό και σταματούν την διαδικασία διαίρεσή τους όταν κύτταρα που ανήκουν στον συγκεκριμένο τύπο έχουν αναπληρωθεί. Επιπροσθέτως, τα καρκινικά κύτταρα αγνοούν ένα ακόμα σήμα του οργανισμού, αυτό της έναρξης της απόπτωσης, της διαδικασίας δηλαδή του προγραμματισμένου κυτταρικού θανάτου, την οποία χρησιμοποιεί ο οργανισμός για να εξαλείψει γηρασμένα, κατεστραμμένα ή μολυσμένα κύτταρα.

Η πλειονότητα των καρκίνων προκαλείται από μια γενετική μετάλλαξη σε ένα γονίδιο που βοηθά στον έλεγχο της ανάπτυξης και διαίρεσης των κυττάρων. Η ανωμαλία αυτή μπορεί να επηρεάσει ένα μόνο κύτταρο αρχικά. Στη συνέχεια, καθώς το κύτταρο διαιρείται ανεξέλεγκτα, μια ολόκληρη μάζα ανώμαλων κυττάρων σύντομα αναπτύσσεται, σχηματίζοντας ένα μόρφωμα γνωστό ως κακοήθη «όγκο» ή «νεόπλασμα». Ο όγκος γίνεται

προοδευτικά μεγαλύτερος και μπορεί να εισβάλει σε γειτονικούς ιστούς, καταστρέφοντας υγιείς ιστούς και όργανα.

Κάθε καρκίνος έχει έναν μοναδικό συνδυασμό γενετικών μεταλλάξεων. Καθώς ο όγκος συνεχίζει να μεγαλώνει ακόμη περισσότερες μεταλλάξεις θα εμφανιστούν. Ακόμα και στον ίδιο όγκο, διαφορετικά κύτταρα μπορεί να έχουν διαφορετικές γενετικές μεταβολές. Γενικά, τα καρκινικά κύτταρα έχουν περισσότερες μεταλλάξεις του γενετικού υλικού από τα υγιή κύτταρα. Ο μεταλλάξεις αυτές πολύ συχνά μπορεί να είναι προϊόν του καρκίνου και όχι η αιτία εμφάνισής του.

Τα καρκινικά κύτταρα μπορούν επίσης να εγκαταλείψουν τον όγκο και να μεταφερθούν σε άλλες περιοχές του σώματος. Το φαινόμενο αυτό είναι γνωστό ως μετάσταση, και οι καρκίνοι που προκύπτουν με αυτό τον τρόπο ονομάζονται μεταστατικοί ώστε να διαφοροποιούνται από τον πρωτογενή καρκίνο. Οι πιο κοινές περιοχές όπου αναπτύσσονται μεταστάσεις είναι οι πνεύμονες, το ήπαρ, οι λεμφαδένες, τα οστά και ο εγκέφαλος.

Ένας όγκος μπορεί να είναι καλοήθης ή κακοήθης. Οι καλοήθεις όγκοι δεν θεωρούνται επικίνδυνοι για την υγεία. Τα κύτταρα ενός καλοήθους όγκου είναι αρκετά όμοια στην εμφάνιση με τα φυσιολογικά, η ανάπτυξή τους είναι αργή και δεν εισβάλουν σε γειτονικούς ιστούς ή μεταφέρονται σε άλλες περιοχές του σώματος. Οι κακοήθεις όγκοι είναι καρκινικοί και μπορεί να είναι επικίνδυνοι. Εάν δεν ελεγχθούν εγκαίρως, τα κακοήθη κύτταρα μπορεί τελικά να εξαπλωθούν πέραν του πρωτογενούς όγκου σε γειτονικούς ιστούς και να μεταφερθούν σε άλλες περιοχές του σώματος. Σε αυτή την περίπτωση ο καρκίνος ονομάζεται διηθητικός.

1.1.2 Η Σταδιοποίηση του Καρκίνου

Η πρόγνωση για έναν διηθητικό καρκίνο επηρεάζεται σε μεγάλο βαθμό από το στάδιο στο οποίο βρίσκεται η νόσος, δηλαδή από την έκταση της εξάπλωσης του καρκίνου κατά την διάρκεια της πρώτης διάγνωσης. Υπάρχουν δύο κύρια συστήματα σταδιοποίησης του καρκίνου.

1.1.2.1 Το σύστημα σταδιοποίησης TNM

Το σύστημα σταδιοποίησης TNM εκφράζει την ανατομική έκταση της νόσου και βασίζεται στον προσδιορισμό τριών συνιστωσών :

1. **T**-Tumor την έκταση του πρωτοπαθούς όγκου
2. **N**-Nodes την απουσία ή παρουσία λεμφαδενικής προσβολής και τον βαθμό προσβολής των περιοχικών λεμφαδένων
3. **M**-Metastasis την απουσία ή παρουσία απομακρυσμένων μεταστάσεων.

Σε κάθε μία από τις παραπάνω κατηγορίες αποδίδεται ένα αριθμητικό πρόθεμα το οποίο εκφράζει την έκταση της νόσου για την συγκεκριμένη κατηγορία. Αφού οι συνιστώσες T, N και M έχουν βαθμονομηθεί, καθορίζεται και το στάδιο (stage) της νόσου το οποίο βαθμονομείται ως 0, I, II, III, ή IV, με το στάδιο 0 να χαρακτηρίζει καρκίνωμα το οποίο είναι μη διηθητικό (in situ), το στάδιο I διηθητικό καρκίνωμα σε αρχικό στάδιο, και το στάδιο IV να χαρακτηρίζει το πιο προχωρημένο στάδιο της νόσου, με ύπαρξη μεταστάσεων σε απομακρυσμένα όργανα (Πίνακας 1.1).[15]

Πίνακας 1.1 Αντιστοίχηση Σταδίου του καρκίνου του μαστού και των βαθμών των συνιστωσών του συστήματος σταδιοποίησης TNM. [44]

When T is…	And N is…	And M is…	Then the stage group is..
Tis	N0	M0	0
T1	N0	M0	IA
T0	N1mi	M0	IB
T1	N1mi	M0	IB
T0	N1	M0	IIA
T1	N1	M0	IIA
T2	N0	M0	IIA
T2	N1	M0	IIB
T3	N0	M0	IIB
T0	N2	M0	IIIA
T1	N2	M0	IIIA
T2	N2	M0	IIIA
T3	N1	M0	IIIA
T3	N2	M0	IIIA
T4	N0	M0	IIIB
T4	N1	M0	IIIB
T4	N2	M0	IIIB
Any T	N3	M0	IIIC
Any T	Any N	M1	IV

1.1.2.2 Σταδιοποίηση βάση βαθμού διαφοροποίησης

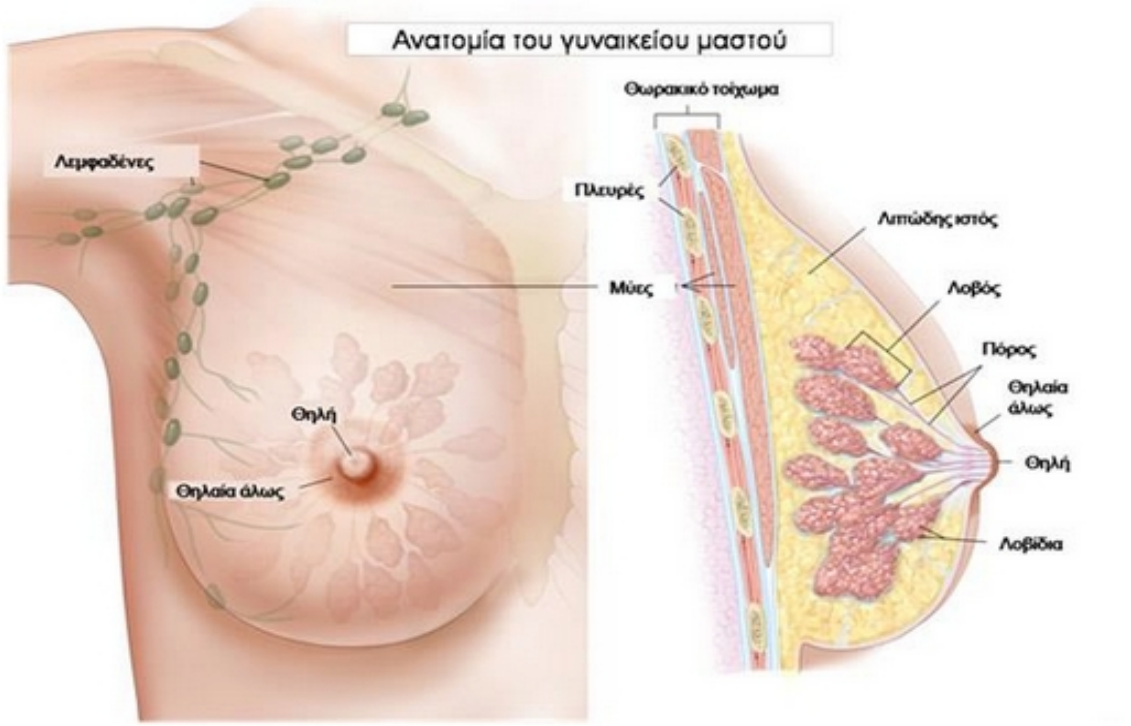
Ένα άλλο σύστημα σταδιοποίησης του καρκίνου είναι ο βαθμός (grade) διαφοροποίησης και αναφέρεται στην μικροσκοπική παθολογοανατομική εικόνα του όγκου δηλώνοντας το βαθμό διαφοροποίησης των νεοπλασματικών κυττάρων και την ομοιότητά τους ως προς τα υγιεί

κύτταρα του ιστού προέλευσης, για παράδειγμα ως προς την μορφολογία του πυρήνα ή την ιστολογική αρχιτεκτονική. Ο υγιείς ιστός συνήθως αποτελείται από πολλούς διαφορετικούς τύπους κυττάρων. Εάν ο καρκίνος μοιάζει με τον υγιεί ιστό και αποτελείται από ομάδες διαφορετικών τύπων κυττάρων και παρουσιάζει μικρή συχνότητα μιτώσεων, τότε χαρακτηρίζεται ως G1 και συχνά αναφέρεται ως Διαφοροποιημένος ή Χαμηλού βαθμού όγκος. Εάν ο καρκινικός ιστός δείχνει πολύ διαφορετικός από τον υγιεί ιστό και παρουσιάζει μεγάλη συχνότητα μιτώσεων χαρακτηρίζεται ως G4 και συχνά αναφέρεται ως Αδιαφοροποίητος ή Υψηλού βαθμού όγκος. Ο βαθμός διαφοροποίησης του καρκίνου είναι ο δείκτης ο οποίος βοηθάει στην πρόβλεψη της ταχύτητας εξάπλωσης του καρκίνου και βαθμονομείται από το 1 έως το 4. Σαν γενικό κανόνα θα μπορούσαμε να πούμε ότι όσο χαμηλότερος είναι ο βαθμός, τόσο καλύτερη είναι η πρόγνωση. Διαφορετικοί τύποι καρκίνου έχουν διαφορετικές μεθόδους ανάθεσης βαθμού διαφοροποίησης.[4], [6]

1.1.3 Καρκίνος του Μαστού

1.1.3.1 Εισαγωγή

Ο όρος ‘καρκίνος του μαστού’ αναφέρεται σε κακοήθη όγκο ο οποίος έχει αναπτυχθεί από κύτταρα εντός του μαστού. Ο καρκίνος του μαστού συνηθέστερα θα ξεκινήσει είτε από τα κύτταρα των γαλακτοπαραγωγικών αδένων, λοβών, ή από τους γαλακτοφόρους πόρους μέσω των οποίων μεταφέρεται το γάλα από τους λοβούς στην θηλή. Ο αυλός των γαλακτοφόρων πόρων και οι κυψελοειδείς χώροι των λοβών του μαστού επαλείφονται από επιθηλιακά κύτταρα. Ο καρκίνος του μαστού που δημιουργείται στα επιθηλιακά κύτταρα των λοβών και των πόρων αποτελεί το 90% των κακοηθών όγκων που εμφανίζονται στον μαστό και ονομάζονται λοβιακό και πορογενές καρκίνωμα του μαστού αντίστοιχα. Πιο σπάνια, ο καρκίνος στο μαστό μπορεί να ξεκινήσει από τους στρωματικούς ιστούς, δηλαδή τους λιπαρούς και ινώδης συνδετικούς ιστούς του μαστού (Εικόνα 1.1).[8]-[10]



Εικόνα 1.1 Ανατομία του γυναικείου μαστού. [7]

1.1.3.2 Σταδιοποίηση του Καρκίνου του μαστού

Η δυναμική ενός καρκίνου του μαστού μπορεί να περιγραφεί από τα ακόλουθα 5 στάδια : [4]-[5], [11]-[15]

- **Στάδιο 0**. Αυτό το στάδιο περιγράφει τον επιτόπιο καρκίνο (in situ). Οι κακοήθεις όγκοι σταδίου 0 εξακολουθούν να εντοπίζονται στην περιοχή του πρωτογενούς όγκου και δεν έχουν εξαπλωθεί σε γειτονικούς ιστούς.

  Οι 3 πιο πιθανοί τύποι καρκινώματος σταδίου 0 είναι:

  - Μη διηθητικό πορογενές καρκίνωμα του μαστού (DCIS)
  - Μη διηθητικό λοβιακό καρκίνωμα του μαστού (LCIS)
  - Νόσος Paget στο δέρμα της θηλής του στήθους

  Αυτό το στάδιο καρκίνου είναι μη διηθητικό και είναι εξαιρετικά ιάσιμο μέσω χειρουργικής αφαίρεσης του όγκου ή μαστεκτομής, συχνά χωρίς την ανάγκη ακτινοθεραπείας. Σε περίπτωση όμως που δεν ανιχνευθεί και αντιμετωπιστεί εγκαίρως μπορεί να εξαπλωθεί στους περιβάλλοντες ιστούς .

- **Στάδιο Ι**. Το στάδιο Ι περιγράφει διηθητικό καρκίνο ο οποίος περιορίζεται στο στήθος ή η ανάπτυξή του έχει εξαπλωθεί μόνο στους διπλανούς λεμφαδένες. Συχνά

αναφέρεται ως πρώιμου σταδίου ή εντοπισμένος καρκίνος και χωρίζεται σε δύο υποκατηγορίες :

- **Στάδιο ΙΑ**. Ο κακοήθης όγκος έχει διάμετρο έως 2 εκατοστά (cm), ο καρκίνος δεν έχει επεκταθεί πολύ σε γειτονικούς ιστούς, παραμένει εντός των ορίων του μαστού και δεν έχει εξαπλωθεί στους λεμφαδένες.
- **Στάδιο ΙΒ**. Στο στάδιο ΙΒ έχουμε δύο υποκατηγορίες. Στην πρώτη, δεν υπάρχει κακοήθης όγκος στον μαστό, αλλά υπάρχουν μικρές ομάδες καρκινικών κυττάρων στους λεμφαδένες (μικρομεταστάσεις), με διάμετρο το πολύ 2 χιλιοστά (mm). Στην δεύτερη, υπάρχει κακοήθης όγκος στον μαστό με διάμετρο έως 2 εκατοστά (cm) ο οποίος δεν έχει επεκταθεί πολύ σε γειτονικούς ιστούς και υπάρχουν μικρές ομάδες καρκινικών κυττάρων στους λεμφαδένες (μικρομεταστάσεις), με διάμετρο το πολύ 2 χιλιοστά (mm).

Στον καρκίνο σταδίου Ι η επέκταση των καρκινικών κυττάρων πέραν των ορίων του λοβού ή πόρου δεν ξεπερνά το 1 χιλιοστό (mm).

Ο καρκίνος σταδίου Ι μπορεί να θεραπευθεί αποτελεσματικά, αλλά χρειάζεται άμεση θεραπεία, συνήθως χειρουργική αφαίρεση του όγκου και των επηρεασμένων λεμφαδένων ή ακτινοβολία ή και τα δύο. Αναλόγως της κατάστασης και των παραγόντων κινδύνου μπορεί να εφαρμοστεί και ορμονοθεραπεία.

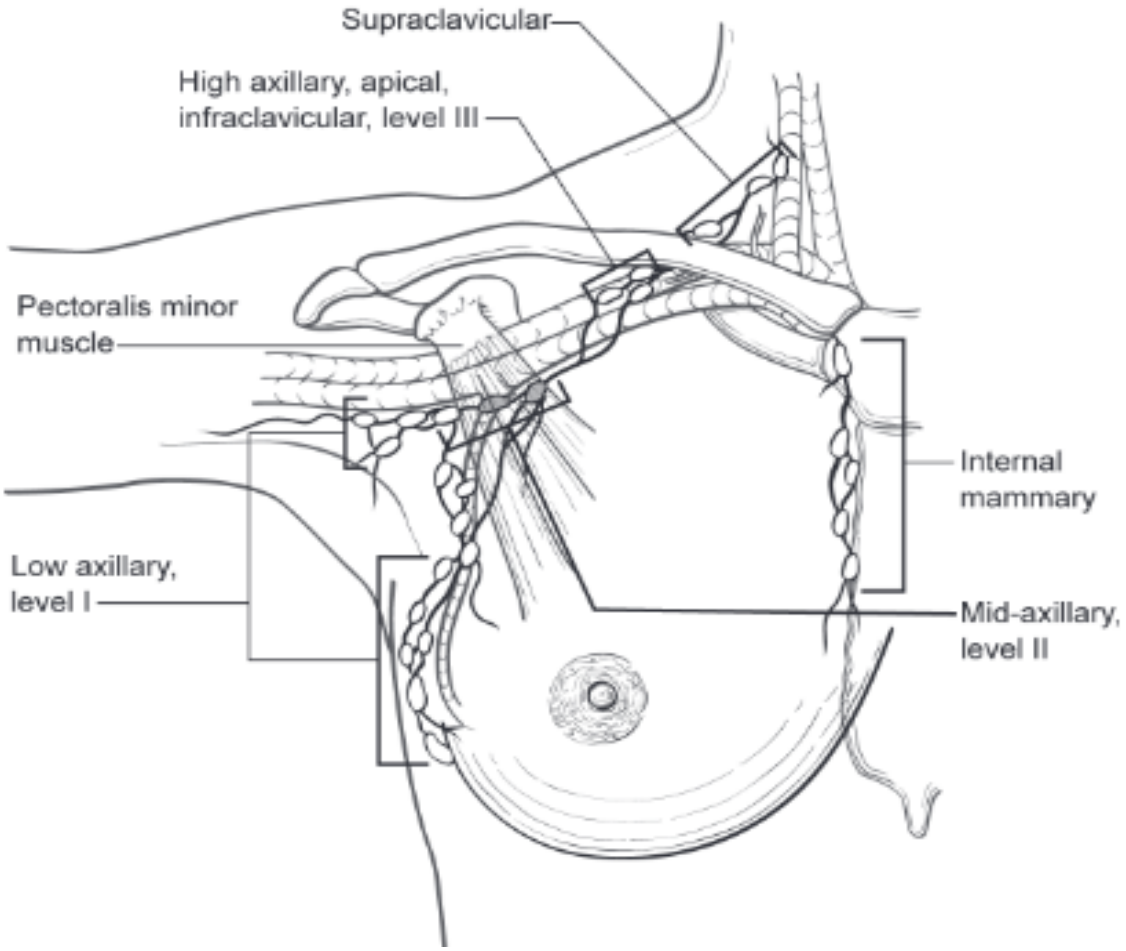
- **Στάδιο ΙΙ**. Το στάδιο ΙΙ περιγράφει διηθητικό καρκίνο ο οποίος αυξάνεται μεν, αλλά περιορίζεται στην περιοχή του μαστού ή η ανάπτυξή του έχει εξαπλωθεί μόνο στους διπλανούς λεμφαδένες και χωρίζεται σε δύο υποκατηγορίες :
  - **Στάδιο ΙΙΑ**. Όταν ισχύει μία από τις ακόλουθες περιγραφές :
    - Απουσία κακοήθους όγκου στον μαστό ή παρουσία όγκου με διάμετρο έως 2 εκατοστά (cm) και ύπαρξη ομάδων καρκινικών κυττάρων με διάμετρο από 0,2 έως 2 χιλιοστά (mm), σε 1 έως 3 μασχαλιαίους λεμφαδένες (low or mid axillary lymph nodes, Εικόνα 1.2) ή στους εσωτερικούς μαστικούς λεμφαδένες (internal mammary lymph nodes, Εικόνα 1.2) ή σε συνδυασμό των δύο.
    - Ο κακοήθης όγκος έχει διάμετρο από 2 έως 5 εκατοστά (cm) και δεν έχει εξαπλωθεί στους λεμφαδένες.

- **Στάδιο ΙΙΒ**. Όταν ισχύει μία από τις ακόλουθες περιγραφές :
  - Υπάρχει κακοήθης όγκος στον μαστό με διάμετρο από 2 έως 5 εκατοστά (cm) και ύπαρξη ομάδων καρκινικών κυττάρων με διάμετρο από 0,2 έως 2 χιλιοστά (mm), σε 1 έως 3 μασχαλιαίους λεμφαδένες ή στους εσωτερικούς μαστικούς λεμφαδένες ή σε συνδυασμό τους.
  - Ο κακοήθης όγκος έχει διάμετρο μεγαλύτερη από 5 εκατοστά (cm) και δεν έχει εξαπλωθεί στους λεμφαδένες.

Ο καρκίνος στο στάδιο ΙΙ ανταποκρίνεται καλά στην θεραπεία και μπορεί να θεραπευθεί αποτελεσματικά. Η θεραπεία που εφαρμόζεται σε αυτές τις περιπτώσεις συνήθως περιλαμβάνει χειρουργική αφαίρεση του όγκου σε συνδυασμό με ακτινοθεραπεία ή μαστεκτομή με ή χωρίς εφαρμογή ακτινοθεραπείας, χειρουργική αφαίρεση των επηρεασμένων λεμφαδένων με πιθανότητες ανάγκης ακτινοθεραπείας στους λεμφαδένες, χημειοθεραπεία, ορμονοθεραπεία ανάλογα με την περίπτωση ή συνδυασμούς των παραπάνω.

- **Στάδιο ΙΙΙ**. Το στάδιο ΙΙ περιγράφει διηθητικό καρκίνο που έχει επεκταθεί από την αρχική περιοχή του όγκου και μπορεί να εισέβαλε στους λεμφαδένες και στους μύες, ωστόσο δεν έχει επεκταθεί σε παρακείμενα όργανα και χωρίζεται σε τρεις υποκατηγορίες :
  - **Στάδιο ΙΙΙΑ**. Όταν ισχύει μία από τις ακόλουθες περιγραφές :
    - Απουσία κακοήθους όγκου στον μαστό ή παρουσία όγκου με διάμετρο έως 5 εκατοστά (cm) και ύπαρξη ομάδων καρκινικών κυττάρων σε 4 έως 9 μασχαλιαίους λεμφαδένες ή στους εσωτερικούς μαστικούς λεμφαδένες, αλλά όχι σε συνδυασμό τους.
    - Υπάρχει κακοήθης όγκος στον μαστό με διάμετρο πάνω 5 εκατοστά (cm) και ύπαρξη ομάδων καρκινικών κυττάρων με διάμετρο από 0,2 έως 2 χιλιοστά (mm), σε 1 έως 3 μασχαλιαίους λεμφαδένες ή στους εσωτερικούς μαστικούς λεμφαδένες ή σε συνδυασμό τους.
  - **Στάδιο ΙΙΙΒ**. Ο όγκος μπορεί να είναι οποιουδήποτε μεγέθους και ο καρκίνος έχει εισβάλλει στο θωρακικό τοίχωμα ή στο δέρμα του μαστού ή και στα δύο με ενδείξεις διογκωμένης φλεγμονής ή έλκη (όπως ο φλεγμονώδης καρκίνος του μαστού). Επίσης, ο καρκίνος μπορεί να έχει εισβάλει σε έως 9 μασχαλιαίους λεμφαδένες ή στους εσωτερικούς μαστικούς λεμφαδένες.

- **Στάδιο IIIC**. Ο όγκος μπορεί να είναι οποιουδήποτε μεγέθους και ο καρκίνος έχει εισβάλλει στο θωρακικό τοίχωμα ή στο δέρμα του μαστού ή και στα δύο, και ο καρκίνος έχει εξαπλωθεί σε :
  - Πάνω από 10 μασχαλιαίους λεμφαδένες ή λεμφαδένες κάτω από το οστό του τραχήλου (high axillary, apical, infraclavicular lymph nodes, Εικόνα 1.2), ή
  - Συνδυασμό άνω των 3 μασχαλιαίων λεμφαδένες και εσωτερικών μαστικών λεμφαδένων, ή
  - Λεμφαδένες πάνω από το οστό του τραχήλου (supraclavicular lymph nodes, Εικόνα 2).



Εικόνα 1.2 Ανατομία του γυναικείου μαστού και των περιβαλλόντων λεμφαδένων . [15]

Στο στάδιο III η θεραπεία είναι πιο περίπλοκη. Εάν ο καρκίνος χαρακτηριστεί ως μη χειρουργήσιμος (συνήθως στο στάδιο IIIC), πρέπει αρχικά να εφαρμοστεί μία μορφή θεραπείας (όπως η χημειοθεραπεία) η οποία να έχει ως αποτέλεσμα την μείωση της έκτασης του καρκίνου πριν η χειρουργική επέμβαση να αποτελεί επιλογή. Μαστεκτομή και ακτινοβολία ως τοπική θεραπεία και ορμονοθεραπεία και

χημειοθεραπεία ως συστηματική θεραπεία είναι κάποιες από τις συνηθέστερες μεθόδους. Η πλειοψηφία των ασθενών που έχουν τη νόσο σε αυτό το στάδιο ανταποκρίνονται καλύτερα όταν εφαρμόζεται ένας συνδυασμός θεραπειών.

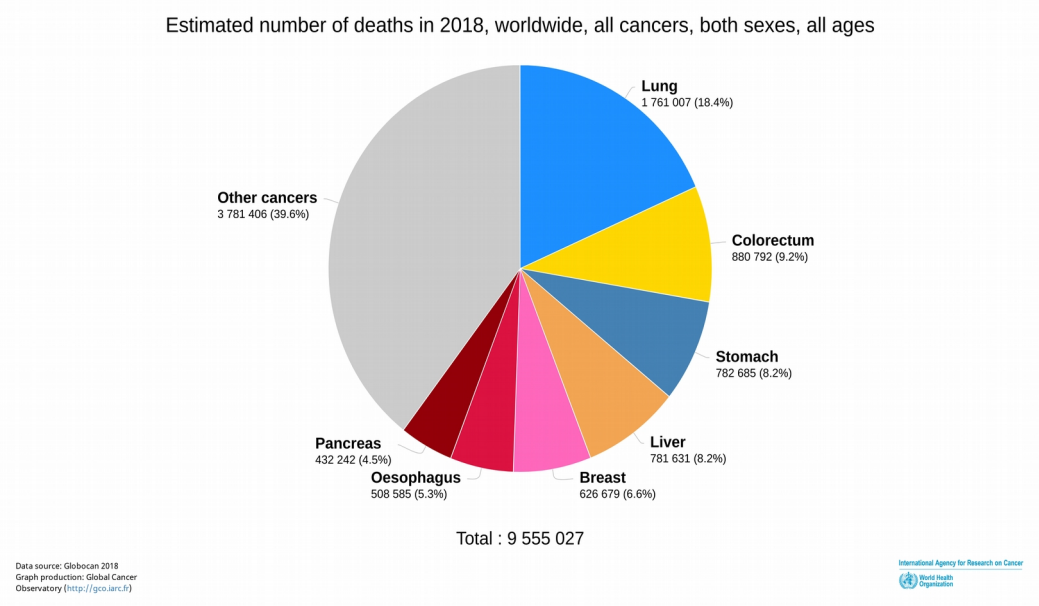
- **Στάδιο IV**. Το στάδιο ΙΙ περιγράφει διηθητικό καρκίνο που έχει εξαπλωθεί σε άλλα όργανα με πιο συνηθισμένες εντοπίσεις μεταστάσεων του καρκίνου του μαστού να είναι τα οστά, το ήπαρ, οι πνεύμονες και ο εγκέφαλος. Από τη στιγμή που καρκινικά κύτταρα έχουν επεκταθεί σε άλλα σημεία του σώματος, η συστηματική θεραπεία είναι μονόδρομος. Περίπου το 6% των γυναικών με καρκίνο του μαστού έχουν μετάσταση τη στιγμή της αρχικής διάγνωσης.

Πέραν των παραπάνω σταδίων πρέπει να αναφέρουμε και την περίπτωση του επανεμφανιζόμενου (recurrent) καρκίνου του μαστού. Σε αυτή την περίπτωση ο καρκίνος επανέρχεται μετά από την θεραπεία. Εάν η επανεμφάνιση συμβεί στην ίδια περιοχή από την οποία ξεκίνησε αρχικά ο καρκίνος, ονομάζεται τοπικής επανεμφάνισης. Εάν επανεμφανιστεί σε ιστούς ή λεμφαδένες κοντά στην περιοχή του αρχικού καρκινώματος, ονομάζεται περιφερειακή επανεμφάνιση. Εάν, τέλος, η επανεμφάνιση παρουσιαστεί σε απομακρυσμένους ιστούς τότε ονομάζεται απομακρυσμένη. Σε κάθε περίπτωση, ο επανεμφανιζόμενος καρκίνος αντιμετωπίζεται σαν να ήταν νέος καρκίνος.

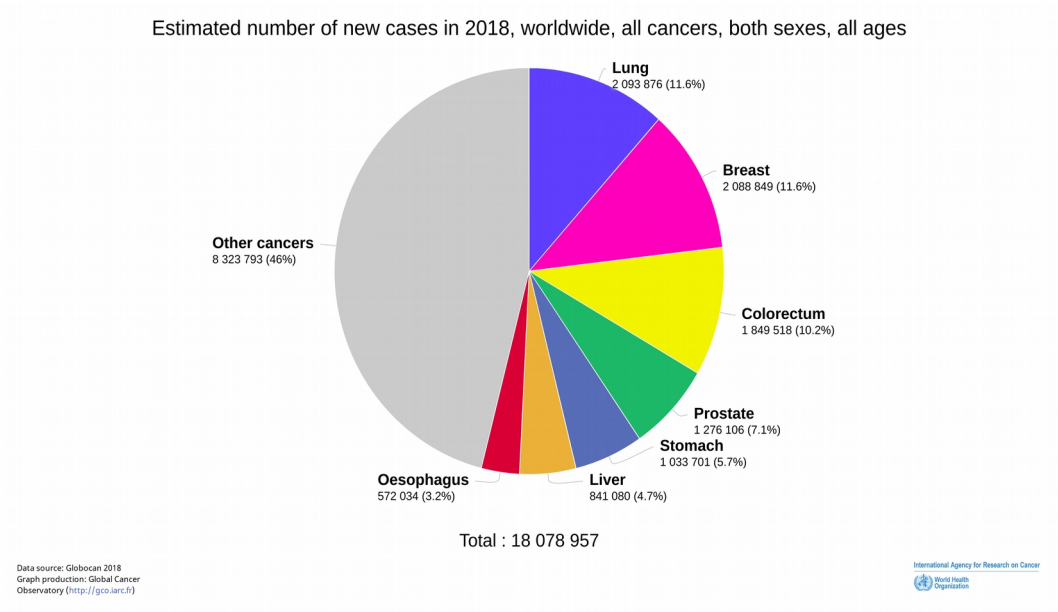
1.2 Επιδημιολογικά στοιχεία

Ο καρκίνος αποτελεί την σημαντικότερη αιτία θανάτου παγκοσμίως, όντας υπεύθυνος για περίπου 9,6 εκατομμύρια θανάτους το 2018. Από αυτούς 626.679 σημειώθηκαν λόγο του καρκίνου του μαστού, καθιστώντας τον καρκίνο του μαστού την πέμπτη κατά σειρά πιο σοβαρή μορφή καρκίνου και για τα δύο φύλα μαζί (Εικόνα 1.3). Κατά την διάρκεια του 2018 τα νέα περιστατικά καρκίνου του μαστού αντιστοιχούν σε 2,09 εκατομμύρια. Ο καρκίνος του μαστού είναι δεύτερος, σε εμφάνιση και στα δύο φύλα, κατά μόλις 5.027 περιστατικά παγκοσμίως μετά τον καρκίνο του πνεύμονα (Εικόνα 1.4). Για τον γυναικείο πληθυσμό ο καρκίνος του μαστού αποτελεί την πιο συχνή μορφή καρκίνου και ταυτόχρονα εκείνη με την μεγαλύτερη θνησιμότητα, όντας υπεύθυνος για το περίπου 15% των θανάτων λόγο καρκίνου επί του γυναικείου πληθυσμού (Εικόνα 1.5). Τα ποσοστά καρκίνου του μαστού είναι

μεγαλύτερα μεταξύ γυναικών σε πιο ανεπτυγμένες περιοχές του πλανήτη, παρόλα αυτά τα ποσοστά έχουν αυξητικές τάσης παγκοσμίως.[1]

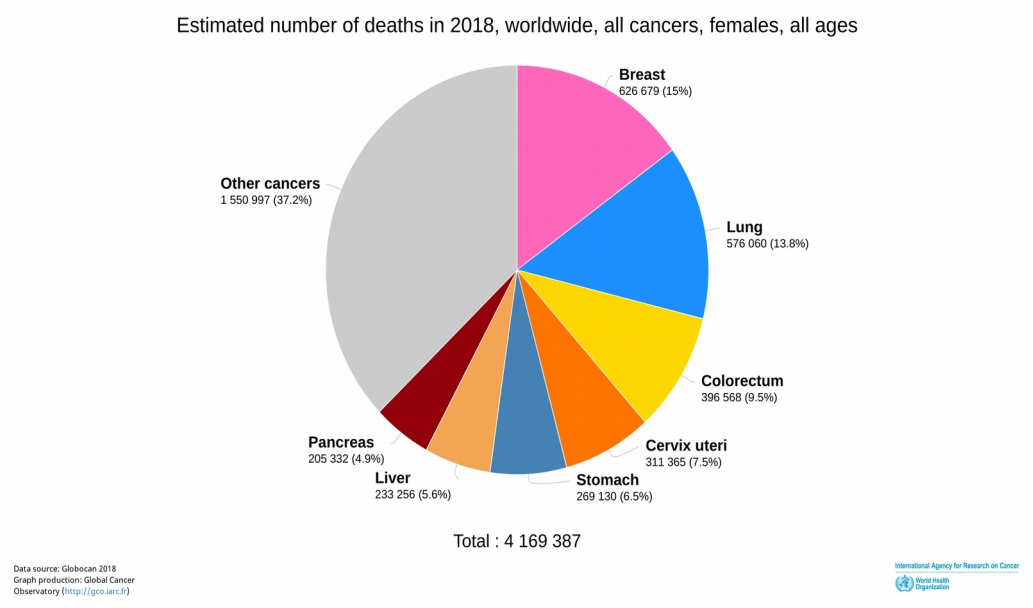


Εικόνα 1.3 Θάνατοι λόγο καρκίνου παγκοσμίως το 2018 ανά τύπο καρκίνου, για όλα τα φύλα.[1]



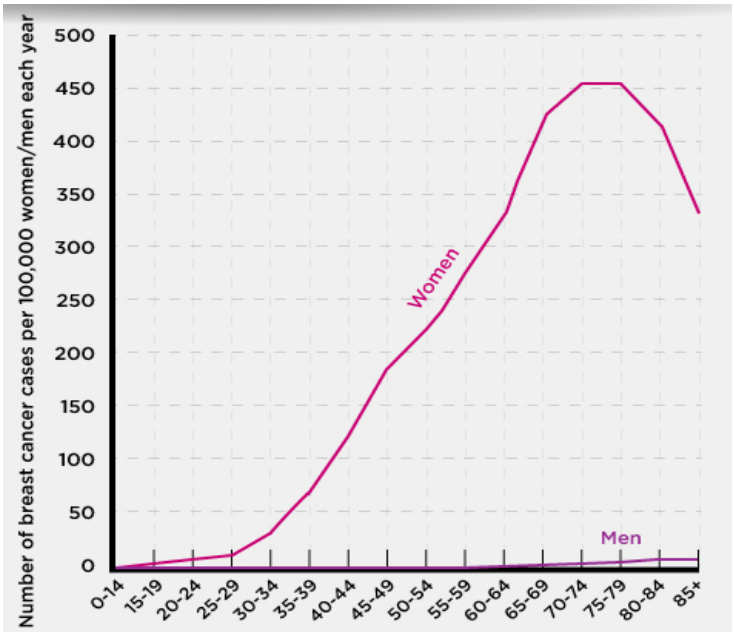
Εικόνα 1.4 Νέα περιστατικά καρκίνου παγκοσμίως το 2018 ανά τύπο, για όλα τα φύλα.[1]

Στην Ελλάδα κατά το 2018 υπήρξαν 7.734 νέα περιστατικά καρκίνου του μαστού, αριθμός που αναλογεί στο 27,2% των νέων περιστατικών καρκίνου επί του γυναικείου πληθυσμού και το 11,5% επί του συνολικού πληθυσμού. Την ίδια χρονιά 2.207 γυναίκες απεβίωσαν λόγο του καρκίνου του μαστού.



Εικόνα 1.5 Θάνατοι λόγο καρκίνου παγκοσμίως το 2018 ανά τύπο καρκίνου, για τον γυναικείο πληθυσμό.[1]

Η ηλικία επηρεάζει σημαντικά την εμφάνιση της νόσου. Σε γυναίκες ηλικίας κάτω των 35 ετών ο καρκίνος του μαστού εμφανίζεται σπάνια. Από αυτή την ηλικία και έως και την εμμηνόπαυση η νόσος παρουσιάζει μεγάλη αυξητική τάση εμφάνισης, η οποία συνεχίζει να παρατηρείται καθώς η ηλικία αυξάνει. Μετά την εμμηνόπαυση υπάρχει πτώση στον ρυθμό αύξησης της εμφάνισης, όπως μπορούμε να δούμε γραφικά στην Εικόνα 1.6, κάτι που ενισχύει τον ρόλο των αναπαραγωγικών ορμονών στην αιτιολογία της νόσου. [2]



Εικόνα 1.6 Αριθμός περιπτώσεων καρκίνου του μαστού ανά 100.000 γυναίκες/ άντρες ανά χρόνο κατά ηλικιακή ομάδα στον πληθυσμό των Η.Π.Α. [3]

Ο καρκίνος του μαστού επηρεάζει και τον ανδρικό πληθυσμό, αλλά με 100 φορές μικρότερη συχνότητα από ότι τον γυναικείο πληθυσμό. [2] Ο καρκίνος του μαστού αντιστοιχεί σε ποσοστό μικρότερο του 1% επί των συνολικών περιπτώσεων καρκίνου στους άνδρες. Τα ποσοστά επιβίωσης μεταξύ των δύο φύλων είναι ίδια.

Παρά το γεγονός ότι ο καρκίνος του μαστού θεωρείται μία νόσος του ανεπτυγμένου κόσμου, περίπου το 50% των περιπτώσεων καρκίνου του μαστού και 58% των θανάτων λόγο της ασθένειας λαμβάνουν χώρα σε λιγότερο ανεπτυγμένα κράτη. Η συχνότητα εμφάνισης ανά την υφήλιο διαφοροποιείται σημαντικά, από 19,3 ανά 100.000 γυναίκες στην Ανατολική Αφρική σε 89,7 ανά 100.000 γυναίκες στην Δυτική Ευρώπη. Στις περισσότερες από τις αναπτυσσόμενες περιοχές του πλανήτη η συχνότητα εμφάνισης είναι μικρότερη από 40 ανά 100.000 γυναίκες. Οι μικρότερες τιμές στην συχνότητα εμφάνισης συναντώνται στις περισσότερες Αφρικανικές χώρες, αλλά ακόμα και εκεί η συχνότητα εμφάνισης του καρκίνου του μαστού αυξάνεται.

Τα ποσοστά επιβίωσης από τον καρκίνο του μαστού ανά την υφήλιο διαφοροποιούνται εξίσου σημαντικά, από 80% και πλέον σε χώρες όπως η Βόρεια Αμερική, η Σουηδία και η Ιαπωνία σε περίπου 60% σε χώρες με μεσαίο μέσο κατά κεφαλήν εισόδημα και σε λιγότερο από 40% σε χώρες με χαμηλό μέσο κατά κεφαλήν εισόδημα. Τα χαμηλότερα ποσοστά επιβίωσης στα λιγότερο ανεπτυγμένα κράτη μπορούν να αναχθούν στην έλλειψη προγραμμάτων έγκαιρης ανίχνευσης, έχοντας ως αποτέλεσμα το ποσοστό των γυναικών που θα διαγνωστούν με την νόσο ήδη σε προχωρημένο στάδιο να είναι πολύ μεγάλο, καθώς επίσης και στην έλλειψη επαρκών ιατρικών εγκαταστάσεων διάγνωσης και θεραπείας. [1]

1.3 Η ανίχνευση του Καρκίνου του Μαστού μέσω των κλασικών μεθόδων

Η έγκαιρη και ακριβής ανίχνευση και διάγνωση καρκίνου του μαστού αποτελεί ένα πολύ σημαντικό όπλο στην αποτελεσματική αντιμετώπιση της νόσου καθώς, κατά γενικό κανόνα, σε όσο πιο αρχικό στάδιο διαγνωστεί ο καρκίνος τόσο μεγαλύτερες είναι και οι πιθανότητες επιβίωσης (Πίνακας 1.2). Επίσης, η έγκαιρη ανίχνευση παρέχει στον ασθενή την δυνατότητα επιλογής ανάμεσα σε περισσότερες διαφορετικές θεραπείες, όπως μικρότερης έκτασης χειρουργική επέμβαση, την χρήση χημειοθεραπείας με λιγότερες παρενέργειες ή ακόμα και αποχή από χημειοθεραπεία.

Πίνακας 1.2 Πιθανότητα 5 ετούς επιβίωσης ανά στάδιο καρκίνου του μαστού.[22]

Stage	5-year Relative Survival Rate
0	100%
I	100%
IIA	92%
IIB	81%
IIIA	67%
IIIB	54%
IV	20%

Ο καρκίνος του μαστού συνήθως ανιχνεύεται είτε μέσω ειδικής εξέτασης ή αφού μία γυναίκα παρατηρήσει έναν όγκο στον μαστό. Οι περισσότεροι όγκοι που θα ανιχνευθούν μέσω μαστογραφίας είναι καλοήθεις. Όταν υπάρχει υποψία ότι ο όγκος μπορεί να είναι κακοήθης, τότε για να γίνει διάγνωση πρέπει να πραγματοποιηθεί εξέταση στο μικροσκόπιο του μαστικού ιστού ή ογκιδίου ώστε να καθοριστεί ο τύπος, το στάδιο και ο βαθμός του καρκίνου. Η εξέταση αυτή αναφέρεται συχνά ως ιστολογική εξέταση, ή βιοψία. Ο ιστός για την πραγματοποίηση βιοψίας λαμβάνεται μέσω παρακέντησης με ειδική βελόνα (λεπτή ή μεγαλύτερης διαμέτρου) ή με χειρουργική επέμβαση. Η μικροσκοπική ανάλυση του ιστού πραγματοποιείται σήμερα από εκπαιδευμένο παθολόγο, Παθολογοανατόμο.

1.3.1 Μαστογραφία

Το κύριο εργαλείο για την ανίχνευση του καρκίνου του μαστού αυτή τη στιγμή είναι η μαστογραφία. Σε όλες τις ανεπτυγμένες χώρες υπάρχουν εθνικά προληπτικά προγράμματα κατά του καρκίνου του μαστού, τα οποία περιλαμβάνουν την διενέργεια προληπτικής μαστογραφίας. Στην πιο γενική περίπτωση τα προγράμματα αυτά περιλαμβάνουν ετήσια ή ανά διετία προληπτική μαστογραφία για όλες τις γυναίκες μετά την ηλικία των 40 ετών.
Η μαστογραφία είναι μία ακτινογραφία χαμηλής δόσης η οποία επιτρέπει την απεικόνιση της εσωτερικής δομής του μαστού. Υπάρχουν τρία κύρια είδη μαστογραφίας :

- **Αναλογική μαστογραφία:** αφορά στη παλαιότερη τεχνική της μαστογραφίας κατά τη οποία η απορρόφηση της ακτινοβολίας μετά τη διέλευσή της από το μαστό γίνεται από ακτινολογικό φιλμ. Θεωρείται ότι χορηγεί μεγαλύτερη δόση ακτινοβολίας και επιτυγχάνει λιγότερο λεπτομερή απεικόνιση του παρεγχύματος. Είναι ικανοποιητική ως εξέταση σε γυναίκες με λιπώδεις μαστούς, των οποίων η περιεκτικότητα σε πυκνά ινοαδενικά στοιχεία είναι μικρή. Υπάρχουν μελέτες όπου η αναλογική μαστογραφία

φαίνεται να έχει λίγο μεγαλύτερη ακρίβεια από την ψηφιακή μαστογραφία στην ομάδα γυναικών άνω 65 με λιπώδης μαστούς.[22]

- **Ψηφιακή Μαστογραφία:** αποτελεί την τεχνολογική εξέλιξη της αναλογικής. Η ακτινοβολία που διαπερνά το μαστό μετράται με ψηφιακό τρόπο και ακολούθως εκτυπώνεται σε φιλμ ή προβάλλεται σε οθόνη. Χορηγεί σχετικά μικρότερες δόσεις ακτινοβολίας και δίνει πιο λεπτομερή απεικόνιση. Με τον τρόπο αυτό, παρέχει παρέχει βελτιωμένη ευαισθησία (sensitivity) σε γυναίκες με πυκνά ινοαδενικά στοιχεία καθώς και σε γυναίκες κάτω των 50 ετών.[22]-[24]

- **Ψηφιακή Τομοσύνθεση:** οι νεότερης γενιάς μαστογράφοι έχουν τη δυνατότητα να χορηγούν την ίδια περίπου δόση ακτινοβολίας σε πολλαπλά διαφορετικά επίπεδα στο μαστό, που απέχουν μεταξύ τους κατά περίπου 1 εκατοστό, και να λαμβάνουν έτσι διαδοχικές μαστογραφικές “φέτες” του μαστού οι οποίες μπορούν να χρησιμοποιηθούν ώστε να κατασκευαστεί μία τρισδιάστατη εικόνα του μαστού. Έρευνες δείχνουν ότι η χρήση ψηφιακής τομοσύνθεσης σε συνδυασμό με δισδιάστατη μαστογραφία μπορεί να συμβάλει στην μείωση την λανθασμένων θετικών διαγνώσεων (false positives) και να παρέχει μικρή αύξηση της ευαισθησίας (sensitivity) συγκριτικά με την χρήση μόνο δισδιάστατης μαστογραφίας.[26]-[27] Παρόλα αυτά, εάν η εξέταση για την λήψη δισδιάστατης μαστογραφίας γίνει σε διαφορετική στιγμή από την ψηφιακή τομοσύνθεση, τότε η γυναίκα λαμβάνει περίπου την διπλάσια ποσότητα ακτινοβολίας, κάτι που σχετίζεται με αυξημένο κίνδυνο ανάπτυξης καρκίνου του μαστού.[25]

Σε μελέτη που πραγματοποιήθηκε στις ΗΠΑ το 2016[45] με βάση 1.683.504 εξετάσεις που διενεργήθηκαν από το 2007 έως το 2013, εκτίθηκε ότι η απόδοση της σύγχρονης ψηφιακής μαστογραφίας έχει μέση τιμή ευαισθησίας (Sensitivity) 86.9% και ειδικότητας (Specificity) 88,9%.

Η διαγνωστική ακρίβεια της ψηφιακής και της αναλογικής μαστογραφίας είναι παραπλήσιες επί του συνολικού πληθυσμού (Πίνακας 1). Επίσης, η διαγνωστική ακρίβεια δεν διαφέρει σημαντικά ανάμεσα στις δύο τεχνικές ως προς την ομάδα επικινδυνότητας ή την φυλή, καθώς και σε γυναίκες 50 ετών ή μεγαλύτερες, γυναίκες με λιπώδης ή μικρή πυκνότητα σε ινοαδενικά στοιχεία και σε γυναίκες μετά την εμμηνόπαυση.[28]

Πίνακας 1.3 Διαγνωστική ακρίβεια της ψηφιακής και της αναλογικής μαστογραφίας με χρήση BIRADS Score μετά από 365 παρακολούθησης. [28]

	Ψηφιακή Μαστογραφία	Αναλογική Μαστογραφία
Συνολικός πληθυσμός		
Ευαισθησία (Sensitivity)	0.7	0.66
Ειδικότητα (Specificity)	0.92	0.92
AUC	0.78	0.74
Γυναίκες <50 ετών		
Ευαισθησία (Sensitivity)	0.78	0.51
Ειδικότητα (Specificity)	0.9	0.9
AUC	0.84	0.69
Γυναίκες πριν την εμμηνόπαυση		
Ευαισθησία (Sensitivity)	0.72	0.51
Ειδικότητα (Specificity)	0.9	0.9
AUC	0.82	0.67
Γυναίκες με ετερογενώς πυκνό ή εξαιρετικά πυκνό μαστό		
Ευαισθησία (Sensitivity)	0.7	0.55
Ειδικότητα (Specificity)	0.91	0.9
AUC	0.78	0.68

Παρόλα αυτά, η επίδοση της ψηφιακής μαστογραφίας είναι σημαντικά καλύτερη σε σχέση με την αναλογική σε γυναίκες κάτω των 50 ετών, σε γυναίκες με ετερογενώς πυκνό ή εξαιρετικά πυκνό μαστό και σε γυναίκες πριν την εμμηνόπαυση (Πίνακας 1.3).[28]

Αποτελέσματα από οργανωμένα προγράμματα μαστογραφίας στην Ευρώπη και τον Καναδά δείχνουν ότι οι θάνατοι από καρκίνο του μαστού μειώθηκαν κατά πάνω από 40% στις γυναίκες οι οποίες υποβλήθηκαν σε τακτικό μαστογραφικό έλεγχο.[29]-[31] Περίπου το 12,6% των γυναικών που συμμετέχουν σε προγράμματα τακτικής προληπτικής μαστογραφίας παρουσιάζουν ευρήματα βάση των οποίων πρέπει υποβληθούν σε βιοψία. Από αυτές, το 27,5% βρέθηκαν θετικές σε καρκίνο μετά την βιοψία.[32] Η συχνή διενέργεια μαστογραφίας και η αθροιστική έκθεση στην ακτινοβολία την οποία συνεπάγεται μπορεί να οδηγήσει σε μικρή αύξηση του κινδύνου εμφάνισης καρκίνου του μαστού.[25] Παρόλα αυτά η ποσότητα της ακτινοβολίας στην οποία υποβάλλεται η γυναίκα κατά την διάρκεια της εξέτασης θεωρείται ότι είναι σχετικά μικρή, και έτσι το όφελος της εξέτασης ξεπερνά τις αρνητικές επιδράσεις. Η μείωση της έκθεσης στην ακτινοβολία μέσω πιο αποτελεσματικών απεικονιστικών τεχνικών αποτελεί αντικείμενο έρευνας.

1.3.2 Μαγνητική Μαστογραφία

Η μαγνητική μαστογραφία είναι η απεικόνιση του εσωτερικού ιστού του μαστού βάση της αρχής του πυρηνικού μαγνητικού συντονισμού και δεν κάνει χρήση ιονίζουσας ακτινοβολίας. Η μαγνητική τομογραφία χρησιμοποιεί μαγνητικά πεδία για να παράγει εικόνες εγκάρσιας τομής της δομής του εσωτερικού ιστού, παρέχοντας πολύ καλή αντίθεση του μαλακού ιστού. Η αντίθεση ανάμεσα στους ιστούς του μαστού (λιπώδη ιστό, αδένες, συνδετικό ή ινώδη ιστό κ.λπ.) εξαρτάται από την κινητικότητα και την συγκέντρωση των ατόμων υδρογόνου σε αυτούς. Η αλληλεπίδραση των ατόμων υδρογόνου με τον μαγνητικό πεδίο και μαγνητικό παλμό τον οποίο δημιουργεί ο μαγνητικός τομογράφος δημιουργεί το μετρήσιμο σήμα που καθορίζει τη φωτεινότητα των ιστών στην απεικόνιση. Συχνά, προκειμένου να επιτευχθεί υψηλότερη διακριτική ικανότητα, ιδιαίτερα στο διαχωρισμό της σύστασης των ιστών, εγχέονται ενδοφλεβίως σκιαγραφικοί παράγοντες με βάση το Γαδολίνιο (Gadolinium Based Contrast Agents- GBCAs). Οι εικόνες που λαμβάνονται με αυτή την μέθοδο είναι ενισχυμένες (contrast enhanced MRI) και βοηθούν στην αξιόπιστη ανίχνευση καρκίνων και άλλων αλλοιώσεων. Η μαγνητική μαστογραφία ενισχυμένης αντίθεσης έχει αυξημένη ευαισθησία (sensitivity) και μπορεί να ανιχνεύει όγκους μικρότερου μεγέθους σε σχέση με την μαστογραφία. Επίσης η ευαισθησία της δεν επηρεάζεται από την πυκνότητα του μαστού, την παρουσία ουλώδη ιστού έπειτα από εγχείρηση, ακτινοβολία ή την παρουσία ενθεμάτων. [33]-[37]

Η υψηλή της ευαισθησία της μεθόδου, η οποία κυμαίνεται από 90-100%, συνοδεύεται όμως από χαμηλή ειδικότητα (50-70% με βάση τις περισσότερες μελέτες), κάτι το οποίο έχει ως αποτέλεσμα μεγάλο αριθμό ψευδώς θετικών αποτελεσμάτων. Για τον λόγο αυτό η μαγνητική μαστογραφία δεν μπορεί, προς το παρόν, να αντικαταστήσει την μαστογραφία ως την κύρια εξέταση ρουτίνας για το μαστό. Λόγω του υψηλού ποσοστού ψευδώς θετικών και επειδή οι γυναίκες με υψηλότερο κίνδυνο εμφάνισης καρκίνου του μαστού έχουν πολύ περισσότερες πιθανότητες να ωφεληθούν, η εξέταση συστήνεται κυρίως σε γυναίκες που έχουν περισσότερο από 15% πιθανότητα εμφάνισης καρκίνου του μαστού καθώς και σε γυναίκες με μεγάλη πυκνότητα του μαστού.[38]

1.3.3 Υπερηχογράφημα μαστού

Το υπερηχογράφημα είναι μία μέθοδος απεικόνισης των εσωτερικών ιστών η οποία βασίζεται στην αλληλεπίδραση των υπερήχων με τους βιολογικούς ιστούς. Είναι μη επεμβατική εξέταση και δεν κάνει χρήση ιονίζουσας ακτινοβολίας. Το υπερηχογράφημα μαστού χρησιμοποιείται μερικές φορές για την αξιολόγηση μη φυσιολογικών ευρημάτων

μετά από μαστογραφία ή φυσική εξέταση. Το υπερηχογράφημα βοηθάει στο να καθοριστεί αν το εύρημα είναι κύστη ή ένας συμπαγής όγκος, αλλά δεν συνίσταται η χρήση του χωρίς την πραγματοποίηση μαστογραφίας. Επίσης, επειδή επιτρέπει να καθοριστεί η θέση και το μέγεθος του ευρήματος χρησιμοποιείται ως οδηγός κατά την εκτέλεση βιοψίας.

Για γυναίκες με μαστογραφικά πυκνό μαστό, ο υπερηχογράφος σε συνδυασμό με τη μαστογραφία μπορεί να προσφέρει μεγαλύτερη ευαισθησία από τη μαστογραφία μόνο. Ωστόσο, αυξάνει επίσης την πιθανότητα ψευδώς θετικών αποτελεσμάτων. [39], [40]

Εκτός από τη χρήση για τον προσδιορισμό της φύσης μιας ανωμαλίας του μαστού, μπορεί επίσης να πραγματοποιηθεί υπερηχογράφημα μαστού σε γυναίκες που θα πρέπει να αποφεύγουν την ακτινοβολία, όπως γυναίκες ηλικίας κάτω των 25 ετών, εγκύους, γυναίκες που θηλάζουν.

1.3.4 Κλινική εξέταση μαστού

Η κλινική εξέταση του μαστού (ψηλάφηση) από ειδικό μαστολόγο αποτελεί μία προτεινόμενη μέθοδο προληπτικής εξέτασης του μαστού σε συνδυασμό και συμπληρωματικά με την μαστογραφία. Συστήνεται να πραγματοποιείται προληπτικά μία φορά το χρόνο σε γυναίκες άνω των 30 ετών ή δύο φορές το χρόνο αν υπάρχει οικογενειακό ιστορικό καρκίνου του μαστού.

Πρόσφατες έρευνες δείχνουν απουσία ξεκάθαρων οφελών από την πραγματοποίηση της κλινικής εξέτασης του μαστού σε ασυμπτωματικές γυναίκες που βρίσκονται σε ομάδες μέσης επικινδυνότητας εμφάνισης του καρκίνου του μαστού. Επιπλέον, υπάρχουν ενδείξεις ότι η προσθήκη της κλινικής εξέτασης του μαστού στον προγραμματισμένο προληπτικό μαστογραφικό έλεγχο αυξάνει την πιθανότητα εμφάνισης ψευδών θετικών αποτελεσμάτων. [41]

Παρόλα αυτά, σε άλλη έρευνα η οποία πραγματοποιήθηκε το 2016 από το σύνολο των ασθενών που εξετάστηκαν 36,5% των καρκίνων ανιχνεύθηκαν με χρήση μόνο μαστογραφίας, 54,8% από συνδυασμό μαστογραφίας και κλινικής εξέτασης και 8,7% από την διενέργεια μόνο κλινικής εξέτασης του μαστού. Αυτά τα αποτελέσματα δείχνουν ότι ένας σημαντικός αριθμός καρκίνων δεν θα είχαν διαγνωστεί αν δεν είχε εκτελεστεί κλινική εξέταση. Επίσης, σε σύγκριση με τους καρκίνους που εντοπίστηκαν μόνο με μαστογραφία, εκείνες που ανιχνεύθηκαν σε συνδυασμό με κλινική εξέταση είχαν πιο επιθετικά χαρακτηριστικά, συμπεραίνοντας ότι η κλινική εξέταση του μαστού είναι μια εξέταση πολύ χαμηλού κόστους που συμβάλει στην ανίχνευση του καρκίνου του μαστού και συστήνει να συνεχίσει να αποτελεί μέρος της προγραμματισμένης προληπτικής εξέτασης του μαστού

ειδικά για γυναίκες μικρότερες των 50 ή μεγαλύτερες των 69 ετών, γυναίκες με μέτρια ή μεγάλη πιθανότητα εμφάνισης καρκίνου του μαστού και σε γυναίκες που είχαν καρκίνο του μαστού στο παρελθόν.[42]

1.3.5 Αυτοεξέταση Μαστού

Η αυτοεξέταση των μαστών θεωρήθηκε ως μια αποτελεσματική και οικονομικά συμφέρουσα μέθοδος πληθυσμιακού ελέγχου για την έγκαιρη ανίχνευση του καρκίνου του μαστού. Ωστόσο, φάνηκε ότι η αυτοεξέταση των μαστών δεν αυξάνει την επιβίωση. Η Ομάδα Υπηρεσιών Πρόληψης των ΗΠΑ (U.S. Preventive Services Task Force-USPSTF) έφτασε στο συμπέρασμα ότι τα αποδεικτικά στοιχεία από μεγάλες τυχαιοποιημένες μελέτες οδηγούν κατά της εκμάθησης της αυτοεξέταση των μαστών. Αυτή η σύσταση δεν είναι γνωστή σε μεγάλο ποσοστό των επαγγελματιών υγείας και των ασθενών ενώ η σημασία της αυτοεξέταση των μαστών έχει αμβλυνθεί και από άλλες επιστημονικές εταιρείες. Ωστόσο, σύγχρονες βιβλιογραφίες αναφέρονται ακόμα στην αξία της αυτοεξέτασης των μαστών. Η Αμερικανική Εταιρεία του Καρκίνου (American Cancer Society) δεν συστήνει πλέον τη μηνιαία αυτοεξέταση των μαστών, τονίζει όμως ότι είναι σημαντικό όλες οι γυναίκες να είναι εξοικειωμένες με την εμφάνιση και αίσθηση των μαστών τους και να αναφέρουν εγκαίρως οποιεσδήποτε μεταβολές στον γιατρό. Τέλος, καλό είναι οι γυναίκες να ενημερώνονται για τους περιορισμούς, τις δυσμενείς επιδράσεις (ψευδώς θετικά ευρήματα και αυξημένος αριθμός βιοψιών) αλλά και τα δυνητικά οφέλη που σχετίζονται με την αυτοεξέταση των μαστών.[43]

Κεφάλαιο 2

Βιβλιογραφική Επισκόπηση Μεθόδων Ανίχνευσης, Πρόγνωσης και Διάγνωσης του Καρκίνου του Μαστού

2.1 Εισαγωγή

Τα τελευταία χρόνια η τεχνητή νοημοσύνη και η μηχανική μάθηση έχουν αρχίσει να δείχνουν ότι μπορούν να παίξουν σημαντικό ρόλο στο μέλλον της ανίχνευσης, της διάγνωσης αλλά και της πρόγνωσης του καρκίνου του μαστού. Από τα τέλη της δεκαετίας του ’80 σημαντική έρευνα έχει πραγματοποιηθεί για την εφαρμογή τεχνικών μηχανικής μάθησης στην διαγνωστική και προγνωστική διαδικασία. Αναλύοντας δεδομένα ή εικόνες από δείγματα όγκων που είχαν ληφθεί μέσω παρακέντησης, τα μοντέλα μηχανικής μάθησης καταφέρνουν να διακρίνουν καλοήθεις από κακοήθεις όγκους με ακρίβεια έως και 99%. Η χρήση της μηχανικής μάθησης φαίνεται να μπορεί να συμβάλει αποτελεσματικά και στην ανίχνευση του καρκίνου του μαστού, είτε μέσω ανάλυσης εικόνων ψηφιακής μαστογραφίας ή με ανάλυση αιματολογικών δεικτών και σωματομετρικών δεδομένων.

2.2 Χρήση τεχνητής νοημοσύνης για διάγνωση και πρόγνωση του καρκίνου του μαστού

Πληθώρα ερευνών έχουν πραγματοποιηθεί σχετικά με την εφαρμογή της τεχνητής νοημοσύνης και της μηχανικής μάθησης στην διαδικασία της διάγνωσης του καρκίνου του μαστού.

Εκτενής έρευνα έχει γίνει πάνω σε τρεις βάσεις δεδομένων οι οποίες συλλέχθηκαν από το πανεπιστήμιο του Wisconsin κατά την δεκαετία του 1990. Στην αρχική βάση δεδομένων Breast Cancer Wisconsin (Original) Data Set [46], η οποία έγινε δημοσίως διαθέσιμη το 1992, παραθέτονται 11 κυτταρολογικά χαρακτηριστικά όγκων του μαστού από 699 συμμετέχοντες. Τα δείγματα συλλέχθηκαν με την μέθοδο παρακέντησης λεπτής βελόνας

(fine needle aspirate , FNA) ενώ τα δεδομένα βαθμονομήθηκαν σε κλίμακα από το 1 έως το 10 την στιγμή της συλλογής του δείγματος. Τα παραπάνω δεδομένα χρησιμοποιήθηκαν αρχικά για την κατηγοριοποίηση των δειγμάτων ως καλοήθεις ή κακοήθεις όγκους μέσω μαθηματικών μεθόδων [47]. Στην παραπάνω βάση πραγματοποιήθηκε, εκτός άλλων, και συγκριτική μελέτη απόδοσης, [48], μεταξύ των αλγορίθμων μηχανικής μάθησης Support Vector Machine (SVM), K-Nearest Neighbors (k-NN), Decision Tree (C4.5) και Naive Bayes (NB). Βέλτιστα αποτελέσματα σε αυτή τη μελέτη παρουσίασε ο αλγόριθμος SVM με ακρίβεια 97.13%.

Εν συνεχεία, το 1995 δημοσιεύτηκαν άλλες δύο βάσεις δεδομένων σχετικά με τον καρκίνο του μαστού από το πανεπιστήμιο του Wisconsin, η μία με διαγνωστικό και η δεύτερη με προγνωστικό αντικείμενο. Η βάση Breast Cancer Wisconsin (Diagnostic) Data Set [49] παρέχει δεδομένα από χαρακτηριστικά όγκων (357 καλοήθεις – 212 κακοήθεις) του μαστού από δείγματα τα οποία συλλέχθηκαν μέσω παρακέντησης λεπτής βελόνας (fine needle aspirate , FNA) από 569 συμμετέχοντες. Από τα δείγματα λήφθηκαν ψηφιακές εικόνες μέσω σάρωσης και από αυτές υπολογίστηκαν οι τιμές για 10 χαρακτηριστικά των πυρήνων των κυττάρων των όγκων. Τα δεδομένα χρησιμοποιήθηκαν για την κατηγοριοποίηση των δειγμάτων ως καλοήθεις ή κακοήθεις όγκους μέσω τεχνικών μηχανικής μάθησης σε διάφορες έρευνες. Οι Aličković et al.[50] χρησιμοποίησαν γενετικό αλγόριθμο για την επιλογή χαρακτηριστικών και εν συνεχεία συνέκριναν την ακρίβεια κατηγοριοποίησης μεταξύ των αλγορίθμων Logistic Regression, Decision Tree, Random Forest, Bayesian Network, Multilayer Perceptron (MLP), Radial Basis Function Networks (RBFN), Support Vector Machines (SVM) και Rotation Forest, πετυχαίνοντας βέλτιστα αποτελέσματα με Rotation Forest και ακρίβεια 99,48%. Οι Albrecht et al[64] πέτυχαν ακρίβεια 98,80% χρησιμοποιώντας αλγόριθμο μάθησης που συνδύαζε λογαριθμική Simulated Annealing με νευρώνες Perceptron. Οι Goodman et al.[65] χρησιμοποίησαν τρεις μεθόδους, Artificial Immune Recognition System (AIRS), Learning Vector Quantization (LVQ) και Optimized Learning Vector Quantization, για να επιτύχουν 97,2 , 96,8 και 96,7% ακρίβεια, αντίστοιχα. Οι Sahan και Polat[66] χρησιμοποίησαν μια πρωτότυπη υβριδική τεχνική βασισμένη σε Fuzzy-artificial Immune Systems και k-NN, πετυχαίνοντας ακρίβεια κατηγοριοποίησης 99,14%. Οι Μαγκλογιάννης, Ζαφειρόπουλος και Αναγνωστόπουλος[55] συνέκριναν 3 διαφορετικούς αλγορίθμους κατηγοριοποίησης: SVM, Naive Bayes και τεχνητά νευρωνικά δίκτυα (ANN) καταλήγοντας σε καλύτερα αποτελέσματα ακρίβειας, ευαισθησίας και ειδικότητας με τον αλγόριθμο SVM. Οι Stoean R. και Stoean C.[67] χρησιμοποιήθηκε SVM πετυχαίνοντας ακρίβεια 97%, ενώ με χρήση εξελικτικού αλγορίθμου

πραγματοποιείτε επεξήγηση της διαδικασίας λήψης της απόφασης διάγνωσης. Οι Koloseni et al.[68] χρησιμοποίησαν διαφορικό εξελικτικό ταξινομητή με βέλτιστα μέτρα απόστασης λαμβάνοντας μέση τιμή κατηγοριοποίησης 93,64%. Οι Saez et al.[69] πρότειναν αμοιβαία πληροφόρηση (mutual information, MI) μεταξύ των χαρακτηριστικών ως παράγοντα στάθμισης του βάρους των χαρακτηριστικών που εισάγονται σε αλγόριθμο k-NN και η ακρίβεια κατηγοριοποίησης ήταν 96,14%. Οι Chen et al.[70] πρότειναν ένα σύστημα βελτιστοποίησης του SVM βασισμένο σε βελτιστοποίηση παράλληλης χρονικής παραλλαγής σμήνους σωματιδίων (parallel time-variant particle swarm optimization, PTVPSO) για παράλληλη βελτιστοποίηση παραμέτρων και επιλογή χαρακτηριστικών και πετυχαίνοντας ακρίβεια κατηγοριοποίησης 98,44%. Οι Zheng et al.[71] πρότειναν ένα υβριδικό μοντέλο διάγνωσης χρησιμοποιώντας επιβλεπόμενη (SVM) και μη επιβλεπόμενη μάθηση (K-means) και η ακρίβεια κατηγοριοποίησης ήταν 97,38%.

Βλέπουμε ότι πολλά από τα μοντέλα που δημιουργήθηκαν και εκπαιδεύτηκαν πάνω στα δεδομένα των παραπάνω βάσεων δεδομένων του Wisconsin καταφέρνουν να έχουν υψηλή ακρίβεια κατηγοριοποίησης. Ένα αρνητικό αυτών τον μεθόδων είναι ότι τα δεδομένα τα οποία χρειάζονται για να λάβουν απόφαση είναι κυτταρολογικά και για την απόκτησή τους χρειάζεται να πραγματοποιηθεί παρακέντηση και εξέταση των κυττάρων του όγκου. Η διαδικασία αυτή είναι επεμβατική για τον ασθενή και χρησιμοποιείται για την διενέργεια της βιοψίας η οποία είναι η εξέταση κατά την οποία ο Παθολογοανατόμος, μέσω μικροσκοπικής κυτταρολογικής ανάλυσης του ιστού του όγκου, καταλήγει σε διάγνωση για το εάν ο όγκος είναι καλοήθης ή κακοήθης και σε περίπτωση κακοήθιας καθορίζει τον τύπο, τον βαθμό και το στάδιο του καρκίνου. Επομένως, η χρήση τους θα περιορίζεται ως μία δεύτερη γνώμη ώς προς την ύπαρξη ή μη κακοήθιας κατά την διενέργεια της βιοψίας.

Τέλος, η βάση Breast Cancer Wisconsin (Prognostic) Data Set [52] παρέχει δεδομένα παρακολούθησης σε βάθος χρόνου ασθενών με καρκίνο του μαστού. Πίο συγκεκριμένα, συμμετείχαν οι 198 από τις συμμετέχοντες στην έρευνα για την διαγνωστική βάση δεδομένων [49], οι οποίες είχαν διαγνωστεί με διηθητικό (επιθετικό) τύπο καρκίνου και δεν υπήρχαν στοιχεία εμφάνισης μετάστασης. Αφού ο κακοήθεις όγκος αφαιρέθηκε χειρουργικά οι ασθενείς συνέχισαν να παρακολουθούνται έως την επανεμφάνιση καρκίνου ή περισσότερο από 6 χρόνια σε περίπτωση μη επανεμφάνισης. Σε 151 περιπτώσεις δεν υπήρξε επανεμφάνιση καρκίνου του μαστού, ενώ σε 47 υπήρξε. Χαρακτηριστικά παραδείγματα χρήσης της συγκεκριμένης βάσης για την δημιουργία προβλεπτικών μοντέλων επανεμφάνισης καρκίνου του μαστού είναι τα [53] - [55].

Μία ακόμα βάση δεδομένων στην οποία έχει πραγματοποιηθεί έρευνα για την διάγνωση του καρκίνου του μαστού με τεχνικές μηχανικής μάθησης είναι η Breast Cancer Data Set η οποία συλλέχθηκε από το Πανεπιστημιακό Νοσοκομείο της Ljubljana [56] κατά την δεκαετία του 1980, και έγινε δημοσίως διαθέσιμη για έρευνα το 1988. Εδώ, όπως και στην Breast Cancer Wisconsin (Prognostic) Data Set, ελέγχεται η επανεμφάνιση ή μη της νόσου 5 χρόνια μετά από χειρουργική αφαίρεση του κακοήθους όγκου. Στην έρευνα συμμετείχαν 286 ασθενείς από τους οποίους 85 επανεμφάνισαν τη νόσο μέσα στα επόμενα 5 χρόνια, ενώ 201 δεν νόσησαν εκ νέου. Κάθε περίπτωση περιγράφεται από 9 χαρακτηριστικά, τα οποία όμως ως σύνολο δεν υπήρξαν επαρκεί για την εξαγωγή υψηλής ποιότητας συμπερασμάτων, καθώς η ακρίβεια των προβλέψεων κυμαίνονται μεταξύ 65 και 78%, όπως μπορούμε να δούμε στα[57] - [60].

Μία διαφορετική προσέγγιση μπορούμε επίσης να δούμε σε έρευνες οι οποίες έχουν γίνει πάνω σε δεδομένα φασματοσκοπίας ηλεκτρικής αντίστασης μαστικού ιστού. Εργασίες σε μία ακόμη βάση δεδομένων [61], η οποία είναι δημοσίως διαθέσιμη στην βιβλιοθήκη UCI Machine Learning Repository. Σε αυτήν 120 δείγματα ιστού συλλέχθηκαν από 64 ασθενείς κατά την διάρκεια χειρουργικής επέμβασης στον μαστό. Δύο διαθέσιμες έρευνες πάνω στην συγκεκριμένη βάση[62]-[63], δεν χρησιμοποίησαν τεχνικές μηχανικής μάθησης αλλά είχαν στατιστική προσέγγιση. Τα αποτέλεσμα δείχνουν ότι κακοήθης όγκος στον μαστό μπορεί να διακριθεί από άλλες ασθένειες του μαστού μέσω φασματοσκοπίας ηλεκτρικής αντίστασης μαστικού ιστού με ακρίβεια 78 και 86% αντίστοιχα.

2.3 Χρήση τεχνητής νοημοσύνης για ανίχνευση του καρκίνου του μαστού μέσω ανάλυσης εικόνας

Από τις αρχές της δεκαετίας του 1990 έχουν αναπτυχθεί υπολογιστικά συστήματα υποβοήθησης με χρήση υπολογιστή για την ανίχνευση και διάγνωση (computer-aided detection and diagnosis, CAD) του καρκίνου του μαστού, τα οποία μετά από την έγκριση από την Υπηρεσία Φαρμάκων και Τροφίμων (Food and Drug Administration, FDA) των ΗΠΑ το 1998 άρχισαν να εφαρμόζονται στην κλινική πρακτική ώστε να βοηθήσουν τους ακτινολόγους στην ερμηνεία της μαστογραφίας. Τα συστήματα αυτά αναλύουν ψηφιοποιημένα μαστογραφίες και εντοπίζουν ύποπτες περιοχές για αναθεώρηση από τον ακτινολόγο. Δυστυχώς, τα δεδομένα υποδηλώνουν ότι τα αρχικά εμπορικά συστήματα CAD

δεν οδήγησαν σε σημαντική βελτίωση στην ακρίβεια στην ανίχνευση του καρκίνου του μαστού, με αποτέλεσμα να μην υπάρξει ουσιαστική πρόοδος για περισσότερο από μια δεκαετία από τότε που εισήχθησαν.[72]-[73],[152]

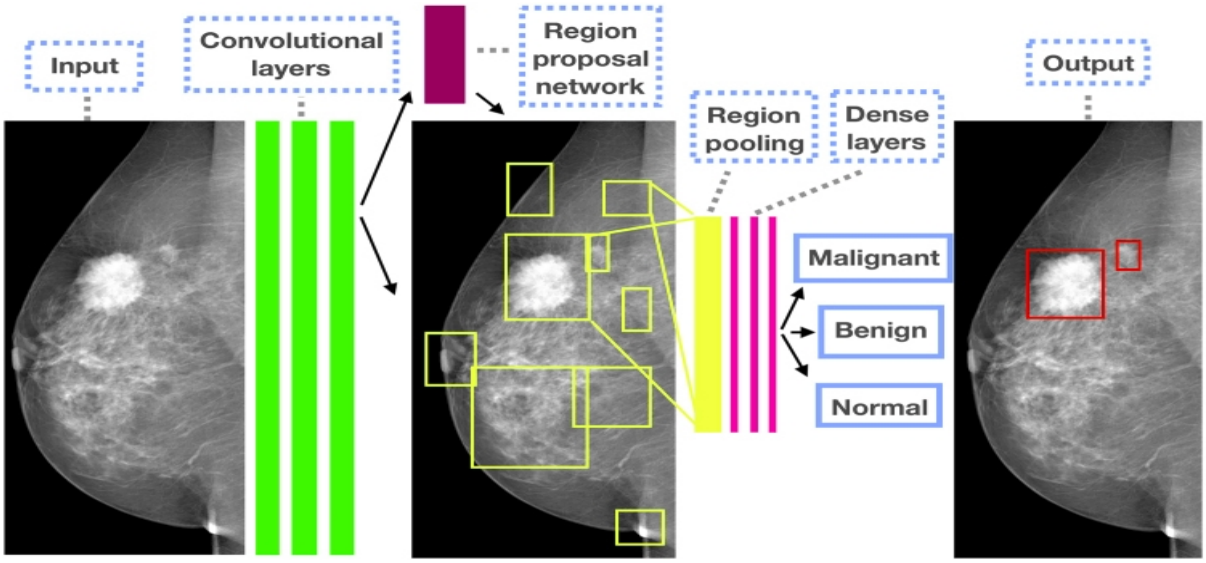
Τα τελευταία χρόνια μοντέλα βασισμένα σε τεχνικές Βαθιάς Μάθησης φαίνεται να υπερτερούν έναντι των κλασσικών τεχνικών μηχανικής μάθησης στον τομέα της αναγνώρισης εικόνας από υπολογιστή. Αυτή η εξέλιξη έχει προκαλέσει μεγάλο ενδιαφέρον για την εφαρμογή Βαθιάς Μάθησης σε προβλήματα συσχετισμένα με τεχνικές ιατρικής απεικόνισης, όπως η ανίχνευση καρκίνου του μαστού με εξέταση εικόνων μαστογραφίας.

Οι Dhungel et al.[74] προτείνουν ένα σύστημα CAD με ελάχιστη παρέμβαση του χρήστη για ανίχνευση, τμηματοποίηση και κατηγοριοποίηση των όγκων σε μαστογραφίες. Η ανίχνευση του όγκου έγινε με χρήση Deep Cascade Learning και μοντέλου Random Forest για ορισμό πιθανών περιοχών ενδιαφέροντος, στις οποίες εφαρμόζεται στην συνέχεια Μπαεσιανή βελτιστοποίηση (Bayesian optimization). Η κατηγοριοποίηση έγινε με χρήση ενός συνελικτικού νευρωνικού δικτύου (convolutional neural network, CNN). Το σύστημα καταφέρνει να ανιχνεύσει το 90% των όγκων, ενώ έχει ακρίβεια τμηματοποίησης 0,85 κατά Dice index στους σωστά ανιχνευθέντες όγκους. Η κατηγοριοποίηση των όγκων πραγματοποιείται με ευαισθησία (Sensitivity) 0,98 και ειδικότητα (Specificity) 0,70.

Οι Wang et al.[75] υλοποίησαν ένα μοντέλο ανίχνευσης καρκίνου του μαστού σε πρώιμο στάδιο. Χρησιμοποιώντας τεχνικές Βαθιάς Μάθησης, η μέθοδος στοχεύει στην ανίχνευση όγκων και μικρο-ασβεστοποιήσεων (microcalcifications) οι οποίες μπορούν να χρησιμοποιηθούν ως δείκτης πρώιμου καρκίνου του μαστού. Η αρχιτεκτονική που χρησιμοποίησαν αποτελούνταν από ιεραρχικά στοιβαγμένους πολλαπλούς αυτόματους κωδικοποιητές (stacked autoencoders, SAE). Η μέθοδος είχε τρία σενάρια: (1) με παρουσία μόνο μικρο-ασβεστοποιήσεων, (2) με παρουσία μικρο-ασβεστοποιήσεων και όγκων μαζί και (3) με παρουσία μόνο όγκων. Η μέθοδος τους επέτυχε ακρίβεια κατηγοριοποίησης σε καλοήθη ή κακοήθη ευρήματα 87,3% στην πρώτη περίπτωση, 89,7% στην δεύτερη και 61,3% στην τρίτη.

Οι Riddli et al.[76] χρησιμοποίησαν μεταφερόμενη μάθηση για να εφαρμόσουν το μοντέλο Faster R-CNN για την ανίχνευση όγκων σε μαστογραφίες και κατηγοριοποίηση αυτών χωρίς καμία ανθρώπινη παρέμβαση (Εικόνα 2.1). Το μοντέλο Faster R-CNN ανίχνευσε κακοήθεις όγκους με ευαισθησία 0,9 και 0,3 λανθασμένα θετικές επισημάνσεις (false positive marks) ανά μαστογραφία, ενώ κατά την διαδικασία της κατηγοριοποίησης στην δημοσίως διαθέσιμη βάση Inbreast είχε AUC 0,95.

Οι Gao et al.[77] παρουσίασαν ένα μοντέλο βασισμένο σε ρηχό-βαθύ CNN (SD-CNN) για την ανίχνευση όγκων και την κατηγοριοποίηση αυτών ψηφιακές μαστογραφίες ενισχυμένης αντίθεσης (CEDM). Μέσω της εφαρμογής του SD-CNN στις εικόνες CEDM η ακρίβεια κατηγοριοποίησης, στην δημοσίως διαθέσιμη βάση Inbreast, ήταν 0,9 με AUC = 0,92.



Εικόνα 2.1 Παράδειγμα ανίχνευσης και κατηγοριοποίησης όγκων σε μαστογραφία μέσω του μοντέλου Faster R-CNN.[34]

Οι Hagos et al.[78] παρουσίασαν ένα multiview CNN για την ανίχνευση όγκων. Η αρχική αναγνώριση των υποψήφιων περιοχών ενδιαφέροντος έγινε βάση της εργασίας των Karssemeijer et al.[79]. Στην συνέχεια έγινε διαλογή μεταξύ των υποψηφίων περιοχών ενδιαφέροντος μέσω της εκμάθησης των συμμετρικών διαφορών μεταξύ των εικόνων από τις μαστογραφίες και των δύο μαστών, πετυχαίνοντας AUC 0,933.

Οι Tuwen et al.[80] συνέκριναν συστήματα αυτοματοποιημένης ανίχνευσης και διάγνωσης όγκου του μαστού με βάση τα συνελικτικά νευρωνικά δίκτυα (convolutional neural networks, CNN). Το καλύτερο μοντέλο ήταν βασισμένο σε FasterRCNN, και είχε ευαισθησία 0,73 και 0,86 με 0,1 και 0,54 λανθασμένα θετικές επισημάνσεις (false positive marks) ανά μαστογραφία.

Οι Jung et al.[81] πρότειναν μοντέλο ανίχνευσης όγκου ενός σταδίου, χρησιμοποιώντας το μοντέλο RetinaNet[82]. Το RetinaNet είναι μια μέθοδος ανίχνευσης αντικειμένων μονού σταδίου που μπορεί να ξεπεράσει το πρόβλημα της ανισορροπίας τάξης και κάνει χρήση μιας νέας συνάρτησης απώλειας, την οποία ονόμασαν Focal Loss, η οποία του επιτρέπει να εστιάσει περισσότερο σε πιο περίπλοκες εικόνες. Μετά από εκπαίδευση στη βάση δεδομένων GURO από το Πανεπιστημιακό Νοσοκομείο GURO της Νοτίου Κορέας, το

μοντέλο RetinaNet δοκιμάστηκε στην βάση δεδομένων INbreast πέτυχε ακρίβεια κατηγοριοποίησης 0.86 με 0,5 λανθασμένα θετικές επισημάνσεις (false positive marks) ανά μαστογραφία.

Έχουν γίνει πολλές προσπάθειες για τη μείωση της ανθρώπινης παρέμβασης και την παραγωγή πλήρως αυτοματοποιημένων CAD. Στην πραγματικότητα, όλες οι μέθοδοι απαιτούν βάσεις δεδομένων από μαστογραφίες με επισημασμένες περιοχές ενδιαφέροντος και τελική διάγνωση από ακτινολόγο για να επικυρώσουν τα ευρήματά τους κατά τη διάρκεια της εκπαίδευσης του μοντέλου και της δοκιμής του. Τα συστήματα Βαθιάς Μάθησης χρειάζονται μεγάλο όγκο δεδομένων για την εκπαίδευσή τους και το γεγονός ότι η πρόσβαση σε επισημασμένες από ακτινολόγους μαστογραφίες είναι δύσκολη αναγκάζει τους ερευνητές να καταφύγουν σε πρωτότυπες λύσεις, όπως η μεταφερόμενη μάθηση.

Από τα παραπάνω μπορούμε να δούμε ότι η εφαρμογή μοντέλων Deep Learning φαίνεται ότι μπορεί να συμβάλει στη βελτίωση των διαγνωστικών επιδόσεων των συστημάτων υποβοήθησης για την ανίχνευση και διάγνωση (computer-aided detection and diagnosis, CAD) του καρκίνου του μαστού με χρήση υπολογιστή, αλλά εξακολουθούν να υπάρχουν προκλήσεις για την κλινική εφαρμογή τέτοιων μεθόδων και απαιτείται περισσότερη έρευνα.[83]

2.4 Χρήση τεχνητής νοημοσύνης για ανίχνευση του καρκίνου του μαστού μέσω αιματολογικών εξετάσεων

Τα τελευταία χρόνια αναφέρονται έρευνες για την ανίχνευση καρκίνου του μαστού λαμβάνοντας υπόψη συγκεντρώσεις βιοδεικτών του καρκίνου του μαστού στο αίμα, σε συνδυασμό με σωματομετρικά δεδομένα όπως η ηλικία και ο δείκτης μάζας σώματος (BMI). Η προσέγγιση αυτή έχει το πλεονέκτημα ότι δεν βασίζεται σε μαστογραφική εικόνα, η οποία χρησιμοποιεί ιονίζουσα ακτινοβολία, ενώ θα μπορούσε να προστεθεί στην λίστα των αιματολογικών εξετάσεων ρουτίνας κάθε γυναίκας, αποφεύγοντας με αυτό τον τρόπο την πραγματοποίηση περαιτέρω εξετάσεων.

2.4.1 Έρευνες στην βάση δεδομένων Coimbra Breast Cancer Dataset

Το 2016 μία μελέτη από τους Crisóstomo J et al. [97], ώστε να εξεταστεί η συσχέτιση του δείκτη μάζας σώματος (BMI), της γλυκαιμίας, της υπερινσουλιναιμίας, της

ινσουλινοαντίστασης και των επιπέδων των αντιποκινών στο αίμα με τον καρκίνο του μαστού, αποτέλεσε τον προπομπό για την δημιουργία μίας νέας προσέγγισης στην εφαρμογή της μηχανικής μάθησης για την ανίχνευση του καρκίνου του μαστού. Τα αποτελέσματα αυτής της έρευνας δείχνουν ότι ο δυσμεταβολισμός της γλυκόζης, η αντίσταση στην ινσουλίνη και οι μεταβολές στην έκκριση των αντιποκινών, και ιδιαίτερα της ρεζιστίνης, μπορεί να εμπλέκονται στην ανάπτυξη και εξέλιξη του καρκίνου του μαστού σε υπέρβαρες / παχύσαρκες γυναίκες πριν και μετά την εμμηνόπαυση. Η βάση δεδομένων της συγκεκριμένης έρευνας έγινε δημοσίως διαθέσιμη το 2018 από την βιβλιοθήκη UCI Machine Learning Repository[85] και χρησιμοποιήθηκε σε διάφορες ερευνητικές μελέτες με σκοπό την ανίχνευση καρκίνου του μαστού με χρήση μοντέλων μηχανικής μάθησης. Σε αυτή την βάση έχουν καταχωρηθεί τα δεδομένα από 116 συμμετέχουσες, εκ των οποίων 64 είχαν προσφάτως διαγνωστεί με καρκίνο του μαστού, ενώ οι 52 αποτέλεσαν την υγιή ομάδα ελέγχου. Για κάθε μία από τις συμμετέχουσες πραγματοποιήθηκαν αιματολογικές εξετάσεις των επιπέδων γλυκόζης, ινσουλίνης, λεπτίνης, αντιπονεκτίνης, ρεζιστίνης και MCP-1, και υπολογίστηκε το μοντέλο εκτίμησης ομοιόστασης (HOMA). Μαζί με αυτά, σωματομετρικά δεδομένα όπως η ηλικία και ο δείκτης μάζας σώματος (BMI) καταχωρήθηκαν για κάθε συμμετέχουσα.

Η πρώτη εργασία ανάπτυξης μοντέλου μηχανικής μάθησης για την πρόβλεψη ύπαρξης καρκίνου του μαστού πάνω στα δεδομένα της Coimbra Breast Cancer Dataset[85] έγινε από τους M. Patrício et al.[98] και δημοσιεύτηκε τον Ιανουάριο του 2018. Οι M. Patrício et al. προχώρησαν στην υλοποίηση και σύγκριση της απόδοσης προβλεπτικών μοντέλων μηχανικής μάθησης τριών διαφορετικών αλγορίθμων κατηγοριοποίησης : Logisic Regression (LR), Random Forest (RF) και Support Vector Machines (SVM).

Ώστε να ορίσουν τους συνδυασμούς των χαρακτηριστικών οι οποίοι θα αποτελούσαν την είσοδο κατά την εκπαίδευση και αξιολόγηση των προβλεπτικών μοντέλων χρησιμοποιήθηκε ο δείκτης Gini ώστε να ληφθεί μία αρχική εκτίμηση της συνεισφοράς καθενός από τους υποψήφιους βιοδείκτες σε ένα προβλεπτικό μοντέλο. Οι βιοδείκτες κατατάχθηκαν ως προς την βαρύτητα της συνεισφοράς τους, από την μέγιστη στην ελάχιστη, ως εξής : V1 γλυκόζη, V2 ρεζιστίνη, V3 ηλικία, V4 BMI, V5 HOMA, V6 λεπτίνη, V7 ινσουλίνη, V8 αντιπονεκτίνης, V9 MCP-1. Βάση αυτής της κατάταξης ορίστηκε ότι οι συνδυασμοί των χαρακτηριστικών θα ήταν οι : V1-V2, V1-V3, V1-V4, V1-V5, V1-V6 και V1-V9.

Για την εκπαίδευση και αξιολόγηση των μοντέλων χρησιμοποιήθηκε διαδικασία Monte Carlo Cross-Validation (MCCV), στην οποία το σύνολο των δεδομένων χωρίστηκε

τυχαία σε ένα σύνολο εκπαίδευσης (training set), το οποίο είχε το 69,8% των δεδομένων, και ένα σύνολο αξιολόγησης (test set), με το υπόλοιπο 30,2% των δεδομένων. Η διαδικασία εκπαίδευσης και αξιολόγησης πραγματοποιήθηκε με 500 διαφορετικούς τυχαίους συνδυασμούς συνόλων, για κάθε έναν από τους τρεις ταξινομητές (LR, RF και SVM) και για κάθε έναν από τους 6 συνδυασμούς χαρακτηριστικών (V1-V2, V1-V3, V1-V4, V1-V5, V1-V6 και V1-V9). Βέλτιστα αποτελέσματα για την ανίχνευση καρκίνου του μαστού επί των συνόλων αξιολόγησης, παρουσίασε το προβλεπτικό μοντέλο βασισμένο στον ταξινομητή Support Vector Machines το οποίο ακολουθώντας διαστήματα εμπιστοσύνης 95% πέτυχε Ευαισθησία [82%, 88%], Ειδικότητα [84%, 90%] και AUC [0.87, 0.91].

Οι Muhammet Fatih Aslan et al. [86] υλοποίησαν και συνέκριναν προβλεπτικά μοντέλα βασισμένα σε : Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks, ANN), Extreme Learning Machine (ELM), Support Vector Machines (SVM) και K-Nearest Neighbor (k-NN). Στην προσέγγισή τους οι Muhammet Fatih Aslan et al. εφάρμοσαν αρχικά ομαλοποίηση τύπου Min-Max Normalization στα δεδομένα, χρησιμοποιώντας την εξίσωση :

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (2.4.1.1)$$

Ως είσοδος στα μοντέλα χρησιμοποιήθηκε το σύνολο και των 9 χαρακτηριστικών (V1-V9). Στη συνέχεια έγινε βελτιστοποίηση των παραμέτρων (hyperparameters) κάθε μοντέλου και η εκτίμηση της απόδοσης έγινε με χρήση διαδικασίας Monte Carlo Cross-Validation (MCCV), στην οποία το σύνολο εκπαίδευσης (training set) είχε περίπου το 80% των δεδομένων και το σύνολο αξιολόγησης (test set) είχε το υπόλοιπο 20% των δεδομένων. Η διαδικασία εκπαίδευσης και αξιολόγησης πραγματοποιήθηκε με 10 διαφορετικούς τυχαίους συνδυασμούς συνόλων, για κάθε ένα από τα τέσσερα προβλεπτικά μοντέλα και η τελική ακρίβεια υπολογίστηκε ως ο μέσος όρος της ακρίβειας όλων των επαναλήψεων. Βέλτιστα αποτελέσματα παρουσίασε το μοντέλο Extreme Learning Machine πετυχαίνοντας Ακρίβεια πρόβλεψης 80%.

Οι Silva Araújo V.J. et al. [87] χρησιμοποίησαν την Coimbra Breast Cancer Dataset ώστε να εξετάσουν την εφαρμογή Ασαφών Νευρωνικών Δικτύων (Fuzzy Neural Networks, FNN) στην πρόβλεψη ύπαρξης καρκίνου του μαστού. Τα τρία νευρωνικά δίκτυα που δοκιμάστηκαν χρησιμοποιούσαν τους εξής τρεις διαφορετικού τύπου νευρώνες : (1) τον unineuron με Γκαουσιανή συνάρτηση συμμετοχής (UniNet), (2) τον andneuron με τριγωνική συνάρτηση συμμετοχής (AndNet) και (3) τον orneuron με Γκαουσιανή συνάρτηση συμμετοχής (OrNet). Εκτός των ασαφών νευρωνικών δικτύων υλοποιήθηκαν και

προβλεπτικά μοντέλα βασισμένα σε αλγόριθμους Naive Bayes (NB), Zero R (ZR) και Random Tree (RT), τον C 4.5 ο οποίος είναι ένας ακόμα αλγόριθμο βασισμένο στα Δέντρα Απόφασης και τέλος ένα πολυεπίπεδο Perceptron (Multilayer Perceptron, MLP) το οποίο είναι ένα είδος τεχνητού νευρωνικού δικτύου εμπρόσθιας τροφοδότησης (FeedForward Neural Network, FFNN), ώστε να συγκριθούν με τα μοντέλα FNN. Αρχικά πραγματοποιήθηκε η εκπαίδευση και αξιολόγηση των τριών μοντέλων Ασαφών Νευρωνικών Δικτύων, όπου χρησιμοποιήθηκε η διαδικασία Monte Carlo Cross-Validation (MCCV), στην οποία το σύνολο των δεδομένων χωρίστηκε τυχαία θέτοντας το 70% των δεδομένων στο σύνολο εκπαίδευσης (training set) και το υπόλοιπο 30% στο σύνολο αξιολόγησης (test set). Η διαδικασία εκπαίδευσης και αξιολόγησης πραγματοποιήθηκε με 30 διαφορετικούς τυχαίους συνδυασμούς συνόλων, για κάθε ένα από τα FNN και για κάθε έναν από τους 6 συνδυασμούς χαρακτηριστικών (V1-V2, V1-V3, V1-V4, V1-V5, V1-V6 και V1-V9), όπως ορίστηκαν στην εργασία των M. Patrício et al.[3][98]. Και σε αυτή την εργασία τα καλύτερα αποτελέσματα παρουσιάζονται, και για τα τρία FNN, όταν είσοδο αποτελεί ο συνδυασμός V1-V4. Μετά την αναγνώριση του V1-V4 ως βέλτιστου συνδυασμού χαρακτηριστικών εισόδου, πραγματοποιήθηκε η εκπαίδευση και αξιολόγηση των πέντε μοντέλων (NB, RT, ZR, C 4.5 και MLP) με την ίδια μεθοδολογία που εφαρμόστηκε για τα FNN, αλλά μόνο για τον συνδυασμό V1-V4 ως είσοδο. Οι τελικές τιμές ακρίβειας, ευαισθησίας, ειδικότητας και AUC υπολογίστηκαν ως ο μέσος όρος των αντίστοιχων τιμών όλων των επαναλήψεων. Το Ασαφές Νευρωνικό Δίκτυο OrNet παρουσίασε τα καλύτερα αποτελέσματα πετυχαίνοντας Ακρίβεια 81,04%, Ευαισθησία 81,93% , Ειδικότητα 81,18% και AUC 0,8019.

Η Lara Dular κατά την εκπόνηση της διπλωματικής της εργασίας[88] χρησιμοποίησε την Coimbra Breast Cancer Dataset ως μία από δέκα βάσεις δεδομένων ώστε να συγκρίνει την απόδοση τεσσάρων αλγορίθμων που χρησιμοποιούνται στην μηχανική μάθηση, οι οποίοι ήταν οι : Λογιστική Παλινδρόμηση (LR), Αναδρομική διαμέριση και Παλινδρομικά Δέντρα (Recursive Partitioning and Regression Trees, RPART), k-Nearest Neighbors (k-NN) και Random Forest (RF). Οι αλγόριθμοι εφαρμόστηκαν με τις αρχικές του παραμέτρους, με μόνες διαφοροποιήσεις τον ορισμό του αριθμού των γειτόνων στον k-NN ίσο με 5, ορίζοντάς τον έτσι ως 5-NN, και στον RF ο αριθμός των φύλων ορίστηκε στα 500 και ο αριθμός των μεταβλητών σε κάθε διαχωρισμό ως 3. Για την εκπαίδευση και αξιολόγηση των μοντέλων χρησιμοποιήθηκε διαδικασία 5-fold Cross-Validation και ο διαχωρισμός των δεδομένων σε σύνολα εκπαίδευσης και αξιολόγησης ήταν ο ίδιος για όλους τους αλγορίθμους. Ως είσοδος χρησιμοποιήθηκε το συνδυασμός V1-V4, όπως αυτός ορίστηκε στην εργασία των M. Patrício et al.[3][98]. Οι αλγόριθμοι συγκρίθηκαν ως προς την ακρίβεια και το AUC και οι τελικές τιμές

τους υπολογίστηκαν ως ο μέσος όρος των αντίστοιχων τιμών των 5 διαφορετικών συνόλων. Το μοντέλο βασισμένο σε RF ήταν εκείνο με την καλύτερη απόδοση, παρουσιάζοντας Ακρίβεια 75,00% και AUC 0,805.

Οι K. Polat et al.[89] προτείνουν ένα υβριδικό μοντέλο τριών σταδίων για την πρόβλεψη ύπαρξης καρκίνου του μαστού. Αρχικά εφαρμόζεται στα δεδομένα ομαλοποίηση απόκλισης απολύτου διαμέσου (median absolute deviation) MAD Normalization βάση της εξίσωσης :

$$MAD = median(|Yi - median(Yi)|) \quad (2.4.1.2)$$

Στην συνέχεια, και σε αντίθεση με τις υπόλοιπες προσεγγίσεις, αντί να γίνει επιλογή μεταξύ των χαρακτηριστικών, εφαρμόστηκε στάθμιση των χαρακτηριστικών βασισμένη σε K-means clustering όπως προτάθηκε από τον K. Polat[90]. Η χρήση αυτής της μεθόδου έγινε ώστε να μετατραπούν οι δύο κλάσεις (ασθενής – υγιής) της βάσης δεδομένων από γραμμικά μη διαχωρίσιμες σε γραμμική διαχωρίσιμες. Για το στάδιο της κατηγοριοποίησης επιλέχθηκε ο αλγόριθμος Adaptive Boosting (AdaBoost.M1). Για την εκπαίδευση και αξιολόγηση του μοντέλου χρησιμοποιήθηκε διαδικασία 10-fold Cross-Validation και οι τελικές τιμές των μέτρων αξιολόγησης υπολογίστηκαν ως ο μέσος όρος των αντίστοιχων τιμών των 10 διαφορετικών συνόλων. Το μοντέλο σημείωσε πολύ καλή επίδοση στην πρόβλεψη ύπαρξης ή μη καρκίνου του μαστού έχοντας Ακρίβεια 91,37% , Ευαισθησία 0,914 , Ειδικότητα 0,923 , AUC 0,938 , F-Measure 0,914 , Kappa Value 0,8276 , TP Rate 91,40% , FP Rate 7,7% και Precision 0,919.

Τέλος, οι Mohaimenul Islam et al.[96], το 2019, υλοποίησαν και συνέκριναν προβλεπτικά μοντέλα βασισμένα σε : decision tree (DT), random forest (RF), K-nearest neighbors (KNN), support vector machine (SVM), Λογιστική Παλινδρόμηση (LR) και Τεχνητά Νευρωνικά Δίκτυα (ANN). Αρχικά, για την επιλογή των χαρακτηριστικών που θα χρησιμοποιούσαν στην δημιουργία των μοντέλων, χρησιμοποίησαν το πακέτο Boruta το οποίο είναι ένας αλγόριθμος κατάταξης και επιλογής χαρακτηριστικών με βάση τον αλγόριθμο Random Forest. Βάση των αποτελεσμάτων αποφάσισαν να χρησιμοποιήσουν 5 από τα διαθέσιμα χαρακτηριστικά, τα οποία ήταν τα : Γλυκόζη, Ρεζιστίνη, BMI, HOMA και Ινσουλίνη. Για την εκπαίδευση και αξιολόγηση των μοντέλων χρησιμοποιήθηκε διαδικασία 10-fold Cross-Validation και η διαδικασία επαναλήφθηκε 5 φορες. Οι αλγόριθμοι συγκρίθηκαν ως προς την ακρίβεια, την ευαισθησία, την ειδικότητα και το AUC και οι τελικές τιμές τους υπολογίστηκαν ως ο μέσος όρος των αντίστοιχων τιμών των 50

διαφορετικών συνόλων. Το μοντέλο βασισμένο σε k-NN ήταν εκείνο με την καλύτερη απόδοση, παρουσιάζοντας Ακρίβεια 86%, Ευαισθησία 80%, Ειδικότητα 91% και AUC 0,95.

2.4.2 Άλλες προσεγγίσεις

Οι Hwa H. L. et al.[84] το 2008 υλοποίησαν ένα μοντέλο μηχανικής μάθησης, χρησιμοποιώντας τον αλγόριθμο της Λογιστική Παλινδρόμηση (Logistic Regression), για να προβλέψουν την ύπαρξη καρκίνου στον μαστό καθώς και την εξάπλωση ή μη των καρκινικών κυττάρων στους λεμφαδένες. Ως εισόδους για την εκπαίδευση και αξιολόγηση του μοντέλου τους χρησιμοποίησαν κλινικοπαθολογικά δεδομένα και τις συγκεντρώσεις καρκινικών δεικτών στον ορό αίματος 94 γυναικών, εκ των οποίων 55 έπασχαν από καρκίνο του μαστού και οι 39 αποτέλεσαν την υγιή ομάδα ελέγχου. Οι καρκινικοί δείκτες οι οποίοι δοκιμάστηκαν στο μοντέλο ήταν : το καρκινοεμβρυονικό αντιγόνο CEA (carcinoembryonic antigen), το καρκινικό αντιγόνο CA15.3, το ειδικό αντιγόνο – πολυπεπτίδιο ιστού TPS (tissue polypeptide-specific antigen), ο υποδοχέας μεμβράνης sIL-2R (soluble interleukin-2 receptor και η πρωτεΐνη IGFBP-3 (insulin-like growth factor binding protein-3). Για την πρόβλεψη παρουσίας καρκίνου του μαστού, το μοντέλο λογιστικής παλινδρόμησης απέδωσε βέλτιστα με την χρήση των TPS, CA15-3 και IGFBP-3 ως εισόδους, πετυχαίνοντας AUC 0.86, ευαισθησία 85% και ειδικότητα 62%. Ο συνδυασμός των βιοδεικτών sIL-2R και TPS έδωσε τα καλύτερα αποτελέσματα ως είσοδος του μοντέλου για την πρόβλεψη της μετάστασης στους λεμφαδένες με ευαισθησία 69%.

Το 2018 οι Cohen JD et al.[91] περιγράφουν μία μέθοδο ανίχνευσης και εντοπισμού οκτώ διαφορετικών τύπων καρκίνου βάση των συγκεντρώσεων πρωτεϊνικών βιοδεικτών στο αίμα και μεταλλάξεων στο cell-free DNA με χρήση μηχανικής μάθησης. Στην έρευνα συμμετείχαν 1005 ασθενείς με καρκίνο, εκ των οποίων 209 είχαν καρκίνο του μαστού, και 812 υγιείς συμμετέχοντες αποτέλεσαν την ομάδα ελέγχου. Από τις 41 διαφορετικές πρωτεΐνες οι οποίες δοκιμάστηκαν, οι 8 βρέθηκαν να έχουν σημαντική συμβολή στην διαδικασία της ανίχνευσης καρκίνου, οι οποίες είναι οι : καρκινικό αντιγόνο 125 (CA-125), καρκινοεμβρυονικό αντιγόνο (CEA), καρκινικό αντιγόνο 19-9 (CA19-9), προλακτίνη (PRL), αυξητικός παράγοντας των κυττάρων του ύπατος (HGF), οστεοποντίνη (OPN), μυελοϋπεροξειδάση (MPO), και ιστικός αναστολές μεταλλοπρωτεϊνασών 1 (TIMP-1). Επίσης, σημαντικός δείκτης ήταν και η παρουσία στο cell-free DNA του πλάσματος τουλάχιστον μίας μετάλλαξης ανάμεσα σε 1933 διαφορετικές θέσεις στο γονιδίωμα, η οποία παίρνει πραγματικές τιμές και ορίζεται ως τιμή μετάλλαξης Ω (mutation omega score). Με αυτά τα χαρακτηριστικά ως εισόδους, δημιουργήθηκε το μοντέλο που ονομάστηκε

CancerSeek και βασίζεται στον αλγόριθμο της Λογιστικής Παλινδρόμησης (LR). Για την εκπαίδευση και αξιολόγηση του μοντέλου χρησιμοποιήθηκε διαδικασία 10-fold Cross-Validation η οποία επαναλήφθηκε 10 φορές και οι τελικές τιμές ευαισθησίας και ειδικότητας υπολογίστηκαν ως ο μέσος όρος των αντίστοιχων τιμών των 100 διαφορετικών συνόλων. Το μοντέλο είχε AUC περίπου 0,91 ως προς την ανίχνευση ύπαρξης καρκίνου επί του συνόλου των συμμετεχόντων, όμως η ανίχνευση καρκίνου μεταξύ των συμμετεχόντων που έπασχαν από καρκίνο του μαστού υπήρξε η λιγότερο επιτυχημένη διαδικασία για το μοντέλο, έχοντας Ευαισθησία μόλις 33,49%. Παρόλα αυτά το μοντέλο είχε πολύ καλή Ειδικότητα έχοντας μόλις 7 λανθασμένα θετικά αποτελέσματα μεταξύ των 812 υγιών συμμετεχόντων, δηλαδή περίπου 99,14%. Σε μία δεύτερη λειτουργία, μοντέλο βασισμένο στον αλγόριθμο Random Forest, δεχόμενο ως εισόδους το Ω score, τις τιμές συγκεντρώσεων στο αίμα όλων των 41 πρωτεϊνικών βιοδεικτών συν το φύλο του ασθενούς, αλλά αυτή την φορά μόνο από τους 626 ασθενείς οι οποίοι είχαν προηγουμένως βρεθεί ορθώς θετικοί σε καρκίνο από το CancerSeek, προσπαθεί να εντοπίσει τον τύπο του καρκίνου (μεταξύ των καρκίνων των ωοθηκών, του ήπατος, του στομάχου, του παγκρέατος, του οισοφάγου, του παχέος εντέρου, του πνεύμονα και του μαστού) από τον οποίο πάσχει ο ασθενής. Για την εκπαίδευση και αξιολόγηση του μοντέλου εντοπισμού χρησιμοποιήθηκε διαδικασία 10-fold Cross-Validation η οποία επαναλήφθηκε 10 φορές και οι τελική τιμή ακρίβειας υπολογίστηκε ως ο μέσος όρος της ακρίβειας των 100 διαφορετικών συνόλων. Σε αυτή του την λειτουργία, το μοντέλο κατάφερε να εντοπίσει τον καρκίνου του μαστού στο 63% των περιπτώσεων.

Το 2019 οι Wong KC et al.[92] εργάστηκαν στα ίδια δεδομένα, καταφέρνοντας να βελτιώσουν σημαντικά τα αποτελέσματα, δοκιμάζοντας και συγκρίνοντας διαφορετικούς αλγορίθμους ως βάση για τα μοντέλα τους, δύο μοντέλα βασισμένα σε Aggregating One-Dependence Estimators (AODE)[93]-[94], ο οποίος αποτελεί παραλλαγή του Naive Bayes, τα οποία ονόμασαν CancerA1DE και CancerA2DE, τρία μοντέλα Βαθιάς Μάθησης, ένα μοντέλο Naive Bayes και ένα μοντέλο βασισμένο στον αλγόριθμο τύπου δέντρου απόφασης J48. Τα παραπάνω μοντέλα δοκιμάστηκαν για την ανίχνευση και για τον εντοπισμό του καρκίνου, ενώ η μεθοδολογία που ακολούθησαν ως προς την επιλογή χαρακτηριστικών εισόδου και εκπαίδευσης και αξιολόγησης υπήρξε η ίδια με αυτή των Cohen JD et al.[91]. Σε όλες τις περιπτώσεις τα μοντέλα που ήταν βασισμένα σε AODE παρουσίασαν την βέλτιστη απόδοση. Ως προς την ανίχνευση ύπαρξης καρκίνου επί του συνόλου των συμμετεχόντων το μοντέλο είχε AUC περίπου 0,99. Η ανίχνευση καρκίνου μεταξύ των συμμετεχόντων που έπασχαν από καρκίνο του μαστού υπήρξε και εδώ η λιγότερο επιτυχημένη διαδικασία για όλα τα μοντέλα, με την βέλτιστη επίδοση να σημειώνετε από το CancerA1DE έχοντας

Ευαισθησία περίπου 67,5% και Ειδικότητα 99%. Ως προς την διαδικασία του εντοπισμού του καρκίνου το CancerA1DE κατάφερε να εντοπίσει τον καρκίνου του μαστού στο περίπου 70% των περιπτώσεων.

Οι Cristiano Stephen et al.[95], επίσης το 2019, παρουσίασαν άλλη μία προσέγγιση ανίχνευσης και εντοπισμού διαφορετικών τύπων καρκίνου βάση της παρουσίας μεταλλάξεων στο cell-free DNA με χρήση μηχανικής μάθησης. Στην έρευνα συμμετείχαν 236 ασθενείς με καρκίνο του μαστού (54), του παχέος εντέρου (27), του πνεύμονα (12), των ωοθηκών (28), του παγκρέατος (34), του στομάχου (27) και του χοληφόρου πόρου (26) και 245 υγιή άτομα αποτέλεσαν την ομάδα ελέγχου. Οι Cristiano Stephen et al.[95] δημιούργησαν μία μέθοδο ανίχνευσης, ανάλυσης και συσχέτισης των ανωμαλιών που παρατηρούνται στον κατακερματισμό του cell-free DNA του πλάσματος των καρκινοπαθών σε σχέση με των υγιών ατόμων, την οποία ονόμασαν ‘DNA evaluation of fragments for early interception’ (DELFI). Το προφίλ κατακερματισμού κάθε συμμετέχοντα στην έρευνα αποτέλεσε την είσοδο για την εκπαίδευση και αξιολόγηση ενός μοντέλου μηχανικής μάθησης βασισμένου σε stochastic gradient boosting (gbm), το οποίο κατηγοριοποιεί τους συμμετέχοντες σε υγιείς ή καρκινοπαθείς. Για την εκπαίδευση και αξιολόγηση του μοντέλου χρησιμοποιήθηκε διαδικασία 10-fold Cross-Validation η οποία επαναλήφθηκε 10 φορές και οι τελικές τιμές των μέτρων αξιολόγησης υπολογίστηκαν ως ο μέσος όρος των αντίστοιχων τιμών των 100 διαφορετικών συνόλων. Το μοντέλο είχε Ευαισθησία 79,81% όταν η Ειδικότητα ήταν ορισμένη στο 95% ως προς την ανίχνευση ύπαρξης καρκίνου επί του συνόλου των συμμετεχόντων, ενώ στην ανίχνευση καρκίνου μεταξύ των συμμετεχόντων που έπασχαν από καρκίνο του μαστού το μοντέλο είχε Ευαισθησία 70,37% όταν η Ειδικότητα ήταν ορισμένη στο 95%. Για τον εντοπισμό του τύπου του καρκίνου από τον οποίο πάσχει ο ασθενής μεταξύ των 174 ασθενών οι οποίοι είχαν προηγουμένως βρεθεί ορθώς θετικοί σε καρκίνο από το μοντέλο ανίχνευσης (έχοντας την Ειδικότητα ορισμένη στο 90%), αναπτύχθηκε μοντέλο μηχανικής μάθησης βασισμένο ξανά σε stochastic gradient boosting (gbm) και δεχόμενο τα ίδια χαρακτηριστικά με το μοντέλο ανίχνευσης ως εισόδους. Για την εκπαίδευση και αξιολόγηση του μοντέλου εντοπισμού χρησιμοποιήθηκε διαδικασία 10-fold Cross-Validation η οποία επαναλήφθηκε 10 φορές και οι τελική τιμή ακρίβειας υπολογίστηκε ως ο μέσος όρος της ακρίβειας των 100 διαφορετικών συνόλων. Σε αυτή του την λειτουργία, το μοντέλο κατάφερε να εντοπίσει τον καρκίνου του μαστού στο 76.19% των περιπτώσεων.

Κεφάλαιο 3

Ανάπτυξη Μοντέλου Ανίχνευσης του Καρκίνου του Μαστού

3.1 Δεδομένα

Η συλλογή των δεδομένων για την ανάπτυξη του παρόντος μοντέλου πραγματοποιήθηκε στο Πανεπιστημιακό Νοσοκομείο της Coimbra στην Πορτογαλία μεταξύ 2009 και 2013. Τα δεδομένα χρησιμοποιήθηκαν αρχικά στην έρευνα των J. Crisóstomo et al.[97], στην οποία αποδεικνύεται συσχέτιση μεταξύ δυσμεταβολισμού της γλυκόζης, της αντίστασης στην ινσουλίνη καθώς και σε μεταβολές στα επίπεδα έκκρισης αντιποκινών με την ανάπτυξη και εξέλιξη του καρκίνου του μαστού σε υπέρβαρες γυναίκες.

3.1.1 Συμμετέχοντες

Συνολικά 154 γυναίκες επιλέχτηκαν από το Πανεπιστημιακό Νοσοκομείο της Coimbra για να συμμετέχουν στην έρευνα. Διαχωρίστηκαν σε 4 ομάδες ανάλογα με την παρουσία ή μη καρκίνου του μαστού και εάν ήταν υπέρβαρες ή όχι, βάση του Δείκτη Μάζας Σώματος (BMI). Από τον συγκεκριμένο πληθυσμό αποκλείστηκαν 38 συμμετέχοντες, είτε διότι ο Δείκτης Μάζας Σώματός του ήταν μεγαλύτερος από 40 kg/m^2 είτε διότι έλλειπαν τα δεδομένα για τουλάχιστον μία από τις ελεγχόμενες μεταβλητές.[98]

Οι συμμετέχοντες στην υγιή ομάδα ελέγχου επιλέχτηκαν εφόσον δεν είχαν διαγνωστεί στο παρελθόν με οποιαδήποτε μορφή καλοήθους ή κακοήθους καρκίνου και δεν είχαν οικογενειακό ιστορικό καρκίνου του μαστού. Ως ασθενείς συμμετέχοντες επιλέχτηκαν γυναίκες οι οποίες είχαν προσφάτως διαγνωστεί θετικές σε καρκίνο του μαστού από μαστογραφία, ενώ η διάγνωση επιβεβαιώθηκε από ιστολογική εξέταση. Καμία από τις ασθενείς δεν είχε υποβληθεί σε θεραπεία κατά του καρκίνου πριν την λήψη του αιματολογικού δείγματος για την υλοποίηση της έρευνας. Καμία από τις συμμετέχοντες δεν έπασχε από μόλυνση ή οξεία ασθένεια κατά την χρονική περίοδο της επιλογής των συμμετεχόντων. [97]

Κλινικά δεδομένα (προσωπικό και οικογενειακό ιατρικό ιστορικό), καθώς και δημογραφικά και ανθρωπομετρικά (ηλικία, ύψος και βάρος) δεδομένα συλλέχθηκαν από

όλες της συμμετέχοντες κατά την διάρκεια της πρώτης εξέτασης, υπό τις ίδιες συνθήκες και από τον ίδιο ερευνητή ιατρό. Δείγματα αίματος συλλέχθηκαν την ίδια ώρα της ημέρας, ενώ όλες οι συμμετέχοντες είχαν προηγουμένως υποβληθεί σε ολονύκτια νηστεία (τουλάχιστον 8 ωρών). Όλες οι συμμετέχοντες κατέθεσαν εγγράφως την συναίνεσή τους. [97]

3.1.2 Ανάλυση Δειγμάτων

Τα δείγματα αίματος υποβλήθηκαν σε φυγοκέντριση (2500 g) στους 4 ºC αποθηκεύτηκαν στους -80 ºC για βιοχημική ανάλυση και καθορισμό παραμέτρων του ορού και του πλάσματος του αίματος. [97]

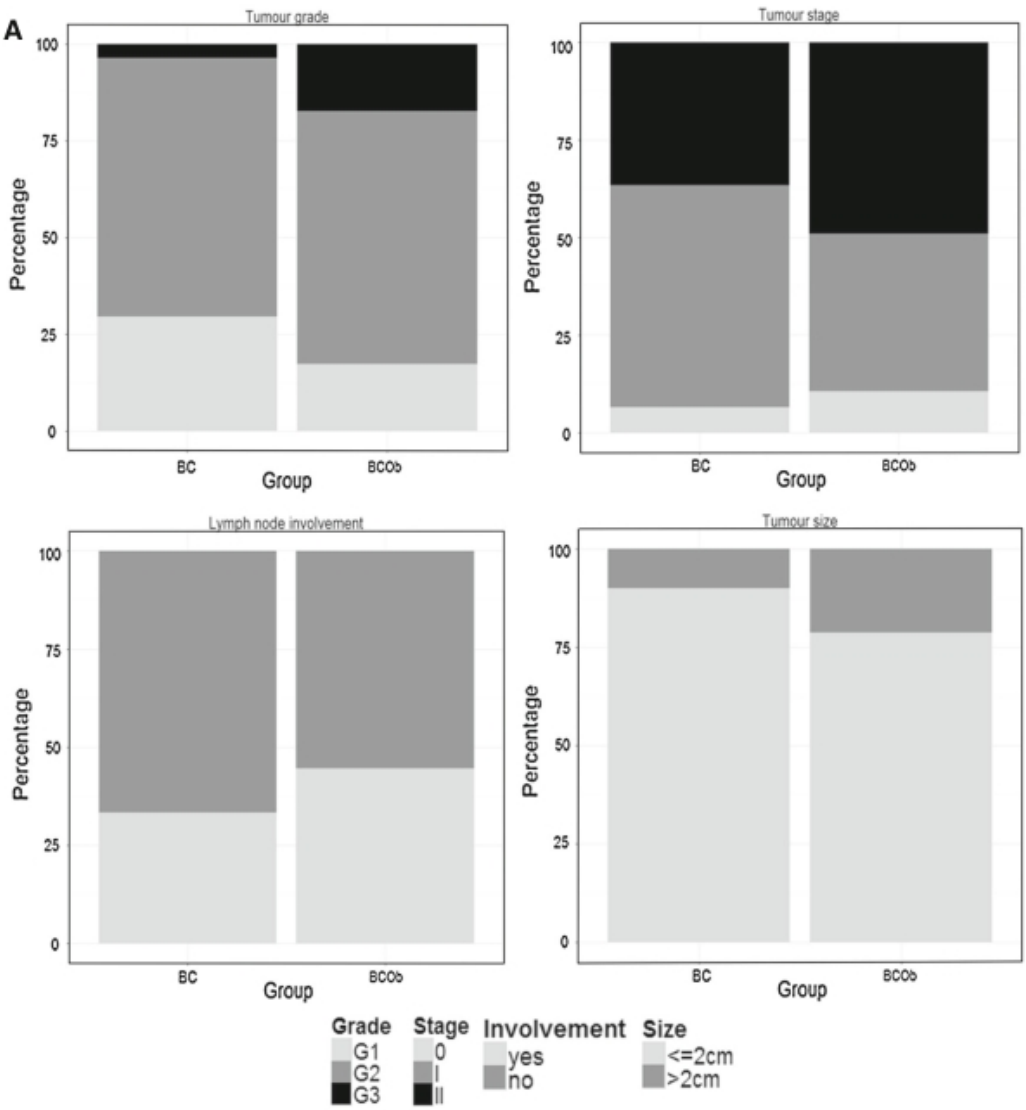
Οι τιμές για τις ελεγχόμενες λιποκίνες ορού (λεπτίνη, αντιπονεκτίνη, ρεζιστίνη) εκτιμήθηκαν με χρήση των ακόλουθων εμπορικών διαγνωστικών κιτ ανοσοενζυμικής μεθόδου ELISA : Duo Set ELISA Development System Human Leptin, Duo Set ELISA Development System Human Adiponectin, Duo Set ELISA Development System Human Resistin της R&D System, UK. Οι τιμές της μονοκυτταρικής χημειοτακτικής πρωτεΐνης 1 MCP-1 (Chemokine Monocyte Chemoattractant Protein 1) ορού εκτιμήθηκαν με χρήση του ακόλουθου εμπορικού διαγνωστικού κιτ ανοσοενζυμικής μεθόδου ELISA : Human MCP-1 ELISA Set της BD Biosciences Pharmingen, CA, EUA. Οι τιμές ινσουλίνης πλάσματος εκτιμήθηκαν επίσης με χρήση εμπορικού διαγνωστικού κιτ ανοσοενζυμικής μεθόδου ELISA, Mercodia Insulin ELISA της Mercodia AB, Sweden. Οι τιμές γλυκόζης πλάσματος εκτιμήθηκαν από αυτόματο αναλυτή με χρήση του εμπορικού κιτ : Olympus Diagnóstica Portugal της Produtos de Diagnóstico SA, Portugal.

Η αντίσταση ινσουλίνης εκτιμήθηκε μέσω του μοντέλο εκτίμησης ομοιόστασης (HOMA), το οποίο υπολογίστηκε ως ο λογάριθμος του λόγου του επιπέδου ινσουλίνης νηστείας σε μU/mL πολλαπλασιασμένου επί το επίπεδο της γλυκόζης νηστείας σε mmol/L προς 22.5 (log((Insulin x Glucose)/22.5)). Ο δείκτης μάζας σώματος (BMI) υπολογίστηκε ως ο λόγος του βάρους σε kg προς το τετράγωνο του ύψους σε μέτρα (kg/m^2). Τέλος, ο λόγος λεπτίνης / αντιπονεκτίνης (L/A) υπολογίστηκε ως ο λόγος της συγκέντρωσης λεπτίνης σε ng/ mL προς την συγκέντρωση αντιπονεκτίνης σε μg/mL. Ο τελευταίος προβλεπτικός παράγοντας προστέθηκε στους υπό εξέταση παράγοντες έχοντας λάβει υπόψιν τα αποτελέσματα της εργασίας των JG. Santillán-Benítez et al. [99].

Τέλος, η λήψη του ιστού του όγκου από τις ασθενείς έγινε είτε μέσω μαστεκτομής ή μέσω αφαίρεσης του όγκου. Η εκτίμηση του τύπο, του βαθμού και του μεγέθους του όγκου αλλά και η εξάπλωση στους λεμφαδένες έγιναν από Παθολογοανατόμο στο τμήμα Ανατομικής Παθολογίας του Πανεπιστημιακού Νοσοκομείο της Coimbra. Περαιτέρω

πληροφορίες για την ογκολογική εκτίμηση έχουν προηγουμένως αναφερθεί στην βιβλιογραφία, [97]. Σχετικά με τον τύπο των όγκων μπορούμε να συνοψίσουμε ότι η πλειονότητα των καρκινοπαθών συμμετεχόντων διαγνώστηκαν με διηθητικό πορογενή καρκίνωμα (invasive ductal carcinoma), το 80% των μη υπέρβαρων ασθενών και το 79% των υπέρβαρων, με BMI > 25 kg/m^2. Στην ομάδα των μη υπέρβαρων ασθενών (BC group), το 10% είχε μη διηθητικό πορογενές καρκίνωμα (ductal carcinoma in situ) και το 10% διηθητικό λοβιακό (invasive lobular carcinoma). Στην ομάδα των υπέρβαρων ασθενών (BCOb group), το 13% είχε μη διηθητικό πορογενές καρκίνωμα (ductal carcinoma in situ) ενὼ 4% διαγνώστηκαν με διηθητικό λοβιακό (invasive lobular carcinoma) και επίσης 4% με βλεννώδες διηθητικό καρκίνωμα. Άλλα χαρακτηριστικά που μετρήθηκαν είναι ο βαθμός διαφοροποίησης, το μέγεθος, η ύπαρξη καρκινικών κυττάρων στους λεμφαδένες (lymph node involvement) και το στάδιο στο οποίο βρίσκεται ο καρκίνος (Εικόνα 3.1). Εδώ παρατηρείται διαφορά στις ομάδες υπέρβαρων ή μη ασθενών, με την ομάδα των υπέρβαρων να παρουσιάζει μεγαλύτερα ποσοστά στις κατηγορίες αυξημένου κινδύνου, χωρίς όμως η στατιστική διαφορά μεταξύ των 2 ομάδων να χαρακτηρίζεται ως σημαντική. Πιο αναλυτικά, στην ομάδα των υπέρβαρων ασθενών έχουμε καρκινώματα $3^{ου}$ βαθμού σε ποσοστό 17%, μέγεθος όγκου μεγαλύτερο των 2 cm 21%, ύπαρξη καρκινικών κυττάρων στους λεμφαδένες σε ποσοστό 45% και στάδιο II καρκίνου 49%, με τα αντίστοιχα ποσοστά για την ομάδα μη υπέρβαρων ασθενών να είναι 3, 7, 33 και 34%. [97]

Μία παρατήρηση που θα μπορούσαμε να κάνουμε είναι ότι το στάδιο των καρκινωμάτων του δείγματός μας φαίνεται είναι κατά κύριο λόγο σε πρώιμο στάδιο, όντας το 56.25% των περιπτώσεων σε στάδιο 0 ή I, ενώ οι πιο προχωρημένες περιπτώσεις είναι σταδίου II, κάτι που σημαίνει ότι ο καρκίνος δεν έχει επεκταθεί σημαντικά και παραμένει στην περιοχή του μαστού. Επίσης, το μέγεθος των καρκινωμάτων είναι κατά κύριο λόγο μικρότερο των 2 cm, το ποσοστό των καρκινικών όγκων $3^{ου}$ βαθμού διαφοροποίησης είναι συγκριτικά σημαντικά μικρότερο από αυτό των $1^{ου}$ και $2^{ου}$ βαθμού και η ύπαρξη εξάπλωσης στους λεμφαδένες είναι σε ποσοστό κάτω του 50% και στις δύο ομάδες. Συνοψίζοντας, το δείγμα μας αποτελείται είτε από καρκίνο πρώιμου σταδίου ή από καρκίνο ο οποίος βρίσκεται ακόμα σε στάδιο κατά το οποίο, στην πλειονότητα των περιπτώσεων, ανταποκρίνεται θετικά στην θεραπεία.



Εικόνα 3.1 Ιστοπαθολογικά χαρακτηριστικά όγκων. [97]

3.2 Ανάλυση βιολογικής σημασίας Βιοδεικτών και Χαρακτηριστικών

Βιοδείκτης (biomarker) είναι ένας όρος που χρησιμοποιείται συχνά τόσο στην τοξικολογία όσο και στην ιατρική, και αναφέρεται σε μετρήσιμες ουσίες ενός βιολογικού συστήματος, που οι διαφορές στην συγκέντρωσή τους αντανακλά σε διαταραχές στην φυσιολογική λειτουργία του συστήματος. Ακολούθως αναλύεται η συσχέτιση των υπο μελέτη βιοδεικτών με την εμφάνιση του καρκίνου του μαστού.

3.2.1 Γλυκόζη

Η γλυκόζη είναι απλός μονοσακχαρίτης ο οποίος χρησιμοποιείται από τα κύτταρα ως την πρωταρχική πηγή ενέργειας και ως μέσο μεταβολισμού. Στον ανθρώπινο οργανισμό, η

γλυκόζη παράγεται από πρωτεΐνες, λίπη και κατά κύριο λόγο υδατάνθρακες. Ο οργανισμός ρυθμίζει αυστηρά τα επίπεδα γλυκόζης στο αίμα ως μέρος της μεταβολικής ομοιόστασης. Χαμηλότερα του φυσιολογικού επίπεδα γλυκόζης στο αίμα (υπογλυκαιμία) μπορεί να προκαλέσουν σύγχυση, άγχος, ή άλλες νευρολογικές διαταραχές. Υψηλότερα επίπεδα γλυκόζης αίματος (υπεργλυκαιμία) μπορεί να οδηγήσει σε γλυκοζυλίωση των ιστών του σώματος. Η υπεργλυκαιμία είναι χαρακτηριστική του σακχαρώδους διαβήτη.

Σύμφωνα με κλινικές μελέτες[101]-[103], τα αυξημένα επίπεδα γλυκόζης αίματος σχετίζονται με αυξημένο συνολικό κίνδυνο εμφάνισης καρκίνου κατά 20% έως 31%. Επιδημιολογικές μελέτες δείχνουν ότι αυξημένα επίπεδα της γλυκόζης ορού αυξάνουν τον κίνδυνο εμφάνισης του καρκίνου του μαστού. Σε μια μελέτη από τους Bi Y et al.[100] φαίνεται ότι οι γυναίκες με διαβήτη και μη φυσιολογικούς δείκτες γλυκόζης, οι οποίοι μετρήθηκαν 3, 5 ή και 10 χρόνια πριν από την διάγνωση καρκίνου του μαστού, είχαν λόγο εμφάνισης καρκίνου του μαστού 1,56 (95% CI: 1,21-2,00). Σε αντίστοιχη έρευνα από τους Parekh N et al.[104] φαίνεται ότι τα αυξημένα επίπεδα γλυκόζης σχετίζονται με αυξημένο κίνδυνο εμφάνισης καρκίνου του μαστού, αν και η συσχέτιση δεν ήταν στατιστικά σημαντική σε όλες τις περιπτώσεις που εξετάστηκαν. Η εργασία των Okumura M. Et al.[108] δείχνει ότι η συγκέντρωση γλυκόζης μπορεί να είναι ένας σημαντικός παράγοντας στον πολλαπλασιασμό των κυττάρων του καρκίνου του μαστού και ότι η εξάπλωση του καρκίνου του μαστού είναι υψηλή σε διαβητικούς ασθενείς. Τέλος, οι J. Crisóstomo et al.[97] παρατήρησαν ότι οι υπέρβαρες γυναίκες με καρκίνο του μαστού είχαν ένα ανεπιθύμητο μεταβολικό προφίλ το οποίο περιλάμβανε διαταραχή του μεταβολισμού της γλυκόζης.

3.2.2 Αντιποκίνες

Οι αντιποκίνες είναι ορμόνες οι οποίες αποτελούνται από πολυπεπτίδια όπως και άλλα μη πρωτεϊνικά προϊόντα και εκφράζονται κυρίως στο λιπώδη ιστό. Η πρώτη αντιποκίνη που ανακαλύφθηκε είναι η λεπτίνη το 1994 και ακολούθησε η ανακάλυψη της αντιπονεκτίνης το 1995. Όλες οι αντιποκίνες εμπλέκονται σημαντικά με την ρύθμιση της όρεξης και του κορεσμού, στην κατανομή του λίπους, στην ευαισθησία στην ινσουλίνη, στην κατανάλωση ενέργειας, στη φλεγμονή, στην αρτηριακή πίεση στην αιμόσταση και στην ενδοθηλιακή λειτουργία. Έρευνες υποδεικνύουν ότι οι περισσότερες αντιποκίνες προάγουν την εξέλιξη των καρκινικών κυττάρων μέσω της ενίσχυσης των κυτταρικών πολλαπλασιασμών και των μεταναστευτικών, φλεγμονωδών και αντί-αποπτωτικών οδών, οι οποίες στη συνέχεια μπορούν να προκαλέσουν καρκινικές μεταστάσεις.

3.2.2.1 Λεπτίνη

Η λεπτίνη εκκρίνεται κυρίως από το λιπώδη ιστό, άλλα βρίσκεται και σε χαμηλότερα ποσοστά στον πλακούντα, στις ωοθήκες, στους σκελετικούς μύες, στη γαστρική βλεννώδη μεμβράνη, στον υποθάλαμο και στο μαστικό επιθήλιο. Η λεπτίνη έχει φλεγμονώδη δράση, διαδραματίζει σημαντικό ρόλο στην ενεργειακή ομοιόσταση και έχει διαπιστωθεί ότι περιορίζει την όρεξη, αυξάνει την κατανάλωση της ενέργειας και είναι σημαντικός ρυθμιστής της ευαισθησίας του οργανισμού στην ινσουλίνη και του μεταβολικού ρυθμού του οργανισμού. Αύξηση των επιπέδων λεπτίνης σχετίζονται με αύξηση του δείκτη μάζας σώματος (BMI). Μείωση ή απώλεια της λειτουργίας της λεπτίνης οδηγεί στην παχυσαρκία. [105]-[106]

Η σύνδεση της λεπτίνης με τον καρκίνο έχει συσχετιστεί με τον πολλαπλασιασμό, την κυτταρική επιβίωση, την αγγειογένεση και την επακόλουθη ανάπτυξη του καρκίνου.[110] Η λεπτίνη είναι απαραίτητη για την ανάπτυξη του φυσιολογικού μαστικού αδένα σε τρωκτικά και ανθρώπους, και έχει ανιχνευτεί σε μαστικά επιθηλιακά κύτταρα, καρκινικά και μη, αλλά η έκφρασή της φαίνεται να είναι κατά πολύ μεγαλύτερη στα καρκινικά κύτταρα.[107], [109] Έρευνες υποδεικνύουν ότι η λεπτίνη είναι συσχετισμένη με την καρκινογένεση στον μαστικό ιστό, μέσω της λειτουργίας της ως αυξητικής ορμόνης.[108] Στα κύτταρα του καρκίνου του μαστού, έχει αποδειχθεί ότι η ινσουλίνη διεγείρει την έκφραση του υποδοχέα της λεπτίνης. Με τον τρόπο αυτό η αύξηση της συγκέντρωσης της ινσουλίνης σε συνδυασμό με την λεπτίνη λειτουργούν αθροιστικά προκαλώντας πρόοδο του καρκίνου.[111] Οι ασθενείς με καρκίνο του μαστού με υψηλά επίπεδα λεπτίνης ορού είναι πιο επιρρεπείς στην περαιτέρω ανάπτυξη των καρκινικών κυττάρων.[112]

3.2.2.2 Αντιπονεκτίνη

Η αντιπονεκτίνη εκκρίνεται κυρίως από τα λιποκύτταρα σαν απάντηση σε μεταβολικούς παράγοντες προκειμένου να ευαισθητοποιήσει το ήπαρ και τους μυς στη δράση της ινσουλίνης και διαδραματίζει σημαντικό ρόλο στην ενεργειακή ομοιόσταση. Υπάρχει μια αντίστροφη σχέση της αντιπονεκτίνης καθώς και των υποδοχέων της, με την αντίσταση στην ινσουλίνη. Τα επίπεδα της αντιπονεκτίνης σχετίζονται αντίστροφα με το δείκτη μάζας σώματος (BMI), το ποσοστό του λίπους στον οργανισμό, την ινσουλίνη νηστείας και τα επίπεδα τριγλυκεριδίων αίματος. Η αντιπονεκτίνη έχει την δυνατότητα να αυξάνει την επεξεργασία των λιπαρών οξέων από λεπτούς ιστούς και να καταστέλλει την παραγωγή της γλυκόζης από το ήπαρ, με συνέπεια τη μείωση της γλυκόζης του αίματος και του σωματικού βάρους. Μελέτες έχουν δείξει ότι η αντιπονεκτίνη έχει αντιφλεγμονώδεις δράσεις και

αναστέλλει τον κυτταρικό πολλαπλασιασμό και την μετανάστευση ενδοθηλιακών κυττάρων, εμποδίζοντας το σχηματισμό νεόπλαστου έσω χιτώνα μετά από τραυματισμό των αγγείων. [105]-[106]

Υπάρχουν δύο τρόποι με τους οποίους η ορμόνη μπορεί να προκαλέσει επιβράδυνση του καρκίνου: είτε άμεσα στα καρκινικά κύτταρα, καθώς πολλές καρκινικές γραμμές εκφράζουν υποδοχείς αντιπονεκτίνης ή μέσω της ευαισθητοποίησης των κυττάρων στην ινσουλίνη. Σε κάθε περίπτωση, η μείωση της συγκέντρωσης της αντιπονεκτίνης έχει συσχετιστεί με καρκίνο του μαστού, του ενδομητρίου, του οισοφάγου και του ήπατος μεταξύ πολλών άλλων. Η δράση της αντιπονεκτίνης στις γραμμές καρκίνου του μαστού όπου τα καρκινικά κύτταρα εκφράζουν τους υποδοχείς της περιλαμβάνει τη μείωση της εισβολής του καρκίνου σε άλλα κύτταρα. Η έρευνα δείχνει ότι η αντιπονεκτίνη είναι αντί-καρκινογόνος σε καρκινικά κύτταρα του μαστού, συμπεριλαμβανομένων των MCF-7, MDA-MB-231 και T47D μέσω των αντί-πολλαπλασιαστικών ιδιοτήτων της. Και στις τρεις καρκινικές κυτταρικές γραμμές που αναφέρθηκαν παραπάνω, η αντιπονεκτίνη αυξάνει την ενεργοποίηση της απόπτωσης των κυττάρων και αναστέλλει τον κυτταρικό ρυθμιστικό κύκλο. [112]

3.2.2.3 Ρεζιστίνη

Η ρεζιστίνη είναι μία πρωτεΐνη η οποία ανακαλύφθηκε το 2001 και παράγεται στα λιποκύτταρα, στα μυϊκά κύτταρα, στα παγκρεατικά κύτταρα, στα μονοπύρηνα κύτταρα, όπως και στα μακροφάγα. Ονομάστηκε ρεζιστίνη λόγω της αντίστασης στην ινσουλίνη που εμφάνιζαν ποντίκια που είχαν ενεθεί με αυτή. Η δράση της στον άνθρωπο φαίνεται να εμπλέκεται στην ανοσία, στη φλεγμονή και στην ινσουλινοαντίσταση. Η έκφραση της συνδέεται με το ποσοστό του λίπους και είναι αυξημένη σε καταστάσεις παχυσαρκίας και διαβήτη, ενώ εμπλέκεται και στην ευαισθησία στην ινσουλίνη, αφού προκαλεί ηπατική ινσουλινοαντίσταση και επηρεάζει των μεταβολισμό των λιπαρών οξέων από τους σκελετικούς μύες. Η ρεζιστίνη είναι ικανή να εμποδίσει την διαφοροποίηση των προ-λιποκυττάρων. Υπάρχουν ακόμα πολλά κενά σχετικά με τη σημασία και τον ακριβή ρόλο της ρεζιστίνης στο σώμα και η έρευνα συνεχίζεται. [105]-[106]

Πρόσφατες μελέτες σε ανθρώπους έχουν δείξει ότι σε αντίθεση με τα τρωκτικά, η ρεζιστίνη εκφράζεται σε μεγάλο βαθμό σε περιφερειακά μονοπύρηνα κύτταρα, ενώ εκφράζεται ελάχιστα σε λιποκύτταρα και προαδιποκύτταρα. Βασισμένοι στο γεγονός ότι η ρεζιστίνη εκφράζεται περισσότερο σε ανοσιακά κύτταρα που διεισδύουν στον λιπώδη ιστό, οι Lerkhe et al. [113] έδειξαν ότι τα επίπεδα της ρεζιστίνης πιθανότατα σχετίζονται με την

φλεγμονώδη κατάσταση του ατόμου. Για το λόγο αυτό έχει υποτεθεί ότι η ρεζιστίνη μπορεί να είναι μία από τις αντιποκίνες που επηρεάζουν έντονα την ανάπτυξη του καρκίνου. Έχει αποδειχθεί ότι η ρεζιστίνη μπορεί να αποτελεί τον σύνδεσμο μεταξύ παχυσαρκίας και μίας αυξημένης φλεγμονώδους κατάστασης και της επίδρασης που έχει αυτή η φλεγμονή στην ανάπτυξη όγκων. [112] Οι J. Crisóstomo et al.[97] καταλήγουν ότι φαίνεται να υπάρχει ισχυρή σχέση μεταξύ της ρεζιστίνης και της φλεγμονής και της επιρροής της στην καρκινογένεση, ενώ συμπεραίνουν ότι τα υψηλά επίπεδα ρεζιστίνης που βρέθηκαν στην ομάδα υπέρβαρων ασθενών με καρκίνου του μαστού δεν μπορούσε να σχετίζεται μόνο με το σωματικό βάρος, υποδεικνύοντας συσχέτιση της υπερεζιστιναιμίας με τον καρκίνο του μαστού. Τέλος, μελέτες δείχνουν την ύπαρξη συσχέτισης μεταξύ των επιπέδων ρεζιστίνης και την ύπαρξη καρκίνου του μαστού στις μετ εμμηνοπαυσιακές γυναίκες[114],[115], αλλά και συσχέτιση της ρεζιστίνης με δείκτες φλεγμονής, το μέγεθος του όγκου, τον βαθμό και το στάδιο του καρκίνου.[114]

3.2.2.4 MCP-1

Η μονοκυτταρική χημειοτακτική πρωτεΐνη 1 MCP-1 (Chemokine Monocyte Chemoattractant Protein 1) είναι μία χημειοκίνη η οποία στρατολογεί πρόδρομα μακροφάγα (μονοκύτταρα) στο λιπώδη ιστό σε καταστάσεις παχυσαρκίας, όπου η έκφραση της είναι αυξημένη. Έχει αποδειχθεί ότι παίζει σημαντικό ρόλο στη ρύθμιση της φλεγμονής, ρυθμίζοντας τη διακίνηση μονοκυττάρων / μακροφάγων κατά την επούλωση πληγών, των λοιμώξεων, αυτοάνοσων παθήσεων και του καρκίνου.[119]

Οι Lebrecht A et al.[117] αναφέρουν ότι η MCP-1 δεν φαίνεται να λειτουργεί ως δείκτης διαφοροποίησης μεταξύ κακοήθων και καλοήθων όγκων του μαστού. Τα δεδομένα όμως ενδέχεται να υποδεικνύουν ότι η MCP-1 επηρεάζει την καρκινογένεση του μαστού διευκολύνοντας την ανάπτυξη του όγκου και την μεταστατική εξάπλωση, τροποποιώντας έτσι τον βιολογικό φαινότυπο της νόσου. Επίσης, σύμφωνα με τους Dutta P. Et al. [118] τα υψηλά επίπεδα της MCP-1 σχετίζεται με τον τριπλά αρνητικό καρκίνο του μαστού και έχει τη δυνατότητα να επηρεάσει την εισβολή των κυττάρων ενεργοποιώντας το μονοπάτι της MAP κινάσης με αυτοκρινή τρόπο. Επίσης, οι J. Crisóstomo et al.[97] συμπεραίνουν ότι η MCP-1 αποτελεί σημαντικό σύνδεσμο μεταξύ ρεζιστίνης και καρκίνου του μαστού.

3.2.3 Ινσουλίνη

Η ινσουλίνη είναι μια ορμόνη που αποτελείται από 51 αμινοξέα και παράγεται από ειδικά κύτταρα του παγκρέατος. Η ορμόνη αυτή ρυθμίζει το μεταβολισμό των υδατανθράκων, των

λιπών και των πρωτεϊνών, προωθώντας την απορρόφηση των υδατανθράκων, και ιδιαίτερα της γλυκόζης, από το αίμα στο ήπαρ, το λίπος και τα σκελετικά μυϊκά κύτταρα.

Η ινσουλίνη, μέσω της ενεργοποίησης υποδοχέων ινσουλίνης που εκφράζονται σε φυσιολογικά αλλά και σε καρκινικά κύτταρα, μπορεί να ενισχύσει την αναβολική κατάσταση η οποία είναι απαραίτητη για την κυτταρική ανάπτυξη, αυξάνοντας την διαθεσιμότητα ουσιών, όπως η γλυκόζη και τα αμινοξέα, στα κύτταρα. Με τον τρόπο αυτό η ινσουλίνη μπορεί να παίξει σημαντικό ρόλο στην ανάπτυξη του καρκίνου αλλά και στην πρόκληση μίας φαινοτυπικής τροποποίησης του καρκινώματος μετατρέποντας το σε περισσότερο κακοήθες.[122] Τα αυξημένα επίπεδα ορού ινσουλίνης φαίνεται να έχουν σημαντική συσχέτιση με τον κίνδυνο εμφάνισης καρκίνου του μαστού.[120] Μελέτες ενισχύουν την άποψη ότι η ινσουλίνη έχει δυνητικά άμεσες επιπτώσεις στο αποτέλεσμα του καρκίνου του μαστού. Σε μια ομάδα γυναικών με καρκίνο του μαστού σε πρώιμη φάση, οι οποίες δεν είχαν προηγουμένως διαγνωστεί με σακχαρώδη διαβήτη, τα επίπεδα ινσουλίνης νηστείας συσχετίστηκαν με μετάσταση του καρκίνου σε μακρινά όργανα και θάνατο.[121]

3.2.4 Μοντέλο Εκτίμησης Ομοιόστασης (HOMA)

Το Μοντέλο Εκτίμησης της Ομοιόστασης (HOMA) είναι ένα μοντέλο υπολογισμού της ινσουλινοαντίστασης και υπολογίζεται ως ο λογάριθμος του λόγου του επιπέδου ινσουλίνης νηστείας σε μU/mL πολλαπλασιασμένου επί το επίπεδο της γλυκόζης νηστείας σε mmol/L προς 22.5 (log((Insulin x Glucose)/22.5)).

Μελέτες δείχνουν ότι η αντίσταση στην ινσουλίνη, όπως εκτιμάται από το HOMA, μπορεί να είναι ένας δείκτης κινδύνου εμφάνισης του καρκίνου του μαστού[125], και θα μπορούσε να είναι χρήσιμος σε προληπτικά προγράμματα για την ανίχνευση του καρκίνου του μαστού.[126] Επίσης, υψηλές τιμές του HOMA σε γυναίκες με καρκίνο του μαστού σχετίζονται με δυσμενή πρόγνωση για εμφάνιση μεταστατικού καρκίνου[126], ενώ η αντίσταση στην ινσουλίνη συνδέεται, σε γυναίκες με καρκίνο του μαστού σε πρώιμο στάδιο, και με αυξημένο κίνδυνο επανεμφάνισης και θανάτου, ακόμη και αν δεν υπάρχει διαβήτης.[128]

3.2.5 Λόγος Λεπτίνης / Αντιπονεκτίνης (L/A-ratio)

Ο λόγος λεπτίνης / αντιπονεκτίνης υπολογίστηκε ως ο λόγος της συγκέντρωσης λεπτίνης σε ng/mL προς την συγκέντρωση αντιπονεκτίνης σε μg/mL. Λόγο του αντικαρκινικού ρόλου της αντιπονεκτίνης και της συσχέτισης της λεπτίνης με την ανάπτυξη του καρκίνου του μαστού, ο λόγος λεπτίνης / αντιπονεκτίνης αναμένεται ότι μπορεί να παίξει σημαντικό ρόλο

στην εκτίμηση του κινδύνου ανάπτυξης του καρκίνου του μαστού.[129]-[131] Μάλιστα υπάρχουν έρευνες στις οποίες αναφέρεται ότι ο λόγος L/A μπορεί να είναι πιο σημαντικός για την ανάπτυξη του καρκίνου του μαστού από τις απόλυτες τιμές των συγκεντρώσεων αυτών των δύο αντιποκινών[130] , και μελέτες έδειξαν ότι ο λόγος L/A είναι σημαντικά αυξημένος στους ασθενείς με καρκίνο του μαστού σε σύγκριση με την υγιεί ομάδα ελέγχου. [129],[131]

3.2.6 Δείκτης Μάζας Σώματος (BMI)

Πολλές μελέτες έχουν αποδείξει ότι η παχυσαρκία αποτελεί παράγοντα κινδύνου για την ανάπτυξη καρκίνου του μαστού σε μετ εμμηνοπαυσιακές γυναίκες. Συγκεκριμένα, ένα μεγάλο εύρος στοιχείων υποδηλώνει ότι η μετ εμμηνοπαυσιακή παχυσαρκία, που αξιολογείται από τον BMI, συνδέεται με αυξημένο κίνδυνο εμφάνισης ορμονο-εξαρτώμενου καρκίνου του μαστού. Η επίπτωση του υπερβολικού σωματικού βάρους στον καρκίνο του μαστού στις γυναίκες πριν από την εμμηνόπαυση εξακολουθεί να είναι ασαφής. Παρόλα αυτά δεδομένα δείχνουν ότι η παχυσαρκία του ανώτερου σώματος αυξάνει τον κίνδυνο καρκίνου του μαστού, ανεξάρτητα από το εάν η γυναίκα βρίσκεται πριν ή μετά την εμμηνόπαυση. Επιπλέον, ο υψηλός BMI είναι σημαντικά συσχετισμένος με αυξημένο κίνδυνο εμφάνισης του φλεγμονώδους καρκίνου του μαστού τόσο σε προ εμμηνοπαυσιακούς όσο και σε μετ εμμηνοπαυσιακούς πληθυσμούς.[110]

3.2.7 Ηλικία

Η ηλικία είναι ο μεγαλύτερος παράγοντας κινδύνου ανάπτυξης καρκίνου. Μεταλλάξεις και άλλες αλλαγές στο γονιδίωμα αποτελούν τη βασική αιτία της εμφάνισης καρκινικών κυττάρων στον οργανισμό. Κατά τον πολλαπλασιασμό των κυττάρων, η διαδικασία της αντιγραφής του DNA μπορεί να περιλαμβάνει τυχαία σφάλματα τα οποία αποτελούν μεταλλάξεις. Επίσης, όσο μεγαλύτερη είναι η ηλικία ενός ανθρώπου τόσο παρατείνεται η έκθεσή του σε μεταλλαξιογόνους παράγοντες. Ως αποτέλεσμα, κατά την διάρκεια της ζωής τα κύτταρα συσσωρεύουν μεταλλάξεις, αυξάνοντας με τον τρόπο αυτό την πιθανότητα εμφανίσεως μεταλλάξεων οι οποίες οδηγούν σε εμφάνιση καρκίνου. Μεταλλάξεις που διαταράσσουν γονίδια που ρυθμίζουν την κυτταρική διαίρεση και ανάπτυξη μπορεί να αποτελέσουν έναυσμα ώστε κάποια κύτταρα να αρχίσουν να αναπτύσσονται και να πολλαπλασιάζονται ανεξέλεγκτα σχηματίζοντας όγκους. Επιπρόσθετες μεταλλάξεις μπορεί να απενεργοποιήσουν της πρωτεΐνες που καταστέλλουν τον όγκο, βοηθώντας περαιτέρω την εξέλιξη και την επιθετικότητα του όγκου. Επιπλέον, κατά την γήρανση συμβαίνουν αλλαγές στους ιστούς και τα όργανα που καθιστούν το μικροπεριβάλλον των κυττάρων πιο ευνοϊκό

για την ανάπτυξη του καρκίνου. Άλλοι παράγοντες για τους οποίους η εμφάνιση καρκίνου είναι συσχετισμένη με τη γήρανση περιλαμβάνουν τις μακροπρόθεσμες επιδράσεις της χρόνιας φλεγμονής, τις καρκινογόνες μεταλλάξεις που προκαλούνται από ελεύθερες ρίζες οξυγόνου, τον λιγότερο αποτελεσματικό μηχανισμό αποκατάστασης βλαβών του DNA και της εξασθένησης του ανοσοποιητικού συστήματος έχοντας ως αποτέλεσμα την λιγότερο αποτελεσματική ανίχνευση και αντιμετώπιση των καρκινικών κυττάρων.

Η πιθανότητα εμφάνισης του καρκίνου του μαστού αυξάνει με την ηλικία, και από την ηλικία των 45 ετών αρχίζει να είναι υψηλή. Κατά τα χρόνια πριν την εμμηνόπαυση, ο ρυθμός αύξησης της εμφάνισης του καρκίνου του μαστού είναι περίπου 8 με 9% τον χρόνο. Αυτός ο ρυθμός αύξησης της εμφάνισης διατηρείται σε αυτά τα επίπεδα καθ όλη την διάρκεια της ζωής έως την εμμηνόπαυση. Μετά την εμμηνόπαυση ο ρυθμός αύξησης μειώνεται σε 2 με 3% τον χρόνο. Η συσχέτιση του ρυθμού της εμφάνισης του καρκίνου του μαστού με την εμμηνόπαυση υποδεικνύει τον ρόλο των αναπαραγωγικών ορμονών στην αιτιολογία της ασθένειας.[123]-[124]

3.3 Στατιστική ανάλυση Βιοδεικτών και Χαρακτηριστικών

Κατά την στατιστική ανάλυση που πραγματοποιήθηκε σε προηγουμένη μελέτη από τους M. Patrício et al.[98] εκτιμήθηκαν η μέση (mean) και η ενδιάμεση (median) τιμή, το ενδοτεταρτημοριακό εύρος (interquartile range) και η τιμή σημαντικότητας (p-value) για κάθε έναν από τα τους υποψήφιους βιοδείκτες, και παρατηρήθηκε ότι δεν υπάρχει σημαντική στατιστική διαφορά στην ηλικία και τον δείκτη μάζας σώματος μεταξύ των δύο ομάδων των συμμετεχόντων, αφού έχουμε τιμή σημαντικότητας και στις δύο περιπτώσεις μεγαλύτερη του 0,05 . Ως προς τους μεταβολικούς παράγοντες, σημαντική στατιστική διαφορά παρουσιάζουν οι : γλυκόζη, ρεζιστίνη, ινσουλίνη και το μοντέλο εκτίμησης ομοιόστασης HOMA έχοντας όλες τους μεγαλύτερη μέση τιμή για την ομάδα των ασθενών σε σχέση με την υγιή ομάδα ελέγχου, ενώ οι : λεπτίνη, αντιπονεκτίνη και MCP-1 δεν παρουσιάζουν σημαντική στατιστική διαφορά ανάμεσα στις δύο ομάδες (Πίνακας 3.1).

Τέλος, ο λόγος L/A δεν παρουσιάζει σημαντική διαφορά μεταξύ των δύο ομάδων, καθώς είναι έχει μέση τιμή 4,05 για την υγιεί ομάδα ελέγχου και 3,91 για την ομάδα με καρκίνο του μαστού.

Πίνακας 3.1 Η μέση τιμή μαζί με το ενδοτεταρτημοριακό εύρος, σε παρένθεση, και η τιμή σημαντικότητας (p-value) για κάθε ένα από τα χαρακτηριστικά. [98]

	Patients	Controls	*p*-value
Age (years)	53.0 (23.0)	65.0 (33.2)	0.479
BMI (kg/m^2)	27.0 (4.6)	28.3 (5.4)	0.202
Glucose (mg/dL)	105.6 (26.6)	88.2 (10.2)	<0.001
Insulin (μU/mL)	12.5 (12.3)	6.9 (4.9)	0.027
HOMA	3.6 (4.6)	1.6 (1.2)	0.003
Leptin (ng/mL)	26.6 (19.2)	26.6 (19.3)	0.949
Adiponectin (μg/mL)	10.1 (6.2)	10.3 (7.6)	0.767
Resistin (ng/mL)	17.3 (12.6)	11.6 (11.4)	0.002
MCP-1(pg/dL)	563.0 (384.0)	499.7 (292.2)	0.504

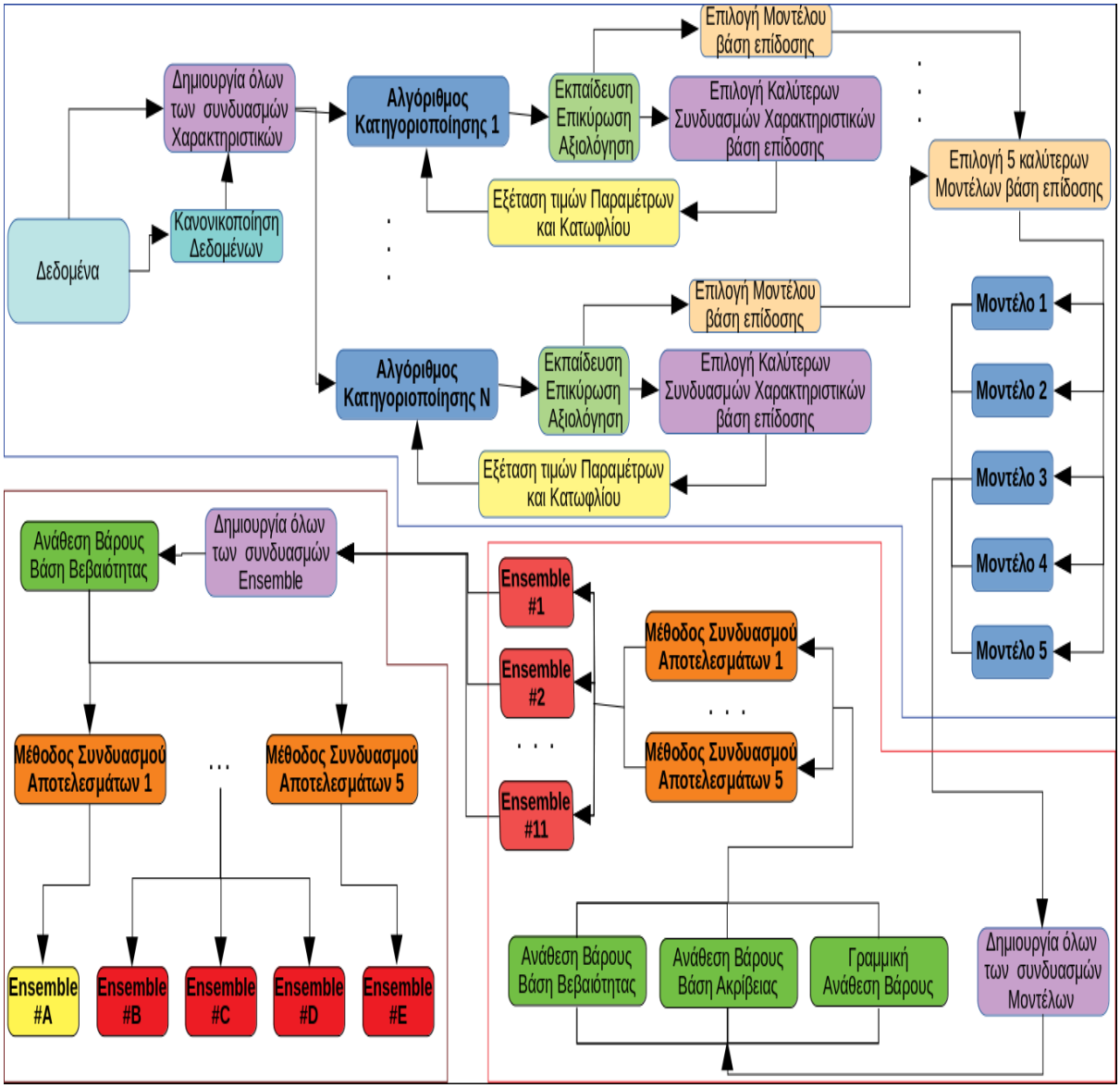
Παρατηρούμε εδώ ότι η μέση τιμή του λόγου L/A είναι μεγαλύτερη για την ομάδα ελέγχου κάτι το οποίο φαίνεται ασυνεπές με την βιολογική ερμηνεία του λόγου L/A. Συν αυτού, το γεγονός ότι η μέσες τιμές Αντιπονεκτίνης και Λεπτίνης δεν παρουσιάζουν σημαντική απόκλιση μεταξύ των δύο ομάδων φαίνεται παράξενο βάση του ρόλου των δύο αυτών ορμονών στην ανάπτυξη και εξέλιξη του καρκίνου του μαστού. Παρόλα αυτά, το γεγονός ότι η μέση τιμή του BMI είναι μεγαλύτερη στην ομάδα ελέγχου (28,3) σε σχέση με την ομάδα των ασθενών (27,0) δείχνει προς μία εξήγηση σε αυτά τα παράδοξα. Μειωμένα επίπεδα Αντιπονεκτίνης σχετίζονται με παχυσαρκία και ύπαρξη καρκίνου, ενώ φυσιολογικά επίπεδα Αντιπονεκτίνης σχετίζονται με φυσιολογικό BMI και μειωμένο κίνδυνο ανάπτυξης καρκίνου, κάτι που οδηγεί σε ισορροπία των επιπέδων Αντιπονεκτίνης μεταξύ των δύο ομάδων. Με παρόμοια λογική μπορούμε να εξηγήσουμε και την ισορροπία στα επίπεδα της Λεπτίνης μεταξύ των ομάδων, καθώς υψηλά επίπεδα Λεπτίνης σχετίζονται με παχυσαρκία και ύπαρξη καρκίνου, ενώ φυσιολογικά επίπεδα Λεπτίνης σχετίζονται με φυσιολογικό BMI και μειωμένο κίνδυνο ανάπτυξης καρκίνου. Βάση των παραπάνω, εξηγούνται και οι μέσες τιμές του λόγου L/A. Στα πλαίσια της εργασίας αυτής δεν θεωρείται σκόπιμο να αναλύσουμε περαιτέρω τις παραπάνω σχέσεις, ενώ είναι γνωστό ότι δεν αποτελούν τις μοναδικές σχέσεις αλληλεξάρτησης των υπό εξέταση χαρακτηριστικών, και ότι το πρόβλημα της εξήγησης αυτών των τιμών είναι πολύ πιο περίπλοκο από τον τρόπο με τον οποίο παρουσιάστηκε. Ο

λόγος της παραπάνω παρατήρησης είναι ώστε να επισημανθεί ότι το γεγονός ότι υπάρχει ομοιότητα των μέσων τιμών και ότι δεν υπάρχει σημαντική στατιστική διαφορά σε ορισμένα χαρακτηριστικά μεταξύ των δύο ομάδων δεν σημαίνει ότι τα χαρακτηριστικά αυτά δεν συνεχίζουν να έχουν διακριτική ικανότητα μεταξύ των ομάδων όταν συνδυαστούν με άλλα χαρακτηριστικά.

3.4 Μεθοδολογία

Για την υλοποίηση του τελικού προτεινόμενου προβλεπτικού μοντέλου συλλογικής μάθησης ακολουθήθηκε η διαδικασία που περιγράφεται σχηματικά στο παρακάτω διάγραμμα ροής (Εικόνα 3.2).

Αρχικά δημιουργήθηκαν μοντέλα μηχανικής μάθησης ελέγχοντας όλους τους συνδυασμούς βιοδεικτών ως χαρακτηριστικά εισόδου για κάθε έναν από τους εξεταζόμενους αλγορίθμους κατηγοριοποίησης. Στην συνέχεια προχωρήσαμε σε βελτιστοποίηση των παραμέτρων των 16 μοντέλων με την καλύτερη απόδοση για κάθε έναν από τους αλγορίθμους κατηγοριοποίησης. Ταυτόχρονα εξετάστηκαν και διαφορετικές τιμές κατωφλίου. Μετά την ολοκλήρωση της φάσης της βελτιστοποίησης παραμέτρων τα μοντέλα με την καλύτερη προβλεπτική ικανότητα για κάθε αλγόριθμο συγκρίθηκαν μεταξύ τους βάση μετρικών απόδοσης καταλήγοντας σε 5 μοντέλα μηχανικής μάθησης βέλτιστης απόδοσης. Εξετάζοντας όλους τους συνδυασμούς των εξόδων των 5 αυτών μοντέλων μηχανικής μάθησης, εφαρμόζοντας τεχνικές συλλογικής μάθησης, και επιλέγοντας το μοντέλο με την βέλτιστη απόδοση για κάθε μέθοδο συνδυασμού των αποτελεσμάτων και ανάθεσης βάρους, καταλήξαμε στην δημιουργία 11 μοντέλων συλλογικής μάθησης. Τελικά, τα αποτελέσματα των 11 μοντέλων συλλογικής μάθησης βέλτιστης απόδοσης συνδυάστηκαν εκ νέου εφαρμόζοντας τεχνικές συλλογικής μάθησης ώστε να καταλήξουμε στο προτεινόμενο μοντέλο συλλογικής μάθησης Ensemble A.



Εικόνα 3.2 Διάγραμμα Ροής υλοποίησης μοντέλου συλλογικής μηχανικής μάθησης.

Η ανάπτυξη του κώδικα για την υλοποίηση όλων των μοντέλων μηχανικής μάθησης, έγινε σε γλώσσα προγραμματισμού Python 3.7.3, πάνω σε λειτουργικό σύστημα Ubuntu 18.04.3 LTS.

3.4.1 Αλγόριθμοι Κατηγοριοποίησης

Στο πλαίσιο της εργασίας υλοποιήθηκαν μοντέλα επιβλεπόμενης μηχανικής μάθησης βασισμένα στους εξής αλγορίθμους κατηγοριοποίησης :

3.4.1.1 k-Nearest Neighbors (k-NN)

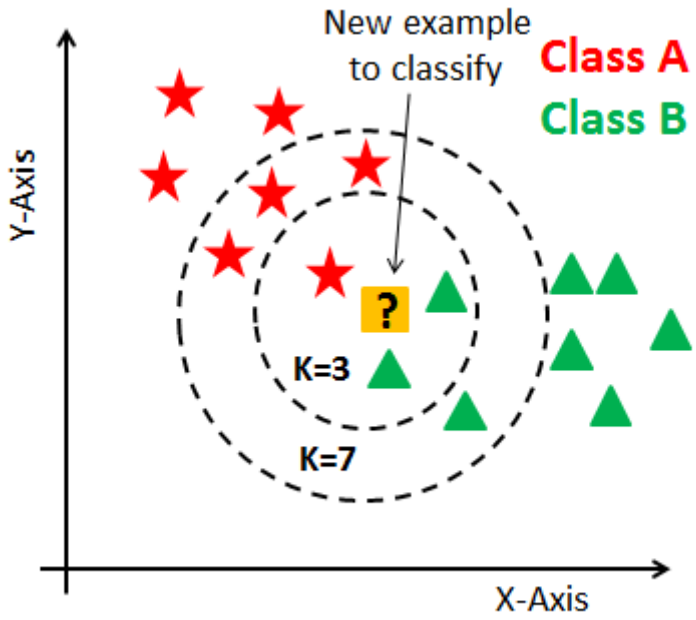
Η μέθοδος k-Nearest Neighbors (k-NN) [135] είναι ένας αλγόριθμος της οικογένειας των Κατηγοριοποιητών Βασισμένων σε Στιγμιότυπα (Instance Based Classifiers, IBC). Οι κατηγοριοποιητές της οικογένειας IBC δεν έχουν στάδιο εκπαίδευσης και δεν παράγουν κάποιο μοντέλο, μέχρι να χρειαστεί να κατηγοριοποιηθεί ένα στιγμιότυπο. Όταν χρειαστεί να κατηγοριοποιήσουν ένα νέο στιγμιότυπο, το συγκρίνουν με όλα τα στιγμιότυπα του συνόλου εκπαίδευσης, των οποίων η κατηγορία είναι γνωστή. Αυτό απαιτεί την αποθήκευση όλου, ή τουλάχιστον μέρους, του συνόλου εκπαίδευσης. Για τον λόγο αυτό, οι κατηγοριοποιητές IBC καλούνται και «οκνηροί» (lazy classifiers).

Ο k-NN θεωρηθεί κάθε στιγμιότυπο ως ένα σημείο στον χώρο ν διαστάσεων, όπου ν είναι ο αριθμός των χαρακτηριστικών. Μέσα στον ν-διάστατο χώρο, το στιγμιότυπο Χ απέχει από ένα άλλο στιγμιότυπο Υ κάποια απόσταση d(X,Y). Ο k-NN υπολογίζει, κάθε φορά που θα χρειαστεί να κατηγοριοποιήσει, την απόσταση d(X,Y) του υπό κατηγοριοποίηση στιγμιότυπου Χ από όλα τα στιγμιότυπα Υ που ανήκουν στο σύνολο εκπαίδευσης. Η απόσταση d(X,Y) μπορεί να υπολογιστεί ως η απόσταση Minkowski η οποία ορίζεται από την εξίσωση :

$$d_p = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} \qquad (3.4.1.1.1)$$

Η απόσταση Minkowski για p = 1 είναι η απόσταση Manhattan και για p = 2 είναι η Ευκλείδεια απόσταση.

Στον αλγόριθμο k-NN, ο k αποτελεί μία παράμετρο, η τιμή της οποίας καθορίζετε από τον χρήστη, και συμβολίζει τον αριθμό των πλησιέστερων γειτόνων του νέου στιγμιότυπου οι οποίοι θα ληφθούν υπόψη για την κατηγοριοποίησή του. Ο αλγόριθμος αναζητά μέσα στον ν-διάστατο χώρο τα k στιγμιότυπα που βρίσκονται πλησιέστερα στο νέο στιγμιότυπο. Ο κατηγοριοποιητής καταχωρεί το νέο στιγμιότυπο στην κλάση που πλειοψηφεί μεταξύ των k πλησιέστερων γειτόνων. Όπως φαίνεται στην Εικόνα 3.3 διαφορετικές τιμές του k μπορεί να συνεπάγονται διαφορετικές κατηγοριοποιήσεις.



Εικόνα 3.3 Παράδειγμα λειτουργίας ενός κατηγοριοποιητή k-NN για διαφορετικές τιμές του k. [134]

Υπάρχουν διάφορες παραμετροποιήσεις του k-NN, όπως για παράδειγμα η απόφαση κατηγοριοποίησης να μην λαμβάνεται από ισότιμη ψηφοφορία μεταξύ των επιλεγμένων πλησιέστερων γειτόνων, αλλά από αυτούς να συνεισφέρουν περισσότερο τα σημεία τα οποία είναι πλησιέστερα στο νέο στιγμιότυπο. Ένας απλός τρόπος για να επιτευχθεί αυτό είναι να εκχωρηθούν συντελεστές βαρύτητας ψήφου στα επιλεγμένα σημεία. Επίσης, μπορούν να εκχωρηθούν συντελεστές βαρύτητας και μεταξύ των χαρακτηριστικών, μετατρέποντας έτσι την (3.4.1.1) σε :

$$d_p = \left(\sum_{i=1}^{n} w_i * \left| x_i - y_i \right|^p \right)^{1/p} \quad (3.4.1.1.2)$$

Κατά την διαδικασία βελτιστοποίησης παραμέτρων και επιλογής χαρακτηριστικών για την επιλογή του μοντέλου k-NN με την μέγιστη διακριτική ικανότητα δοκιμάστηκαν όλοι οι συνδυασμοί χαρακτηριστικών ως είσοδοι, ενώ οι τιμές των παραμέτρων κυμάνθηκαν ως εξής :

n_neighbors	3 – 10
leaf_size	1 – 40
weights	uniform και distance
algorithm	auto, ball_tree, kd_tree και brute
p	1 – 5

Καταλήγοντας στο συμπέρασμα ότι το μοντέλο k-NN με την καλύτερη διακριτική ικανότητα δέχεται ως εισόδους τα χαρακτηριστικά :

Ηλικία, BMI, Γλυκόζη, Ρεζιστίνη και Αντιπονεκτίνη

έχοντας λαμβάνοντας υπόψιν 5 γείτονες (n_neighbors = 5), leaf_size = 30, algorithm = auto με βάρη βασισμένα σε απόσταση (weights = distance), metric = minkowski και p = 1.

3.4.1.2 Naive Bayes (NB)

Η απλοϊκή κατηγοριοποίηση κατά Bayes (Naive Bayes)[136] βασίζεται στον κανόνα του Bayes. Στον αλγόριθμο Naive Bayes θεωρείται ότι η συνεισφορά όλων των χαρακτηριστικών των δεδομένων εκπαίδευσης είναι ανεξάρτητη και ότι το κάθε ένα από τα χαρακτηριστικά συνεισφέρει εξίσου στην διαδικασία της κατηγοριοποίησης. Ο όρος Naive, αφελής, οφείλεται σε αυτή την θεώρηση της μη ύπαρξης συσχέτισης μεταξύ των τιμών των διαφόρων χαρακτηριστικών.

Ο κανόνας του Bayes μαθηματικά ορίζεται ως εξής [132]: Δεδομένης μιας υπόθεσης h_k από ένα σύνολο m πιθανών υποθέσεων και μιας ομάδας χαρακτηριστικών $X = (x_1, x_2, ..., x_n)$, η πιθανότητα εμφάνισης της υπόθεσης h_k δεδομένου ενός χαρακτηριστικού x_i είναι :

$$P(h_k|x_i) = \frac{P(x_i|h_k)P(h_k)}{P(x_i)} \quad \text{όπου} \quad P(x_i) = \sum_{j=1}^{m} P(x_i|h_j)P(h_j) \qquad (3.4.1.2.1)$$

Το $P(h_k|x_i)$ ονομάζεται εκ των υστέρων πιθανότητα, ενώ το $P(h_k)$ είναι η εκ των προτέρων πιθανότητα που σχετίζεται με την υπόθεση h_k. Το $P(x_i)$ εκφράζει την πιθανότητα το δεδομένο με τιμή x_i να πραγματοποιηθεί και το $P(x_i|h_k)$ είναι η υπό συνθήκη πιθανότητα να ικανοποιείται από το υπό εξέταση στιγμιότυπο η δεδομένη υπόθεση.

Δοθέντος ενός συνόλου εκπαίδευσης, ο αλγόριθμου Naive Bayes αρχικά εκτιμά την εκ των προτέρων πιθανότητα $P(C_j)$ για κάθε κατηγορία, μετρώντας πόσο συχνά εμφανίζεται κάθε κατηγορία στα δεδομένα εκπαίδευσης. Για κάθε χαρακτηριστικό x_i , μπορεί να υπολογιστεί ο αριθμός των εμφανίσεων κάθε τιμής του χαρακτηριστικού αυτού, προκειμένου να καθοριστεί η αντίστοιχη πιθανότητα $P(x_i)$. Με τον τρόπο αυτό υπολογίζεται η πιθανότητα $P(x_i|C_j)$, μετρώντας πόσο συχνά εμφανίζεται κάθε τιμή στην εξεταζόμενη κατηγορία στα δεδομένα εκπαίδευσης. Η διαδικασία αυτή επαναλαμβάνεται για κάθε χαρακτηριστικό και για κάθε τιμή των χαρακτηριστικών. Αφού ολοκληρωθεί ο υπολογισμός αυτών των πιθανοτήτων, τότε οι ίδιες αυτές πιθανότητες μπορούν να χρησιμοποιηθούν για την κατηγοριοποίηση ενός νέου στιγμιοτύπου.

Για ένα νέο στιγμιότυπο t_i με p διαφορετικές τιμές χαρακτηριστικών (x_{i1}, x_{i2}, ... , x_{ip}), τότε γνωρίζοντας ήδη την πιθανότητα $P(x_{ik}|C_j)$ για κάθε κατηγορία C_j και γνώρισμα x_{ik}, μπορεί να υπολογιστεί η $P(t_i|C_j)$ από την σχέση :

$$P(t_i|C_j) = \prod_{k=1}^{p} P(x_{ik}|C_j) \qquad (3.4.1.2.2)$$

Ο υπολογισμός της $P(t_i)$ γίνεται αθροίζοντας τις επιμέρους τιμές πιθανοφάνειας το υπό εξέταση στιγμιότυπο να ανήκει σε καθεμία από τις κατηγορίες. Τέλος, η εκ των υστέρων πιθανότητα $P(C_j|t_i)$, δηλαδή η πιθανότητα το στιγμιότυπο t_i να ανήκει στην κατηγορία C_j, υπολογίζεται για κάθε κατηγορία και το υπό εξέταση στιγμιότυπο κατηγοριοποιείται στην κλάση με την μεγαλύτερη πιθανότητα $P(C_j|t_i)$.

Ένα πλεονέκτημα του αλγορίθμου Naive Bayes είναι ότι τα μοντέλα κατηγοριοποίησης που βασίζονται σε αυτόν χρειάζονται έναν μικρό σχετικά αριθμό στιγμιοτύπων για εκπαίδευση για να κάνουν εκτίμηση των απαραίτητων παραμέτρων για την κατηγοριοποίηση. Συνεπώς, η εκπαίδευσή τους ολοκληρώνεται γρήγορα σε σύγκριση με άλλους κατηγοριοποιητές και έχουν την δυνατότητα να αποδώσουν αρκετά καλά χωρίς την ύπαρξη πολλών δεδομένων για την εκπαίδευσή τους. Αντιθέτως, η απόδοση του αλγορίθμου δεν είναι καλή όταν του χορηγηθεί μεγάλος αριθμός χαρακτηριστικών.

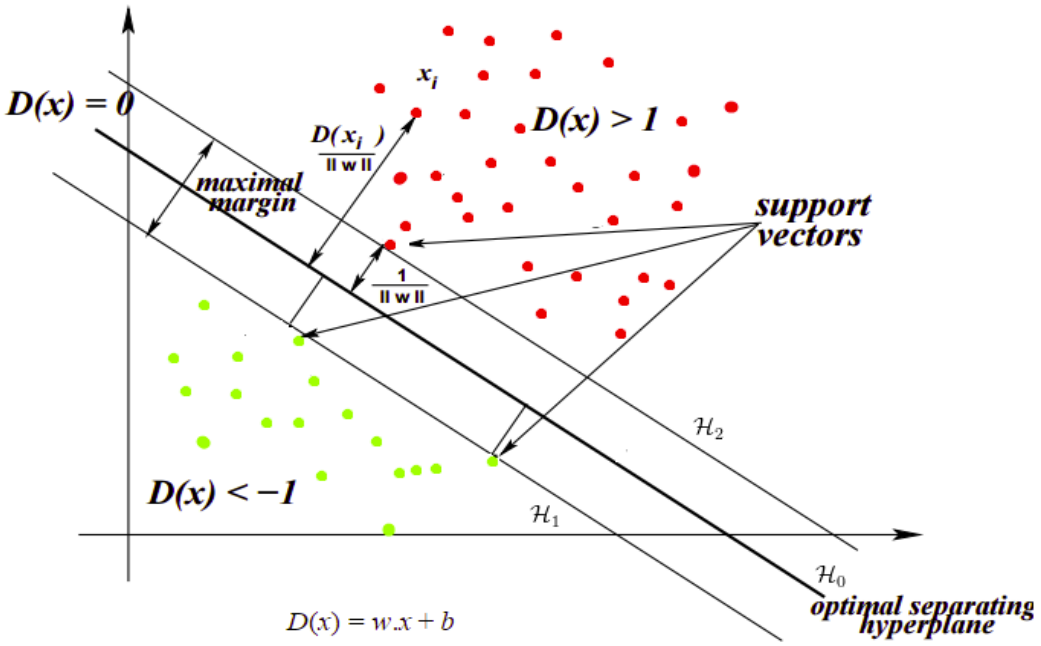
Κατά την αρχική δοκιμή των μοντέλων Naive Bayes δοκιμάστηκαν όλοι οι συνδυασμοί χαρακτηριστικών ως είσοδοι, ενώ ως τιμές παραμέτρων διατηρήθηκαν οι προεπιλογές του μοντέλου GaussianNB της βιβλιοθήκης sklearn. Η διακριτική ικανότητα όλων των παραγόμενων μοντέλων μέσω του αλγορίθμου NB υπολείπονταν σημαντικά έναντι των υπόλοιπων αλγορίθμων. Για τον λόγο αυτό δεν υπήρξε διαδικασία βελτιστοποίησης παραμέτρων και επιλογής χαρακτηριστικών για τον αλγόριθμο NB.

3.4.1.3 Support Vector Machines (SVM)

Ο αλγόριθμος SVM[137] λειτουργεί αναπαριστώντας το σύνολο των δεδομένων εκπαίδευσης ως σημεία σε κάποιο χώρο. Η τοπολογία των σημείων αυτών καθορίζεται έτσι ώστε τα στιγμιότυπα που ανήκουν σε διαφορετικές κατηγορίες να διαχωρίζονται μεταξύ τους χωρικά κατά όσο το δυνατών μεγαλύτερη απόσταση. Ο χώρος που διαχωρίζει τα σημεία διαφορετικών κατηγοριών μεταξύ τους ονομάζεται υπέρ-επιφάνεια απόφασης (separating hyperplane).

Κατά την φάση της εκπαίδευσης, ο SVM προσπαθεί να εντοπίσει την υπέρ-επιφάνεια βέλτιστου διαχωρισμού των διαφορετικής κατηγορίας δεδομένων, η οποία

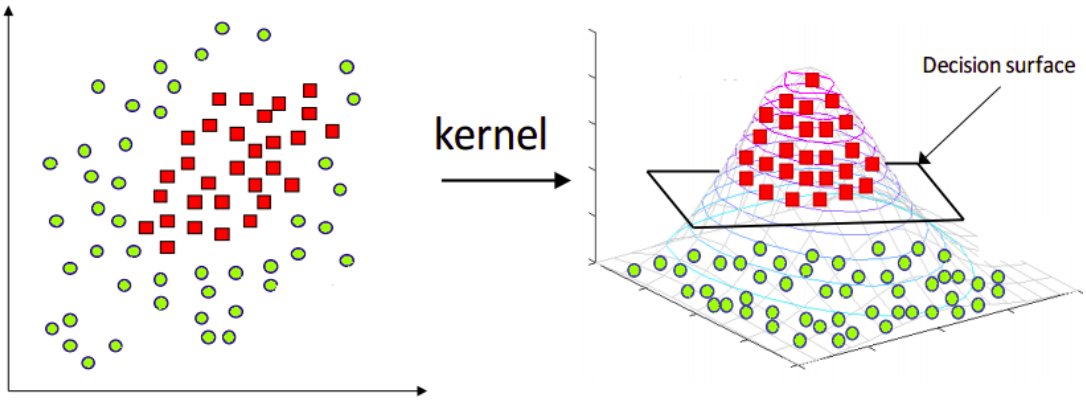
ορίζεται ως η υπέρ-επιφάνεια μέγιστου εύρους (maximum margin hyperplane). Ουσιαστικά προσπαθεί να εντοπίσει την υπέρ-επιφάνεια που διαχωρίζει τα δεδομένα του συνόλου



Εικόνα 3.4 Δυαδική κατηγοριοποίηση μέσω SVM. [133]

εκπαίδευσης με τη μεγαλύτερη δυνατή απόσταση. Για να επιτύχει αυτό τον στόχο, ο SVM επιλέγει παράλληλες υπέρ-επιφάνειες (tangential hyperplane), οι οποίες είναι τέτοιες ώστε ανάμεσά τους να μην υπάρχουν δεδομένα του συνόλου εκπαίδευσης και άρα να το διαχωρίζουν τέλεια. Στις παράλληλες υπέρ-επιφάνειες αυτές εφάπτονται κάποια από τα στιγμιότυπα της κάθε διαφορετικής κατηγορίας, μία κατηγορία στιγμιοτύπων ανά υπέρ-επιφάνεια, τα οποία ονομάζονται support vectors. Η απόσταση των παράλληλων υπέρ-επιφανειών είναι το εύρος (margin), το οποίο και ο αλγόριθμος επιχειρεί να μεγιστοποιήσει. Η υπέρ-επιφάνεια απόφασης ορίζεται στο μέσο των παράλληλων υπέρ-επιφανειών μεγίστου εύρους (Εικόνα 3.4).

Η κατηγοριοποίηση των στιγμιοτύπων κατά την φάση της επικύρωσης γίνεται με γνώμονα σε ποια πλευρά της υπέρ-επιφάνεια απόφασης εκπίπτει. Αν και η υπέρ-επιφάνεια αυτή είναι γραμμική ο αλγόριθμος SVM μπορεί να πετύχει την κατηγοριοποίηση και μη γραμμικά διαχωρίσιμων δεδομένων, αναπαριστώντας τα σε ένα χώρο διαφορετικής διάστασης από τον αρχικό, στον οποίο μπορεί να ορίσει μία γραμμική υπέρ-επιφάνεια η οποία να διαχωρίζει τα δεδομένα. Αυτή η τεχνική αναφέρεται και ως μέθοδος πυρήνα ή kernel trick (Εικόνα 3.5) .



Εικόνα 3.5 Παράδειγμα εφαρμογής της μεθόδου του πυρήνα.[133]

Στην συγκεκριμένη εργασία δοκιμάστηκαν 4 (‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’) διαφορετικοί πυρήνες (kernels) που προσφέρονται ως επιλογή στην βιβλιοθήκη Scikit_Learn.

- Γραμμικός Πυρήνας (linear kernel) :

$$K\left(\vec{x}_j,\vec{x}_i\right)=\left(\vec{x}_i\cdot\vec{x}_j\right) \quad (3.4.1.3.1)$$

- Πολυωνυμικός Πυρήνας (polynomial ‘poly’ kernel) :

$$K\left(\vec{x}_j,\vec{x}_i\right)=\left(\overrightarrow{\gamma\cdot x}_i\cdot\vec{x}_j+1\right)^d,\gamma>0 \quad (3.4.1.3.2)$$

- Πυρήνας ακτινωτής βάσης (Radial Basis Function ‘rbf’ kernel):

$$K\left(\vec{x}_j,\vec{x}_i\right)=\exp\left(-\gamma\cdot\left\|\vec{x}_i-\vec{x}_j\right\|^2\right),\gamma>0 \quad (3.4.1.3.3)$$

- Σιγμοειδής Πυρήνας (sigmoid kernel) :

$$K\left(\vec{x}_j,\vec{x}_i\right)=\tan h\left(\gamma\cdot\vec{x}_i\cdot\vec{x}_j+r\right) \quad (3.4.1.3.4)$$

Κατά την διαδικασία βελτιστοποίησης παραμέτρων και επιλογής χαρακτηριστικών για την επιλογή του μοντέλου SVM με την μέγιστη διακριτική ικανότητα δοκιμάστηκαν όλοι οι συνδυασμοί χαρακτηριστικών ως είσοδοι και όλοι οι διαφορετικοί πυρήνες. Καθώς μοντέλα με πυρήνα Γραμμικό, Πολυωνυμικό και Ακτινωτής Βάσης έφεραν ικανοποιητικά αποτελέσματα με τις προεπιλεγμένες τιμές χαρακτηριστικών εφαρμόστηκε διαδικασία

βελτιστοποίησης παραμέτρων και επιλογής χαρακτηριστικών και για τον αλγόριθμο SVM και με τους 3 αυτούς πυρήνες. Οι τιμές των παραμέτρων κυμάνθηκαν ως εξής:

	gamma	$2^{-14} – 2^{8}$, auto & scale
	C	$2^{-10} – 2^{9}$
	degree	2 – 4 (μόνο για kernels : Poly & Linear)

Τα μοντέλα με την καλύτερη διακριτική ικανότητα με κάθε διαφορετικό πυρήνα αλγορίθμου SVM είχαν και διαφορετικό συνδιασμό χαρακτηριστικών και παραμέτρων. Έτσι τα βέλτιστα μοντέλα ανά πυρήνα είναι τα εξής :

Γραμμικός Πυρήνας (linear kernel)

Χαρακτηριστικά :	Ηλικία, BMI, Γλυκόζη, Ρεζιστίνη, Λεπτίνη και Ινσουλίνη	
Παράμετροι :	gamma	0.000244140625
	C	1.2
	degree	3

Πολυωνυμικός Πυρήνας (polynomial ‘poly’ kernel)

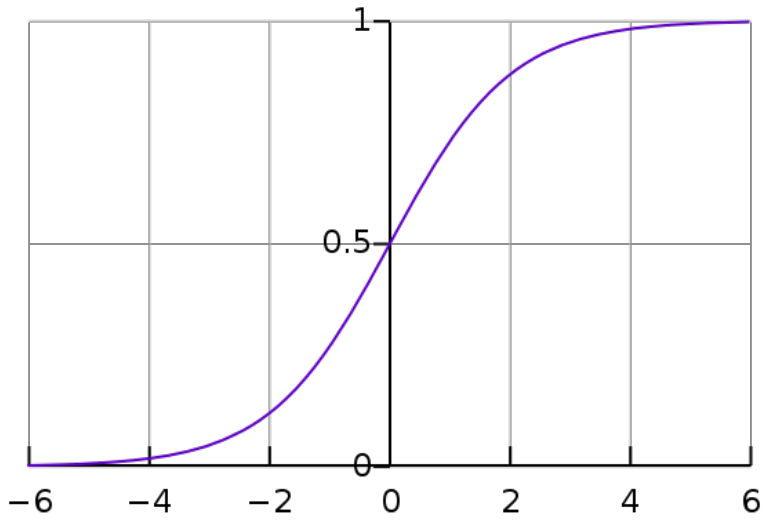
Χαρακτηριστικά :	Ηλικία, BMI, Γλυκόζη, Ρεζιστίνη, Λόγος Λεπτίνης/Αντιπονεκτίνης και Ινσουλίνη	
Παράμετροι :	gamma	0.0009765625
	C	2
	degree	2

Πυρήνας ακτινωτής βάσης (Radial Basis Function ‘rbf’ kernel)

Χαρακτηριστικά :	Ηλικία, BMI, Γλυκόζη και Ρεζιστίνη	
Παράμετροι :	gamma	0.00038
	C	0.99
	degree	3 (προεπιλογή)

3.4.1.4 Λογιστική Παλινδρόμηση (Logistic Regression, LR)

Η λογιστική παλινδρόμηση[138] χρησιμοποιείται για την πρόβλεψη της πιθανότητας εμφάνισης ενός γεγονότος προσαρμόζοντας τα δεδομένα της μελέτης στην εξίσωση της σιγμοειδούς καμπύλης (Εικόνα 3.6).



Εικόνα 3.6 Γραφική παράσταση της σιγμοειδούς συνάρτησης.

Η δυαδική λογιστική παλινδρόμηση αποτελεί μια διωνυμική εξίσωση στην οποία η μεταβλητή απόκρισης Y είναι το τυχαίο αποτέλεσμα εμφάνισης μιας από δύο δυνητικές εκβάσεις του τύπου επιτυχία ή αποτυχία. Η εξίσωση της δυαδικής λογιστικής παλινδρόμησης έχει την μορφή :

$$f(z)=\frac{e^z}{1+e^z}=\frac{1}{1+e^{-z}} \quad (3.4.1.5)$$

όπου z είναι η μεταβλητή εισόδου και f(z) το αποτέλεσμα αυτής. Στα πλεονεκτήματα της εξίσωσης συγκαταλέγεται και το γεγονός ότι η μεταβλητή εισόδου λαμβάνει θετικές και αρνητικές τιμές ενώ το αποτέλεσμα αυτής f(z) περιορίζεται σε εύρος τιμών μεταξύ 0 και 1.

Αναλυτικότερα, η μεταβλητή z εκπροσωπεί τη δράση μιας ομάδας ανεξάρτητων χαρακτηριστικών, ενώ η f(z) προσδιορίζει την πιθανότητα το συγκεκριμένο στιγμιότυπο να ανήκει σε μία από δύο κλάσεις. Η μεταβλητή z εκφράζει επίσης το μέτρο της ολικής συνεισφοράς όλων των χαρακτηριστικών στο μοντέλο και ορίζεται ως :

$$z = \beta_0+\beta_1X_1+\beta_2X_2+ \dots +\beta_vX_v \quad (3.4.1.5)$$

όπου β_0 είναι το ύψος της κλίσης της γραμμής παλινδρόμησης και ισούται με την τιμή z όταν οι τιμές όλων των χαρακτηριστικών ισούνται με 0, ενώ β_i είναι οι συντελεστές

παλινδρόμησης καθένας εκ των οποίων εκφράζει το βαθμό συνεισφοράς του αντίστοιχου χαρακτηριστικού. Θετική τιμή του συντελεστή δηλώνει ότι το χαρακτηριστικό αυξάνει την πιθανότητα το στιγμιότυπο να ανήκει στην κλάση 1 (θετική έκβαση), ενώ αρνητική τιμή σημαίνει ότι το χαρακτηριστικό μειώνει την πιθανότητα το στιγμιότυπο να ανήκει στην κλάση 1, άρα αυξάνει την πιθανότητα να ανήκει στην κλάση 0 (αρνητική έκβαση). Υψηλή τιμή του συντελεστή σημαίνει ότι το χαρακτηριστικό επηρεάζει πολύ ισχυρά την πιθανότητα το στιγμιότυπο να ανήκει σε μία από τις κλάσεις, ενώ χαμηλή τιμή δηλώνει μικρή επίδραση του στιγμιότυπου στην πιθανότητα αυτή.

Κατά την φάση της εκπαίδευσης του, ένα μοντέλο λογιστικής παλινδρόμησης ρυθμίζει τους συντελεστές του βαθμού συνεισφοράς των χαρακτηριστικών έτσι ώστε να μεγιστοποιήσει την ακρίβεια κατηγοριοποίησης.

Κατά την διαδικασία βελτιστοποίησης παραμέτρων και επιλογής χαρακτηριστικών για την επιλογή του μοντέλου Λογιστικής Παλινδρόμισης με την μέγιστη διακριτική ικανότητα δοκιμάστηκαν όλοι οι συνδυασμοί χαρακτηριστικών ως είσοδοι, ενώ οι τιμές των παραμέτρων κυμάνθηκαν ως εξής :

solver	lbfgs και liblinear
C	0.001 – 1000

Καταλήγοντας στο συμπέρασμα ότι το μοντέλο Λογιστικής Παλινδρόμισης με την καλύτερη διακριτική ικανότητα δέχεται ως εισόδους τα χαρακτηριστικά :

Ηλικία, BMI, Γλυκόζη, Ρεζιστίνη και HOMA

χρησιμοποιόντας τον αλγόριθμο βελτιστοποίησης lbfgs (Large-scale Bound-constrained Optimization) και C = 0.18 .

3.4.1.5 Random Forest (RF)

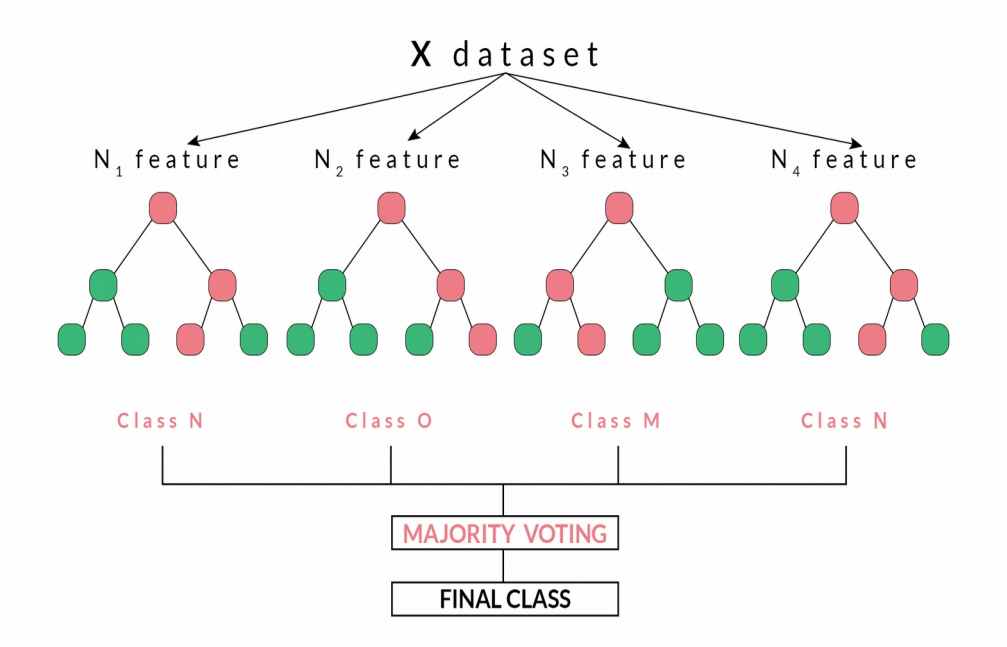
Ο αλγόριθμος Random Forest[139] είναι ένας αλγόριθμος που βασίζεται στην λογική της συλλογικής μάθησης (ensemble learning). Βάση της μεθοδολογίας του ensemble learning, μπορούμε να κατασκευάσουμε ισχυρά προβλεπτικά μοντέλα, συνδυάζοντας απλούστερες μεθόδους προβλέψεων. Υπάρχουν διαφορετικές προσεγγίσεις εφαρμογής του ensemble learning. Ο αλγόριθμος Random Forest κάνει χρήση της μεθόδου Bootstrap aggregating ή Bagging[140] ώστε να δημιουργήσει πολλαπλά, τυχαία υποσύνολα του συνόλου εκπαίδευσης και από κάθε ένα από αυτά αναπτύσσει ένα δέντρο απόφασης. Στη συνέχεια, για κάθε δέντρο ο Random Forest επιλέγει ένα τυχαίο υποσύνολο των χαρακτηριστικών, από το σύνολο των χαρακτηριστικών της εισόδου, το οποίο χρησιμοποιείται για τον διαχωρισμό (split) κάθε κόμβου σε ένα δέντρο κατά την φάση της ανάπτυξης, σε αντίθεση με τον

αλγόριθμο Bagging όπου όλα τα χαρακτηριστικά θεωρούνται για τον διαχωρισμό ενός κόμβου. Με τον τρόπο αυτό αποτρέπεται η πιθανότητα τα δέντρα που αναπτύσσονται να είναι συσχετισμένα μεταξύ τους, όπως μπορεί να συμβεί κατά την λειτουργία του Bagging, όπου μία μικρή ομάδα ισχυρών χαρακτηριστικών μπορεί να υπερκεράσει τα υπόλοιπα με συνέπεια να επιλέγεται στην πλειονότητα των δεντρων.

Ο αλγόριθμος Random Forest δημιουργεί ένα δάσος δέντρων απόφασης :

$$\{h(\mathbf{x}, \Theta_k) , k = 1, 2, ...\} \quad (3.4.1.6)$$

όπου τα Θ_k είναι ομοιόμορφα κατανεμημένα, ανεξάρτητα μεταξύ τους, τυχαία διανύσματα χαρακτηριστικών και το **x** αποτελεί το διάνυσμα που αντιστοιχεί στο υπό κατηγοριοποίηση στιγμιότυπο εισόδου. Αφού ένας μεγάλος αριθμός δέντρων καταλήξουν σε απόφαση για την κλάση στην οποία ανήκει το στιγμιότυπο **x**, προχωρούν σε ψηφοφορία (voting) ώστε να αποφασιστεί ποια είναι η δημοφιλέστερη κλάση για το στιγμιότυπο αυτό, η οποία είναι και η τελική πρόβλεψη του Random Forest για την κλάση του **x**.



Εικόνα 3.7 Παράδειγμα λειτουργίας του αλγορίθμου Random Forest.[141]

Ο Random Forest είναι πολύ ανθεκτικός στο overfitting, λόγο του ότι βασίζει την απόφασή του σε πολύ μεγάλο αριθμό δέντρων απόφασης και της τυχαιότητας και της ανεξαρτησίας του διανύσματος των χαρακτηριστικών που επιδρούν σε κάθε ξεχωριστό δέντρο. Επίσης, έχει καλή απόδοση και με μη ισορροπημένα ή/και ελλείποντα δεδομένα. Τέλος, λόγο του γεγονότος ότι για κάθε δέντρο χρησιμοποιείται μόνο κάποιο υποσύνολο των

χαρακτηριστικών εισόδου, ο Random Forest έχει μειωμένο χρόνο εκπαίδευσης σε σύγκριση με αλγορίθμους Bagging.

Κατά την διαδικασία βελτιστοποίησης παραμέτρων και επιλογής χαρακτηριστικών για την επιλογή του μοντέλου Random Forest με την μέγιστη διακριτική ικανότητα δοκιμάστηκαν όλοι οι συνδυασμοί χαρακτηριστικών ως είσοδοι, ενώ οι τιμές των παραμέτρων κυμάνθηκαν ως εξής :

n_estimators	4 – 200
max_depth	3 – 25
min_samples_split	2 – 10
min_samples_leaf	0.0001 – 0.076
max_features	0.5 – 1 & auto, sqrt, None

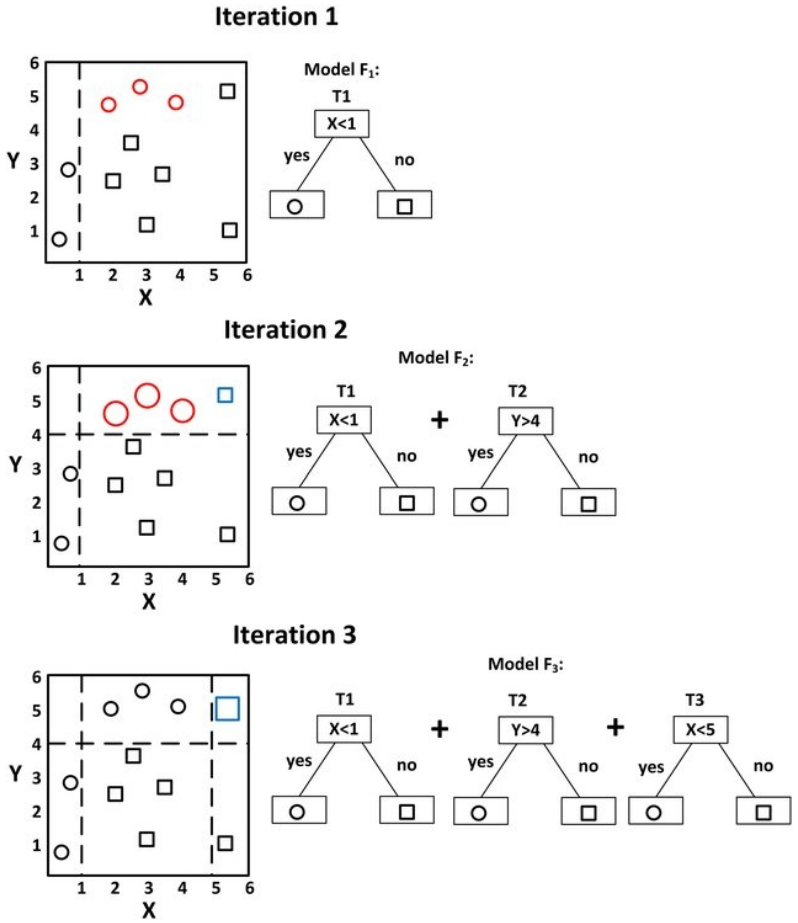
Καταλήγοντας στο συμπέρασμα ότι το μοντέλο Random Forest με την καλύτερη διακριτική ικανότητα δέχεται ως εισόδους τα χαρακτηριστικά :

Ηλικία, BMI, Γλυκόζη, Ρεζιστίνη, Λεπτίνη και Αντιπονεκτίνη

χρησιμοποιόντας 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 0.001, 'min_samples_split': 7 και 'n_estimators': 18.

3.4.1.6 XGBoost (XGB)

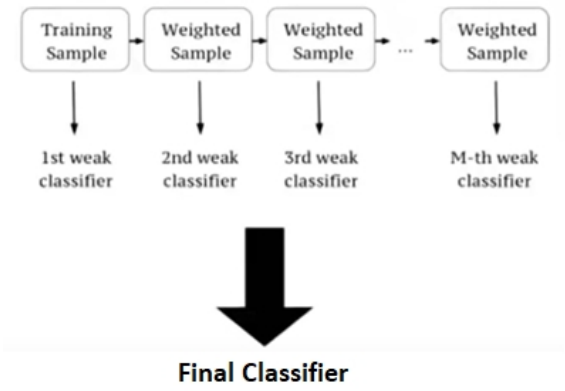
Ο αλγόριθμος XGBoost (Extreme Gradient Boosting)[142] είναι ένας ακόμα αλγόριθμος που βασίζεται στην λογική της συλλογικής μάθησης (ensemble learning). Η ensemble learning προσέγγιση που έχει επιλεγεί στην περίπτωση του XGBoost είναι το Boosting[143]. Ο XGBoost ουσιαστικά αποτελεί εξέλιξη του Gradient Boosting[144], ο οποίος κάνει χρήση της μεθόδου βελτιστοποίησης Gradient Descent ώστε να ελαχιστοποιήσει τα σφάλματα σε μία συλλογή αδύναμων μοντέλων τα οποία δημιουργούνται σειριακά. (Εικόνα 3.8)



Εικόνα 3.8 Παράδειγμα λειτουργίας του Gradient Boosting.[147]

Σύμφωνα με την μέθοδο ensemble learning, μία συλλογή από αδύναμους ταξινομητές (weak learners) μπορεί να αποτελέσει την βάση για την δημιουργία ενός ισχυρού ταξινομητή. Ως αδύναμος ταξινομητής ορίζεται κάθε ταξινομητής του οποίου η απόδοση είναι έστω και ελάχιστα καλύτερη από αυτή της τυχαίας επιλογής. Σε αντίθεση με την μέθοδο Bagging, κατά την οποία οι αδύναμοι ταξινομητές δημιουργούνται παράλληλα ώστε να καταλήξουν σε αποφάσεις οι οποίες θα συμμετέχουν σε ψηφοφορία για την επιλογή της επικρατούσας, στην μέθοδο Boosting οι αδύναμοι ταξινομητές δημιουργούνται σειριακά. Στην περίπτωση του XGBoost οι αδύναμοι ταξινομητές είναι στην γενική περίπτωση δέντρα

απόφασης, αλλά γενικά υπάρχει και η επιλογή να είναι ένας εκ των Tree, DART, Linear ή Tweedie Regression. Σε κάθε αδύναμο ταξινομητή ορίζεται κάποιο βάρος το οποίο είναι σχετισμένο με την ακρίβεια της πρόβλεψής του. Επίσης, μετά από κάθε επανάληψη ορίζεται ένα βάρος και σε κάθε στιγμιότυπο. Εάν το στιγμιότυπο δεν ταξινομηθεί σωστά το βάρος του αυξάνεται. Η επιλογή των στιγμιοτύπων που θα συμμετέχουν σε κάθε επανάληψη εξαρτάται από το βάρος κάθε στιγμιοτύπου. Με αυτό τον τρόπο υπάρχουν περισσότερες επαναλήψεις της διαδικασίας με τα στιγμιότυπα τα οποία είχαν ταξινομηθεί λανθασμένα από τους προηγούμενους αδύναμους ταξινομητές, σε μια προσπάθεια του μοντέλου να επιμείνει στην επίλυση των πιο δύσκολων περιπτώσεων. Μετά την προσθήκη κάθε νέου αδύναμου ταξινομητή, τα βάρη κάθε αδύναμου ταξινομητή και κάθε στιγμιοτύπου υπολογίζονται εκ νέου. (Εικόνα 3.9)



Εικόνα 3.9 Βασική λειτουργία του XGBoost.[147]

Για την αποφυγή του overfitting ο XGBoost χρησιμοποιεί μία μέθοδο η οποία είναι αντίστοιχη του ρυθμού εκμάθησης (learning rate) του Stochastic Gradient Boosting[145] και εδώ αναφέρεται ως συρρίκνωση (shrinkage). Η λειτουργία της συρρίκνωσης μειώνει την επιρροή κάθε αδύναμου ταξινομητή, κλιμακώνοντας τα προστιθέμενα βάρη βάση ενός συντελεστή n, μετά από κάθε βήμα της ενίσχυσης. Με τον τρόπο αυτό αφήνει χώρο στους μελλοντικούς αδύναμους ταξινομητές να βελτιώσουν περαιτέρω το μοντέλο.

Για την υλοποίηση των μοντέλων με βάση τους αλγορίθμους 1 έως 5 χρησιμοποιήθηκαν οι εκδοχές των αλγορίθμων όπως υπάρχουν διαθέσιμες στην βιβλιοθήκη Scikit-learn, ενώ έγινε χρήση και της βιβλιοθήκης xgboost η οποία παρείχε τον αλγόριθμο XGBoost.

Κατά την διαδικασία βελτιστοποίησης παραμέτρων και επιλογής χαρακτηριστικών για την επιλογή του μοντέλου XGboost με την μέγιστη διακριτική ικανότητα δοκιμάστηκαν όλοι οι συνδυασμοί χαρακτηριστικών ως είσοδοι, ενώ οι τιμές των παραμέτρων κυμάνθηκαν ως εξής :

max_depth	2 – 20
min_child_weight	1 – 4
gamma	0 – 1
subsample	0.3 – 1
colsample_bytree	0.8 – 1
reg_alpha	e^{-5} – 1
learning_rate	0.001 – 3
n_estimators	10 – 200

Καταλήγοντας στο συμπέρασμα ότι το μοντέλο XGBoost με την καλύτερη διακριτική ικανότητα δέχεται ως εισόδους τα χαρακτηριστικά :

Ηλικία, BMI, Γλυκόζη, Ρεζιστίνη, Λόγος Λεπτίνης/Αντιπονεκτίνης και Αντιπονεκτίνη

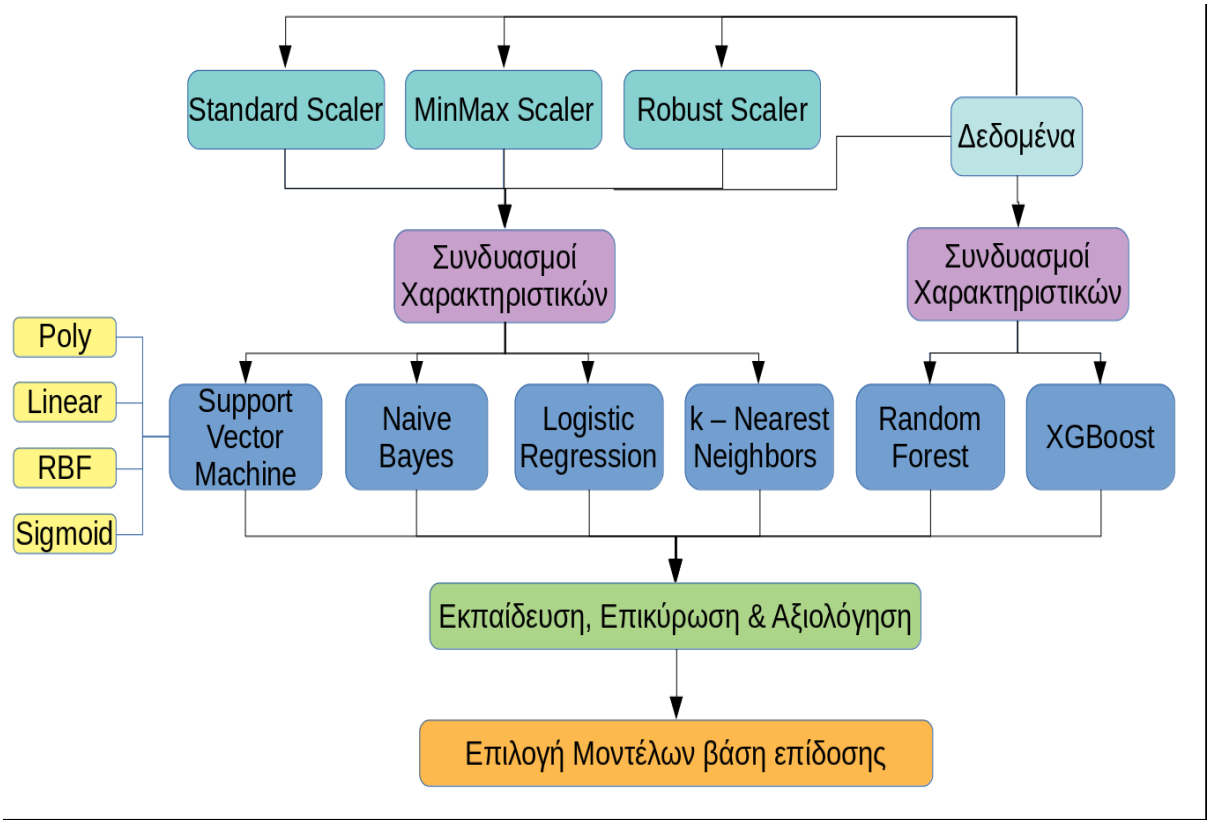
χρησιμοποιόντας 'colsample_bytree': 0.84, 'gamma': 0.1125, 'learning_rate': 0.2, 'max_depth': 3, 'min_child_weight': 2, 'n_estimators': 42, 'reg_alpha': 0.96 και 'subsample': 0.77.

3.4.2 Υλοποίηση μοντέλων μηχανικής μάθησης

Θεωρώντας ότι ο αριθμός των υποψήφιων χαρακτηριστικών (n=10) δεν είναι απαγορευτικός για το εγχείρημα και ο αριθμός των δειγμάτων (ν=116) δεν απαιτεί αυξημένη υπολογιστική ισχύ , αποφασίστηκε να δοκιμαστούν όλοι οι συνδυασμοί χαρακτηριστικών σε κάθε αλγόριθμο. Ο συνολικός αριθμός των συνδυασμών των χαρακτηριστικών είναι :

$$2^{10} – 1 = 1023$$

Από αυτούς αφαιρέθηκαν όλοι οι 128 συνδυασμοί οι οποίο είχαν μαζί και τα 3 χαρακτηριστικά Λεπτίνη, Αντιπονεκτίνη και λόγο L/A, αφού το γεγονός ότι ο λόγος L/A είναι αποτέλεσμα απλής διαίρεσης της τιμής της Λεπτίνης προς την τιμή της Αντιπονεκτίνης καθιστά την παρουσία και των τριών μαζί στο ίδιο μοντέλο περιττή. Έτσι ο συνολικός αριθμός των συνδυασμών οι οποίοι εφαρμόστηκαν ως είσοδοι σε όλα τα μοντέλα ήταν : 895.

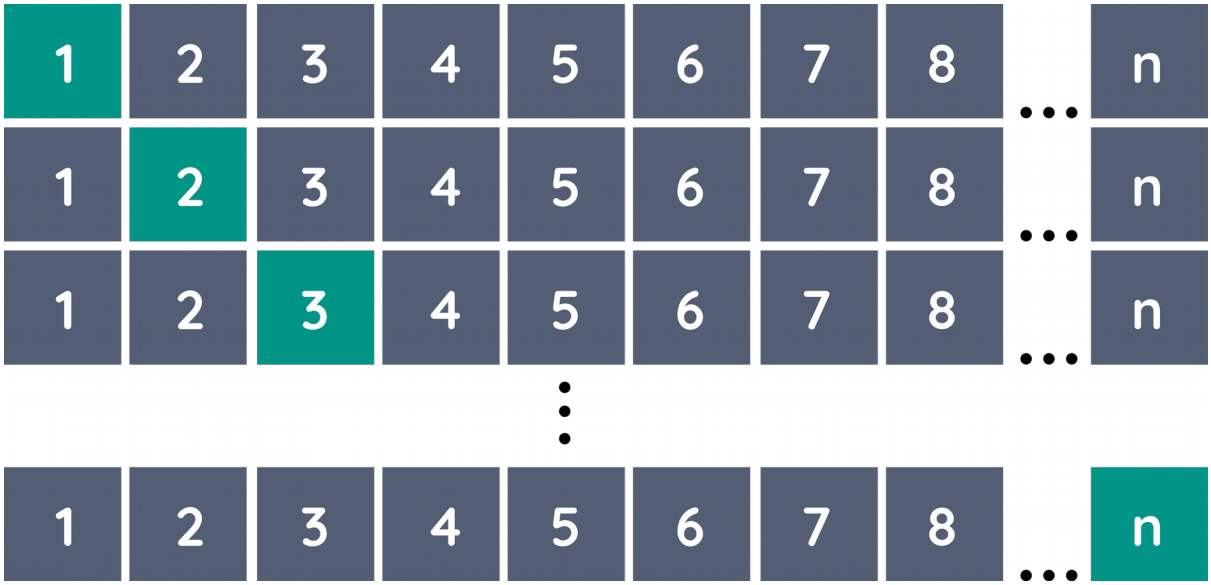


Εικόνα 3.10 Μεθοδολογία υλοποίησης μοντέλων μηχανικής μάθησης.

3.4.3 Εκπαίδευση και Επικύρωση

Για την εκπαίδευση και επικύρωση των μοντέλων χρησιμοποιήθηκε η μέθοδος Leave One Out Cross Validation (LOOCV). Ο κύριος λόγος για τον οποίο επιλέχθηκε η χρήση αυτής της μεθόδου στην συγκεκριμένη εργασία είναι διότι τα στιγμιότυπα που έχουμε στην διάθεσή μας είναι περιορισμένα (ν = 116) και η μέθοδος LOOCV μας επιτρέπει να χρησιμοποιήσουμε περισσότερα δεδομένα για την εκπαίδευση των μοντέλων μας από οποιαδήποτε άλλη μέθοδο εκπαίδευσης και επικύρωσης.

Κατά την μέθοδο LOOCV τα δεδομένα μας χωρίζονται σε δύο ομάδες, μία ομάδα εκπαίδευσης (training set) και μία ομάδα επικύρωσης (validation set) οι οποίες είναι ξένες μεταξύ τους. Η ομάδα εκπαίδευσης αποτελείται από το σύνολο των δεδομένων, με εξαίρεση ένα στιγμιότυπο το οποίο αποτελεί την ομάδα επικύρωσης. Πιο απλά, κρατάμε ένα στιγμιότυπο για επικύρωση, και τα υπόλοιπα δεδομένα χρησιμοποιούνται για την εκπαίδευση του μοντέλου. Στην συνέχεια στο μοντέλο εισάγονται τα δεδομένα από τα χαρακτηριστικά του στιγμιότυπου της ομάδας επικύρωσης, ώστε να προβλέψει την κλάση την οποία ανήκει το στιγμιότυπο, στην συγκεκριμένη περίπτωση έχουμε 2 πιθανές κλάσεις επομένως το πρόβλημα είναι δυαδικής κατηγοριοποίησης. Η διαδικασία εκπαίδευσης και επικύρωσης θα



Εικόνα 3.11 Leave One Out Cross Validation.

επαναληφθεί τόσες φορές όσο είναι το σύνολο των στιγμιοτύπων (Εικόνα 3.11). Με τον τρόπο αυτό καταλήγουμε να έχουμε μία πρόβλεψη της πιθανότητας, για κάθε ένα από τα στιγμιότυπα, να ανήκει σε κάθε μία από τις κλάσεις. Η πρόβλεψη αυτή συγκρίνεται στην συνέχεια με την πραγματική τιμή της κλάσης (0 για υγιείς και 1 για ασθενείς) κάθε στιγμιότυπου βάση ενός ορισθέντος κατωφλίου (threshold), ώστε να κατατάξουμε την απόκριση του μοντέλου για αυτό το στιγμιότυπο ως :

- Αληθώς θετική (true positive, **TP**)
  Το μοντέλο ταξινομεί σωστά το στιγμιότυπο στην κλάση που ορίστηκε ως θετική. Στην συγκεκριμένη περίπτωση το μοντέλο αποκρίνεται σωστά ότι η συμμετέχουσα πάσχει από καρκίνο του μαστού.

- Αληθώς αρνητική (true negative, **TN**)
  Το μοντέλο ταξινομεί σωστά το στιγμιότυπο στην κλάση που ορίστηκε ως αρνητική. Στην συγκεκριμένη περίπτωση το μοντέλο αποκρίνεται σωστά ότι η συμμετέχουσα δεν πάσχει από καρκίνο του μαστού.

- Ψευδώς θετική (false positive, **FP**)
  Το μοντέλο ταξινομεί λανθασμένα το στιγμιότυπο στην κλάση που ορίστηκε ως θετική. Στην συγκεκριμένη περίπτωση το μοντέλο αποκρίνεται ότι η συμμετέχουσα πάσχει από καρκίνο του μαστού, όμως στην πραγματικότητα εκείνη είναι υγιείς.

- Ψευδώς αρνητική (false negative, **FN**)

  Το μοντέλο ταξινομεί λανθασμένα το στιγμιότυπο στην κλάση που ορίστηκε ως αρνητική. Στην συγκεκριμένη περίπτωση το μοντέλο αποκρίνεται ότι η συμμετέχουσα είναι υγιείς, όμως στην πραγματικότητα εκείνη πάσχει από καρκίνο του μαστού.

Στην συγκεκριμένη περίπτωση, κάθε ένα από τα 116 στιγμιότυπα αποτελεί για μία φορά την ομάδα επικύρωσης ενώ είναι 115 φορές ένα από τα 115, διαφορετικά κάθε φορά, στιγμιότυπα τα αποτελούν την ομάδα εκπαίδευσης.

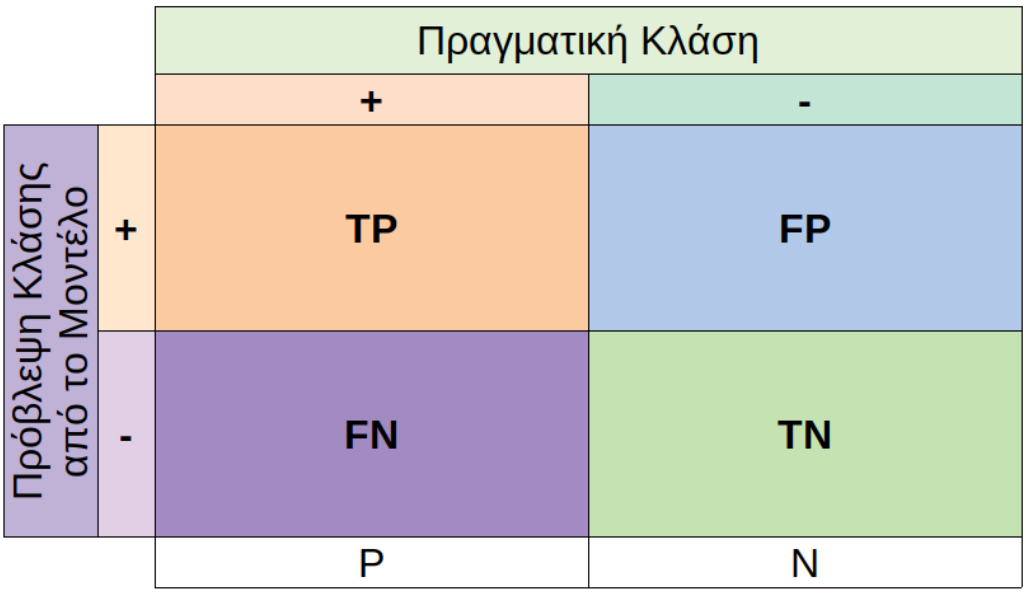
3.4.4 Αξιολόγηση

Μετά την ολοκλήρωση όλων επαναλήψεων της φάσης εκπαίδευσης και επικύρωσης, έχουμε στην διάθεσή μας το σύνολο των προβλέψεων (ως πιθανότητα το υπό εξέταση στιγμιότυπο να είναι θετικό) σε μία λίστα καθώς και τα συνολικά αποτελέσματα TP, FP, TN και FN του ταξινομητή για την συγκεκριμένη λίστα χαρακτηριστικών. Βάση αυτών, μπορούμε πλέον να υπολογίσουμε διάφορα μέτρα αξιολόγησης του συγκεκριμένου μοντέλου δυαδικής κατηγοριοποίησης.

Χρησιμοποιώντας τα TP, FP, TN και FN μπορούμε να κατασκευάσουμε τον πίνακα σύγχυσης (confusion matrix) και να υπολογίσουμε τις τιμές διαφόρων μέτρων αξιολόγησης δυαδικής κατηγοριοποίησης. Στην συγκεκριμένη εργασία, για την επιλογή βέλτιστων συνδυασμών χαρακτηριστικών και παραμέτρων, ως μέτρα αξιολόγησης χρησιμοποιήθηκαν κυρίως η Ακρίβεια, το AUC, η Ευαισθησία και η Ειδικότητα. Μαζί με αυτά θα ορίσουμε στην συνέχεια και την Θετική Προβλεπτική Αξία και το F-measure τα οποία χρησιμοποιούνται επίσης στην βιβλιογραφία.

Ο **confusion matrix** βοηθάει στην οπτικοποίηση της απόδοσης του μοντέλου και στην περίπτωση πού το πρόβλημα είναι η κατηγοριοποίηση των στιγμιοτύπων σε 2 κλάσεις, έχει την μορφή που φαίνεται στον πίνακα 3.2.

Πίνακας 3.2 Πιθανά αποτελέσματα για την πρόβλεψη μεταξύ δύο κλάσεων.



Η Ακρίβεια (**Accuracy**) εκφράζει το ποσοστό των σωστά ταξινομημένων στιγμιοτύπων από το μοντέλο, και μπορεί να υπολογιστεί από τον τύπο :

$$Accuracy = \frac{TP+TN}{P+N} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3.4.4.1)$$

όπου P = TP + FN , είναι το σύνολο των θετικών στιγμιοτύπων της βάσης εξέτασης και

N = TN + FP , είναι το σύνολο των αρνητικών στιγμιοτύπων της βάσης εξέτασης

Η Ευαισθησία (**Sensitivity**) εκφράζει την ικανότητα του μοντέλου να ταξινομεί σωστά τα θετικά στιγμιότυπα της βάσης εξέτασης, υποδεικνύει την ικανότητα του μοντέλου να εντοπίζει την ύπαρξη καρκίνου του μαστού στις συμμετέχουσες, και μπορεί να υπολογιστεί από τον τύπο :

$$Sensitivity = \frac{TP}{P} = \frac{TP}{TP+FN} \quad (3.4.4.2)$$

Η Ειδικότητα (**Specificity**) εκφράζει την ικανότητα του μοντέλου να ταξινομεί σωστά τα αρνητικά στιγμιότυπα της βάσης εξέτασης, υποδεικνύει την ικανότητα του μοντέλου να εντοπίζει τις συμμετέχουσες οι οποίες είναι υγιείς και μπορεί να υπολογιστεί από τον τύπο :

$$Specificity = \frac{TN}{N} = \frac{TN}{TN+FP} \quad (3.4.4.3)$$

Βάση των παραπάνω σχέσεων μπορούμε να συμπεράνουμε ότι η Ακρίβεια αποτελεί γραμμικό συνδυασμό της Ευαισθησίας και της Ειδικότητας, όπως φαίνεται στον τύπο :

$$Accuracy = Sensitivity * \frac{P}{P+N} + Specificity * \frac{N}{P+N} \quad (3.4.4.4)$$

Η θετική προβλεπτική αξία (**Precision)** εκφράζεται ως ο λόγος των σωστά ταξινομημένων θετικών στιγμιοτύπων επί του συνολικού αριθμού των θετικά ταξινομημένων στιγμιοτύπων από το μοντέλο, δηλαδή στην συγκεκριμένη περίπτωση είναι το ποσοστό των συμμετεχόντων οι οποίες προβλέφθηκε από το μοντέλο ότι πάσχουν από καρκίνο του μαστού και είναι όντως ασθενείς, επομένως υποδεικνύει την αξιοπιστία της διάγνωσης ασθένειας από το μοντέλο, και μπορεί να υπολογιστεί από τον τύπο :

$$Precision = \frac{TP}{TP+FP} \quad (3.4.4.5)$$

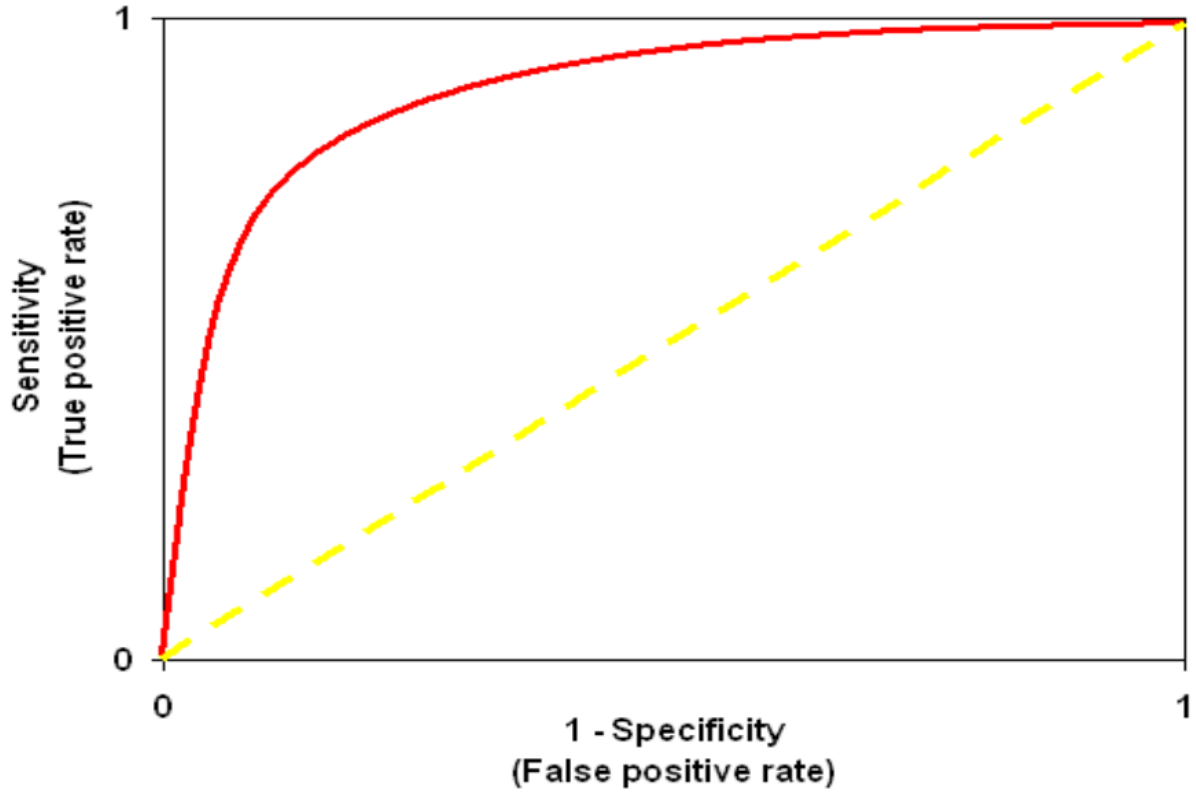
Τέλος, το **F-measure** αποτελεί τον αρμονικό μέσω των sensitivity και precision, και υπολογίζεται από τον τύπο :

$$F - measure = \frac{2 * Precision * Sensitivity}{Precision + Sensitivity} = \frac{2 * TP}{2 * TP + FP + FN} \quad (3.4.4.6)$$

Χρησιμοποιώντας την λίστα των προβλέψεων του μοντέλου μπορούμε να κατασκευάσουμε την χαρακτηριστική καμπύλη λειτουργίας δέκτη (Receiver Operatic Characteristic, ROC curve) και βάση αυτής να υπολογίσουμε το εμβαδό κάτω από την καμπύλη ROC (Area Under the ROC Curve – **AUC**). Η καμπύλη ROC απεικονίζει τη διαγνωστική ικανότητα ενός μοντέλου δυαδικής κατηγοριοποίησης καθώς η τιμή του κατωφλίου του μεταβάλλεται και κατασκευάζεται απεικονίζοντας στον κάθετο άξονα του γραφήματος τις τιμές True Positive Rate (Sensitivity) του μοντέλου και στον οριζόντιο άξονα τις τιμές του False Positive Rate (1-Specificity) του μοντέλου καθώς το κατώφλι απόφασης μεταβάλλεται από 0 έως 1 (Εικόνα 3.12). Το εμβαδό κάτω από την καμπύλη ROC (AUC) αποτελεί ένα αριθμητικό μέτρο της απόδοσης ενός μοντέλου και είναι 1 στην ιδανική περίπτωση και 0.5 στην περίπτωση της τυχαίας κατηγοριοποίησης. Για τον υπολογισμό του εμβαδό κάτω από την καμπύλη ROC χρησιμοποιήσαμε την συνάρτησης

roc_auc_score(y_test_list, proba_list)

της βιβλιοθήκης sklearn.



Εικόνα 3.12 Παράδειγμα καμπύλης λειτουργίας δέκτη (Receiver Operatic Characteristic, ROC curve).

3.4.5 Χρόνος Εκτέλεσης

Από την στιγμή που έχουμε 116 στιγμιότυπα στη διάθεσή μας, η διαδικασία της εκπαίδευσης και επικύρωσης επαναλαμβάνεται 116 φορές για κάθε συνδυασμό χαρακτηριστικών. Σε συνδυασμό τώρα με το γεγονός ότι οι υπό εξέταση συνδυασμοί για κάθε διαφορετικό ταξινομητή είναι 895, υπολογίζουμε ότι για την εκτίμηση της απόδοσης κάθε διαφορετικού αλγορίθμου χρειάστηκαν να εκπαιδευτούν συνολικά 103.820 μοντέλα. Αυτό είναι και το κύριο αρνητικό στοιχείο της μεθόδου LOOCV. Πρακτικά, λόγο του μικρού όγκου των δεδομένων ο χρόνος εκτέλεσης των πειραμάτων δεν υπήρξε αποτρεπτικός, εκτός από την περίπτωση του SVM linear, όπου βλέποντας ότι μετά από 4 ημέρες είχε ολοκληρώσει μόνο 443 από τις 895 συνολικά δοκιμές συνδυασμών, εγκαταλείψαμε την επιλογή του συγκεκριμένου πυρήνα. Ενδεικτικά, ο αμέσως πιο απαιτητικός αλγόριθμος σε χρόνο εκτέλεσης ήταν ο Random Forest ο οποίος χρειάστηκε 10 ώρες για να ολοκληρώσει

την εκπαίδευση 103.820 μοντέλων, ενώ ο πιο γρήγορος ήταν ο Naive Bayes ο οποίος χρειάστηκε περίπου 2 λεπτά.

3.4.6 Κανονικοποίηση των δεδομένων (data normalization ή scaling)

Η κανονικοποίηση ή ομαλοποίηση είναι μια τεχνική που συχνά εφαρμόζεται ως μέρος της προετοιμασίας των δεδομένων για την δημιουργία ενός μοντέλου μηχανικής μάθησης. Ο στόχος της κανονικοποίησης είναι να μεταβληθούν οι τιμές των χαρακτηριστικών στο σύνολο δεδομένων σε μια κοινή κλίμακα, χωρίς όμως να διαταράσσονται οι διαφορές στο εύρος των τιμών.

Τα μοντέλα βασισμένα σε αλγορίθμους που βασίζονται σε δέντρα απόφασης, όπως ο Random Forest και ο XGBoost, ταξινομούν βάση ενός σημείου διάσπασης (split point) σε κάθε ένα από τα χαρακτηριστικά. Το σημείο διάσπασης σε κάθε χαρακτηριστικό καθορίζεται από το ποσοστό των περιπτώσεων που ταξινομούνται σωστά χρησιμοποιώντας αυτό το σημείο διάσπασης σε αυτό το χαρακτηριστικό. Με αυτό τον τρόπο οι αλγόριθμοι που βασίζονται σε δέντρα απόφασης δεν επηρεάζονται από την διαφορά κλίμακας μεταξύ των χαρακτηριστικών και επομένως δεν χρειάζονται ομαλοποίηση των δεδομένων.

Μοντέλα με αλγορίθμους LR, NB, k-NN, SVM sigmoid, SVM linear, SVM poly και SVM rbf δημιουργήθηκαν ξανά για όλους του συνδυασμούς χαρακτηριστικών, ενώ προηγουμένως εφαρμόστηκαν 3 διαφορετικές μέθοδοι κανονικοποίησης στα δεδομένα τους οποίους παρέχει η βιβλιοθήκη Scikit_Learn:

- **Standard Scaler**

  (Κανονικοποίηση Μέσης Απόλυτης Απόκλισης - MAD Normalization)

  Όταν ο StandardScaler εφαρμοστεί στα δεδομένα τα μετατρέπει, στην γενική περίπτωση, έτσι ώστε σε κάθε χαρακτηριστικό η μέση τιμή να είναι 0 και η τυπική απόκλιση 1. Η νέα τιμή $x_{νέο}$ κάθε δείγματος x υπολογίζεται ως:

  $$x_{νέο} = (x - u) / s \quad (3.4.5.1)$$

  όπου u είναι η μέση τιμή των δειγμάτων εκπαίδευσης, και s είναι η τυπική απόκλιση των δειγμάτων κατάρτισης, ή 1 εάν with_std = False. Δοκιμάστηκαν οι περιπτώσεις για with_std να είναι είτε True ή False.

- **Min-Max Scaler**

  Όταν ο Min-Max Scaler εφαρμοστεί στα δεδομένα ουσιαστικά συρρικνώνει το εύρος τιμών σε κάθε χαρακτηριστικό ορίζοντάς το μεταξύ 0 και 1 ,ή -1 έως 1 εάν υπάρχουν αρνητικές τιμές. Η νέα τιμή $x_{νέο}$ κάθε δείγματος x, ενός διανύσματος χαρακτηριστικού X υπολογίζεται ως:

$$x_{νέο} = (x - \max(X)) / (\max(X) - \min(X)) \quad (3.4.5.2)$$

- **Robust Scaler**

  Ο RobustScaler χρησιμοποιεί παρόμοια μέθοδο με αυτή του Min-Max Scaler, αλλά αντί για την μέγιστη και ελάχιστη τιμή κάθε χαρακτηριστικού, χρησιμοποιεί το ενδοτεταρτημοριακό εύρος. Η νέα τιμή $x_{νέο}$ κάθε δείγματος x, ενός διανύσματος χαρακτηριστικού X υπολογίζεται ως:

$$x_{νέο} = (x - Q_1(X)) / (Q_3(X) - Q_1(X)) \quad (3.4.5.3)$$

  όπου Q_1 είναι το $25^ο$ ποσοστημόριο και Q_3 είναι το $75^ο$ ποσοστημόριο.

3.4.7 Επιλογή Μοντέλων Βάση Επίδοσης

Μέχρι αυτό το σημείο έχουν πραγματοποιηθεί πειράματα για όλους τους αλγορίθμους κατηγοριοποίησης, με όλους τους συνδυασμούς χαρακτηριστικών για κάθε αλγόριθμο. Επιπλέον, η παραπάνω διαδικασία έχει επαναληφθεί για κάθε έναν από τους αλγορίθμους LR, NB, k-NN, SVM sigmoid, SVM linear, SVM poly και SVM rbf οι οποίοι θα μπορούσαν να επωφεληθούν από κανονικοποίηση των δεδομένων. Σε αυτό το σημείο φτάσαμε στο σημείο να κάνουμε μία αρχική επιλογή, ως προς :

1. Ποιοι αλγόριθμοι θα επιλεγούν ώστε να συμμετέχουν στην δημιουργία μοντέλων συλλογικής μάθησης (ensemble learning),
2. Εάν θα γίνει κανονικοποίηση των δεδομένων για αυτά τα μοντέλα, και εάν ναι τι είδους μέθοδος θα χρησιμοποιηθεί,
3. Για κάθε αλγόριθμο που θα επιλεγεί, ποιοι συνδυασμοί χαρακτηριστικών θα επιλεγούν.

Σε αυτό το σημείο θεωρούμε ότι είναι καλό να οπτικοποιηθεί ο τρόπος με τον οποίο αποθηκεύτηκαν και ταξινομήθηκαν τα αποτελέσματα κάθε πειράματος, με ένα παράδειγμα όπως φαίνεται στον Πίνακα 3.3.

Πίνακας 3.3 Παράδειγμα αποτελεσμάτων πειράματος με τον αλγόριθμο της Λογιστικής Παλινδρόμησης.

Model	AUC	Accuracy	Sensitivity	Specificity	Training_time	No of Attributes	Attributes	Attributes	Attributes	Attributes	Attributes	Attributes	Attributes	Attributes
Log Reg lbfgs	80.95	77.59	75.00	80.77	3.58	6	Glucose	Resistin	Age	BMI	L/A Ratio	Insulin		
Log Reg lbfgs	80.95	75.00	75.00	75.00	2.65	5	Glucose	Resistin	Age	BMI	L/A Ratio			
Log Reg lbfgs	80.92	77.59	75.00	80.77	2.44	5	Glucose	Resistin	Age	BMI	HOMA			
Log Reg lbfgs	80.92	77.59	75.00	80.77	3.52	6	Glucose	Resistin	Age	BMI	L/A Ratio	HOMA		
Log Reg lbfgs	80.89	77.59	75.00	80.77	2.75	5	Glucose	Resistin	Age	BMI	Insulin			
Log Reg lbfgs	80.68	75.86	73.44	78.85	5.19	7	Glucose	Resistin	Age	BMI	L/A Ratio	HOMA	Insulin	
Log Reg lbfgs	80.65	77.59	75.00	80.77	3.90	6	Glucose	Resistin	Age	BMI	HOMA	Insulin		
Log Reg lbfgs	80.53	74.14	73.44	75.00	2.26	4	Glucose	Resistin	Age	BMI				
Log Reg lbfgs	80.35	76.72	73.44	80.77	5.17	7	Glucose	Resistin	Age	BMI	Adiponectin	L/A Ratio	HOMA	
Log Reg lbfgs	80.29	75.86	73.44	78.85	4.41	7	Glucose	Resistin	Age	BMI	Adiponectin	L/A Ratio	Insulin	
Log Reg lbfgs	80.17	75.86	73.44	78.85	7.16	8	Glucose	Resistin	Age	BMI	Adiponectin	L/A Ratio	HOMA	Insulin
Log Reg lbfgs	80.14	76.72	75.00	78.85	3.45	6	Glucose	Resistin	Age	BMI	Leptin	Insulin		
Log Reg lbfgs	80.11	77.59	75.00	80.77	4.29	6	Glucose	Resistin	Age	BMI	Leptin	HOMA		
Log Reg lbfgs	80.11	76.72	75.00	78.85	3.28	5	Glucose	Resistin	Age	BMI	Leptin			
Log Reg lbfgs	80.08	75.86	76.56	75.00	5.62	6	Glucose	Resistin	Age	BMI	MCP.1	Insulin		
Log Reg lbfgs	80.05	75.86	76.56	75.00	9.28	7	Glucose	Resistin	Age	BMI	HOMA	MCP.1	Insulin	
Log Reg lbfgs	80.02	77.59	78.13	76.92	6.08	6	Glucose	Resistin	Age	BMI	HOMA	MCP.1		
Log Reg lbfgs	79.96	75.00	73.44	76.92	3.73	6	Glucose	Resistin	Age	BMI	Adiponectin	L/A Ratio		
Log Reg lbfgs	79.93	77.59	75.00	80.77	4.20	6	Glucose	Resistin	Age	BMI	Adiponectin	HOMA		
Log Reg lbfgs	79.93	75.00	75.00	75.00	3.51	6	Glucose	Resistin	Age	BMI	Leptin	L/A Ratio		

Στον Πίνακα 3.3 εμφανίζονται τα πρώτα 20 σε απόδοση μοντέλα, από το σύνολο των 895 μοντέλων που υλοποιήθηκαν βασισμένα στον αλγόριθμο κατηγοριοποίησης Logistic Regression, με λύτη (solver) lbfgs, χωρίς να έχει γίνει κανενός είδους κανονικοποίηση των δεδομένων εισόδου. Η απόδοση των μοντέλων συγκρίθηκε βάση φθίνουσας τιμής AUC, Accuracy, Sensitivity, Specificity και αύξουσας τιμής Αριθμού Χαρακτηριστικών (No of Attributes) και Training Time, και η σειρά αυτή υπήρξε και η σειρά σημαντικότητας κάθε μέτρου αξιολόγησης. Για κάθε μοντέλο μπορούμε να δούμε επίσης και ποιος είναι ο συνδυασμός των χαρακτηριστικών που χρησιμοποιούνται ως είσοδοι.

Ως ένα συγκριτικό μέτρο αξιολόγησης και κατάταξης των πειραμάτων αποφασίσαμε να χρησιμοποιήσουμε τον μέσο όρο των τιμών AUC, Accuracy, Sensitivity και Specificity των πρώτων σε απόδοση 16 μοντέλων κάθε πειράματος, το οποίο θα ονομάσουμε Απόδοση-16. Με τον τρόπο αυτό έχουμε μία αριθμητική τιμή την οποία θα χρησιμοποιήσουμε για να κατατάξουμε την απόδοση κάθε συνδυασμού αλγορίθμου κατηγοριοποίησης και μεθόδου κανονικοποίησης των δεδομένων. Η σειρά κατάταξης παρουσιάζεται στον Πίνακα 3.4.

Πίνακας 3.4 Κατάταξη των αλγορίθμων κατηγοριοποίησης βάση της τιμής Απόδοση-16.

Αλγόριθμος Ταξινόμησης	Κανονικοποίηση	Απόδοση-16	Παρατηρήσεις	Σχόλια
SVM rbf	Standard Scaler	83.20		Τα υπόλοιπα δεδομένα δεν είχαν αποθηκευθεί κατά την διεξαγωγή των πειραμάτων. Είχε βέλτιστη απόδοση με Standard Scaler.
XGBoost	-	80.50		
SVM rbf	-	79.82		
SVM linear	Standard Scaler	79.20		
SVM linear	Robust Scaler	78.92		
SVM linear	-	78.47		Αποκλείεται λόγο πολύ μεγάλου χρόνου εκτέλεσης
SVM poly	Min-Max Scaler	77.84	Spec 66.28	Αποκλείεται λόγο πολύ χαμηλού μέσου όρου Spec
Random Forest	-	77.67		
LogReg	Standard Scaler	77.63		
LogReg	-	77.61		
LogReg	Robust Scaler	77.61		
SVM poly	Standard Scaler	77.46	Spec 53.60	Αποκλείεται λόγο πολύ χαμηλού μέσου όρου Spec
SVM poly	-	77.44	Sens 70.02	
K-NN	-	75.79		
K-NN	Standard Scaler	75.22		
K-NN	Robust Scaler	74.67		
K-NN	Min-Max Scaler	74.56		
NaiveBayes	Min-Max Scaler	68.96		
NaiveBayes	Robust Scaler	68.96		
NaiveBayes	Standard Scaler	68.95		
NaiveBayes	-	68.71		
SVM linear	Min-Max Scaler	68.48		
LogReg	Min-Max Scaler	67.45		
SVM sigmoid	Standard Scaler	66.38		Τα υπόλοιπα δεδομένα δεν είχαν αποθηκευθεί κατά την διεξαγωγή των πειραμάτων. Είχε βέλτιστη απόδοση με Standard Scaler.
SVM poly	Robust Scaler	58.76		

Στον πίνακα 3.4 φαίνεται ότι έχουμε αποκλείσει κάποια από τα αποτελέσματα. Στην μία περίπτωση του SMV linear χωρίς κανονικοποίηση των δεδομένων, η διαδικασία της εκπαίδευσης έπαιρνε πάρα πολύ χρόνο (περίπου 30 λεπτά ανά μοντέλο) στις περιπτώσεις όπου ένα από τα χαρακτηριστικά ήταν το MCP-1, διότι οι τιμές που μπορεί να παίρνει είναι κατά πολύ μεγαλύτερες από ότι στα υπόλοιπα χαρακτηριστικά. Αυτό λύθηκε με εφαρμογή κανονικοποίησης. Να σημειώσουμε μάλιστα, ότι ο χρόνος εκπαίδευσης μειώθηκε αισθητά για όλους του συνδυασμούς χαρακτηριστικών, ακόμα δηλαδή και στους συνδυασμούς στου οποίους δεν συμμετείχε το MCP-1. Επίσης, αποκλείσαμε από τις επιλογές μας τα αποτελέσματα του αλγορίθμου SVM poly με κανονικοποίηση Standard και Min-Max Scaler, διότι είδαμε ότι πέτυχαν γενικά πολύ χαμηλά αποτελέσματα Ειδικότητας.

Με τον τρόπο αυτό καταλήξαμε στην επιλογή 7 αλγορίθμων και 16 συνδυασμούς χαρακτηριστικών για τον κάθε ένα. Οι αλγόριθμοι είναι οι :

- SVM rbf με κανονικοποίηση δεδομένων μέσω της μεθόδου Standard Scaler
- XGBoost
- SVM linear με κανονικοποίηση δεδομένων μέσω της μεθόδου Standard Scaler
- Random Forest

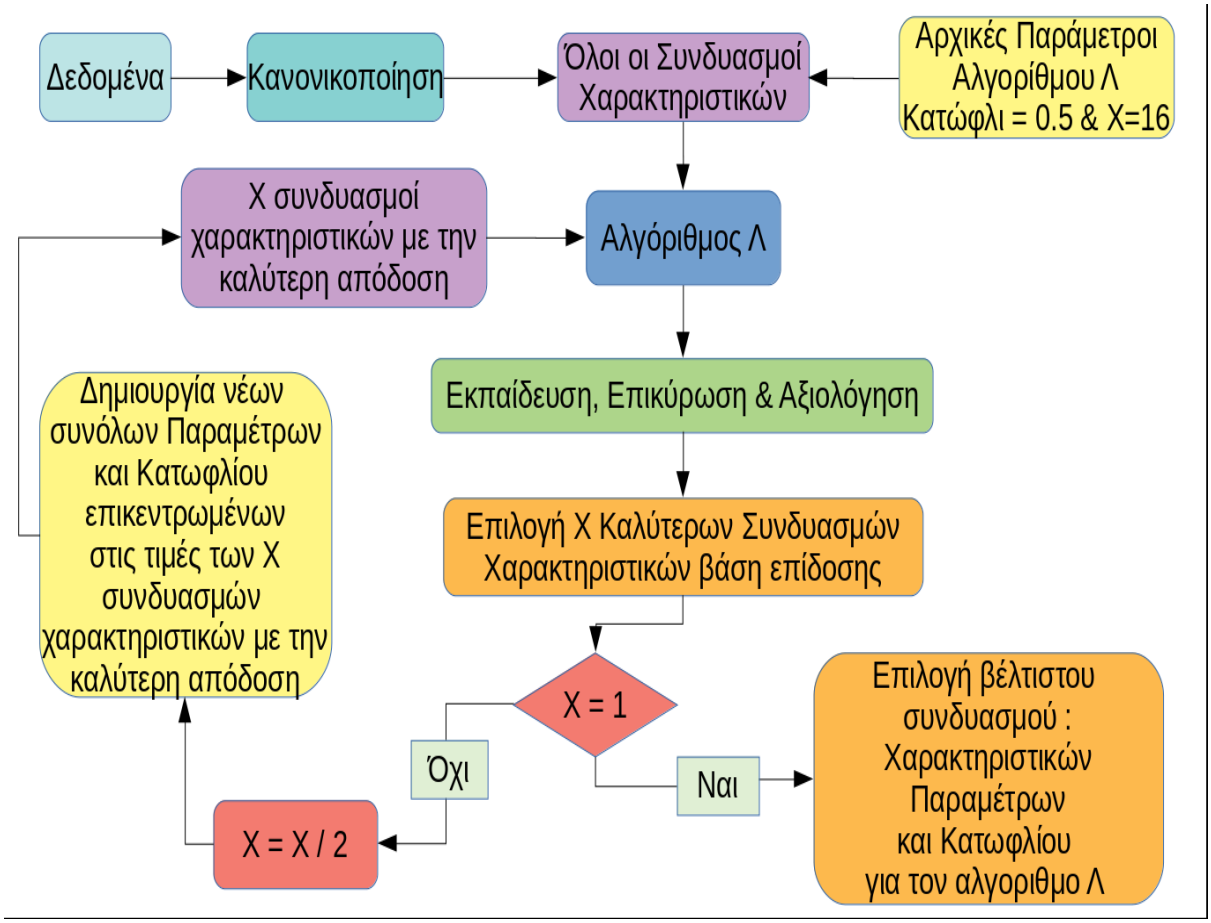
- Logistic Regression με κανονικοποίηση δεδομένων μέσω της μεθόδου Standard Scaler
- SVM poly
- k-NN

Οι 16 επιλεχθέντες συνδυασμοί χαρακτηριστικών για κάθε αλγόριθμο κρίνουμε ότι δεν είναι απαραίτητο να αναφερθούν σε αυτό το σημείο.

3.4.8 Βελτιστοποίηση παραμέτρων και επιλογή χαρακτηριστικών

Επόμενο βήμα είναι η βελτιστοποίηση των επιλεχθέντων, μέχρι αυτό το σημείο, μοντέλων μέσω της βελτιστοποίησης των παραμέτρων των αλγορίθμων τους. Για κάθε έναν από τους 7 αλγόριθμους έχουμε επιλέξει τους Χ = 16 συνδυασμούς χαρακτηριστικών οι οποίοι αποδίδουν καλύτερα με τις παραμέτρους αμετάβλητες. Θα προχωρήσουμε σε παράλληλη βελτιστοποίηση παραμέτρων και επιλογή συνδυασμών χαρακτηριστικών για κάθε αλγόριθμο. Ταυτόχρονα με τις παραμέτρους εξετάστηκαν και διαφορετικές τιμές του κατωφλίου για κάθε μοντέλο. Παρόλο που το κατώφλι δεν θεωρείται ως παράμετρος κάθε αλγορίθμου, αποτελεί μία από τις παραμέτρους του μοντέλου, επομένως όταν αναφερόμαστε στην συνέχεια σε παραμέτρους εννοείται ότι η τιμή του κατωφλίου είναι μία από αυτές. Η διαδικασία που ακολουθήθηκε είναι η εξής :

1. Για τον αλγόριθμο Λ υπολόγισε την απόδοση για Χ συνδυασμούς χαρακτηριστικών, και για σύνολα συνδυασμών των παραμέτρων.
2. Εάν Χ = 1 επιλογή του βέλτιστου συνδυασμού παραμέτρων βάση Accuracy και Sensitivity και έξοδος.
3. Επέλεξε τους Χ = Χ/2 συνδυασμούς χαρακτηριστικών με την καλύτερη απόδοση βάση φθίνουσας τιμής AUC, Accuracy, Sensitivity, Specificity και αύξουσας τιμής Αριθμού Χαρακτηριστικών (No of Attributes)
4. Επέλεξε τις τιμές των παραμέτρων των Χ βέλτιστων συνδυασμών χαρακτηριστικών και παραμέτρων και δημιούργησε νέα σύνολα συνδυασμών παραμέτρων επικεντρωμένα σε αυτές τις τιμές. Σε περίπτωση περίπτωση ίσης απόδοσης επέλεξε τιμές παραμέτρων οι οποίες αυξάνουν την συντηρητικότητα του μοντέλου για αποφυγή overfitting.
5. Επανέλαβε τα βήματα 1-4.



Εικόνα 3.13 Μεθοδολογία επιλογής βέλτιστου συνδυασμού χαρακτηριστικών και κατωφλίου και βελτιστοποίησης παραμέτρων.

Με την εφαρμογή της παραπάνω διαδικασίας καταλήξαμε για κάθε έναν από τους 7 αλγόριθμους σε 1 μοντέλο με τον βέλτιστο συνδυασμό χαρακτηριστικών και παραμέτρων.

Οι παράμετροι κάθε αλγορίθμου είναι διαφορετικές τόσο σε ποιοτικά χαρακτηριστικά, σε εύρος αλλά και σε ποσότητα. Επομένως, η παραπάνω διαδικασία ανάλογα με την ποσότητα των παραμέτρων κάθε αλγορίθμου είχε διαφορετική πολυπλοκότητα και κόστος σε υπολογιστικό χρόνο.

Τα τελικά αποτελέσματα της απόδοσης των βέλτιστων προβλεπτικών μοντέλων φαίνονται στον Πίνακα 3.5.

Πίνακας 3.5 Αποτελέσματα βέλτιστων προβλεπτικών μοντέλων μετά από βελτιστοποίηση παραμέτρων.

Model	AUC	Accuracy	Sensitivity	Specificity	Training_time
XGboost	86.66	88.79	90.63	86.54	1.63
Random Forest	87.86	87.07	89.06	84.62	25.69
KNN	88.34	86.21	90.63	80.77	11.96
SVM poly	87.05	86.21	87.50	84.62	1.94
SVM rbf	88.40	85.34	87.50	82.69	0.83
SVM linear	83.89	80.17	81.25	78.85	0.47
LogReg	80.98	78.45	76.56	80.77	1.48

Οι τιμές των παραμέτρων και του κατωφλίου καθώς και τα χαρακτηριστικά των βέλτιστων προβλεπτικών μοντέλων φαίνονται στον Πίνακα 3.6.

Πίνακας 3.6 Οι παράμετροι, το κατώφλι και τα χαρακτηριστικά των βέλτιστων προβλεπτικών μοντέλων.

Model	Parameters	Threshold	Attributes
XGboost	{'base_score': 0.5, 'booster': 'gbtree', 'colsample_bylevel': 1, 'colsample_bynode': 1, 'colsample_bytree': 0.84, 'gamma': 0.1125, 'learning_rate': 0.2, 'max_delta_step': 0, 'max_depth': 3, 'min_child_weight': 2, 'missing': None, 'n_estimators': 42, 'objective': 'binary:logistic', 'reg_alpha': 0.96, 'reg_lambda': 1, 'scale_pos_weight': 1, 'seed': 50, 'subsample': 0.77, 'verbosity': 1}	0.5	Glucose Resistin Age BMI Adiponectin L/A Ratio
Random Forest	{'bootstrap': True, 'class_weight': None, 'criterion': 'gini', 'max_depth': 4, 'max_features': 'sqrt', 'max_leaf_nodes': None, 'min_impurity_decrease': 0.0, 'min_impurity_split': None, 'min_samples_leaf': 0.001, 'min_samples_split': 7, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 18, 'n_jobs': -1, 'oob_score': False, 'random_state': 50, 'verbose': 0, 'warm_start': False}	0.5	Glucose Resistin Age BMI Leptin Adiponectin
KNN	{'algorithm': 'auto', 'leaf_size': 30, 'metric': 'minkowski', 'metric_params': None, 'n_jobs': -1, 'n_neighbors': 5, 'p': 1, 'weights': 'distance'}	0.5	Glucose Resistin Age BMI Adiponectin
SVM poly	{'C': 2, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': 'ovr', 'degree': 2, 'gamma': 0.0009765625, 'kernel': 'poly', 'max_iter': -1, 'probability': True, 'random_state': 50, 'shrinking': True, 'tol': 0.001, 'verbose': False}	0.5	Glucose Resistin Age BMI L/A Ratio Insulin
SVM rbf	{'C': 0.99, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': 'ovr', 'degree': 3, 'gamma': 0.00038, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': 50, 'shrinking': True, 'tol': 0.001, 'verbose': False}	0.5	Glucose Resistin Age BMI
SVM linear	{'C': 1.2, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': 'ovr', 'degree': 3, 'gamma': 0.000244140625, 'kernel': 'linear', 'max_iter': -1, 'probability': True, 'random_state': 50, 'shrinking': True, 'tol': 0.001, 'verbose': False}	0.46	Glucose Resistin Age BMI Leptin Insulin
LogReg	{'C': 0.18, 'class_weight': None, 'dual': False, 'fit_intercept': True, 'intercept_scaling': 1, 'l1_ratio': None, 'max_iter': 1000, 'multi_class': 'warn', 'n_jobs': -1, 'penalty': 'l2', 'random_state': None, 'solver': 'lbfgs', 'tol': 0.0001, 'verbose': 0, 'warm_start': False}	0.5	Glucose Resistin Age BMI HOMA

Ο λόγος για τον οποίο στο βήμα 2 επιλέγουμε να χρησιμοποιήσουμε το Accuracy ως πρωταρχικό μέτρο κατάταξης είναι διότι σε αυτό το σημείο έχουμε είδη εξετάσει διαφορετικές τιμές του κατωφλίου και επομένως θεωρούμε ότι το AUC δεν έχει να μας προσφέρει παραπάνω πληροφορία για την απόδοση του μοντέλου. Επίσης, επιλέγουμε να δώσουμε μεγαλύτερη βαρύτητα στο Sensitivity διότι θεωρούμε ότι η σωστή ανίχνευση του καρκίνου του μαστού είναι μεγαλύτερης σημαντικότητας από την σωστή γνώση ότι δεν

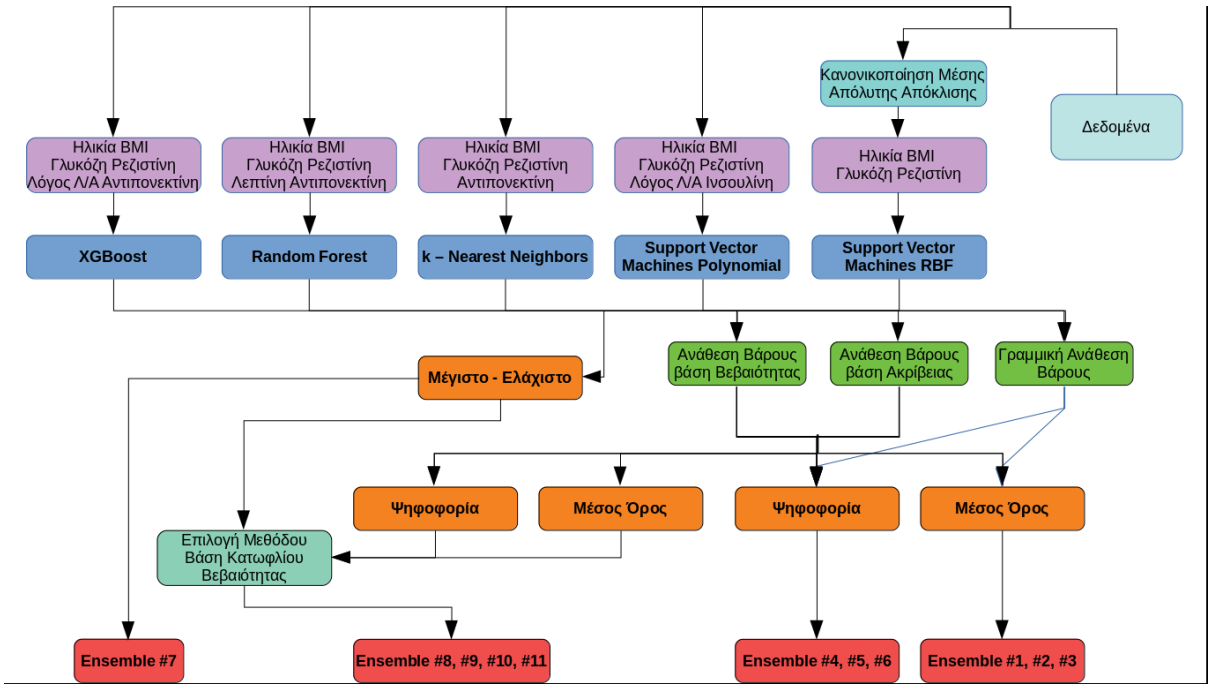
υπάρχει καρκίνος του μαστού. Παρόλα αυτά γνωρίζουμε ότι εάν το Specificity του μοντέλου είναι μικρό έχουμε κατά συνέπεια μεγαλύτερο False Positive Rate, αφού :

False Positive Rate = 1 - Specificity

και στην συγκεκριμένη περίπτωση, αυξημένο False Positive Rate σημαίνει αυξημένες πιθανότητα να χρειαστεί περαιτέρω παρακολούθηση του ατόμου, ή ακόμα και βιοψία, η οποία θα ήταν κανονικά αχρείαστη.

3.4.9 Δημιουργία μοντέλων συλλογικής μάθησης

Ως επόμενο βήμα συνδυάσαμε τα 5 μοντέλα τα οποία μετά την διαδικασία της βελτιστοποίησης είχαν την καλύτερη απόδοση ώστε να δημιουργήσουμε μοντέλα συλλογικής μάθησης (ensembles). Στόχος είναι ο έλεγχος όλων των συνδυασμών των 5 αυτών μοντέλων βάση μεθόδων συλλογικής μάθησης, ώστε τελικά να δημιουργήσουμε μοντέλα με ισχυρότερη προβλεπτική ικανότητα (Εικόνα 3.14).



Εικόνα 3.14 Μεθοδολογία δημιουργίας μοντέλων συλλογικής μάθησης από τα 5 μοντέλα μηχανικής μάθησης με την καλύτερη ικανότητα κατηγοριοποίησης.

Στο σημείο αυτό αναφέρουμε ότι, αφού έχουμε ήδη ολοκληρώσει την εκπαίδευση κάθε ενός από τα μοντέλα, έχουμε τα αποτελέσματα των εκτιμήσεών τους για την κατηγορία στην οποία ανήκει κάθε ένα από τα στιγμιότυπα (σε μορφή πιθανότητας να ανήκει στην κατηγορία 2 η οποία υποδηλώνει ύπαρξη καρκίνου του μαστού) ήδη αποθηκευμένα.

3.4.9.1 Μέθοδοι Ανάθεσης Βάρους

Σε κάθε ensemble κάθε μοντέλο συμμετέχει με ένα βάρος w_{kij}. Χρησιμοποιήθηκαν τρεις μέθοδοι ανάθεσης βάρους.

3.4.9.1.1 Δυναμική Ανάθεση Βάρους βάση Βεβαιότητας

Κατά την Δυναμική ανάθεση βαρών βάση Βεβαιότητας (Dynamic Weighting based on Certainties[148], [149]), σε κάθε μοντέλο k ανατίθεται ένα βάρος w_{kij} για κάθε διαφορετικό στιγμιότυπο j βάση της βεβαιότητας c_{kj} της κατηγοριοποίησης του στιγμιοτύπου j από το μοντέλο k. Η βεβαιότητα κατηγοριοποίησης c_{kj} είναι τόσο αυξημένη όσο πιο κοντά είναι στα 0 ή 1. Το βάρος κάθε στιγμιοτύπου j διαφέρει μεταξύ των ensembles για το κάθε μοντέλο k καθώς εξαρτάται και από τις Βεβαιότητες όλων των m μοντέλων που συμμετέχουν στο κάθε ensemble i. Ο υπολογισμός του w_{kij} γίνεται από τον τύπο :

$$w_{kij} = \frac{c_{kj}}{\sum_{k=1}^{m} c_{kj}} \quad (3.4.8.1.1.2)$$

όπου

$$c_{kj} = \begin{cases} p_{kj}, \text{εάν } p_{kj} \geq 0.5 \\ 1 - p_{kj}, \text{εάν } p_{kj} < 0.5 \end{cases} \quad (3.4.8.1.1.3)$$

m είναι ο συνολικός αριθμός των μοντέλων που συμμετέχουν στην δημιουργία του ensemble i

3.4.9.1.2 Δυναμική Ανάθεση Βάρους βάση Ακρίβειας

Κατά την Δυναμική ανάθεση βαρών βάση της Ακρίβειας, σε κάθε στιγμιότυπο j του μοντέλου k ανατίθεται ένα βάρος w_{kij} το οποίο είναι συνέπεια της Ακρίβειας κατηγοριοποίησης του μοντέλου πάνω στα δεδομένα εκπαίδευσης, εξαιρώντας δηλαδή την απόφαση του μοντέλου k για το στιγμιότυπο j. Ονομάζουμε αυτή την τιμή Ακρίβειας ως $Accuracy_k(j)$. Το βάρος w_{kij} του μοντέλου k διαφέρει σε κάθε ensemble i αφού εξαρτάται από την ελάχιστη τιμή Ακρίβειας $Accuracy_m(j)$ μεταξύ των μοντέλων m που συμμετέχουν σε αυτό. Ο υπολογισμός των βαρών γίνεται από τον τύπο :

$$w_{kij} = Accuracy_k(j) - \left(\min_{m \in i} \left(Accuracy_m(j) \right) - 1 \right) \quad (3.4.8.1.2.1)$$

3.4.9.1.3 Γραμμική Ανάθεση Βάρους

Κατά την γραμμική ανάθεση βάρους, κάθε μοντέλο k συμμετέχει στην λήψη απόφασης του ensemble i με μία βαρύτητα w_{kij} από το 0 έως το 9. Το βάρος w_{kij} παραμένει κοινό για όλα τα στιγμιότυπα j του ensemble i, όμως μεταβάλλεται μεταξύ των ensembles.

3.4.9.2 Μέθοδοι κατηγοριοποίησης με Συλλογική Μάθηση – Υλοποίηση των Ensembles

Οι μέθοδοι οι οποίες χρησιμοποιήθηκαν για τον συνδυασμό των αποτελεσμάτων των 5 καλύτερων μοντέλων για την κατηγοριοποίηση και με αυτό τον τρόπο την δημιουργία των ensembles ήταν οι εξής :

3.4.9.2.1 Κατηγοριοποίηση βάση Μέσου Όρου Πιθανοτήτων

Κατά την Κατηγοριοποίηση βάση Μέσου Όρου Πιθανοτήτων, η λήψη της απόφασης από το ensemble λαμβάνεται μέσω του μέσου όρου όλων των πιθανοτήτων των μοντέλων για την κατηγορία στην οποία ανήκει το στιγμιότυπο. Η πιθανότητα κάθε μοντέλου μπορεί να συμμετέχει στην λήψη απόφασης από το ensemble με διαφορετικό βάρος.

Ο υπολογισμός της πιθανότητας το στιγμιότυπο j να ανήκει στην κατηγορία 2, που υποδηλώνει παρουσία καρκίνου του μαστού, βάση του ensemble i με Μέσο Όρο δίνεται από τον τύπο:

$$\textit{Πιθανότητα Ensemble}_{ij} = \frac{\sum_{k=1}^{m} p_{kj} * w_{kij}}{\sum_{k=1}^{m} w_{kij}} \quad (3.4.8.2.1.1)$$

όπου
- p_{kj} είναι η πιθανότητα του μοντέλου k για το στιγμιότυπο j και παίρνει τιμές από 0 έως 1,
- w_{kij} είναι το βάρος του μοντέλου k για το στιγμιότυπο j στο ensemble i,
- m είναι ο συνολικός αριθμός των μοντέλων που συμμετέχουν στην δημιουργία του ensemble i,
- i είναι ο δείκτης του ensemble και ορίζεται από 1 έως ν, όπου ν είναι το σύνολο των διαφορετικών ensemble.

Χρησιμοποιήσαμε 3 μεθόδους ανάθεσης βάρους για διαφορετικές υλοποιήσεις της Κατηγοριοποίησης βάση Μέσου Όρου:

1. Η ανάθεση βαρών γίνεται με *Δυναμική Ανάθεση Βάρους βάση Βεβαιότητας*, όπως περιγράφεται στο 3.4.8.1.1.

   Η Πιθανότητα Ensemble_{ij} υπολογίζεται από τον τύπο (3.4.8.2.1.1) και συγκρίνεται με το κατώφλι απόφασης του ensemble, ώστε να καταλήξει το κάθε ensemble i στην τελική του απόφαση για την κατηγορία στην οποία ανήκει το στιγμιότυπο j.

   Δημιουργούμε όλους τους συνδυασμούς των 5 μοντέλων, δηλαδή $v = 2^5 = 32$ ensembles.

   Το ensemble με την μεγαλύτερη διακριτική ικανότητα που προέκυψε βάση αυτής της μεθόδου ονομάστηκε Ensemble #1.

2. Η ανάθεση βαρών γίνεται με *Δυναμική Ανάθεση Βάρους βάση Ακρίβειας*, όπως περιγράφεται στο 3.4.8.1.2.

   Η Πιθανότητα Ensemble_{ij} υπολογίζεται από τον τύπο (3.4.8.2.1.1) και συγκρίνεται με το κατώφλι απόφασης του ensemble, ώστε να καταλήξει το ensemble i στην τελική του απόφαση για την κατηγορία στην οποία ανήκει το στιγμιότυπο j.

   Δημιουργούμε όλους τους συνδυασμούς των 5 μοντέλων, δηλαδή $v = 2^5 = 32$ ensembles.

   Το ensemble με την μεγαλύτερη διακριτική ικανότητα που προέκυψε βάση αυτής της μεθόδου ονομάστηκε Ensemble #2.

3. Η ανάθεση βαρών γίνεται με *Γραμμική Ανάθεση Βάρους*, όπως περιγράφεται στο 3.4.8.1.3.

   Η Πιθανότητα Ensemble_{ij} υπολογίζεται από τον τύπο (3.4.8.2.1.1) και συγκρίνεται με το κατώφλι απόφασης του ensemble, ώστε να καταλήξει το ensemble i στην τελική του απόφαση για την κατηγορία στην οποία ανήκει το στιγμιότυπο j.

   Δημιουργούμε όλους τους συνδυασμούς 10 διαφορετικών βαρών και 5 διαφορετικών μοντέλων, δηλαδή $v = 10^5 = 100000$ ensembles.

   Το ensemble με την μεγαλύτερη διακριτική ικανότητα που προέκυψε βάση αυτής της μεθόδου ονομάστηκε Ensemble #3.

3.4.9.2.2 Κατηγοριοποίηση βάση Ψηφοφορίας

Κατά την ψηφοφορία κάθε μοντέλο συμμετέχει στην λήψη απόφασης από το ensemble με μία ψήφο, η οποία είναι η απόφαση του μοντέλου για την κατηγορία στην οποία ανήκει το στιγμιότυπο. Η κάθε ψήφος κάθε μοντέλου μπορεί να έχει διαφορετικό βάρος.

Ο υπολογισμός της πιθανότητας το στιγμιότυπο j να ανήκει στην κατηγορία 2, που υποδηλώνει παρουσία καρκίνου του μαστού, βάση του ensemble i με Ψηφοφορία δίνεται από τον τύπο:

$$\textit{Πιθανότητα Ensemble}_{ij} = \frac{\sum_{k=1}^{m} d_{kj} * w_{kij}}{\sum_{k=1}^{m} w_{kij}} \quad (3.4.8.2.2.1)$$

όπου

- d_{kj} είναι η απόφαση του μοντέλου k για το στιγμιότυπο j και παίρνει τιμές 1 ή 2, βάση του ορισμένου κατωφλίου,
- w_{kij} είναι το βάρος του μοντέλου k για το στιγμιότυπο j στο ensemble i,
- m είναι ο συνολικός αριθμός των μοντέλων που συμμετέχουν στην δημιουργία του ensemble i,
- i είναι ο δείκτης του ensemble και ορίζεται από 1 έως ν, όπου ν είναι το σύνολο των διαφορετικών ensemble.

Χρησιμοποιήσαμε 3 μεθόδους ανάθεσης βάρους για διαφορετικές υλοποιήσεις της Κατηγοριοποίησης βάση Ψηφοφορίας :

1. Η ανάθεση βαρών γίνεται με *Δυναμική Ανάθεση Βάρους βάση Βεβαιότητας*, όπως περιγράφεται στο 3.4.8.1.1.

   Η Πιθανότητα Ensemble_{ij} υπολογίζεται από τον τύπο (3.4.8.2.2.1) και συγκρίνεται με το κατώφλι απόφασης του ensemble, ώστε να καταλήξει το κάθε ensemble i στην τελική του απόφαση για την κατηγορία στην οποία ανήκει το στιγμιότυπο j.

   Δημιουργούμε όλους τους συνδυασμούς των 5 μοντέλων, δηλαδή ν = 2^5 = 32 ensembles.

   Το ensemble με την μεγαλύτερη διακριτική ικανότητα που προέκυψε βάση αυτής της μεθόδου ονομάστηκε Ensemble #4.

2. Η ανάθεση βαρών γίνεται με *Δυναμική Ανάθεση Βάρους βάση Ακρίβεια*, όπως περιγράφεται στο 3.4.8.1.2.

   Η Πιθανότητα $Ensemble_{ij}$ υπολογίζεται από τον τύπο (3.4.8.2.2.1) και συγκρίνεται με το κατώφλι απόφασης του ensemble, ώστε να καταλήξει το ensemble i στην τελική του απόφαση για την κατηγορία στην οποία ανήκει το στιγμιότυπο j.

   Δημιουργούμε όλους τους συνδυασμούς των 5 μοντέλων, δηλαδή $v = 2^5 = 32$ ensembles.

   Το ensemble με την μεγαλύτερη διακριτική ικανότητα που προέκυψε βάση αυτής της μεθόδου ονομάστηκε Ensemble #5.

3. Η ανάθεση βαρών γίνεται με *Γραμμική Ανάθεση Βάρους*, όπως περιγράφεται στο 3.4.8.1.3.

   Η Πιθανότητα $Ensemble_{ij}$ υπολογίζεται από τον τύπο (3.4.8.2.2.1) και συγκρίνεται με το κατώφλι απόφασης του ensemble, ώστε να καταλήξει το ensemble i στην τελική του απόφαση για την κατηγορία στην οποία ανήκει το στιγμιότυπο j.

   Δημιουργούμε όλους τους συνδυασμούς 10 διαφορετικών βαρών και 5 διαφορετικών μοντέλων, δηλαδή $v = 10^5 = 100000$ ensembles.

   Το ensemble με την μεγαλύτερη διακριτική ικανότητα που προέκυψε βάση αυτής της μεθόδου ονομάστηκε Ensemble #6.

3.4.9.2.3 Κατηγοριοποίηση βάση Μέγιστης – Ελάχιστης βεβαιότητας

Βάση αυτής της μεθόδου[150], γίνεται σύγκριση της βεβαιότητας κατηγοριοποίησης c_{kj} κάθε στιγμιοτύπου j μεταξύ όλων των μοντέλων που συμμετέχουν στο ensemble. Το ensemble θα κατηγοριοποιήσει το στιγμιότυπο j στην κλάση που το κατηγοριοποίησε το μοντέλο με την μεγαλύτερη βεβαιότητα. Η βεβαιότητα κατηγοριοποίησης c_{kj} ορίζεται από τον τύπο (3.4.8.1.1.3).

Δημιουργούμε όλους τους συνδυασμούς των 5 μοντέλων, δηλαδή $v = 2^5 = 32$ ensembles.

Το ensemble με την μεγαλύτερη διακριτική ικανότητα που προέκυψε βάση αυτής της μεθόδου ονομάστηκε Ensemble #7.

3.4.9.2.4 Κατηγοριοποίηση με συνδυασμό Μέγιστης – Ελάχιστης βεβαιότητας και Μέσου Όρου

Η μέθοδος αυτή αποτελεί συνδυασμό των μεθόδων Μέγιστης – Ελάχιστης βεβαιότητας και της Κατηγοριοποίησης βάση Μέσου Όρου Πιθανοτήτων. Βάση της μεθόδου αυτής θέτουμε

ένα κατώφλι βεβαιότητας. Εάν υπάρχει έστω κι ένα μοντέλο k το οποίο έχει βεβαιότητα κατηγοριοποίησης c_{kj} μεγαλύτερη ή ίση του κατωφλίου βεβαιότητας τότε το ensemble i κατηγοριοποιεί το στιγμιότυπο j βάση της μεθόδου Μέγιστης – Ελάχιστης βεβαιότητας. Εάν κανένα από τα μοντέλα δεν ξεπερνά το κατώφλι βεβαιότητας που έχουμε θέσει, τότε η κατηγοριοποίηση του στιγμιοτύπου j από το το ensemble i γίνεται με την μέθοδο της *Κατηγοριοποίησης βάση Μέσου Όρου Πιθανοτήτων*. Για την ανάθεση βάρους χρησιμοποιήσαμε 2 μεθόδους :

- *Δυναμική Ανάθεση Βάρους βάση Βεβαιότητας*, όπως περιγράφεται στο 3.4.8.1.1.

  Το ensemble με την μεγαλύτερη διακριτική ικανότητα που προέκυψε βάση αυτής της μεθόδου ονομάστηκε Ensemble #8.
- *Δυναμική Ανάθεση Βάρους βάση Ακρίβειας*, όπως περιγράφεται στο 3.4.8.1.2.

  Το ensemble με την μεγαλύτερη διακριτική ικανότητα που προέκυψε βάση αυτής της μεθόδου ονομάστηκε Ensemble #9.

Εξετάσαμε κατώφλια βεβαιότητας από 0.60 έως και 0.95. Δημιουργούμε όλους τους συνδυασμούς των 5 μοντέλων, δηλαδή $v = 2^5 = 32$ ensembles.

3.4.9.2.5 Κατηγοριοποίηση με συνδυασμό Μέγιστης – Ελάχιστης βεβαιότητας και Ψηφοφορίας

Η μέθοδος αυτή αποτελεί συνδυασμό των μεθόδων Μέγιστης – Ελάχιστης βεβαιότητας και της Κατηγοριοποίησης βάση Ψηφοφορίας. Βάση της μεθόδου αυτής θέτουμε ένα κατώφλι βεβαιότητας. Εάν υπάρχει έστω κι ένα μοντέλο k το οποίο έχει βεβαιότητα κατηγοριοποίησης c_{kj} μεγαλύτερη ή ίση του κατωφλίου βεβαιότητας τότε το ensemble i κατηγοριοποιεί το στιγμιότυπο j βάση της μεθόδου Μέγιστης – Ελάχιστης βεβαιότητας. Εάν κανένα από τα μοντέλα δεν ξεπερνά το κατώφλι βεβαιότητας που έχουμε θέσει, τότε η κατηγοριοποίηση του στιγμιοτύπου j από το το ensemble i γίνεται με την μέθοδο της *Κατηγοριοποίησης βάση Ψηφοφορίας*. Για την ανάθεση βάρους χρησιμοποιήσαμε 2 μεθόδους:

- *Δυναμική Ανάθεση Βάρους βάση Βεβαιότητας*, όπως περιγράφεται στο 3.4.8.1.1.

  Το ensemble με την μεγαλύτερη διακριτική ικανότητα που προέκυψε βάση αυτής της μεθόδου ονομάστηκε Ensemble #10.

- *Δυναμική Ανάθεση Βάρους βάση Ακρίβειας*, όπως περιγράφεται στο 3.4.8.1.2. Το ensemble με την μεγαλύτερη διακριτική ικανότητα που προέκυψε βάση αυτής της μεθόδου ονομάστηκε Ensemble #11.

Εξετάσαμε κατώφλια βεβαιότητας από 0.60 έως και 0.95. Δημιουργούμε όλους τους συνδυασμούς των 5 μοντέλων, δηλαδή $v = 2^5 = 32$ ensembles.

3.4.10 Επικύρωση και αξιολόγηση των μοντέλων συλλογικής μάθησης

Για τις μεθόδους δημιουργίας των ensemble βάση Μέσου Όρου και βάση Ψηφοφορίας, θεωρήσαμε αρχικά το κατώφλι απόφασης των ensembles ίσο με 0.5. Επίσης, στην περίπτωση της δημιουργίας των ensemble βάση Ψηφοφορίας έχουμε ένα ξεχωριστό κατώφλι απόφασης για κάθε ένα από τα 5 μοντέλα τα οποία συμμετέχουν. Αρχικά ορίσαμε τα πέντε αυτά κατώφλια επίσης ίσα με 0.5 .

Κατά την την φάση της επικύρωσης των ensembles η κατηγορία στην οποία τοποθετεί το ensemble κάθε στιγμιότυπο συγκρίνεται με την πραγματική κατηγορία του στιγμιοτύπου. Μετά την ολοκλήρωση όλων των επαναλήψεων της φάσης επικύρωσης, έχουμε στην διάθεσή μας το σύνολο των προβλέψεων (ως πιθανότητα το υπό εξέταση στιγμιότυπο να είναι θετικό) σε μία λίστα καθώς και τα συνολικά αποτελέσματα TP, FP, TN και FN όλων των ensemble που εξετάζονται για την συγκεκριμένη μέθοδο υλοποίησης των ensembles. Βάση αυτών, υπολογίζουμε τα διάφορα μέτρα αξιολόγησης (AUC, Accuracy, Sensitivity και Specificity) για κάθε ensemble.

Η διαδικασία της επικύρωσης και αξιολόγησης πραγματοποιήθηκε για κάθε μία από τις μεθόδους δημιουργίας των ensembles όπως περιγράφηκαν στο κεφάλαιο 3.4.8.2.

3.4.11 Επιλογή βάση επίδοσης

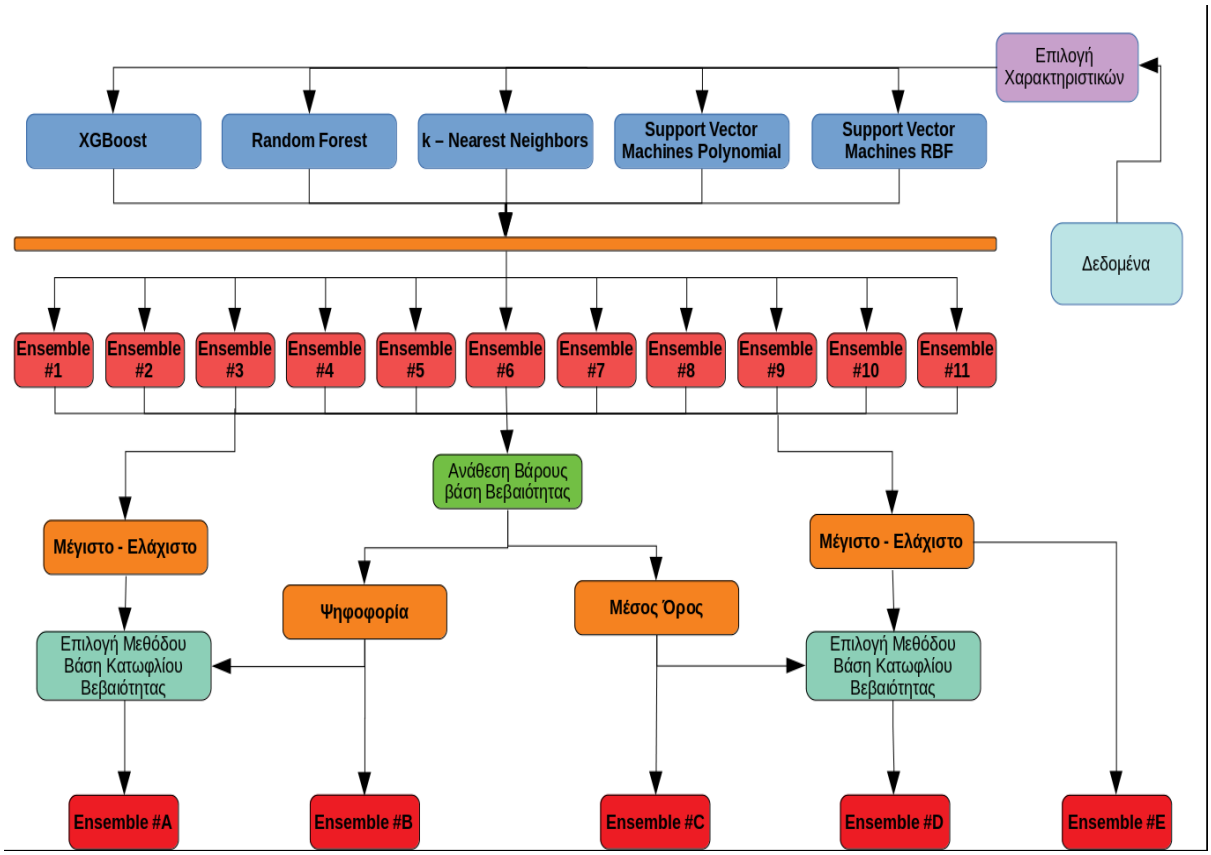
Μετά την ολοκλήρωση της φάσης επικύρωσης και αξιολόγησης, έχουμε ένα αρχείο για κάθε μία από της μεθόδους δημιουργίας των ensembles, το οποίο περιλαμβάνει τις πληροφορίες όλων των ensembles. Σε αυτό το σημείο θέλοντας να εξετάσουμε ακόμα περισσότερους διαφορετικούς συνδυασμούς αποτελεσμάτων αποφασίσαμε να κρατήσουμε τα 8 ensemble με τα καλύτερα αποτελέσματα AUC, Accuracy και Sensitivity κάθε αρχείου και να ελέγξουμε και την απόδοση για διαφορετικές τιμές του κατωφλίου απόφασης του ensemble, αλλά και για διαφορετικές τιμές των κατωφλίων απόφασης σε κάθε ξεχωριστού μοντέλου στην περίπτωση της δημιουργίας των ensemble βάση Ψηφοφορίας.

Στις περίπτωση του συνδυασμού των αποτελεσμάτων βάση Ψηφοφορίας και Μέσου Όρου εξετάστηκαν τιμές του κατωφλίου απόφασης του ensemble από 0.30 έως και 0.70 , ενώ περίπτωση του συνδυασμού των αποτελεσμάτων βάση Ψηφοφορίας εξετάστηκαν και οι συνδυασμοί διαφορετικών τιμών κατωφλίων απόφασης κάθε μοντέλου μηχανικής μάθησης που συμμετείχε στην δημιουργία των ensemble από 0.40 έως και 0.60 .

Μετά την ολοκλήρωση της παραπάνω διαδικασίας επιλέξαμε, για κάθε μία από τις 11 διαφορετικές μεθόδους δημιουργίας μοντέλων συλλογικής μάθησης, το ensemble με την υψηλότερη τιμή Ακρίβειας και ως δεύτερο μέτρο κατάταξης χρησιμοποιήθηκε η Ευαισθησία. Σε περίπτωση ισοβαθμίας η επιλογή ήταν για το μοντέλο με την μικρότερη πολυπλοκότητα. Τα ensemble αυτά ονομάστηκαν Ensemble #1 – 11, όπως φαίνεται στην Εικόνα 3.12.

3.4.12 Συνδυασμός των Ensemble #1-11 για την υλοποίηση μοντέλων συλλογικής μάθησης

Ως τελικό βήμα συνδυάσαμε τα αποτελέσματα των καλύτερων ensembles κάθε μεθόδου για να δημιουργήσουμε νέα μοντέλα συλλογικής μάθησης. Σε αυτό το στάδιο χρησιμοποιήθηκαν και οι πέντε μέθοδοι συνδυασμού των αποτελεσμάτων, αλλά για την ανάθεση βάρους χρησιμοποιήθηκε μόνο η ανάθεση βάρους μέσω βεβαιότητας. Δοκιμάστηκαν όλοι οι συνδυασμοί των Ensemble #1 – 11 και ταυτόχρονα ελέγχθηκαν τιμές κατωφλίου απόφασης από 0.30 έως 0.70, και τιμές κατωφλίου βεβαιότητας από 0.60 έως 0.95 και μετά την ολοκλήρωση της διεξαγωγής των διεργασιών επιλέξαμε, για κάθε μία από τις 5 διαφορετικές μεθόδους δημιουργίας μοντέλων συλλογικής μάθησης, το ensemble με την υψηλότερη τιμή Ακρίβειας και ως δεύτερο μέτρο κατάταξης χρησιμοποιήθηκε η Ευαισθησία. Τα ensemble αυτά ονομάστηκαν Ensemble A, B, C, D και E, και η διαδικασία υλοποίησής τους περιγράφεται σχηματικά στην Εικόνα 3.15.



Εικόνα 3.15 Μεθοδολογία δημιουργίας νέων μοντέλων συλλογικής μάθησης από τα Ensemble #1 - 11.

Κεφάλαιο 4

Αποτελέσματα και Συζήτηση

4.1 Παρουσίαση Αποτελεσμάτων

Σκοπός της παρούσας εργασίας είναι η πρόταση ενός μοντέλου συλλογικής μάθησης για την ανίχνευση του καρκίνου του μαστού βάση ανθρωπομετρικών δεδομένων και αποτελεσμάτων αιματολογικών εξετάσεων. Μετά από την υλοποίηση μοντέλων μηχανικής μάθησης βασισμένων σε 6 διαφορετικούς αλγορίθμους μηχανικής μάθησης, και εξέταση όλων των δυνατών συνδυασμών των διαθέσιμων βιοδεικτών ως χαρακτηριστικά εισόδου των μοντέλων, καταλήξαμε στην επιλογή των πέντε μοντέλων μηχανικής μάθησης με την βέλτιστη απόδοση, όπως αυτά περιγράφονται από τους Πίνακες 4.1 & 4.2, ώστε να αποτελέσουν τα δομικά μέρη για την δημιουργία μοντέλων συλλογικής μάθησης (ensembles).

Πίνακας 4.1 Χαρακτηριστικά μοντέλων βέλτιστης απόδοσης.

Model	Parameters	Threshold	Attributes
XGboost	{'base_score': 0.5, 'booster': 'gbtree', 'colsample_bylevel': 1, 'colsample_bynode': 1, 'colsample_bytree': 0.84, 'gamma': 0.1125, 'learning_rate': 0.2, 'max_delta_step': 0, 'max_depth': 3, 'min_child_weight': 2, 'missing': None, 'n_estimators': 42, 'objective': 'binary:logistic', 'reg_alpha': 0.96, 'reg_lambda': 1, 'scale_pos_weight': 1, 'seed': 50, 'subsample': 0.77, 'verbosity': 1}	0.5	Glucose Resistin Age BMI Adiponectin L/A Ratio
Random Forest	{'bootstrap': True, 'class_weight': None, 'criterion': 'gini', 'max_depth': 4, 'max_features': 'sqrt', 'max_leaf_nodes': None, 'min_impurity_decrease': 0.0, 'min_impurity_split': None, 'min_samples_leaf': 0.001, 'min_samples_split': 7, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 18, 'n_jobs': -1, 'oob_score': False, 'random_state': 50, 'verbose': 0, 'warm_start': False}	0.5	Glucose Resistin Age BMI Leptin Adiponectin
KNN	{'algorithm': 'auto', 'leaf_size': 30, 'metric': 'minkowski', 'metric_params': None, 'n_jobs': -1, 'n_neighbors': 5, 'p': 1, 'weights': 'distance'}	0.5	Glucose Resistin Age BMI Adiponectin
SVM poly	{'C': 2, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': 'ovr', 'degree': 2, 'gamma': 0.0009765625, 'kernel': 'poly', 'max_iter': -1, 'probability': True, 'random_state': 50, 'shrinking': True, 'tol': 0.001, 'verbose': False}	0.5	Glucose Resistin Age BMI L/A Ratio Insulin
SVM rbf	{'C': 0.99, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': 'ovr', 'degree': 3, 'gamma': 0.00038, 'kernel': 'rbf', 'max_iter': -1, 'probability': True, 'random_state': 50, 'shrinking': True, 'tol': 0.001, 'verbose': False}	0.5	Glucose Resistin Age BMI

Πίνακας 4.2 Μετρικές μοντέλων βέλτιστης απόδοσης.

Model	AUC	Accuracy	Sensitivity	Specificity	Training_time
XGboost	86.66	88.79	90.63	86.54	1.63
Random Forest	87.86	87.07	89.06	84.62	25.69
KNN	88.34	86.21	90.63	80.77	11.96
SVM poly	87.05	86.21	87.50	84.62	1.94
SVM rbf	88.40	85.34	87.50	82.69	0.83

Εξετάστηκαν όλοι οι δυνατοί συνδυασμοί των 5 αυτών μοντέλων με διαφορετικές μεθόδους συνδυασμού των αποτελεσμάτων τους όπως αυτές περιγράφηκαν στο 3.4.8.2. Μετά την ολοκλήρωση των διαδικασιών της αξιολόγησης και της επιλογής του βέλτιστου ensemble για κάθε μέθοδο δημιουργίας ensembles, καταλήξαμε στα Ensemble #1 – 11 των οποίων τα χαρακτηριστικά και η απόδοση παρουσιάζονται στον Πίνακα 4.3.

Πίνακας 4.3 Χαρακτηριστικά των Ensemble #1 - 11.

Ensemble	Threshold	Models Used					Model Combining Method	Weighting	AUC	ACC	SENS	SPEC
		XGB	RF	KNN	SVM-Poly	SVM-RBF						
#1	0.32	1	1	0	0	1	Averaging	Certainties	0.891	87.07	89.06	84.62
#2	0.52	0	0	1	1	1	Averaging	Accuracy	0.900	87.93	89.06	86.54
#3	0.50	5	1	0	0	0	Averaging	Linear	0.885	88.79	90.63	86.54
#4	0.63	1	1	1	1	0	Voting	Certainties	0.938	91.38	89.06	94.23
#5	0.34	1	0	1	0	1	Voting	Accuracy	0.913	89.66	95.31	82.69
#6	0.55	4	6	2	2	1	Voting	Linear	0.937	91.38	89.06	94.23
#7	-	1	0	0	0	1	Min – Max	-	0.878	87.07	96.88	75.00
#8	0.45	0	0	1	0	1	Min – Max & Averaging	Certainties	0.888	87.93	96.88	76.92
#9	0.51	1	1	0	0	0	Min – Max & Averaging	Accuracy	0.858	87.93	90.63	84.62
#10	0.63	1	1	1	1	0	Min – Max & Voting	Certainties	0.920	89.66	89.06	90.38
#11	0.40	1	1	0	0	1	Min – Max & Voting	Accuracy	0.929	89.66	92.19	86.54

Στην συνέχεια, εξετάστηκαν όλοι οι δυνατοί συνδυασμοί των 11 αυτών μοντέλων συλλογικής μάθησης, με 5 διαφορετικές μεθόδους συνδυασμού των αποτελεσμάτων τους και ανάθεση βάρους βάση βεβαιότητας. Μετά την ολοκλήρωση των διαδικασιών της αξιολόγησης και της επιλογής του βέλτιστου ensemble για κάθε μία από τις 5 μεθόδους

δημιουργίας μοντέλων συλλογικής μάθησης, καταλήξαμε στα Ensemble A - E των οποίων τα χαρακτηριστικά και η απόδοση παρουσιάζονται στον Πίνακα 4.4.

Πίνακας 4.4 Χαρακτηριστικά των Ensemble A, B, C, D & E.

Ensemble	Threshold	Max Limit	Ensembles Used	Model Combining Method	Weighting	AUC	ACC	SENS	SPEC
A	0.50	0.70	2 4 5 6	Min – Max & Voting	Certainties	0.928	92.24	93.75	90.38
B	0.39		1 4 6	Voting	Certainties	0.919	92.24	90.63	94.23
C	0.50		1 3 5 6 9	Averaging	Certainties	0.894	88.79	90.63	86.54
B	0.55	0.60	3 6 11	Min – Max & Averaging	Certainties	0.889	88.79	90.63	86.54
E	0.52		9 11	Min – Max	-	0.872	87.93	90.63	84.62

4.2 Παρατηρήσεις επί των αποτελεσμάτων των ensembles

Στον Πίνακα 4.3 βλέπουμε ότι μέσω του συνδυασμού των αποτελεσμάτων των μοντέλων μηχανικής μάθησης, καταφέραμε να δημιουργήσουμε μοντέλα με καλύτερη ανιχνευτική ικανότητα από όλα τα αρχικά μοντέλα. Τα μοντέλα αυτά συλλογικής μάθησης είναι τα: Ensemble #3, #4, #5, #6, #10 και #11. Από αυτά, τα Ensemble #4 και Ensemble #6 επιτυγχάνουν την μέγιστη τιμή Ακρίβειας. Και στα 2 ensemble βέλτιστης απόδοσης χρησιμοποιείται η μέθοδος της Κατηγοριοποίησης Βάση Ψηφοφορίας, ενώ τα αποτελέσματα με χρήση Κατηγοριοποίησης Βάση Ψηφοφορίας είναι συνολικά καλύτερα σε σύγκριση με τα αντίστοιχα αποτελέσματα με χρήση Κατηγοριοποίησης Βάση Μέσου Όρου. Παρατηρούμε επίσης, ότι και στις περιπτώσεις όπου έχουμε συνδυασμό της μεθόδου Κατηγοριοποίησης βάση Μέγιστης – Ελάχιστης βεβαιότητας με χρήση Μέσου Όρου ή Ψηφοφορίας, τα ensemble στα οποία χρησιμοποιείται η μέθοδος της Κατηγοριοποίησης Βάση Ψηφοφορίας αποδίδουν καλύτερα από τα ensemble στα οποία χρησιμοποιείται η μέθοδος της Κατηγοριοποίησης Βάση Μέσου Όρου.

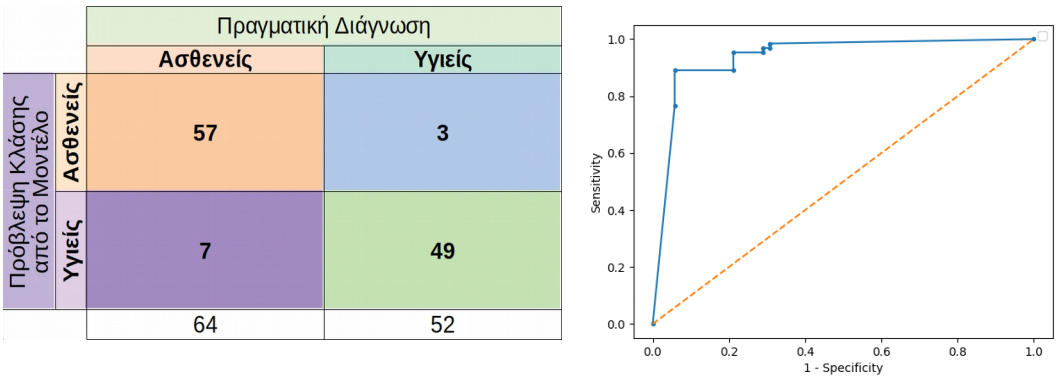
Στον Πίνακα 4.4 βλέπουμε ότι μέσω του συνδυασμού των αποτελεσμάτων των μοντέλων συλλογικής μάθησης, καταφέραμε να δημιουργήσουμε εκ νέου μοντέλα με καλύτερη ανιχνευτική ικανότητα. Τα μοντέλα αυτά συλλογικής μάθησης είναι τα: Ensemble A και Ensemble B. Και στα 2 ensemble βέλτιστης απόδοσης χρησιμοποιείται η μέθοδος της Κατηγοριοποίησης βάση Ψηφοφορίας είτε μόνη είτε σε συνδυασμό με Κατηγοριοποίηση βάση Μέγιστης – Ελάχιστης βεβαιότητας.

4.3 Ανάλυση των βέλτιστων Μοντέλων Συλλογικής Μάθησης και επιλογή του προτεινόμενου

Τα 2 ensemble πρώτης φάσης τα οποία επιτυγχάνουν την μέγιστη τιμή Ακρίβειας 91.38% είναι τα εξής :

- **Ensemble #4**

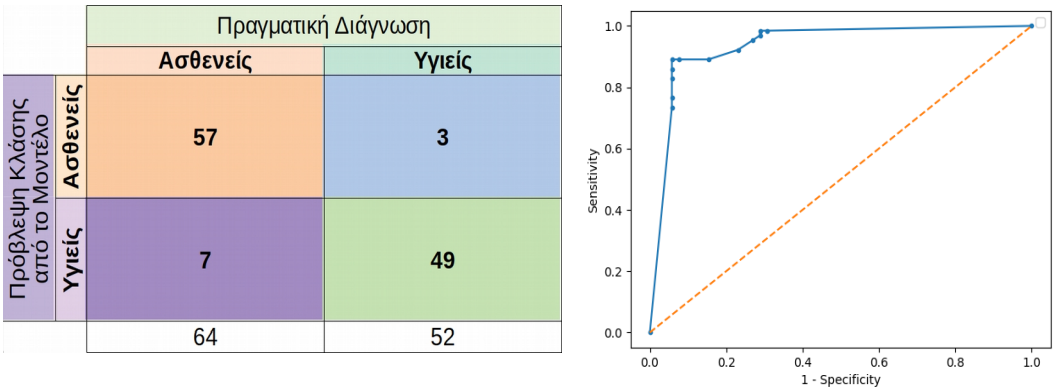
  Αποτελεί συνδυασμό των μοντέλων XGB, RF, k-NN και SVM poly με κατώφλι απόφασης 0.5 για όλα τα μοντέλα. Η ανάθεση βαρών έγινε βάση Βεβαιότητας και η κατηγοριοποίηση βάση Ψηφοφορίας. Το κατώφλι απόφασης του Ensemble #4 είναι 0.63.



Εικόνα 4.1 Confusion Matrix και Καμπύλη ROC του Ensemble #4.

- **Ensemble #6**

  Αποτελεί συνδυασμό των μοντέλων XGB, RF, k-NN, SVM poly και SVM rbf με βάρη 4, 6, 2, 2, 1 αντίστοιχα και κατώφλι απόφασης 0.5 για όλα τα μοντέλα. Έγινε Γραμμική ανάθεση βαρών και η κατηγοριοποίηση βάση Ψηφοφορίας. Το κατώφλι απόφασης του Ensemble #6 είναι 0.5.

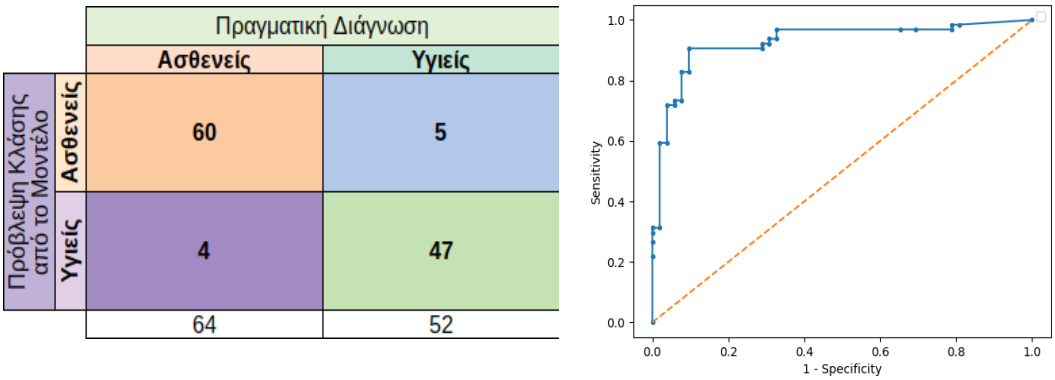


Εικόνα 4.2 Confusion Matrix και Καμπύλη ROC του Ensemble #6.

Τα 2 ensemble δεύτερης φάσης τα οποία επιτυγχάνουν την μέγιστη τιμή Ακρίβειας 92.24% είναι τα εξής :

- **Ensemble A**

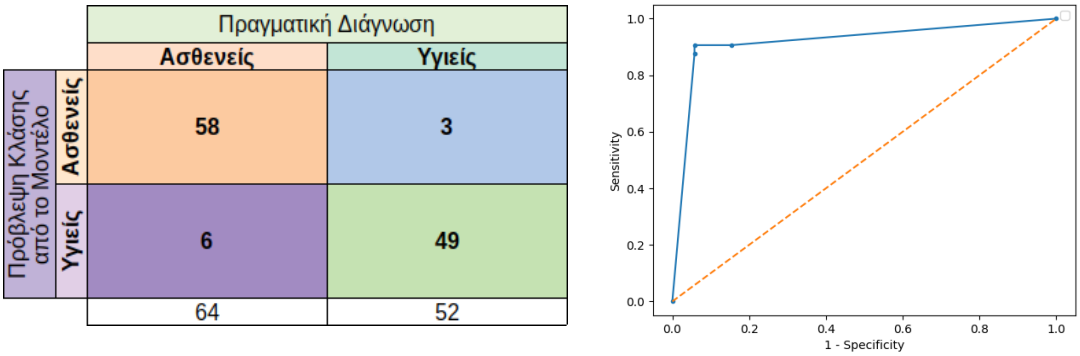
  Αποτελεί συνδυασμό των Ensemble #2, #4, #5 και #6. Η ανάθεση βαρών έγινε βάση Βεβαιότητας και η κατηγοριοποίηση με Συνδυασμό Μέγιστης – Ελάχιστης βεβαιότητας και Ψηφοφορίας, με κατώφλι βεβαιότητας 0.70. Το κατώφλι απόφασης του Ensemble A είναι 0.50.



Εικόνα 4.3 Confusion Matrix και Καμπύλη ROC του Ensemble A.

- **Ensemble B**

  Αποτελεί συνδυασμό των Ensemble #1, #4 και #6. Η ανάθεση βαρών έγινε βάση Βεβαιότητας και η κατηγοριοποίηση βάση Ψηφοφορίας. Το κατώφλι απόφασης του Ensemble B είναι 0.39.



Εικόνα 4.4 Confusion Matrix και Καμπύλη ROC του Ensemble B.

Βλέπουμε, ότι τα Ensemble A & B, παρά το γεγονός ότι έχουν χαμηλότερη τιμή AUC από τα Ensemble #4 & #6, μετά από την εξέταση διαφορετικών τιμών κατωφλίου απόφασης έχουν την μεγαλύτερη ανιχνευτική ικανότητα πετυχαίνοντας την υψηλότερη τιμή Ακρίβειας μεταξύ όλων των μοντέλων. Βάση του ότι η Ακρίβεια Κατηγοριοποίησης των Ensemble A και B είναι η ίδια θεωρούμε ότι το επόμενο κριτήριο για την επιλογή του βέλτιστου μοντέλου κατηγοριοποίησης είναι η τιμή της Ευαισθησίας. Η επιλογή της Ευαισθησίας έγινε διότι θεωρούμε ότι η σωστή ανίχνευση του καρκίνου του μαστού είναι μεγαλύτερης σημαντικότητας από την σωστή γνώση ότι δεν υπάρχει καρκίνος του μαστού.

Βάση αυτής της λογικής καταλήγουμε στην επιλογή του Ensemble A ως το προτεινόμενο μοντέλο συλλογικής μάθησης της προκείμενης εργασίας.

4.4 Σύγκριση με αντίστοιχες εργασίες της βιβλιογραφίας

Τα αποτελέσματα των εργασιών, τις οποίες βρήκαμε κατά την διάρκεια της εκπόνησης της βιβλιογραφικής επισκόπησης, και οι οποίες χρησιμοποίησαν την βάση δεδομένων Coimbra Breast Cancer Dataset για την ανάπτυξη προβλεπτικών μοντέλων μηχανικής μάθησης για την ανίχνευση του καρκίνου του μαστού, παρουσιάζονται συνοπτικά στον Πίνακα 4.5.

Πίνακας 4.5 Εργασίες πάνω στην βάση δεδομένων Coimbra Breast Cancer Dataset.

	Date	Algorithm	AUC	Accuracy	Sensitivity	Specificity	F1-score	Precision
M. Patrício et al.[98]	Jan 2018	Support Vector Machines	[0.87, 0.91]	-	[82%, 88%]	[84%, 90%]	-	-
Muhammet Fatih Aslan et al. [86]	Nov 2018	Extreme Learning Machine	-	80.00%	-	-	-	-
Lara Dular[88]	2018	Random Forest	0.805	75.00%	-	-	-	-
K. Polat et al.[89]	Dec 2018	AdaBoost.M1+ KMCBFW	**0.938**	**91.37%**	**91.40%**	**92.30%**	**0.914**	**0.919**
Silva Araújo V.J. et al.[87]	Jan 2019	Fuzzy Neural Network OrNet	0.802	81.04%	81.93%	81.18%	-	-
Mohaimenul Islam et al.[96]	Aug 2019	k-Nearest Neighbors	0.950	86.00%	80.00%	91.00%	-	-
Ensemble A	Feb 2020	XGBoost + Random Forest + k-NN + SVM	**0.928**	**92.24%**	**93.75%**	**90.38%**	**0.930**	**0.923**

Στον Πίνακα 4.5 βλέπουμε ότι το Ensemble A έχει την καλύτερη ανιχνευτική ικανότητα μεταξύ των υπολοίπων προτεινόμενων μοντέλων. Χαρακτηριστικά παρουσιάζει καλύτερη επίδοση στην Ακρίβεια, την Ευαισθησία, το F1-score και το Precision σε σύγκριση το με υβριδικό μοντέλο AdaBoost.M1 με στάθμιση χαρακτηριστικών βασισμένη σε K-means clustering (KMCBFW) το οποίο πρότειναν οι K. Polat et al.[89] και αποτελεί την State ot the Art προσέγγιση.

Ακόμη περισσότερο, το Ensemble A φαίνεται να έχει καλύτερη Ακρίβεια στην ανίχνευση του καρκίνου του μαστού από τα μοντέλα τα οποία περιγράφηκαν στο 2.4.2 και κάνουν χρήση σωματομετρικών δεδομένων και εξάγουν πληροφορία από το Cell Free DNA ή/και τιμές βιοδεικτών οι οποίοι μπορούν να μετρηθούν μέσω αιματολογικών εξετάσεων. Πρέπει βέβαια εδώ να αναφέρουμε ότι οι βάσεις δεδομένων είναι διαφορετικές επομένως δεν μπορεί να γίνει άμεση σύγκριση.

Τέλος, το Ensemble A φαίνεται ότι έχει καλύτερη Ακρίβεια ανίχνευσης του καρκίνου του μαστού και από τα μοντέλα Βαθιάς Μάθησης που κάνουν ανίχνευση του καρκίνου του μαστού με εξέταση εικόνων μαστογραφίας, τα οποία περιγράφηκαν στο 2.3. Και εδώ όμως πρέπει να αναφερθεί ότι οι βάσεις δεδομένων είναι διαφορετικές επομένως δεν μπορεί να γίνει άμεση σύγκριση μεταξύ των μοντέλων.

4.5 Περαιτέρω ανάλυση των αποτελεσμάτων του προτεινόμενου μοντέλου

Αφού επιλέχθηκε το μοντέλο συλλογικής μάθησης το οποίο θα αποτελέσει και το προτεινόμενο μοντέλο αυτής της εργασίας, προχωρήσαμε μια βαθύτερη ανάλυση της απόδοσης του μοντέλου, εξετάζοντας την απόδοσή του σε διάφορες υποομάδες του πληθυσμιακού δείγματος, όπως παρουσιάζεται στον Πίνακα 4.6.

Πίνακας 4.6 Απόδοση του Ensemble A ανά BMI και ηλικία.

	Accuracy	Sensitivity	Specificity	Confusion Matrix Condition +	Condition -	
Σύνολο Πληθυσμού	92.24%	93.75%	90.38%	60	5	+ Prediction
				4	47	- Prediction
				64	52	Total
Πληθυσμός με BMI > 25	90.79%	92.68%	88.57%	38	4	+ Prediction
				3	31	- Prediction
				41	35	Total
Πληθυσμός με BMI <= 25	95.00%	95.65%	94.12%	22	1	+ Prediction
				1	16	- Prediction
				23	17	Total
Πληθυσμός με Ηλικία < 50	93.62%	92.86%	94.74%	26	1	+ Prediction
				2	18	- Prediction
				28	19	Total
Πληθυσμός με 50<=Ηλικία <= 65	88.00%	100.00%	57.14%	18	3	+ Prediction
				0	4	- Prediction
				18	7	Total
Πληθυσμός με Ηλικία > 65	93.18%	88.89%	96.15%	16	1	+ Prediction
				2	25	- Prediction
				18	26	Total

4.5.1 Απόδοση ανάλογα με τον BMI

Στο πληθυσμιακό δείγμα της βάσης Coimbra Breast Cancer Dataset υπάρχουν 76 περιπτώσεις υπέρβαρων ($BMI > 25$), εκ των οποίων 41 έπασχαν από καρκίνο του μαστού και 35 αποτελούσαν την υγιεί ομάδα ελέγχου και 40 περιπτώσεις με φυσιολογικό σωματικό βάρος ($BMI \leqslant 25$), εκ των οποίων 23 έπασχαν από καρκίνο του μαστού και 17 αποτελούσαν την υγιεί ομάδα ελέγχου.

Όπως βλέπουμε στον Πίνακα 4.6, το Ensemble A έχει καλύτερη απόδοση στην ομάδα των μη παχύσαρκων, κάτι το οποίο αποτελεί μία ευχάριστη έκπληξη, εάν αναλογιστούμε το γεγονός ότι η βάση δεδομένων χρησιμοποιήθηκε αρχικά στην έρευνα των Crisóstomo J et al. [97], τα αποτελέσματα της οποίας δείχνουν ότι ο δυσμεταβολισμός της γλυκόζης, η αντίσταση στην ινσουλίνη και οι μεταβολές στην έκκριση των αντιποκινών, μπορεί να εμπλέκονται στην ανάπτυξη και εξέλιξη του καρκίνου του μαστού σε υπέρβαρες / παχύσαρκες γυναίκες. Παρόλα αυτά, η Ευαισθησία του μοντέλου είναι αυξημένη στον υπέρβαρο πληθυσμό, κάτι που μπορεί να οφείλεται στο γεγονός ότι στο πληθυσμιακό δείγμα μας, τα καρκινώματα επί του υπέρβαρου πληθυσμού ήταν σε σχετικά πιο προχωρημένο στάδιο.

Βλέπουμε επομένως ότι η ικανότητα ανίχνευσης του καρκίνου του μαστού από το προτεινόμενο μοντέλο Ensemble A δεν είναι περιορισμένη στην ομάδα των υπέρβαρων γυναικών.

4.5.2 Απόδοση ανάλογα με την ηλικιακή ομάδα

Στο πληθυσμιακό δείγμα της βάσης Coimbra Breast Cancer Dataset υπάρχουν 47 γυναίκες με ηλικία μικρότερη των 50 ετών, εκ των οποίων 28 έπασχαν από καρκίνο του μαστού και 19 αποτελούσαν την υγιεί ομάδα ελέγχου, 25 γυναίκες με ηλικία από 50 έως και 65 ετών, εκ των οποίων 18 έπασχαν από καρκίνο του μαστού και 7 αποτελούσαν την υγιεί ομάδα ελέγχου και 44 γυναίκες με ηλικία μεγαλύτερη των 65 ετών, εκ των οποίων 18 έπασχαν από καρκίνο του μαστού και 26 αποτελούσαν την υγιεί ομάδα ελέγχου.

Όπως βλέπουμε στον Πίνακα 4.6, το Ensemble A έχει καλύτερη απόδοση στις ηλικιακές ομάδες κάτω των 50 και άνω των 65 ετών. Συγκεκριμένα για την ηλικιακή ομάδα κάτω των 50 ετών το Ensemble A παρουσιάζει την μέγιστη ανιχνευτική ικανότητα έχοντας ακρίβεια 93.62%. Παράλληλα, όπως αναφέρεται από τους Keen JD et al.[151] η μαστογραφία έχει μικρότερη απόδοση στην ηλικιακή αυτή ομάδα.

Από την άλλη μεριά, βλέπουμε ότι το Ensemble A παρουσιάζει μη ισορροπημένη απόδοση στις γυναίκες με ηλικία από 50 έως και 65 ετών, έχοντας 100% Ευαισθησία και μόλις 57.14% Ειδικότητα, κάτι που λογικά οφείλεται στην ανισορροπία των δεδομένων αφού έχουμε 18 καρκινοπαθείς με μόλις 7 υγιείς σε αυτό το ηλικιακό φάσμα. Συν αυτού, το δείγμα μας σε αυτή την περίπτωση είναι εξαιρετικά μικρό.

Κεφάλαιο 5

Συμπεράσματα – Μελλοντικές προτάσεις

5.1 Συμπεράσματα

Στην παρούσα διπλωματική εργασία παρουσιάστηκε η διαδικασία ανάπτυξης ενός μοντέλου μηχανικής μάθησης για την ανίχνευση του καρκίνου του μαστού βάση αποτελεσμάτων αιματολογικών εξετάσεων και ανθρωπομετρικών δεδομένων.

Πιο συγκεκριμένα η διαδικασία αυτή είχε ως αποτέλεσμα ένα μοντέλο συλλογικής μάθησης, το οποίο αποκαλούμε Ensemble A και αποτελεί συνδυασμό τεσσάρων διαφορετικών μοντέλων Συλλογικής Μάθησης, τα οποία με την σειρά τους αποτελούν συνδυασμό πέντε διαφορετικών μοντέλων μηχανικής μάθησης, βασισμένων στους αλγορίθμους XGBoost, Random Forest, k-Nearest Neightbors και Support Vector Machines, με χρήση δύο διαφορετικών πυρήνων.

Για την λήψη απόφασης από το Ensemble A απαιτούνται η Ηλικία, ο Δείκτης Μάζας Σώματος και οι τιμές συγκέντρωσης ορού της Γλυκόζης και των τριών αντιποκινών : Ρεζιστίνη, Λεπτίνη και Αντιπονεκτίνη, η τιμή συγκέντρωσης πλάσματος της Ινσουλίνης και ο υπολογισμός του Λόγου Λεπτίνης/Αντιπονεκτίνης. Τα χαρακτηριστικά εισόδου κάθε Μοντέλου Μηχανικής Μάθησης που συμμετέχει στην δημιουργία των Μοντέλων Συλλογικής Μάθησης είναι :

- XGBoost :
  Ηλικία, BMI, Γλυκόζη, Ρεζιστίνη, Αντιπονεκτίνη και Λόγος Λεπτίνης/Αντιπονεκτίνης
- Random Forest :
  Ηλικία, BMI, Γλυκόζη, Ρεζιστίνη, Λεπτίνη και Αντιπονεκτίνη
- k-Nearest Neightbors :
  Ηλικία, BMI, Γλυκόζη, Ρεζιστίνη και Αντιπονεκτίνη
- Support Vector Machines με Πολυωνυμικό Πυρήνα (Polynomial) :
  Ηλικία, BMI, Γλυκόζη και Ρεζιστίνη, Λόγος Λεπτίνης/Αντιπονεκτίνης και Ινσουλίνη
- Support Vector Machines με Πυρήνα ακτινωτής βάσης (RBF) :
  Ηλικία, BMI, Γλυκόζη και Ρεζιστίνη

Για την εκπαίδευση και αξιολόγηση των Μοντέλων Μηχανικής Μάθησης που συμμετέχουν στο Ensemble A χρησιμοποιήθηκε η μέθοδος Leave One Out Cross Validation πάνω στην βάση Coimbra Breast Cancer Dataset η οποία είναι δημοσίως διαθέσιμη μέσω της βιβλιοθήκης UCI Machine Learning Repository. Το προτεινόμενο μοντέλο έδειξε ότι έχει σημαντική ικανότητα ανιχνεύσεις του καρκίνου του μαστού επιτυγχάνοντας AUC : 0.928 , Ακρίβεια (Accuracy) : 92.24% , Ευαισθησία (Sensitivity) : 93.75% και Ειδικότητα (Specificity) : 90.38% .

Βάση αυτών των αποτελεσμάτων, βλέπουμε ότι το προτεινόμενο ensemble παρουσιάζει καλύτερη επίδοση στην Ακρίβεια, την Ευαισθησία, το F1-score και το Precision σε σύγκριση το με υβριδικό μοντέλο AdaBoost.M1 με στάθμιση χαρακτηριστικών βασισμένη σε K-means clustering (KMCBFW) το οποίο πρότειναν οι K. Polat et al.[89] και αποτελεί την State ot the Art προσέγγιση.

Το Ensemble A δείχνει μάλιστα να έχει καλύτερη ανιχνευτική ικανότητα και από μοντέλα τεχνητής νοημοσύνης τα οποία κάνουν χρήση εικόνων ψηφιακής μαστογραφίας για την λήψη της απόφασής τους, αλλά και από τον μέσο όρο της εξέτασης της σύγχρονης ψηφιακής μαστογραφίας μαστογραφίας από ειδικό ακτινολόγο, η οποία σύμφωνα με τους Lehman CD et al.[45] έχει μέση τιμή ευαισθησίας (Sensitivity) 86.9% και ειδικότητας (Specificity) 88.9%.

Βάση αυτών των αποτελεσμάτων θεωρούμε ότι η χρήση της διαδικασίας που ακολουθήθηκε σε αυτή την εργασία θα μπορούσε να οδηγήσει στην υλοποίηση ενός αντίστοιχου μοντέλου Συλλογικής Μάθησης, το οποίο χειρίζεται ανθρωπομετρικά δεδομένα και τιμές αιματολογικών μεταβολικών δεικτών, το οποίο θα μπορούσε να αποτελέσει προσθήκη στα προγράμματα προληπτικών εξετάσεων για την ανίχνευση του καρκίνου του μαστού.

Ειδικά στις γυναίκες με ηλικία μικρότερης των 50 ετών, για τις οποίες δεν συστήνεται προληπτική εξέταση μαστογραφίας, λόγο της μειωμένης συχνότητας εμφάνισης του καρκίνου του μαστού, του γεγονότος ότι η μαστογραφία εκθέτει την γυναίκα σε ιονίζουσα ακτινοβολία σε συνδυασμό με την μειωμένη ανιχνευτική ικανότητας της μαστογραφίας, θα μπορούσε να αποτελεί την κύρια προληπτική εξέταση για την ανίχνευση του καρκίνου του μαστού. Το ίδιο θα μπορούσε να ισχύει και για τις γυναίκες άνω των 65 ετών, όπου το προτεινόμενο μοντέλο έχει επίσης υψηλή ανιχνευτική ικανότητα, ώστε να μην χρειάζεται να υποβάλλονται στην διαδικασία της μαστογραφίας.

Στα θετικά σημεία του μοντέλου συγκαταλέγεται και το χαμηλό κόστος, αφού εάν θεωρήσουμε ότι οι αιματολογικές εξετάσεις αυτού του είδους θα πραγματοποιούνται σε

κλίμακα τέτοια ώστε τα set των διαγνωστικών εξετάσεων να αγοράζονται σε πακέτα των 96, τα υλικά για τις εξέτασης των τιμών Γλυκόζης, Ινσουλίνης, Ρεζιστίνης, Λεπτίνης και Αντιπονεκτίνης κοστίζουν περίπου 10 ευρώ.

Επίσης, η πλειονότητα των ασθενών συμμετεχόντων διαγνώστηκαν με καρκίνο πρώιμου σταδίου, όντας το 56.25% των περιπτώσεων σε στάδιο 0 ή I, ενώ οι πιο προχωρημένες περιπτώσεις είναι σταδίου II, κάτι που σημαίνει ότι ο καρκίνος δεν έχει επεκταθεί σημαντικά και παραμένει στην περιοχή του μαστού και στην πλειονότητα των περιπτώσεων, ανταποκρίνεται θετικά στην θεραπεία. Επομένως, η μέθοδος έχει την ικανότητα να ανιχνεύσει έγκαιρα τον καρκίνο του μαστού.

Το γεγονός αυτό το κάνει μία προσιτή επιλογή ως κύρια προληπτική εξέταση ανίχνευσης του καρκίνου του μαστού σε λιγότερο ανεπτυγμένες χώρες, όπου η έλλειψη οικονομικής δυνατότητας από το κράτος οδηγεί σε ανίχνευση του καρκίνου του μαστού σε εξελιγμένο στάδιο. Επίσης, σε περιοχές όπου η πρόσβαση είναι εξαιρετική δύσκολη με αποτέλεσμα η μεταφορά ογκώδους ιατρικού εξοπλισμού, όπως ο μαστογράφος, να μην αποτελεί επιλογή, η ύπαρξη μίας διαγνωστικής εξέτασης η οποία στηρίζεται στην λήψη και ανάλυση δείγματος αίματος αποτελεί μία πολύ ευέλικτη επιλογή.

Επίσης, οι εξετάσεις των απαραίτητων βιοδεικτών για το μοντέλο θα μπορούσαν απλά να προστεθούν στην λίστα των δεικτών που ελέγχονται κατά τις αιματολογικές εξετάσεις ρουτίνας των γυναικών από κάποια ηλικία και μετά ώστε να ελαχιστοποιηθεί με αυτό τον τρόπο η επεμβατικότητα, ενώ ταυτόχρονα έχουμε μείωση των συνολικά απαιτούμενων εξετάσεων και του αριθμού των μετακινήσεων του ατόμου σε διαγνωστικά κέντρα.

Τέλος, αποτελεί μία ασφαλή μέθοδο, καθώς η λήψη δείγματος αίματος, όταν πραγματοποιείται από ειδικά εκπαιδευμένο προσωπικό, δεν εκθέτει το άτομο σε παράγοντες οι οποίοι μπορεί να λειτουργούν βλαπτικά στον οργανισμό.

Ως αρνητικό σημείο αυτής της εργασίας πρέπει να αναφέρουμε το μικρό μέγεθος του δείγματος. Το πρόβλημα αντιμετωπίστηκε με την εφαρμογή της μεθόδου LOOCV κατά την εκπαίδευση, επικύρωση και αξιολόγηση του μοντέλου ώστε να είναι διαθέσιμα όσο το δυνατών περισσότερα στιγμιότυπα κατά την διαδικασία εκπαίδευσης. Παρόλα αυτά, το μέγεθος του δείγματος δεν μας επιτρέπει να προτείνουμε το μοντέλο Ensemble A ως μια μέθοδο ανίχνευσης του καρκίνου του μαστού η οποία θα μπορούσε να εφαρμοστεί άμεσα, καθώς για να γίνει αυτό θα πρέπει να δοκιμαστεί η απόδοσή του σε μεγαλύτερο πληθυσμιακό δείγμα. Παρόλα αυτά, μπορούμε να προτείνουμε την μέθοδο που ακολουθήθηκε σε αυτή την εργασία ως βάση για την δημιουργία ενός μοντέλου Συλλογικής Μάθησης για την ανίχνευση

του καρκίνου του μαστού μέσων ανθρωπομετρικών δεδομένων, τις τιμές αίματος Γλυκόζης, Ινσουλίνης, Ρεζιστίνης, Λεπτίνης, Αντιπονεκτίνης και του λόγου Λεπτίνης/Αντιπονεκτίνης.

5.2 Παρατηρήσεις και Προτάσεις για Μελλοντικές Έρευνες

Ολοκληρώνοντας την εργασία θα μπορούσαμε να αναφέρουμε κάποιες παρατηρήσεις που κάναμε κατά την διάρκεια της εκπόνησής της, από τις οποίες θα μπορούσαν να επωφεληθούν μελλοντικές έρευνες καθώς προτάσεις για περαιτέρω μελλοντική έρευνα :

- Στην βάση Coimbra Breast Cancer Dataset δεν υπάρχει πληροφορία για την εμμηνοπαυσιακή φάση στην οποία βρίσκεται η γυναίκα. Η εμμηνοπαυσιακή φάση και η ηλικία στην οποία ξεκινά η εμμηνόπαυση φαίνεται να είναι συσχετισμένα με τον κίνδυνο ανάπτυξης του καρκίνου του μαστού[56]. Μία τέτοια πληροφορία ίσως να μπορούσε να αυξήσει την ανιχνευτική ικανότητα του μοντέλου, αλλά εκτός αυτού θα βοηθούσε και στην ανάλυση των αποτελεσμάτων του μοντέλου. Με παρόμοια λογική θα μπορούσαν να συμπεριληφθούν και πληροφορίες για τον αριθμό τέκνων και την ηλικία έναρξης της εμμηνόρροιας. Όλες αυτές οι πληροφορίες θα ήταν εύκολο να καταγραφούν κατά την διάρκεια της συνέντευξης των συμμετεχόντων.
- Τα χαρακτηριστικά του καρκινώματος (στάδιο, βαθμός, μέγεθος και εξάπλωση στους λεμφαδένες) κάθε ασθενούς, ενώ ήταν γνωστά την αρχική έρευνα, δεν ήταν διαθέσιμα μέσω της UCI Machine Learning Repository στην βάση Coimbra Breast Cancer Dataset. Η πληροφορία αυτή θα μας επέτρεπε να αναλύσουμε περαιτέρω την αποτελεσματικότητα της μεθόδου στην ανίχνευση του καρκίνου του μαστού σε αρχικό στάδιο. Επίσης, θα μπορούσε να τροποποιηθεί το μοντέλο ώστε να εκτιμά εκτός από την ύπαρξη του καρκίνου του μαστού και χαρακτηριστικά του καρκινώματος.
- Όπως έχει ήδη αναφερθεί, το μοντέλο πρέπει να δοκιμαστεί σε μεγαλύτερης έκτασης δείγμα. Επίσης, ειδικά για τις ηλικίες από 50 έως 65 οι οποίες αποτελούν μία ηλικιακή ομάδα κατά την οποία τα ποσοστά εμφάνισης του καρκίνου του μαστού είναι πολύ υψηλά, το δείγμα ήταν εξαιρετικά μικρό. Μία ακόμα παρατήρηση είναι ότι θα ήταν καλό το δείγμα να είναι πιο ισορροπημένο μεταξύ των κλάσεων (ασθενείς – υγιείς) ως προς την ηλικία και τον BMI. Επίσης, ο μέσος όρος του BMI όλου του

δείγματος θα πρέπει να είναι πιο κοντά στο όριο του φυσιολογικού καθώς στην περίπτωση αυτή ήταν αρκετά υψηλό, όντας 27.58.

- Σε περίπτωση επέκτασης του δείγματος ενδιαφέρον θα ήταν και η προσθήκη και δείγματος ανδρικού πληθυσμού.
- Ενδιαφέρων θα παρουσίαζε η δυνατότητα του ελέγχου της αποτελεσματικότητας της μεθόδου ανάλογα με την πυκνότητα του μαστού. Οι γυναίκες με υψηλή πυκνότητα ινοαδενικών στοιχείων έχουν μεγαλύτερη πιθανότητα εμφάνισης του καρκίνου του μαστού και λόγο της πυκνότητας του μαστού η μαστογραφία έχει μειωμένη ανιχνευτική ικανότητα.
- Συνεχείς παρακολούθηση των υγειών συμμετεχόντων κατά την διάρκεια των προληπτικών εξετάσεων για την ανίχνευση του καρκίνου του μαστού, ώστε να εξεταστεί εάν το μοντέλο θα μπορούσε να χρησιμοποιηθεί και για πρόβλεψη.

Πηγές και Βιβλιογραφία

[1] International Agency for Research on Cancer
https://www.iarc.fr/

[2] Λάγιου Α., 2008, Επιδημιολογία και πρόληψη του καρκίνου του μαστού, Αρχεία Ελληνικής Ιατρικής 2008, 25(6):742-748.

[3] https://ww5.komen.org/BreastCancer/GettingOlder.html

[4] https://www.cancer.net/navigating-cancer-care/diagnosing-cancer/stages-cancer

[5] https://www.cancer.nsw.gov.au/understanding-cancer/what-are-the-different-stages-of-cancer

[6] https://fyssas.gr/eidi-karkinou-mastou/

[7] https://www.metropolitan-hospital.gr/el/υπηρεσίες/πρωτοποριακές-υπηρεσίες/ακτινοθεραπευτική-ογκολογία/υπηρεσιεσ/καρκίνος-μαστού#αναπτυξη-του-καρκινου-του-μαστου

[8] https://www.breastcancer.org/symptoms/understand_bc/what_is_bc

[9] https://www.natasapazaiti.gr/τι-είναι-ο-καρκίνος/

[10] http://www.chios-medical.gr/breast%20cancer.htm

[11] https://www.cancer.net/cancer-types/breast-cancer/stages

[12] https://www.cancer.ca/en/cancer-information/cancer-type/breast/staging/?region=on

[13] https://www.breastcancer.org/symptoms/diagnosis/staging

[14] https://www.nationalbreastcancer.org/about-breast-cancer/breast-cancer-staging

[15] Edge SB, Byrd DR, Compton CC, Fritz AG, Greene FG, Trotti A. AJCC Cancer Staging Manual. New York: Springer, 2010.
https://cancerstaging.org/references-tools/deskreferences/Documents/AJCC%207th%20Ed%20Cancer%20Staging%20Manual.pdf

[21] Lehman, C. D. et al. National Performance Benchmarks for Modern Screening Digital Mammography: Update from the Breast Cancer Surveillance Consortium. Radiol. 283, 49–58 (2016).
https://pubs.rsna.org/doi/10.1148/radiol.2016161174

[22] Pisano ED, Hendrick RE, Yaffe MJ, et al. Diagnostic accuracy of digital versus film mammography: exploratory analysis of selected population subgroups in DMIST. Radiology. 2008;246: 376-383.

[23] Kerlikowske K, Hubbard R A, Miglioretti DL, et al. Comparative effectiveness of digital versus film-screen mammography in community practice in the United States: a cohort study. Ann Intern Med. 2011;155: 493-502.
[24] Souza FH, Wendland EM, Rosa MI, Polanczyk CA. Is full-field digital mammography more accurate than screen-film mammography in overall population screening? A systematic review and meta-analysis. Breast. 2013.
https://www.thebreastonline.com/article/S0960-9776(13)00046-5/fulltext
[25] Miglioretti DL, Lange J, van den Broek JJ, et al. Radiation-Induced Breast Cancer Incidence and Mortality From Digital Mammography Screening: A Modeling Study. Ann Intern Med. 2016;16 4(4):205-214.
[26] Conant EF, Barlow WE, Herschorn SD, et al. Association of Digital Breast Tomosynthesis vs Digital Mammography With Cancer Detection and Recall Rates by Age and Breast Density. JAMA Oncol. 2019;28:28.
[27] Marinovich ML, Hunter KE, Macaskill P, Houssami N. Breast Cancer Screening Using Tomosynthesis or Mammography: A Meta-analysis of Cancer Detection and Recall. J Natl Cancer Inst. 2018;110(9):942-949.
[28] Pisano ED, Gatsonis C, Hendrick E, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. N Engl J Med 2005;353:1773-1783
https://www.nejm.org/doi/full/10.1056/nejmoa052911
[29] Coldman A, Phillips N, Wilson C, et al. Pan-Canadian study of mammography screening and mortality from breast cancer. J Natl Cancer Inst. 2 014 ;10 6 (11) .
[30] Paci E, Broeders M, Hofvind S, Puliti D, Duffy SW. European breast cancer service screening outcomes: a first balance sheet of thebenefits and harms. Cancer Epidemiol Biomarkers Prev. 2014;23(7):1159-1163.
[31] Tabar L, Dean PB, Chen TH, et al. The incidence of fatal breast cancer measures the increased effectiveness of therapy in women participating in mammography screening. Cancer. 2019;125(4):515-523.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6588008/
[32] Lehman CD, Arao RF, Sprague BL, et al. National Performance Benchmarks for Modern Screening Digital Mammography: Update from the Breast Cancer Surveillance Consortium. Radiology. 2017;283(1):49-58.
[33] Heywang Kobrunner SH, Bick U, Bradley WG, Jr, et al. International investigation of breast MRI: results of a multicentre study (11 sites) concerning diagnostic parameters for

contrast enhanced MRI based on 519 histopathologically correlated lesions. *Eur Radiol* 2001; 11: 531– 546.

[34] Gilles R, Guinebretiere JM, Toussaint C, et al. Locally advanced breast cancer: contrast enhanced subtraction MR imaging of response to preoperative chemotherapy. *Radiology* 1994; 191: 633– 638.

[35] Harms SE, Flamig DP, Hesley KL, et al. MR imaging of the breast with rotating delivery of excitation off resonance: clinical experience with pathologic correlation. *Radiology* 1993; 187: 493– 501.

[36] Boetes C, Barentsz JO, Mus RD, et al. MR characterization of suspicious breast lesions with a gadolinium enhanced TurboFLASH subtraction technique. *Radiology* 1994; 193: 777– 781.

[37] Liu PF, Debatin JF, Caduff RF, et al. Improved diagnostic accuracy in dynamic contrast enhanced MRI of the breast by combined quantitative and qualitative analysis. *Br J Radiol* 1998; 71: 501– 509.

[38] Saslow D, Boetes C, Burke W, et al. American Cancer Society guidelines for breast screening with MRI as an adjunct to mammography. CA Cancer J Clin. 20 07;57(2):75 -89. https://acsjournals.onlinelibrary.wiley.com/doi/full/10.3322/canjclin.57.2.75

[39] Rebolj M, Assi V, Brentnall A, Parmar D, Duffy SW. Addition of ultrasound to mammography in the case of dense breast tissue: systematic review and meta-analysis. Br J Cancer. 2018;118(12):1559-1570.

[40] Lee JM, Arao RF, Sprague BL, et al. Performance of Screening Ultrasonography as an Adjunct to Screening Mammography in Women Across the Spectrum of Breast Cancer Risk. JAMA Intern Med. 2019;18:18.

[41] Oeffinger KC, Fontham ET, Etzioni R, et al. Breast Cancer Screening for Women at Average Risk: 2015 Guideline Update From the American Cancer Society. JAMA. 2015;314(15):1599-1614. https://jamanetwork.com/journals/jama/fullarticle/2463262

[42] Provencher L, Hogue JC, Desbiens C, et al. Is clinical breast examination important for breast cancer detection? Curr Oncol. 2016;23:e332–9. doi: 10.3747/co.23.2881. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4974039/

[43] Ιατράκης, Γ. Γυναικολογικά προβλήματα και λύσεις. [ηλεκτρ. βιβλ.] Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. 2015; 41:41. https://repository.kallipos.gr/bitstream/11419/1865/1/02_chapter%2003.pdf

[44] M.B. Amin et al. (eds.), AJCC Cancer Staging Manual, Eighth Edition. 2017; 625:625.

https://cancerstaging.org/references-tools/deskreferences/Documents/AJCC%20Breast%20Cancer%20Staging%20System.pdf

[45] Lehman CD, Arao RF, Sprague BL, Lee JM, Buist DS, Kerlikowske K, Henderson LM, Onega T, Tosteson AN, Rauscher GH, Miglioretti DL. National Performance Benchmarks for Modern Screening Digital Mammography: Update from the Breast Cancer Surveillance Consortium. Radiology. 2017 Apr;283(1):49-58. doi: 10.1148/radiol.2016161174. Epub 2016 Dec 5. PMID: 27918707; PMCID: PMC5375631.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5375631/

[46] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29

[47] Wolberg WH, Mangasarian OL. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proc Natl Acad Sci U S A. 1990;87(23):9193–6.
https://www.pnas.org/content/pnas/87/23/9193.full.pdf

[48] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," Procedia Computer Science, vol. 83, pp. 1064-1069, 2016
https://core.ac.uk/download/pdf/82813624.pdf

[49] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

[50] Aličković E, Subasi A. Breast cancer diagnosis using GA feature selection and rotation Forest. Neural Comput & Applic. 2017;28(4):753–63.
https://www.researchgate.net/publication/284233577_Breast_cancer_diagnosis_using_GA_feature_selection_and_Rotation_Forest

[51] R. Alyami, J. Alhajjaj, B. Alnajrani, I. Elaalami, A. Alqahtani, N. Aldhafferi, T. O. Owolabi, and S. O. Olatunji, "Investigating the effect of Correlation based Feature Selection on breast cancer diagnosis using Artificial Neural Network and Support Vector Machines." IEEE International Conference on Informatics, Health & Technology (ICIHT), pp. 1-7, 2017

[52] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Prognostic%29

[53] W. N. Street, O. L. Mangasarian, and W.H. Wolberg. An inductive learning approach to prognostic prediction. In A. Prieditis and S. Russell, editors, Proceedings of the Twelfth International Conference on Machine Learning, pages 522--530, San Francisco, 1995. Morgan Kaufmann.
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.53.7874&rep=rep1&type=pdf

[54] O.L. Mangasarian, W.N. Street and W.H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4), pages 570-577, July-August 1995.

https://www.aaai.org/Papers/Symposia/Spring/1994/SS-94-01/SS94-01-019.pdf

[55] Maglogiannis I, Zafiropoulos E, Anagnostopoulos I. An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers. Appl Intell. 2009;30(1):24–36.

https://www.researchgate.net/publication/220204912_An_intelligent_system_for_automated_breast_cancer_diagnosis_and_prognosis_using_SVM_based_classifiers

[56] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer

[57] Michalski,R.S., Mozetic,I., Hong,J., & Lavrac,N. (1986). The Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains. In Proceedings of the Fifth National Conference on Artificial Intelligence, 1041-1045, Philadelphia, PA: Morgan Kaufmann.

[58] Clark,P. & Niblett,T. (1987). Induction in Noisy Domains. In Progress in Machine Learning (from the Proceedings of the 2nd European Working Session on Learning), 11-30, Bled, Yugoslavia: Sigma Press.

[59] Tan, M., & Eshelman, L. (1988). Using weighted networks to represent classification knowledge in noisy domains. Proceedings of the Fifth International Conference on Machine Learning, 121-134, Ann Arbor, MI.

[60] Cestnik,G., Konenenko,I, & Bratko,I. (1987). Assistant-86: A Knowledge-Elicitation Tool for Sophisticated Users. In I.Bratko & N.Lavrac (Eds.) Progress in Machine Learning, 31-45, Sigma Press.

[61] https://archive.ics.uci.edu/ml/datasets/Breast+Tissue

[62] Jossinet J (1996) Variability of impedivity in normal and pathological breast tissue. Med. & Biol. Eng. & Comput, 34: 346-350.

https://link.springer.com/article/10.1007/BF02520002

[63] Silva JE, Marques de Sá JP, Jossinet J (2000) Classification of Breast Tissue by Electrical Impedance Spectroscopy. Med & Bio Eng & Computing, 38:26-30.

https://www.academia.edu/25462249/Classification_of_breast_tissue_by_electrical_impedance_spectroscopy

[64] Albrecht AA, Lappas G, Vinterbo SA, Wong CK, Ohno-MachadoL (2002) Two applications of the LSA machine. In: 9th international conference on neural information processing, pp 184–189.

[65] Goodman D, Boggess L, Watkins A (2002) Artificial immune system classification of multiple-class problems. In: Intelligent engineering systems through artificial neural

networks: smart engineering system design: neural networks, fuzzy logic, evolutionary programming, complex systems and artificial life, vol 12,755pp 179–184.

[66] Sahan S, Polat K (2007) A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis. Comput Biol Med 3:415–423

[67] Stoean R, Stoean C (2013) Modeling medical decision making by support vector machines, explaining by rules of evolutionary algorithms with feature selection. Expert Syst Appl40:2677–2686

[68] Koloseni D, Lampinen J, Luukka P (2013) Differential evolution based nearest prototype classifier with optimized distance measures for the features in the data sets. Expert Syst Appl40(10):4075–4082

[69] Saez JA, Derrac J, Luengo J, Herrera F (2014) Statistical computation of feature weighting schemes through data estimation for nearest neighbor classifiers. Pattern Recogn 47(12):3941–3948

[70] Chen HL, Yang B, Wang SJ, Liu DY, Li HZ, Wen BL (2014) Towards an optimal support vector machine classifier using a parallel particle swarm optimization strategy. Appl Math Comput239:180–197

[71] Zheng B, Yoon SW, Lam SS (2014) Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Syst Appl41(4):1476–1482

[72] Fenton, J. J. et al. Influence of Computer-Aided Detection on Performance of Screening Mammography. New Engl. J. Medicine 356, 1399–1409 (2007).
https://www.nejm.org/doi/full/10.1056/NEJMoa066099

[73] Cole, E. B. et al. Impact of Computer-Aided Detection Systems on Radiologist Accuracy With Digital Mammography. Am. J. Roentgenol. 203, 909–916 (2014).
https://www.ajronline.org/doi/10.2214/AJR.12.10187

[74] Dhungel N, Carneiro G, Bradley A. A deep learning approach for the analysis of masses in mammograms with minimal user intervention. Med Image Anal. 2017 Dec;37:114–28.
https://cs.adelaide.edu.au/~carneiro/publications/automated_classification_mia.pdf

[75] Wang J, Yang X, Cai H, Tan W, Jin C, Li L. Discrimination of breast cancer with microcalcifications on mammography by deep learning. Sci Rep. 2016 Dec 7;6:27327.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4895132/?report=reader#!po=28.2609

[76] Ribli D, Horváth A, Unger Z, Pollner P, Csabai I. Detecting and classifying lesions in mammograms with deep learning. Sci Rep. 2018 Mar 15;8(1):4165.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5854668/?report=reader

[77] Gao F, Wu T, Li J, Zheng B, Ruan L, Shang D, Patel B. SD-CNN: a shallow-deep CNN for improved breast cancer diagnosis. Comput Med Imaging Graph. 2018 Dec;70:53–62.
https://arxiv.org/pdf/1803.00663.pdf
[78] Hagos YB, Mérida AG, Teuwen J. Improving Breast Cancer Detection Using Symmetry Information with Deep Learning. Proceedings of the Image Analysis for Moving Organ, Breast, and Thoracic Images; RAMBO'18; September 16, 2018; Granada, Spain. 2018. pp. 90–7.
https://arxiv.org/pdf/1808.08273.pdf
[79] Karssemeijer, Nico. ”Local orientation distribution as a function of spatial scalefor detection of masses in mammograms.” Biennial International Conference onInformation Processing in Medical Imaging. Springer, Berlin, Heidelberg, 1999.
[80] Teuwen J, van de Leemput S, Gubern-Mérida A, Rodriguez-Ruiz A, Mann R, Bejnordi B. Soft Tissue Lesion Detection in Mammography Using Deep Neural Networks for Object Detection. Proceedings of the 1st Conference on Medical Imaging with Deep Learning; MIDL'18; July 4-6, 2018; Amsterdam, The Netherlands. 2018. pp. 1–9.
https://openreview.net/pdf?id=rycG7Zhof
[81] Jung H, Kim B, Lee I, Yoo M, Lee J, Ham S, Woo O, Kang J. Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network. PLoS One. 2018;13(9):e0203355.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6143189/?report=reader
[82] Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. arXiv preprint arXiv:170802002. 2017.
https://ieeexplore.ieee.org/document/8417976
[83] Gardezi, Syed Jamal Safdar et al. “Breast Cancer Detection and Diagnosis Using Mammographic Data: Systematic Review.” *Journal of medical Internet research* vol. 21,7 e14464. 26 Jul. 2019.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6688437/#!po=52.0408
[84] Hwa H. L., Kuo W. H., Chang L. Y., Wang M. Y., Tung T. H., Chang K. J. and Hsieh F. J. (2008), Prediction of breast cancer and lymph node metastatic status with tumour markers using logistic regression models. Journal of Evaluation in Clinical Practice, 14: 275-280. doi:10.1111/j.1365-2753.2007.00849.x
https://onlinelibrary.wiley.com/doi/full/10.1111/j.1365-2753.2007.00849.x
[85] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra

[86] Muhammet Fatih Aslan, Yunus Celik, Kadir Sabanci, and Akif Durdu. "Breast cancer diagnosis by different machine learning methods using blood analysis data."International Journal of Intelligent Systems and Applications in Engineering6, no. 4 (2018): 289-293.
https://pdfs.semanticscholar.org/cba4/afdd6ff231d7350c030f3990913991aca9d0.pdf
[87] Silva Araújo V.J., Guimarães A.J., de Campos Souza P.V., Silva Rezende T., Souza Araújo V. Using resistin, glucose, age and BMI and pruning fuzzy neural network for the construction of expert systems in the prediction of breast cancer.Mach. Learn. Knowl. Extr.2019,1, 466–482.
https://pdfs.semanticscholar.org/eedc/a69d0b13e4c629ec0bd20dabb47f1b8225d6.pdf
[88] Lara Dular, STATISTICAL COMPARISON OF MACHINE LEARNING ALGORITHMS WITH RESPECT TO MULTIPLE PERFORMANCE MEASURES. Ljubljana, 2018.
https://repozitorij.uni-lj.si/Dokument.php?id=112727&lang=slv
[89] K. Polat, U. Sentürk, "A Novel ML Approach to Prediction of Breast Cancer: Combining of mad normalization, KMC based feature weighting and AdaBoostM1 classifier," 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, 2018, pp. 1-4. doi: 10.1109/ISMSIT.2018.8567245
https://ieeexplore.ieee.org/document/8567245
[90] Polat, K., Similarity-based attribute weighting methods via clustering algorithms in the classification of imbalanced medical datasets, Neural Computing and Applications 30 (3), 987–1013, 2018.
[91] Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, Douville C, Javed AA, Wong F, Mattox A, Hruban RH, Wolfgang CL, Goggins MG, Dal Molin M, Wang TL, Roden R, Klein AP, Ptak J, Dobbyn L, Schaefer J, Silliman N, Popoli M, Vogelstein JT, Browne JD, Schoen RE, Brand RE, Tie J, Gibbs P, Wong HL, Mansfield AS, Jen J, Hanash SM, Falconi M, Allen PJ, Zhou S, Bettegowda C, Diaz LA Jr, Tomasetti C, Kinzler KW, Vogelstein B, Lennon AM, Papadopoulos N. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science. 2018 Feb 23;359(6378):926-930. doi: 10.1126/science.aar3247.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6080308/
[92] Wong KC, Chen J, Zhang J, Lin J, Yan S, Zhang S, Li X, Liang C, Peng C, Lin Q, Kwong S, Yu J. Early Cancer Detection from Multianalyte Blood Test Results. *iScience*. 2019 May 31;15:332-341. doi: 10.1016/j.isci.2019.04.035.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6548890/

[93] Webb G.I., Boughton J.R., Wang Z. Not so naive Bayes: Aggregating one-dependence estimators. Mach. Learn. 2005;58:5–24.
https://link.springer.com/content/pdf/10.1007/s10994-005-4258-6.pdf
[94] Webb G.I., Boughton J.R., Zheng F., Ting K.M., Salem H. Learning by extrapolation from marginal to full-multivariate probability distributions: decreasingly naive bayesian classification. Mach. Learn. 2012;86:233–272.
https://link.springer.com/content/pdf/10.1007/s10994-011-5263-6.pdf
[95] Cristiano Stephen, Leal Alessandro, Phallen Jillian, Fiksel Jacob, Adleff Vilmos, Bruhm Daniel, Jensen Sarah, Medina Jamie, Hruban Carolyn, White James, Palsgrove Doreen, Niknafs Noushin, Anagnostou Valsamo, Forde Patrick, Naidoo Jarushka, Marrone Kristen, Brahmer Julie, Woodward Brian, Husain Hatim, Velculescu Victor. (2019). Genome-wide cell-free DNA fragmentation in patients with cancer. Nature. 570. 10.1038/s41586-019-1272-6.
https://www.researchgate.net/publication/333476109_Genome-wide_cell-free_DNA_fragmentation_in_patients_with_cancer
[96] Mohaimenul Islam,Tahmina Narin Poly. (2019). Machine Learning Models of Breast Cancer Risk Prediction. ResearchGate. DOI: 10.1101/723304
https://www.researchgate.net/publication/334904141_Machine_Learning_Models_of_Breast_Cancer_Risk_Prediction
[97] J. Crisóstomo, P. Matafome, D. Santos-Silva, A. L. Gomes, M. Gomes, M. Patrício, L. Letra, A. B. Sarmento-Ribeiro, L. Santos and R. Seiça, “Hyperresistinemia and metabolic dysregulation: a risky crosstalk in obese breast cancer,” International Journal of Basic and Clinical Endocrinology, vol. 53, no. 2, pp. 433-442, 2016.
https://link.springer.com/article/10.1007/s12020-016-0893-x
[98] M. Patrício, J. Pereira, J. Crisóstomo, P. Matafome, M. Gomes, R. Seiça, and F. Caramelo, “Using Resistin, glucose, age and BMI to predict the presence of breast cancer,” BMC cancer, vol. 18, no. 1, pp. 18-29, 2018.
https://bmccancer.biomedcentral.com/articles/10.1186/s12885-017-3877-1
[99] JG. Santillán-Benítez, H. Mendieta-Zerón, LM. Gómez-Oliván, JJ. Torres-Juárez, JM. González-Bañales, LV. Hernández-Peña, A. Ordóñez-Quiroz, “The tetrad BMI, leptin, leptin/ adiponectin (L/A) ratio and CA 15-3 are reliable biomarkers of breast cancer.”, Journal of Clinical Laboratory Analysis 27: pp 12-20, 2013
[100] Bi Y, Lu J, Wang W, Mu Y, Zhao J, Liu C, Chen L, Shi L, Li Q, Wan Q, Wu S, Yang T, Yan L, Liu Y, Wang G, Luo Z, Tang X, Chen G, Huo Y, Gao Z, Su Q, Ye Z, Wang Y, Qin

G, Deng H, Yu X, Shen F, Chen L, Zhao L, Zhang J, Sun J, Dai M, Xu M, Xu Y, Chen Y, Lai S, Bloomgarden ZT, Li D, Ning G.(2014), Cohort profile: risk evaluation of cancers in Chinese diabetic individuals: a longitudinal REACTION study. Journal of Diabetes, 6: 147-157. doi:10.1111/1753-0407.12108
https://www.ncbi.nlm.nih.gov/pubmed/24237858/
[101] Jee SH, Ohrr H, Sull JW, Yun JE, Ji M, Samet JM. Fasting serum glucose level and cancer risk in Korean men and women. 2005;293:194–202.
https://jamanetwork.com/journals/jama/fullarticle/200151
[102] Stattin P, Bjor O, Ferrari P, Lukanova A, Lenner P, Lindahl B, et al. Prospective study of hyperglycemia and cancer risk. Diabetes Care 2007;30:561–7.
https://care.diabetesjournals.org/content/30/3/561?
ijkey=957f7fade3571d790bb7d776390a113094441dd9&keytype2=tf_ipsecsha
[103] Rapp K, Schroeder J, Klenk J, Ulmer H, Concin H, Diem G, et al. Fasting blood glucose and cancer risk in a cohort of more than 140,000 adults in Austria. Diabetologia 2006;49:945–52.
https://link.springer.com/article/10.1007%2Fs00125-006-0207-6
[104] Parekh N, Lin Y, Vadiveloo M, Hayes RB, Lu-Yao GL. Metabolic dysregulation of the insulin-glucose axis and risk of obesity-related cancers in the Framingham heart study-offspring cohort (1971-2008). Cancer Epidemiol Biomarkers Prev. 2013 Oct; 22(10):1825-36.
https://cebp.aacrjournals.org/content/22/10/1825.long#sec-12
[105] Καρβέλα Αλεξία. Η επίδραση των κυτταροκινών/ορμονών σε λιπώδη ιστό παχύσαρκων και φυσιολογικών παιδιών: In vitro συγκριτική μελέτη. Πάτρα 2010. 25-44.
https://nemertes.lis.upatras.gr/jspui/bitstream/10889/4959/1/Nimertis_Karvela(i).pdf
[106] Αρβανίτη Βασιλική. ΚΛΙΝΙΚΉ ΜΕΛΈΤΗ ΤΗΣ ΜΗ ΑΛΚΟΟΛΙΚΗΣ ΣΤΕΑΤΟΗΠΑΤΙΤΙΔΑΣ. Πάτρα 2018. 82-105.
https://nemertes.lis.upatras.gr/jspui/bitstream/10889/11604/1/%CE%94%CE%99%CE%91%CE%A4%CE%A1%CE%99%CE%92%CE%97.pdf
[107] Neville MC, McFadden TB, Forsyth I. 2002. Hormonal regulation of mammary differentiation and milk secretion. J Mammary Gland Biol Neoplasia 7(1): 49–66.
[108] Okumura M, Yamamoto M, Sakuma H, Kojima T, Maruyama T, Jamali M, Cooper DR, Yasuda K. 2002. Leptin and high glucose stimulate cell proliferation in MCF 7 human breast cancer cells: Reciprocal involvement of PKC alpha and PPAR expression. Biochim Biophys Acta 1592(2): 107–116.

http://cel.webofknowledge.com/InboundService.do?customersID=atyponcel&smartRedirect=yes&mode=FullRecord&IsProductCode=Yes&product=CEL&Init=Yes&Func=Frame&action=retrieve&SrcApp=literatum&SrcAuth=atyponcel&SID=D1puDgE6nd84JebnzD5&UT=WOS%3A000178821700001

[109] Ishikawa M, Kitayama J, Nagawa H. 2004. Enhanced expression of leptin and leptin receptor (OB R) in human breast cancer. Clin Cancer Res 10(13): 4325–4331.

[110] Garofalo C, Surmacz E. Leptin and cancer. J Cell Physiol 2006;207:12–22. https://onlinelibrary.wiley.com/doi/full/10.1002/jcp.20472

[111] Bartella V, Cascio S, Fiorio E, Auriemma A, Russo A, Surmacz E. Insulin-dependent leptin expression in breast cancer cells. Cancer Res 2008;68:4919–27. https://cancerres.aacrjournals.org/content/canres/68/12/4919.full.pdf

[112] Booth, A., Magnuson, A., Fouts, J., et al. (2015). Adipose tissue, obesity and adipokines: role in cancer promotion. Hormone Molecular Biology and Clinical Investigation, 21(1), pp. 57-74. Retrieved 24 Jan. 2020, doi:10.1515/hmbci-2014-0037 https://www.degruyter.com/view/j/hmbci.2015.21.issue-1/hmbci-2014-0037/hmbci-2014-0037.xml

[113] Lehrke M, Reilly MP, Millington SC, Iqbal N, Rader DJ, Lazar MA. An inflammatory cascade leading to hyperresistinemia in humans. PLos Med 2004;1:161–8. https://journals.plos.org/plosmedicine/article/file?type=printable&id=10.1371/journal.pmed.0010045

[114] Dalamaga M, Sotiropoulos G, Karmaniolas K, Pelekanos N, Papadavid E, Lekka A. (2013) Serum resistin: a biomarker of breast cancer in postmenopausal women? Association with clinicopathological characteristics, tumor markers, inflammatory and metabolic parameters. Clinical biochemistry 46:584-590 https://www.sciencedirect.com/science/article/pii/S0009912013000076

[115] Assiri AM, Kamel HF, Hassanien MF. Resistin, visfatin, adiponectin, and leptin: risk of breast cancer in pre- and postmenopausal saudi females and their possible diagnostic and predictive implications as novel biomarkers. Dis Markers. 2015;2015:253519. doi:10.1155/2015/253519 https://pdfs.semanticscholar.org/3e39/0f3c62839e9f036ca9007a4e58f544070040.pdf

[116] Booth, A., Magnuson, A., Fouts, J., Foster, M.: Adipose tissue, obesity and adipokines: role in cancer promotion. Hormon. Mol. Biol. Clin. Invest. 21(1), 57–74 (2015) https://www.degruyter.com/view/j/hmbci.2015.21.issue-1/hmbci-2014-0037/hmbci-2014-0037.xml

[117] Lebrecht A, Grimm C, Lantzsch T, Ludwig E, Hefler L, Ulbrich E, Koelbl H. Monocyte chemoattractant protein-1 serum levels in patients with breast cancer. Tumour Biol. 2004 Jan-Apr;25(1-2):14-7.
https://www.ncbi.nlm.nih.gov/pubmed/15192307
[118] Dutta P, Sarkissyan M, Paico K, Wu Y, Vadgama JV. MCP-1 is overexpressed in triple-negative breast cancers and drives cancer invasiveness and metastasis. Breast Cancer Res Treat. 2018;170(3):477–486. doi:10.1007/s10549-018-4760-8
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6022526/
[119] Li, S., Lu, J., Chen, Y. et al. MCP-1-induced ERK/GSK-3β/Snail signaling facilitates the epithelial–mesenchymal transition and promotes the migration of MCF-7 human breast carcinoma cells. Cell Mol Immunol 14, 621–630 (2017). *https://doi.org/10.1038/cmi.2015.106*
[120] Del Guidice ME, Fantus IG, Ezzat S, Kckeown-Eyssen G, PageD, Goodwin PJ. Insulin and related factors in premenopausalbreast cancer risk. Breast Cancer Res Treat. 1998;47(2):111-120
[121] Goodwin PJ, Ennis M, Pritchard KI, et al. Fasting insulin andoutcome in early-stage breast cancer: results of a prospectivecohort study.J Clin Oncol. 2002;20(1):42-51.
[122] Boyd DB Insulin and cancer. Integr Cancer Ther. 2003 Dec;2(4):315-29.
https://journals.sagepub.com/doi/pdf/10.1177/1534735403259152
[123] Moolgavkar SH, Day NE, Stevens RG. Two-stage model for carcinogenesis: Epidemiology of breast cancer in females. J Natl Cancer Inst. 1980 Sep;65(3):559-69.
[124] Graham A. Colditz, Bernard Rosner, Cumulative Risk of Breast Cancer to Age 70 Years According to Risk Factor Status: Data from the Nurses' Health Study, American Journal of Epidemiology, Volume 152, Issue 10, 15 November 2000, Pages 950–964, *https://doi.org/10.1093/aje/152.10.950*
[125] D Horakova, K Bouchalova, L Stepanek, O Holy, A Petrakova, L Jurickova, H Kollarova, Insulin resistance as a potential risk factor for breast cancer: Dagmar Horakova, *European Journal of Public Health*, Volume 27, Issue suppl_3, November 2017, ckx186.172, *https://doi.org/10.1093/eurpub/ckx186.172*
[126] Capasso I, Esposito E, Pentimalli F, Montella M, Crispo A, Maurea N, D'Aiuto M, Fucito A, Grimaldi M, Cavalcanti E, Esposito G, Brillante G, Lodato S, Pedicini T, D'Aiuto G, Ciliberto G, Giordano A. Homeostasis model assessment to detect insulin resistance and identify patients at high risk of breast cancer development: National Cancer Institute of

Naples experience. *J Exp Clin Cancer Res*. 2013 Mar 14;32(1):14. doi: 10.1186/1756-9966-32-14.

[127] Alessandra Gennari, Matteo Puntoni, Oriana Nanni, Andrea De Censi, Paolo Bruzzi, Laura Paleari, Andrea Freschi, Laura Amaducci, Alessandra Bologna, Lorenzo Gianni, Dino Amadori. Impact of insulin resistance (IR) on the prognosis of metastatic breast cancer (MBC) patients treated with first-line chemotherapy (CT). Journal of Clinical Oncology 2014 32:15_suppl, 514-514. DOI: 10.1200/jco.2014.32.15_suppl.514

[128] Goodwin, P. Insulin resistance in breast cancer: relevance and clinical implications. Breast Cancer Res 13**,** O7 (2011). *https://doi.org/10.1186/bcr3006*

[129] Chen D, Chung Y, Yeh Y, Chaung H, Kuo F, Fu O, Chen H, HouM, Yuan SF (2006) Serum adiponectin and leptin levels in Tai-wanese breast cancer patients. Cancer Lett 237:109–114

[130] Cleary MP, Ray A, Rogozina OP, Dogan S, Grossmann ME(2009) Targeting the adiponectin: leptin ratio for postmenopausalbreast cancer prevention. Front Biosci (schol Ed) 1:329–357

[131] K. Hancke, D. Grubeck, N. Hauser, R. Kreienberg, J. M. Weiss. Adipocyte fatty acid-binding proteinas a novel prognostic factor in obese breast cancer patients. Breast Cancer Research and Treatment,Springer Verlag, 2009, 119 (2), pp.367-377.
https://hal.archives-ouvertes.fr/hal-00535403/document#lCR36

[132] Λύρας Δ. Παραμετροποίηση στοχαστικών μεθόδων εξόρυξης γνώσης από δεδομένα, μετασχηματισμών συμβολοσειρών και τεχνικών συμπερασματικού λογικού προγραμματισμού. Πανεπιστήμιο Πατρών 2010, pp 235-237.
https://nemertes.lis.upatras.gr/jspui/bitstream/10889/4144/1/%CE%94%CE%B9%CE%B4%CE%B1%CE%BA%CF%84%CE%BF%CF%81%CE%B9%CE%BA%CE%AE%20%CE%94%CE%B9%CE%B1%CF%84%CF%81%CE%B9%CE%B2%CE%AE%20-%20%CE%9B%CF%8D%CF%81%CE%B1%CF%82%20%CE%94%CE%B7%CE%BC%CE%AE%CF%84%CF%81%CE%B9%CE%BF%CF%82.pdf

[133] https://www.hackerearth.com/blog/developers/simple-tutorial-svm-parameter-tuning-python-r

[134] *https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn*

[135] Cover, T.M., & Hart, P.E. (1967). Nearest neighbor pattern classification. IEEE Trans. Information Theory, 13, 21-27.
https://pdfs.semanticscholar.org/0efb/841403aa6252b39ae6975c1cc5410554ef7b.pdf

[136] Domingos, Pedro M. and Michael J. Pazzani. “Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier.” ICML (1996).
https://pdfs.semanticscholar.org/1745/d064f2b600fce59c074faf56ef98c0f48911.pdf
[137] Jakkula, V. Tutorial on Support Vector Machine (SVM). Washington State University.
https://pdfs.semanticscholar.org/7cc8/3e98367721bfb908a8f703ef5379042c4bd9.pdf
[138] Πετρίδης, Δ. Ανάλυση πολυμεταβλητών τεχνικών. [ηλεκτρ. βιβλ.] Αθήνα, 2015:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, pp:90-91.
https://repository.kallipos.gr/bitstream/11419/2128/1/04_chapter03.pdf
[139] Breiman L. Bagging Predictors. *Machine Learning* **24,** 123–140 (1996). *https://doi.org/10.1023/A:1018054314350*
[140] Breiman L. Random Forests. Machine Learning 45**,** 5–32 (2001). *https://doi.org/10.1023/A:1010933404324*
[141] *https://blog.quantinsti.com/random-forest-algorithm-in-python/*
[142] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’16). Association for Computing Machinery, New York, NY, USA, 785–794. DOI:https://doi.org/10.1145/2939672.2939785
https://www.kdd.org/kdd2016/papers/files/rfp0697-chenAemb.pdf
[143] *Leo Breiman (1998). "Arcing classifier (with discussion and a rejoinder by the author)". Ann. Stat. 26 (3): 801–849. doi:10.1214/aos/1024691079*
https://projecteuclid.org/download/pdf_1/euclid.aos/1024691079
[144] Friedman Jerome. Greedy function approximation: A gradient boosting machine. Ann. Statist. 29 (2001), no. 5, 1189--1232. doi:10.1214/aos/1013203451.
https://statweb.stanford.edu/~jhf/ftp/trebst.pdf
[145] J. Friedman. Stochastic gradient boosting. ComputationalStatistics & Data Analysis, 38(4):367–378, 2002
https://statweb.stanford.edu/~jhf/ftp/stobst.pdf
[146] Zhang Z, Mayer G, Dauvilliers Y, Plazzi G, Pizza F, Fronczek R, Santamaria J, Partinen M, Overeem S, Peraita-Adrados R, da Silva AM, Sonka K, Rio-Villegas RD, Heinzer R, Wierzbicka A, Young P, Högl B, Bassetti CL, Manconi M, Feketeova E, Mathis J, Paiva T, Canellas F, Lecendreux M, Baumann CR, Barateau L, Pesenti C, Antelmi E, Gaig C, Iranzo A, Lillo-Triguero L, Medrano-Martínez P, Haba-Rubio J, Gorban C, Luca G, Lammers GJ, Khatami R. Exploring the clinical features of narcolepsy type 1 versus narcolepsy type 2 from European Narcolepsy Network database with machine learning. *Sci*

Rep. 2018 Jul 13;8(1):10628. doi: 10.1038/s41598-018-28840-w. PMID: 30006563; PMCID: PMC6045630.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6045630/
[147] *https://www.geeksforgeeks.org/ml-xgboost-extreme-gradient-boosting/*
[148] R. Valdovinos, J. Sánchez, and R. Barandela, "Dynamic and static weighting in classifier fusion," Pattern Recognit. Image Anal, vol. 3523, pp. 59–66, 2005.
[149] D. Jimenez, "Dynamically weighted ensemble neural networks for classification," in Proc. IEEE World Congr. Comput. Intell./IEEE Int. Joint Neural Netw., Anchorage, AK, USA, May 1998, pp. 753–756.
[150] K. Zarkogianni , M. Athanasiou, A.C. Thanopoulou, K. S. Nikita, "Comparison of Machine Learning Approaches Toward Assessing the Risk of Developing Cardiovascular Disease as a Long-Term Diabetes Complication," IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 22, NO. 5, SEPTEMBER 2018 , pp 1637-47
[151] Keen JD, Keen JE. How does age affect baseline screening mammography performance measures? A decision model. *BMC Med Inform Decis Mak*. 2008 Sep 21;8:40. doi: 10.1186/1472-6947-8-40. PMID: 18803871; PMCID: PMC2563001.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2563001/
[152] Andreas S. Panayides, Amir Amini, Nenad Filipovic, Ashish Sharma, Sotirios Tsaftaris, Alistair Young, David J. Foran, Nhan Do, Spyretta Golemati, Tahsin Kurc, Kun Huang, Konstantina S. Nikita, Ben P. Veasey, Michalis Zervakis, Joel H. Saltz, Constantinos S. Pattichis. "AI and Medical Imaging Informatics: Current Challenges and Future Directions". *IEEE Journal of Biomedical and Health Informatics*, 2020.
[153] Giuseppe Fico, Liss Hernanzez, Jorge Cancela, Arianna Dagliati, Lucia Sacchi, Antonio Martinez-Millana, Jorge Posada, Lidia Manero, Jose Verdú, Andrea Facchinetti, Manuel Ottaviano, Konstantia Zarkogianni, Konstantina Nikita, Leif Groop, Rafael Gabriel-Sanchez, Luca Chiovato, Vicente Traver, Juan Francisco Merino-Torres, Claudio Cobelli, Riccardo Bellazzi, Maria Teresa Arredondo, "What do healthcare professionals need to turn risk models for type 2 diabetes into usable computerized clinical decision support systems? Lessons learned from the MOSAIC project", *BMC medical informatics and decision making*, vol. 19, pp. 1-16, 2019
[154] K. Dalakleidi, K. Zarkogianni, A. Thanopoulou, and K. Nikita, "Comparative Assessment of Statistical and Machine Learning Techniques Towards Estimating the Risk of Developing Type 2 Diabetes and Cardiovascular Complications", Expert Systems, 2017; e12214. https://doi.org/10.1111/exsy.

[155] K. Zarkogianni, E. Litsa, K. Mitsis, P. Wu, C.D. Kaddi, C. Cheng, M. D. Wang, K. S. Nikita, "A Review of Emerging Technologies for the Management of Diabetes Mellitus," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 12, pp.2735-2749, 2015.

[156] K. Zarkogianni, K. Mitsis, E. Litsa, MT Arredondo, G. Fico, A. Fioravanti, K. S. Nikita, “Comparative assessment of glucose prediction models for Patients with Type 1 Diabetes Mellitus applying sensors for glucose and physical activity monitoring”, *Medical & Biological Engineering & Computing*, vol. 53, no. 12, pp. 1333-1343, 2015

[157] K. Zarkogianni, A. Vazeou, S.G. Mougiakakou, A. Prountzou, K.S. Nikita, "An insulin infusion advisory system based on autotuning nonlinear model-predictive control," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 9, pp. 2467-77, 2011.

[158] S.G. Mougiakakou, C. Bartsocas, E. Bozas, N. Chaniotakis, D. Iliopoulou, I. Kouris, S. Pavlopoulos, A. Prountzou, M. Skevofylakas, A. Tsoukalis, K. Varotsis, A. Vazeou, K. Zarkogianni and K. S. Nikita, "SMARTDIAB: A Communication and Information Technology Approach for the Intelligent Monitoring, Management and Follow-up of Type 1 Diabetes Patients", *IEEE Transactions on Information Technology in Biomedicine*, *Special Issue: New and Emerging Trends in Bioinformatics and Bioengineering*, vo. 14, no. 3, pp. 622 – 633, 2010.

[159] Konstantia Zarkogianni and Konstantina S. Nikita, “Personal Health Systems for Diabetes Management, Early Diagnosis and Prevention”, Handbook of Research on Trends in the Diagnosis and Treatment of Chronic Conditions, IGI Global Dessiminator of Knoweledge, ed. Dimitrios Fotiadis, pp. 465-494, 2015

[160] Κ. Ζαρκογιάννη, “Ευφυή Συστήματα Υποστήριξης Εξατομικευμένων Ιατρικών Αποφάσεων για τη Διαχείριση του Σακχαρώδους Διαβήτη”, Διδακτορική Διατριβή, Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Εθνικό Μετσόβιο Πολυτεχνείο, Αθήνα, 2011.

[162] Aimilia Gastounioti, Vassileios D Kolias, Spyretta Golemati, Nikolaos N Tsiaparas, Aikaterini Matsakou, John S Stoitsis, Nikolaos P.E. Kadoglou, Christos Gkekas, John D. Kakisis, Christos D. Liapis, Petros Karakitsos, Sarafis, Ioannis, Angelidis, Pantelis, Konstantina S Nikita. “CAROTID: A web-based platform for optimal personalized management of atherosclerotic patients”. *Computers Method and Programs in Biomedicine*, vol. 114, pp. 183 - 193, 2014.