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

Επισκόπηση Ευφυών Μοντέλων και Τεχνικών Μηχανικής Μάθησης για την πρόβλεψη, διαχείριση και αντιμετώπιση του καρκίνου του μαστού

Διπλωματική Εργασία

ΜΙΤΟΓΛΟΥ ΠΑΡΑΣΚΕΥΗ

Επιβλέπων: Δημήτριος-Διονύσιος Κουτσούρης,
Καθηγητής Ε.Μ.Π.
Συνεπιβλέπουσα: Ουρανία Πετροπούλου
Ειδικό Τεχνικό και Εργαστηριακό Προσωπικό

Αθήνα, Ιούνιος, 2021

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

Επισκόπηση Ευφυών Μοντέλων και Τεχνικών Μηχανικής Μάθησης για την πρόβλεψη, διαχείριση και αντιμετώπιση του καρκίνου του μαστού

Διπλωματική Εργασία

ΜΙΤΟΓΛΟΥ ΠΑΡΑΣΚΕΥΗ

Επιβλέπων: Δημήτριος-Διονύσιος Κουτσούρης,
Καθηγητής Ε.Μ.Π.
Συνεπιβλέπουσα: Ουρανία Πετροπούλου
Ειδικό Τεχνικό και Εργαστηριακό Προσωπικό

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 9η Ιουνίου 2021.

.....	.....	.....
Δημήτριος-Διονύσιος Κουτσούρης	Γεώργιος Ματσόπουλος	Παναγιώτης Τσανάκας
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούνιος, 2021

……..

ΜΙΤΟΓΛΟΥ ΠΑΡΑΣΚΕΥΗ
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών, Ε.Μ.Π.

Περίληψη

Ο καρκίνος είναι μία ομάδα ασθενειών που σχετίζονται με την αφύσικη κυτταρική ανάπτυξη, με πιθανότητα εισβολής και διάδοσης σε άλλα μέρη του σώματος. Τα τελευταία χρόνια, παρατηρείται μια ραγδαία αύξηση της εμφάνισης του καρκίνου του μαστού, καθώς αποτελεί τον πιο συχνό καρκίνο στις γυναίκες παγκοσμίως. Η σοβαρότητα της νόσου, το μεγάλο πλήθος ασθενών και η ανάγκη για συνεχή παρακολούθηση τέτοιων περιστατικών, αποτελούν το τρίπτυχο που ανέδειξε την ανάγκη ανάπτυξης ευφυών μοντέλων και τεχνικών για την πρόβλεψη, διαχείριση και αντιμετώπιση του καρκίνου του μαστού σε πρώιμο στάδιο. Με στόχο την αύξηση της πιθανότητας επιβίωσης και της καλύτερης διαχείρισης των ατόμων που έχουν προσβληθεί από την ασθένεια, μελετάται ενδελεχώς η δυνατότητα ενσωμάτωσης αλγορίθμων μηχανικής μάθησης στην κλινική πράξη. Η πρακτική αυτή είναι ικανή να βελτιώσει την πρόγνωση των ασθενών και να συμβάλει σε μεγαλύτερη ακρίβεια στο κομμάτι της διάγνωσης, της πρόβλεψης μετάστασης και επιβίωσης από τη νόσο. Σκοπός της παρούσας διπλωματικής εργασίας είναι η επισκόπηση ευφυών μοντέλων και τεχνικών μηχανικής μάθησης με εφαρμογή στην αντιμετώπιση του καρκίνου του μαστού. Η επισκόπηση βασίστηκε στην αναζήτηση σε επιστημονικές βάσεις δεδομένων, όπως το PubMed, Science Direct, Scopus και Google Scholar, με χρήση των κατάλληλων λέξεων κλειδιών (π.χ. «μηχανική μάθηση και καρκίνος του μαστού» ή «μηχανική μάθηση και διάγνωση του καρκίνου του μαστού»). Στη συνέχεια, επιλέχθηκαν και μελετήθηκαν πρόσφατες έρευνες (από το 2016 και μετά), γραμμένες στην αγγλική γλώσσα. Βασικές επιδιώξεις της εργασίας είναι να αποτυπωθεί η τεχνολογία αιχμής στον συγκεκριμένο επιστημονικό κλάδο, καταδεικνύοντας τα αποδοτικότερα μοντέλα, να επισημανθούν τυχόν εμπόδια και περιορισμοί που υπάρχουν, και να γίνει αναφορά στις μελλοντικές επεκτάσεις της έρευνας.

Λέξεις κλειδιά

Τεχνητή Νοημοσύνη, Μηχανική Μάθηση, Βαθιά Μάθηση, Καρκίνος του Μαστού, Διάγνωση, Μετάσταση, Επιβίωση.

Abstract

Cancer is a group of diseases associated with abnormal cell growth, with the possibility of invading and spreading to other parts of the body. In recent years, there has been a rapid increase in the cases of breast cancer, as it is the most common cancer in women worldwide. The severity of the disease, the large number of patients and the need for continuous monitoring of such cases, constitute the three most crucial factors that highlighted the need to develop intelligent models and techniques for predicting, managing and treating breast cancer at an early stage. In order to increase the chances of survival and better treatment of people infected with the disease, the possibility of integrating machine learning algorithms into clinical practice is thoroughly studied. This practice is capable of improving patient prognosis and contributing to greater accuracy in the area of diagnosing, anticipating metastasis and surviving the disease. The purpose of this thesis is to review the said intelligent models and machine learning techniques with application in the treatment of breast cancer. The review was based on searching scientific databases such as PubMed, Science Direct, Scopus and Google Scholar, using the appropriate keywords (e.g. "machine learning and breast cancer" or "machine learning and breast cancer diagnosis"). Recent surveys (from 2016 and onwards), written in English, were then selected and studied. The main aims of the work are to capture cutting-edge technology in this scientific field, demonstrating the most efficient models, highlighting any obstacles and limitations that exist, and referring to future extensions of research.

Keywords

Artificial Intelligence, Machine Learning, Deep Learning, Breast Cancer, Diagnosis, Metastasis, Survival.

Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε κατά το ακαδημαϊκό έτος 2020-2021 στα πλαίσια των δραστηριοτήτων του Εργαστηρίου Βιοϊατρικής Τεχνολογίας του τομέα Συστημάτων Μετάδοσης Πληροφορίας και Τεχνολογίας Υλικών της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου, υπό την επίβλεψη του κύριου Δημήτριου-Διονύσιου Κουτσούρη, Καθηγητή Ε.Μ.Π, τον οποίο θέλω να ευχαριστήσω θερμά για την εμπιστοσύνη που μου έδειξε και την ευκαιρία που μου έδωσε.

Θα ήθελα επίσης να ευχαριστήσω την κυρία Ράνια Πετροπούλου, Ε.ΔΙ.Π. Ε.Μ.Π. και τον κύριο Σαραφίδη Μιχαήλ Υποψήφιο Διδάκτορα Ε.Μ.Π., για την συνεχή υποστήριξη και καθοδήγησή τους, καθώς και για τις πολύτιμες συμβουλές τους, που συνετέλεσαν καθοριστικά στην επιτυχή διεκπεραίωση της παρούσας διπλωματικής εργασίας.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου, ιδιαιτέρως τους γονείς μου Δημήτριο και Μαρία, για την αγάπη και την ανιδιοτελή υποστήριξή τους στην πραγματοποίηση όλων μου των στόχων, ένας εκ των οποίων είναι οι σπουδές μου στο Εθνικό Μετσόβιο Πολυτεχνείο, καθώς επίσης και τον αγαπημένο μου σύντροφο Α.Α. που είναι δίπλα μου σε κάθε βήμα της ζωής μου.

Παρασκευή Μιτόγλου

Περιεχόμενα

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΠΙΝΑΚΑΣ ΣΥΝΤΟΜΟΓΡΑΦΙΩΝ

DALY	Disability-adjusted life year
ΤΝ	Τεχνητή Νοημοσύνη
AI	Artificial Intelligence
DSRPAI	Darthmouth Summer Research Project on Artificial Intelligence
ΜΜ	Μηχανική Μάθηση
CNN	Convolutional Neural Networks
ΒΜ	Βαθιά Μάθηση
DL	Deep Learning
NN	Neural Networks
ΝΔ	Νευρωνικά Δίκτυα
ANN	Artificial Neural Networks
FDA	Food and Drug Administration
NLP	Natural Language Processing
SVM	Support Vector Machines
AUC-ROC	Area Under the Curve – Receiver Operating Characteristics Curve
AUC	Area Under the Curve
DNN	Deep Neural Networks
DBN	Deep Belief Networks
RNN	Recurrent Neural Networks
RBM	Restricted Boltzmann Machine
AE	Autoencoders
ΓΠ	Γραμμική Παλινδρόμηση
LR	Logistic Regression
SSL	Semi-Supervised Learning
SL	Supervised Learning
USL	Unsupervised Learning
RL	Reinforcement Learning
DT	Decision Trees
CART	Classification and Regression Algorithm
ASM	Attribute Selection Measure
K-NN	K-Nearest Neighbor
CAD	Computer Aided Detection
WDBC	Wisconsin Diagnostic Breast Cancer Dataset
CFS	Correlation based Feature Selection
RFE	Recursive Feature Elimination
PCA	Principal Component Analysis
LDA	Linear Discriminant Analysis
NBC	Naïve Bayes Classifier

RF	Random Forest
MP	Multi-layer Perceptron
K-SVM	K-Support Vector Machines
FNA	Fine Needle Aspirate
ER	Estrogen Receptor
HER2	Human epidermal growth factor receptor 2
PR	Progesterone Receptor
PPV	Positive Predictive Value
NPV	Negative Predictive Value
LR-	Negative Likelihood Ration
LR+	Positive Likelihood Ratio
BCCD	Breast Cancer Coimbra Dataset
ΔΜΣ	Δείκτης Μάζας Σώματος
HOMA	HOmeostasis Model Assessment
RMSE	Root Mean Square Error
ΜΟ	Μέσος Όρος
ELM	Extreme Learning Machines
ROI	Region Of Interest
DQN	Deep Q-Network
FROC	Free Response Operating Curve
RL-Det	Reinforcement Learning Detection
DL-MSC	Deep Learning Multi-Scale Cascade
C-SL	Clustering and Structure Learning
ΧΚΜ	Χρόνος Κατασκευής Μοντέλου
ΣΤΠ	Σωστά Ταξινομημένες Περιπτώσεις
ΛΤΠ	Λανθασμένα Ταξινομημένες Περιπτώσεις
BCS	Breast Cancer Subtype
2D-DIGE	2-Dimensional, Differential in-Gel Electrophoresis
ALN	Axillary Lymph Node
DCE-MRI	Dynamic Contrast-Enhanced Magnetic Resonance Imaging
MRI	Magnetic Resonance Imaging
LOOCV	Leave-One-Out Cross Validation
CV	Cross Validation
mRNA	messenger RNA
miRNA	microRNA
ncRNA	noncoding RNA
RNA-seq	RNA sequencing
SILAC	Stable Isotope Labeling with Amino Acids in Cell Culture
SVDD	Support Vector Data Description
TNBC	Triple-Negative Breast Cancer

v-SVM	Variant of Support Vector Machine
MI	Mutual Information
BN	Bayesian Networks
FCBF	Fast Correlation Based Filter
K-S	Kolmogorov-Smirnov
PSO	Practical Swarm Optimization
ABC	Artificial Bee Colony
PA	PSO-ABC
JELM	Java optimized Extreme Learning Machines
IG	Information Gain
BDF	Binary Dragonfly
TLBO	Teaching Learning-Based Optimization
MIM	Mutual Information Maximization
AGA	Adaptive Genetic Algorithm
MRMR	Minimum Redundancy and Maximum Relevance
FPA	Flower Pollination Algorithm
RLR	Randomized Logistic Regression
PCC	Pearson Corellation Coefficient
ClusterQGA	Cluster Quantum Genetic Algorithm
QMI	Qualitative Mutual Information
GWO	Gray Wolf Optimization
PGBM	Point-wise Gated Boltzmann Machine
RBM	Restricted Boltzmann Machine
AAAM	Ανίχνευση Αντικειμένου Αλλοιώσεων Μαστού
ABUS	Automated Breast Ultrasound
FCN	Fully Convolutional Network
GAN	Generative Adversarial Networks
VAE	Variational AutoEncoder
PROBAST	Prediction Model Risk Of Bias Assessment Tool

Κεφάλαιο 1 – Εισαγωγή

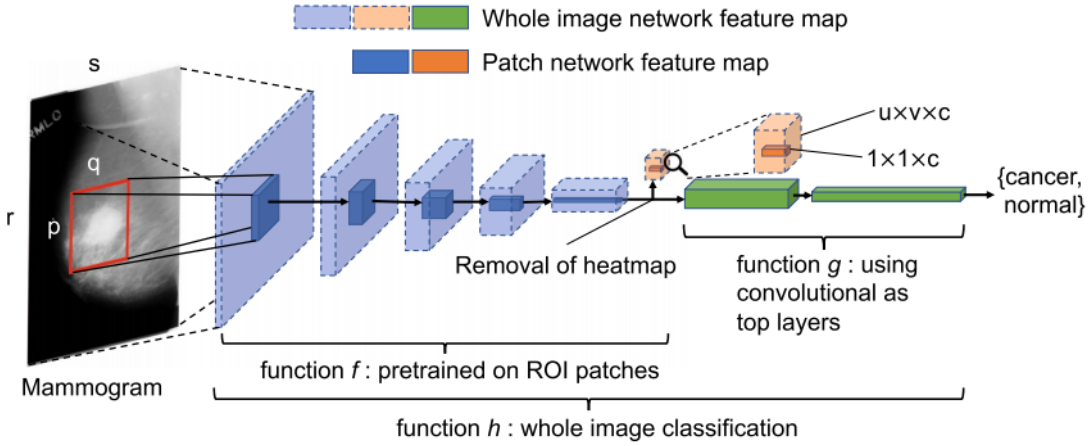
Η θνησιμότητα από καρκίνο του μαστού άλλαξε ελάχιστα από τη δεκαετία του 1930 έως τη δεκαετία του 1970. Οι βελτιώσεις στην επιβίωση ξεκίνησαν τη δεκαετία του 1980, σε χώρες με προγράμματα έγκαιρης ανίχνευσης, σε συνδυασμό με διαφορετικούς τρόπους θεραπείας, για την εξάλειψη των επιθετικών ασθενειών. Με την πάροδο των χρόνων, η συχνότητα εμφάνισης του έχει αυξηθεί κατά πολύ. Το 2020 διαγνώστηκαν 2,3 εκατομμύρια γυναίκες με καρκίνο του μαστού και είχαμε 685.000 θανάτους παγκοσμίως. Από το 2015 μέχρι το τέλος του 2020, υπήρχαν 7,8 εκατομμύρια ζωντανές γυναίκες που διαγνώστηκαν με καρκίνο του μαστού, καθιστώντας αυτού του είδους τον καρκίνο ως τον πιο διαδεδομένο στον κόσμο. Στην Ελλάδα, κάθε χρόνο, καταγράφονται περίπου 1500 νέες περιπτώσεις καρκίνου του μαστού. Είναι εξίσου σημαντικό να αναφερθεί ότι υπάρχουν περισσότερα χαμένα χρόνια ζωής λόγω αναπηρίας (Disability-adjusted life year - DALY) από τον καρκίνο του μαστού, παγκοσμίως, σε σχέση με οποιονδήποτε άλλο τύπο καρκίνου, και εμφανίζεται σε κάθε χώρα του κόσμου σε γυναίκες οποιαδήποτε ηλικίας, μετά την εφηβεία. Ωστόσο, η πιθανότητα να εμφανιστεί σε μεγαλύτερη ηλικία είναι υψηλότερη [1], [2], [3].

Η επιβίωση του καρκίνου του μαστού για τουλάχιστον 5 χρόνια μετά τη διάγνωση, κυμαίνεται σε ποσοστό πάνω του 90% σε χώρες υψηλού εισοδήματος, έως 66% στην Ινδία και 40% στη Νότια Αφρική. Η θνησιμότητα από καρκίνο του μαστού στις χώρες υψηλού εισοδήματος μειώθηκε κατά 40% μεταξύ της δεκαετίας του 1980 και του 2020. Οι χώρες που κατάφεραν να μειώσουν τη θνησιμότητα, κατάφεραν να επιτύχουν και ετήσια μείωση της θνησιμότητας κατά 2-4% ετησίως. Στόχος της Παγκόσμιας Πρωτοβουλίας του ΠΟΥ είναι η μείωση της παγκόσμιας θνησιμότητας κατά 2,5% ετησίως, αποτρέποντας έτσι 2,5 εκατομμύρια θανάτους, μεταξύ 2020 και 2040. Η μείωση της παγκόσμιας θνησιμότητας κατά 2,5% ετησίως θα αποτρέψει το 25% των θανάτων από καρκίνο του μαστού μέχρι το 2030 και το 40% μέχρι το 2040 μεταξύ των γυναικών ηλικίας κάτω των 70 ετών. Οι τρεις πυλώνες για την επίτευξη αυτών των στόχων είναι: η προαγωγή της υγείας για έγκαιρη ανίχνευση, έγκαιρη διάγνωση και ολοκληρωμένη διαχείριση του καρκίνου του μαστού [1].

Ο συνδυασμός έγκαιρης διάγνωσης και εφαρμογής κατάλληλης θεραπείας στους ασθενείς, έχει αποδειχθεί επιτυχής σε χώρες υψηλού εισοδήματος και θα πρέπει να εφαρμόζεται και στις χώρες με περιορισμένους πόρους. Ο συνδυασμός αυτός, μπορεί να πετύχει πιθανότητες επιβίωσης στο 90% και παραπάνω. Επομένως, είναι σημαντικό να είναι διαθέσιμη η κατάλληλη τεχνολογία και κάθε γυναίκα να έχει πρόσβαση σε αυτήν, προκειμένου να πραγματοποιήσει προληπτικό έλεγχο, αφού αυτό μπορεί να μειώσει την πιθανότητα του θανάτου της από τη νόσο σε σημαντικό βαθμό. Οι στρατηγικές για τη βελτίωση της αντιμετώπισης του καρκίνου του μαστού, εξαρτώνται από τη θεμελιώδη ενίσχυση του συστήματος υγείας,

προκειμένου να παρέχονται τα απαραίτητα εργαλεία σε ιατρούς και ασθενείς. Η ενίσχυση αυτή μπορεί να γίνει με ποικίλους τρόπους, ένας εκ των οποίων είναι η εκμετάλλευση της εξέλιξης της τεχνολογίας, που μπορεί να βοηθήσει ουσιαστικά, προκειμένου να μειωθούν τα ποσοστά θνησιμότητας. Επομένως, είναι επιτακτική η ανάγκη της εύρεσης, απόκτησης και χρήσης νέων τεχνολογιών, που θα βοηθήσουν τους ιατρούς στην αντιμετώπιση της νόσου. Υπολογιστικά συστήματα, τα οποία θα ενισχύσουν τις ήδη υπάρχουσες μεθόδους διάγνωσης (όπως μία απλή αιματολογική εξέταση, υπέρηχος μαστού, μαστογραφία, 3-D μαστογραφία, μαγνητική τομοσύνθεση και άλλες), θα δώσουν στους ιατρούς τη δυνατότητα να μειώσουν τον χρόνο που απαιτείται για να γίνει μια διάγνωση, να κάνουν γρήγορη επιβεβαίωση της διάγνωσής τους, με αποτέλεσμα την αύξηση του πλήθους των διαγνώσεων ανά θεράποντα ιατρό, και να προβλέπεται η πιθανότητα επιβίωσης του ασθενούς.

Η ύπαρξη μηχανών, που έχουν την ικανότητα να μιμούνται την ανθρώπινη συμπεριφορά, μπορούν να το κάνουν πραγματικότητα. Η μηχανική μάθηση και η βαθιά μάθηση είναι δύο πολλά υποσχόμενες τεχνολογίες και είναι υποσύνολα της Τεχνητής Νοημοσύνης. Μπορούν να βοηθήσουν τους οργανισμούς περίθαλψης να ανταποκριθούν στις, όλο και περισσότερες, απαιτήσεις που προκύπτουν στα συστήματα περίθαλψης, βελτιώνοντας τις λειτουργίες τους, τα συστήματά τους και μειώνοντας το κόστος και το χρόνο. Έρευνες, που πραγματοποιήθηκαν τα τελευταία χρόνια, απέδειξαν πως, τέτοια συστήματα, μπορούν να διαχειριστούν τον απαραίτητο όγκο δεδομένων, προκειμένου να λειτουργήσουν σωστά, να ενισχύσουν και ανεβάσουν την απόδοση, τη χρησιμότητα και την αναγκαιότητα της ύπαρξής όλων των ήδη υπαρχουσών διαγνωστικών εργαλείων για την καταπολέμηση της νόσου. Οι αλγόριθμοι μηχανικής μάθησης, έχουν τη δυνατότητα να ανιχνεύουν ανωμαλίες που υπάρχουν σε μια εικόνα, πέρα από αυτές που μπορεί να εντοπίσει το ανθρώπινο μάτι. Η δυνατότητα αυτή, όχι μόνο βοηθάει στη διάγνωση και τη θεραπεία, αλλά δίνει τη δυνατότητα να εντοπιστεί η νόσος από νωρίς και, κατά συνέπεια, να μειωθούν τα ποσοστά θνησιμότητας.



Εικόνα 1. Διάγραμμα αλγορίθμου Βαθιάς Μάθησης για την αναγνώριση αλλοίωσης στο μαστό[4].

Κεφάλαιο 2 – Καρκίνος

2.1 Καρκίνος

Ο ανθρώπινος οργανισμός αποτελείται από κύτταρα. Προκειμένου να διατηρηθεί η υγεία του οργανισμού, τα κύτταρα αναπτύσσονται, διαιρούνται – με σκοπό τη δημιουργία θυγατρικών κυττάρων – και στο τέλος πεθαίνουν. Κάποιες φορές, η διαδικασία αυτή αποκλίνει από το φυσιολογικό, με αποτέλεσμα να αναπτύσσονται νέα κύτταρα, υπερβολικά σε πλήθος και χωρίς προγραμματισμό, τα οποία δεν χρειάζεται ο οργανισμός, ενώ, παράλληλα, τα παλιά κύτταρα δεν πεθαίνουν. Επομένως, ο οργανισμός φτάνει σε κάποιο σημείο, κατά το οποίο έχει πλεονάζοντα κύτταρα. Η στιγμή που τα κύτταρα παύουν να είναι φυσιολογικά, είναι η στιγμή της καρκινογένεσης. Αυτό έχει ως αποτέλεσμα τη δημιουργία μιας μάζας, μπορεί να συμβεί σε διάφορα σημεία του σώματος, που ονομάζεται καρκίνος (ή όγκος) και η αιτία του βρίσκεται σε κυτταρικό επίπεδο [5].

Αυτή η ανώμαλη ανάπτυξη των κυττάρων συνήθως ξεκινάει σε ένα σημείο του σώματος. Σε μερικές περιπτώσεις, αυτά τα μη φυσιολογικά κύτταρα εξαπλώνονται σε άλλα σημεία του σώματος, δημιουργώντας με αυτό τον τρόπο δευτερεύοντες όγκους, παρόμοιους με τους αρχικούς. Οι όγκοι αυτοί ονομάζονται μεταστατικοί, ενώ η διαδικασία ονομάζεται μετάσταση. Αξίζει να σημειωθεί πως υπάρχουν πάνω από 200 διαφορετικά είδη καρκίνου, μπορεί να προσβάλλει οποιονδήποτε ιστό του σώματος και να έχει τελείως διαφορετική μορφή σε κάθε σημείο του σώματος. Τις περισσότερες φορές, παίρνει τη μορφή μάζας, εκτός από ορισμένους τύπους καρκίνου όπως η λευχαιμία [5].

Δεν είναι όλοι οι όγκοι επικίνδυνοι. Οι όγκοι που δεν κάνουν μετάσταση και δεν προκαλούν θάνατο, ονομάζονται καλοήθεις. Αντίθετα, οι όγκοι που είναι μεταστατικοί και αποτελούν απειλή για τη ζωή ονομάζονται κακοήθεις, ή καρκίνωμα ή νεόπλασμα. Εάν ο καρκίνος δεν θεραπευθεί, μπορεί τελικά να προκαλέσει το θάνατο. Τέλος, αποτελεί τη δεύτερη αιτία θανάτου στις αναπτυγμένες χώρες παρόλο που η επιβίωση των ασθενών έχει βελτιωθεί σημαντικά τα τελευταία χρόνια [5].

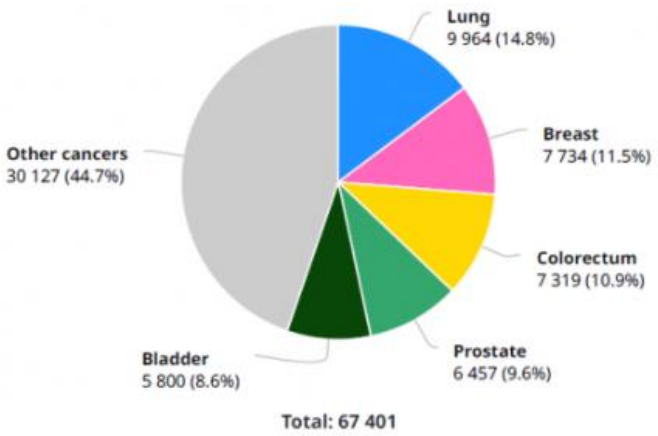
2.2 Καρκίνος του Μαστου

Καρκίνος του μαστού ονομάζεται το καρκίνωμα που δημιουργείται από τα κύτταρα του ιστού του μαστού. Οι περισσότεροι καρκίνοι του μαστού ξεκινάνε στους μαστικούς αδένες ή στα κανάλια που συνδέουν τους μαστικούς αδένες με τη θηλή («in situ»), όπου γενικά δεν προκαλεί συμπτώματα και έχει ελάχιστες δυνατότητες εξάπλωσης. Με την πάροδο του χρόνου, αυτοί οι in situ (στάδιο 0) καρκίνοι μπορεί να προχωρήσουν και να εισβάλουν στον περιβάλλοντα ιστό του μαστού και στη συνέχεια να εξαπλωθούν στους κοντινούς λεμφαδένες ή σε άλλα

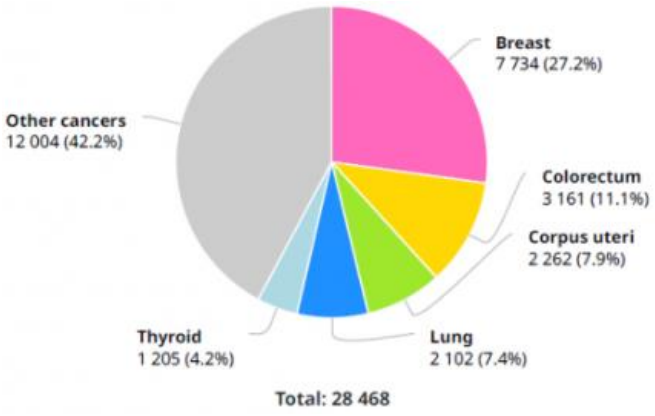
όργανα του σώματος. Αυτός ο τύπος καρκίνου εμφανίζεται κυρίως στις γυναίκες, αποτελεί τον πιο κοινό σε αυτές και είναι η δεύτερη σε σειρά αιτία θανάτου για το γυναικείο φύλο [2], [1].

2.3. Καρκίνος του Μαστού στην Ελλάδα και στον κόσμο

Με βάση την Εικόνα *2* και τα δεδομένα του Παγκόσμιου Οργανισμού Υγείας για τον καρκίνο στην Ελλάδα, βλέπουμε ότι για το 2018, μόνο στον Ελλαδικό χώρο, είχαμε 7.734 νέες περιπτώσεις καρκίνου του μαστού, με 2.207 από αυτούς τους ασθενείς να κατέληξαν [6], ενώ αποτελεί τον πιο συχνά εντοπισμένο καρκίνο στις γυναίκες. Στην Εικόνα 3 παρατηρείται ότι ο καρκίνος του μαστού αποτελεί τον δεύτερο πιο συχνό καρκίνο, στον γενικό πληθυσμό, με πρώτο τον καρκίνο του πνεύμονα. Επιπλέον, το 2018 στην Ελλάδα ήταν η 3^{η} αιτία θανάτου από καρκίνο στο συνολικό πληθυσμό, με πρώτο τον καρκίνο του πνεύμονα (8.343 θάνατοι στα 9.964 περιστατικά) και δεύτερο το παχύ έντερο (2.983 θάνατοι στα 6.013 περιστατικά) [6].

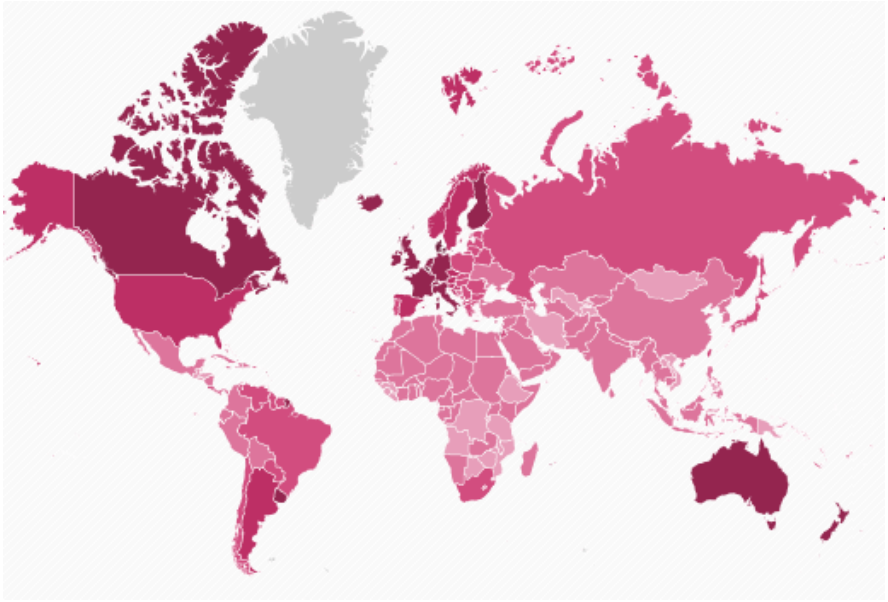


Εικόνα 2. Πλήθος νέων περιστατικών καρκίνου το 2018 και για τα δύο φύλα σε όλων των ηλικιών [6]

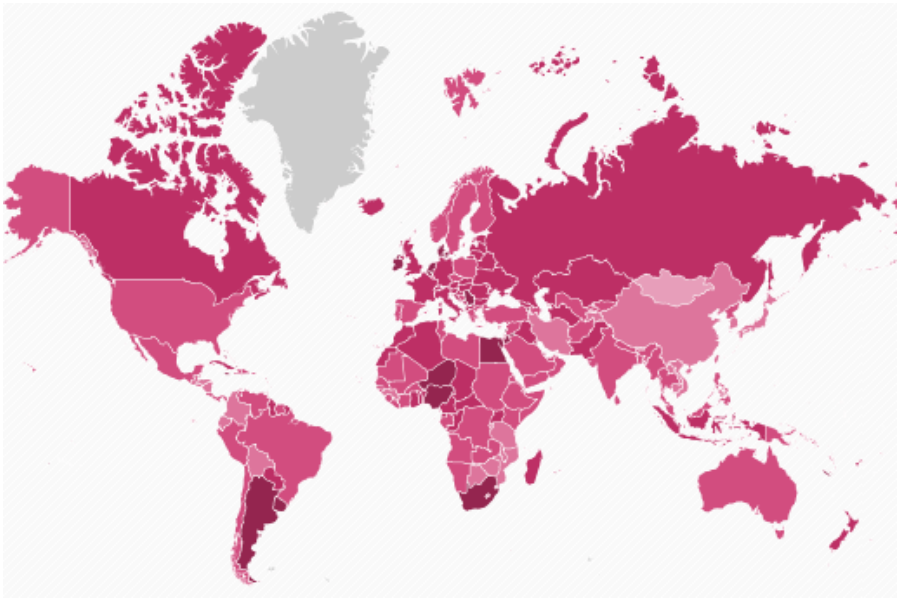


Εικόνα 3. Πλήθος νέων περιστατικών καρκίνου το 2018 σε γυναίκες όλων των ηλικιών [6].

Στην *Εικόνα 4* και στην *Εικόνα 5* βλέπουμε δύο παγκόσμιους χάρτες. Ο πρώτος περιγράφει την εμφάνιση καρκίνου του μαστού ανά 100.000 γυναίκες ανεξαρτήτως ηλικίας, αντίστοιχα ο δεύτερος περιγράφει τα ποσοστά θνησιμότητας ανά 100.000 γυναίκες ανεξαρτήτως ηλικίας. Όσο πιο σκούρο το χρώμα που αντιστοιχεί σε κάθε χώρα, τόσο μεγαλύτερο το ποσοστό. Παρατηρείται ότι στις μη αναπτυγμένες χώρες έχουμε μικρότερα ποσοστά εμφάνισης αλλά μεγαλύτερα ποσοστά θνησιμότητας.



Εικόνα 4. Παγκόσμιος χάρτης εμφάνισης καρκίνου του μαστού [7].



Εικόνα 5. Παγκόσμιος χάρτης θνησιμότητας από τον καρκίνο του μαστού [7].

Κεφάλαιο 3 – Τεχνητή Νοημοσύνη

3.1. Ιστορική Αναδρομή

Η *Τεχνητή Νοημοσύνη – ΤΝ (Artificial Intelligence – AI)* είναι μία από τις πιο μοντέρνες ερευνητικές περιοχές. Στο πρώτο μισό του 20ου αιώνα, ο κόσμος είχε ήδη συμφιλιωθεί με την ιδέα των έξυπνων ρομπότ. Μέχρι την δεκαετία του 1950, υπήρχε ήδη μία νέα γενιά επιστημόνων, οι οποίοι είχαν ήδη αφομοιώσει την έννοια της Τεχνητής Νοημοσύνης (ή AI) στο μυαλό τους [8].

Ο Alan Turing εξερεύνησε τη μαθηματική δυνατότητα της Τεχνητής Νοημοσύνης και, το 1950 στην εργασία του *«Computing Machinery and Intelligence»*, πρότεινε πως να κατασκευάσει έξυπνες μηχανές και τον τρόπο προκειμένου να δοκιμάσει την ευφυία τους. Η ιδέα του κατέρρευσε σύντομα , αφού, πριν το 1949, οι υπολογιστές δεν είχαν τη δυνατότητα να αποθηκεύουν εντολές, αλλά μόνο να τις εκτελούν. Επίσης, το κόστος μίσθωσης ενός υπολογιστή αποτελούσε εμπόδιο, αφού ανερχόταν στα 200.000$ το μήνα. Επομένως, ήταν επιτακτική ανάγκη η ιδέα να υποστηριχθεί από άτομα υψηλού προφίλ, προκειμένου να βρεθούν πηγές χρηματοδότησης [8].

Πέντε χρόνια αργότερα, το 1956, σχεδιάστηκε το πρόγραμμα *«The Logic Theorist»*, με σκοπό να μιμείται τις δεξιότητες επίλυσης προβλημάτων ενός ανθρώπου. Θεωρείται από πολλούς το πρώτο πρόγραμμα τεχνητής νοημοσύνης και παρουσιάστηκε στο *Darthmouth Summer Research Project on Artificial Intelligence* (DSRPAI) που φιλοξένησαν οι John McCarthy και Marvin Minsky [8].

Η ΤΝ είναι πλέον παντού. Εφαρμογές της έχουν βρει γόνιμο έδαφος σε πάρα πολλούς κλάδους όπως η τεχνολογία, οι τραπεζικές συναλλαγές, το μάρκετινγκ, η ψυχαγωγία και η υγεία [8]. Δεν είναι λίγες οι φορές που ερευνητές, από διαφορετικές επιστημονικές περιοχές, καταφεύγουν στην ΤΝ με στόχο την αναζήτηση εργαλείων αυτοματοποίησης των λογικών βημάτων που χρησιμοποιούν στην εργασία τους [9].

3.2. Τι είναι η Τεχνητή Νοημοσύνη

3.2.1. Ορισμός Νοημοσύνης

Στο βιβλίο *«Frames of Mind: The theory of multiple intelligences»* (1983), του Howard Gardner, διακρίνονται 8 διαφορετικοί τύποι νοημοσύνης σε κάθε άνθρωπο (Γλωσσική, Μαθηματική/Λογική, Μουσική, Χωρική, Σωματική, Διαπροσωπική, Ενδοπροσωπική, Φυσιοκρατική) και πρακτικά χρησιμοποιείται ένα μίγμα αυτών. Επομένως, επικοινωνούμε, μαθαίνουμε, λύνουμε προβλήματα με, τουλάχιστον, οχτώ διαφορετικούς τρόπους [9].

Σύμφωνα με το ερμηνευτικό λεξικό του Cambridge, ως νοημοσύνη αναφέρεται *«η ικανότητα για μάθηση, κατανόηση και κρίση ή αιτιολογημένη έκφραση γνώμης»* . Ενώ, στο λεξικό Merriam-Webster, αναφέρεται ότι νοημοσύνη είναι *«η ικανότητα για μάθηση ή κατανόηση ή η αντιμετώπιση νέων ή δύσκολων καταστάσεων».* Επίσης, ο Douglas Hofstadter στο βιβλίο του *«Gödel, Escher, Bach: An Eternal Golden Braid»*, πρότεινε μία λίστα από θεμελιώδεις δυνατότητες νοημοσύνης:

- Ανταπόκριση σε καταστάσεις με ελαστικότητα (αποφυγή μηχανικής συμπεριφοράς)
- Η αντίληψη και η κατανόηση ασαφών ή αντιφατικών μηνυμάτων από τα συμφραζόμενα.
- Η αναγνώριση και ιεράρχηση των καταστάσεων με βάση τη σπουδαιότητα τους.
- Η εύρεση ομοιοτήτων σε φαινομενικά ανόμοιες καταστάσεις.
- Η εύρεση διαφορών σε καταστάσεις που εκ πρώτης όψεως μοιάζουν παρόμοιες.

Οι ικανότητες αυτές αποκτώνται εύκολα από τους ανθρώπους και, συνήθως, βασίζονται σε ένα σύνολο σταθερών και στερεοτυπικών απόψεων που μπορεί να κατέχει οποιοσδήποτε άνθρωπος και αποκαλείται *κοινή λογική* [8]**.**

3.2.2. Ορισμός Τεχνητής Νοημοσύνης

Διάφοροι ορισμοί της ΤΝ έχουν διατυπωθεί κατά καιρούς. Κάποιοι από τους οποίους επικεντρώνονται στη διαδικασία σκέψης και συλλογισμού, ενώ άλλοι στη συμπεριφορά. Οι Russel και Norvig υποστήριξαν ότι αυτοί οι ορισμοί της ΤΝ, ταξινομούνται σε τέσσερις μεγάλες κατηγορίες οι οποίες προσεγγίζουν την περιοχή από διαφορετική σκοπιά ανάλογα τον στόχο της ΤΝ [8].

Στην πρώτη κατηγορία, ο στόχος της ΤΝ είναι η ανάπτυξη συστημάτων που σκέφτονται όπως οι άνθρωποι. Παραδείγματος χάριν, ο Haugeland την ορίζει ως *«Η προσπάθεια να κατασκευάσουμε υπολογιστές με διανοητική ικανότητα με την πλήρη και κυριολεκτική έννοια του όρου».* Η δεύτερη κατηγορία ορίζει την ΤΝ ως την προσπάθεια για ανάπτυξη συστημάτων που σκέφτονται λογικά, με χαρακτηριστικό ορισμό αυτής της κατηγορίας να δίνεται από τον Winston ως «*Η μελέτη των υπολογισμών που καθιστούν εφικτή την αντίληψη, τη λογική σκέψη και την αντίδραση».* Στην τρίτη κατηγορία φαίνεται να ανήκουν, από τον ορισμό που έδωσαν οι Rich και Knight «*Η μελέτη του πώς να κάνουμε τους υπολογιστές να κάνουν πράγματα στα οποία αυτήν τη στιγμή οι άνθρωποι είναι καλύτεροι»*, τα συστήματα που συμπεριφέρονται όπως οι άνθρωποι. Τέλος, στην τέταρτη κατηγορία στόχος είναι η ανάπτυξη συστημάτων που αντιδρούν λογικά, με βάση τον ορισμό που έδωσε ο Luger «*Ο τομέας της επιστήμης των υπολογιστών που ασχολείται με την αυτοματοποίηση της ευφυούς συμπεριφοράς»***.**

Επομένως, ένας γενικότερος ορισμός θα μπορούσε να είναι:

"ΤΝ είναι ο τομέας της Επιστήμης των Υπολογιστών που ασχολείται με την σχεδίαση και την υλοποίηση προγραμμάτων τα οποία είναι ικανά να μιμηθούν τις ανθρώπινες γνωστικές ικανότητες, εμφανίζοντας έτσι χαρακτηριστικά που αποδίδουμε συνήθως σε ανθρώπινη συμπεριφορά, όπως για παράδειγμα η επίλυση προβλημάτων, η αντίληψη μέσω της όρασης, η μάθηση, η εξαγωγή συμπερασμάτων, η κατανόηση φυσικής γλώσσας κτλ." [1].

Κατά συνέπεια, η ΤΝ είναι ένας συνεχώς εξελισσόμενος τομέας της επιστήμης των υπολογιστών. Στόχος είναι να προσφέρει, στην υπηρεσία του ανθρώπου, μηχανές που να μην απαιτούν ειδικές γνώσεις για τον χειρισμό τους, να έχουν τη δυνατότητα προσαρμογής στις ανάγκες του χρήστη, να μαθαίνουν από τα λάθη τους και να επιλύουν πραγματικά, δύσκολα, καθημερινά προβλήματα και όχι μόνο αριθμητικά. Η έρευνα στην ΤΝ καλύπτει ένα ευρύ φάσμα περιοχών και έχει αποδώσει καρπούς σε πολλές από αυτές [1].

3.3. Η Τεχνητή Νοημοσύνη στον τομέα της Υγείας

Η ΤΝ έχει ενσωματωθεί στην καθημερινότητά μας σε διάφορες μορφές, όπως τα ηλεκτρονικά παιχνίδια και η αεροπορία. Ένας άλλος κλάδος, στον οποίο η ΤΝ έχει γνωρίσει μεγάλη ανάπτυξη, είναι η Ιατρική. Στόχος είναι η βελτίωση της περίθαλψης των ασθενών, μέσω της επιτάχυνσης των διαδικασιών, και η επίτευξη μεγαλύτερης ακρίβειας στο κομμάτι της διάγνωσης [10].

Οι δύο πιο σημαντικοί παράγοντες για την επιτυχή φροντίδα των ασθενών είναι η γνώση και η εμπειρία. Όσες περισσότερες γνώσεις έχει ένας γιατρός και όσους περισσότερους ασθενείς φροντίσει, τόσο καλύτερη φροντίδα θα παρέχει στους επόμενους. Η διαδικασία αυτή απαιτεί αρκετό χρόνο και, η κατανόηση της, είναι ζωτικής σημασίας για την χρησιμότητα της ΤΝ στον κλάδο της ιατρικής. Η σωστή λήψη αποφάσεων είναι συνέπεια της εμπειρίας (η οποία προέρχεται από τη φροντίδα ασθενών) και των δεδομένων (από βιβλία, επιστημονικά άρθρα κλπ.) που έχουμε στη διάθεση μας. Όσο περισσότερη η εμπειρία και τα δεδομένα μας, τόσες περισσότερες πιθανότητες έχουμε για τη λήψη μίας σωστής απόφασης [10].

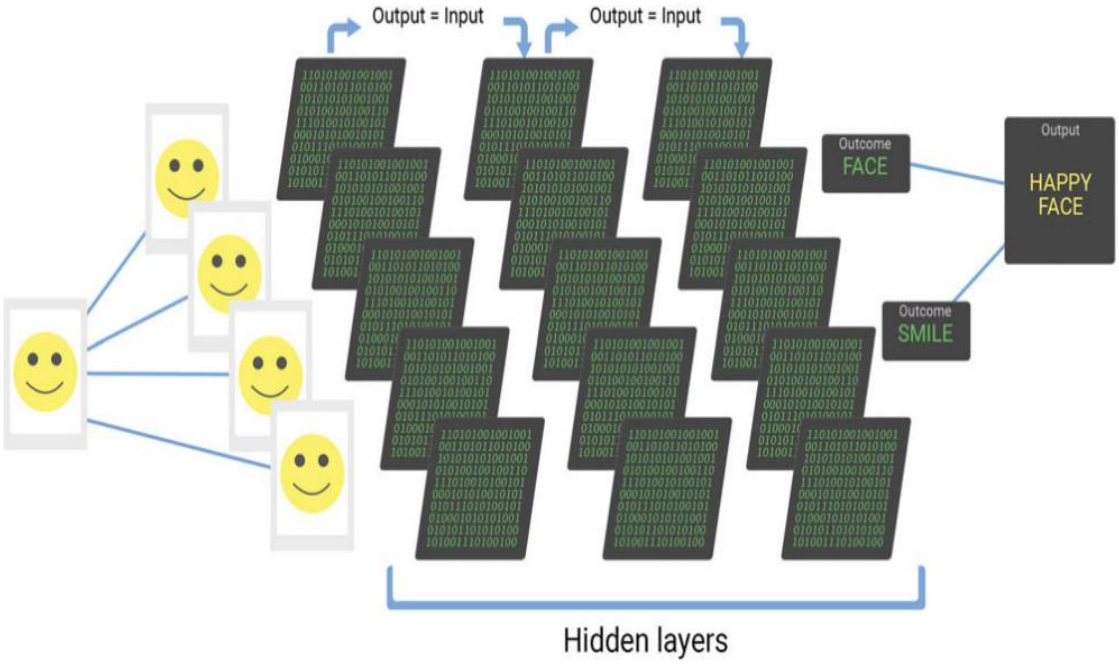
Ο χρόνος αποτελεί κύριο εμπόδιο στην απόκτηση γνώσης και εμπειρίας για το ανθρώπινο μυαλό, δεδομένου ότι, αμφότεροι, απαιτούν πολύ χρόνο. Η δυσκολία αυτή δεν αποτελεί εμπόδιο για τους υπολογιστές, αφού στην εποχή μας μεγάλες ποσότητες δεδομένων των ασθενών μπορούν να προσεγγιστούν, να αποκτηθούν και να αποθηκευτούν. Η κατάλληλη αξιοποίηση αυτών των δεδομένων είναι το σημείο στο οποίο στηρίζεται η ΤΝ. Οι υπολογιστές μπορούν να αποκτήσουν πολύ περισσότερη εμπειρία, σε πολύ μικρότερο χρονικό διάστημα, απ' ότι μπορεί ένας άνθρωπος σε όλη του τη ζωή. Για παράδειγμα, ένας ακτινολόγος μπορεί να εξετάσει, περίπου, 225.000 MRI/CT εξετάσεις κατά τη διάρκεια 40 παραγωγικών ετών εργασίας. Αντίθετα, η ΤΝ έχει τη δυνατότητα να ξεκινήσει με αυτό το νούμερο

και μέσα σε μικρό χρονικό διάστημα να έχει εξετάσει εκατομμύρια σαρώσεων, βελτιώνοντας έτσι την ακρίβεια της στα αποτελέσματα [10].

Ενώ ο όρος της ΤΝ μπορεί να εφαρμοστεί με διάφορες μορφές στην επιστήμη των υπολογιστών, στην Ιατρική μπορεί να επικεντρωθεί κυρίως στους ακόλουθους όρους:

- **Επεξεργασία Εικόνας** – Μαθηματική διαδικασία κατά την οποία εισάγεται μία εικόνα και εξάγεται μία καλύτερα καθορισμένη εικόνα. Αυτή η διαδικασία βελτίωσης γίνεται για λόγους σαφήνειας, ανάκτησης συγκεκριμένων πληροφοριών ή μετρήσεων προτύπων [10]**.**
- **Μηχανική Όραση** – Δέχεται μια εικόνα σαν είσοδο, την επεξεργάζεται και παρέχει σαν έξοδο την ερμηνεία της εικόνας [10]**.**
- **Τεχνητό Νευρωνικό Δίκτυο (Artificial Neural Network – ANN)** – Υπολογιστικά μοντέλα εμπνευσμένα από τον τρόπο λειτουργίας των βιολογικών νευρώνων, οι οποίοι λειτουργούν μέσω πολλαπλών στρωμάτων. Ένα ANN μιμείται τον ανθρώπινο εγκέφαλο, όσον αφορά την επεξεργασία διαφόρων τύπων δεδομένων, και στη δημιουργία προτύπων για χρήση προκειμένου να ληφθούν αποφάσεις μέσω νευρωνικών δικτύων. Το μαθηματικό τους μοντέλο βασίζεται σε μοντέλα με μη γραμμικά στατιστικά δεδομένα, όπου υπάρχουν πολύπλοκες σχέσεις μεταξύ εισόδων και εξόδου. Πιο συγκεκριμένα, εισάγουμε μια είσοδο σε ένα ANN. Αυτή με τη σειρά της εισάγεται σε ένα σύνολο αλγορίθμων. Η έξοδος που θα προκείψει από αυτούς τους αλγόριθμους, θα αποτελεί την είσοδο για ένα νέο σύνολο αλγορίθμων, προκειμένου να φτάσει στο τελικό αποτέλεσμα [10].
- **Μηχανική Μάθηση – ΜΜ (Machine Learning – ML)** – Η ικανότητα ενός υπολογιστή να μάθει από την εμπειρία. Αυτό σημαίνει ότι έχει τη δυνατότητα να τροποποιήσει τον τρόπο που επεξεργάζεται τα δεδομένα του, βασιζόμενο στις νέες πληροφορίες που αποκτήθηκαν. Ένας τρόπος για να πραγματοποιηθεί αυτό, είναι με τη χρήση ενός απλού δέντρου λήψης αποφάσεων ή με αλγορίθμους βαθιάς μάθησης [10].
- **Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks – CNN)** – Συγκεκριμένος τύπος ANN που βασίζεται σε αλγορίθμους βαθιάς μάθησης, με πολλά κρυφά στρώματα σε κάθε CNN, για την ανάλυση δεδομένων. Ονομάζονται συνελικτικά αφού οι σχέσεις μεταξύ των επιπέδων είναι αρκετά πολύπλοκες [10].
- **Βαθιά Μάθηση – ΒΜ (Deep Learning – DL)** – Αποτελεί υποσύνολο της μηχανικής μάθησης**.** Λαμβάνει ταυτόχρονα πολλαπλά σύνολα δεδομένων, αξιολογούνται και επεξεργάζονται κατ' επανάληψη μέχρις ότου να επιτευχθεί ένα αποτέλεσμα. Κάθε αξιολόγηση είναι διαφορετική, πραγματοποιείται σε διαφορετικό επίπεδο και, επομένως, βασίζεται στην έξοδο του προηγούμενου επιπέδου. Οι είσοδοι και οι έξοδοι των επιπέδων υπολογισμού δεν είναι ορατοί και γι' αυτό το λόγο ονομάζονται κρυφά στρώματα. Μόλις η εικόνα προσπελάσει όλα τα επίπεδα, δίνεται το τελικό αποτέλεσμα (*Εικόνα 6*) [10].

- **Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing – NLP)** – Η ικανότητα ενός υπολογιστή να «μάθει» και να «προβλέψει» τη σημασία της ανθρώπινης γλώσσας με χρήση υπολογιστικών αλγορίθμων. Η συγκεκριμένη μέθοδος είναι ευρέως διαδεδομένη στη βιομηχανία και στις επιχειρήσεις. Ωστόσο, η εφαρμογή της στο πεδίο της ιατρικής, για την έρευνα και διαχείριση του καρκίνου του μαστού, είναι αρκετά περιορισμένη, λόγω του ότι οι γιατροί δε γνωρίζουν τις δυνατότητες που μπορεί να προσφέρει η τεχνολογία της ΕΦΓ [11].



Εικόνα 6. Εισαγωγή μίας εικόνας σε ένα CNN[10].

3.4. Εφαρμογές Τεχνητής Νοημοσύνης στον τομέα της Υγείας

Εφαρμογές της ΤΝ έχουν ανθίσει σε πολλούς κλάδους της ιατρικής, όπως είναι η ακτινολογία, η οφθαλμολογία, η καρδιολογία, η χειρουργική, η γαστρεντερολογία, παθολογία, η δερματολογία και η ογκολογία [10] [12]. Η διάγνωση του καρκίνου, καθώς επίσης η δυνατότητα να μπορέσει ένας γιατρός να προσδιορίσει σε ποιο στάδιο βρίσκεται ο ασθενής, είναι τομείς της ιατρικής στους οποίους η εφαρμογή της ΤΝ μπορεί να παρέχει καλύτερα αποτελέσματα από τα ανθρώπινα [10].

Μέσα από διάφορες έρευνες και προσπάθειες που έχουν γίνει, η αξιοπιστία της ΜΜ έχει αποδειχθεί. Σε μια μελέτη, που πραγματοποιήθηκε από τον Bejnordi και την ερευνητική του ομάδα, χρησιμοποιήθηκαν 129 φωτογραφικές διαφάνειες, 49 από τις οποίες είχαν μεταστάσεις στους λεμφαδένες, ενώ οι υπόλοιπες όχι. Όταν αυτές συγκρίθηκαν με 11 παθολογίες, ο αλγόριθμος πέτυχε την καλύτερη διαγνωστική απόδοση, αφού οι παθολόγοι χρειάστηκαν 30 ώρες προκειμένου να

αξιολογήσουν και τις 129 διαφάνειες, ενώ ο χρόνος που χρειάστηκε ο αλγόριθμος θεωρήθηκε αμελητέος [10].

Αποτελεσματικότεροι από έναν άνθρωπο έχουν αποδειχθεί και οι αλγόριθμοι ΤΝ που χρησιμοποιούνται για την ανίχνευση του καρκίνου του πνεύμονα. Σε μελέτη του Yu, με χρήση 2186 χρωματισμένων ιστοπαθολογικών εικόνων ολικής αλλοίωσης του πνευμονικού αδενοκαρκινώματος και πλακώδους καρκινώματος, αποδείχθηκε η ακρίβεια στην διάγνωση που μπορεί να προσφέρει ένας αλγόριθμος ΤΝ για διάγνωση. Άμεση συνέπεια αυτού είναι η πρόγνωση και, επομένως, η βελτίωση της φροντίδας των ασθενών μέσω της θεραπείας που πρέπει να ακολουθήσουν [10].

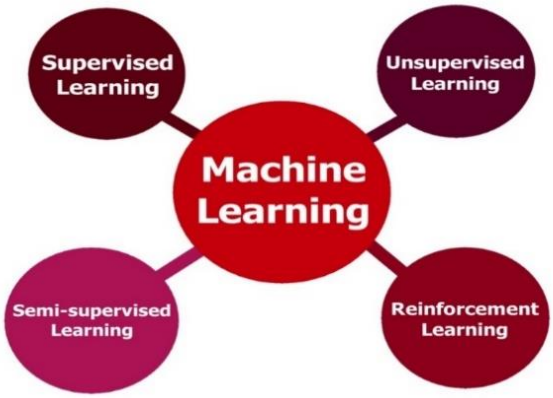
Η ΤΝ έχει κερδίσει έδαφος και στη δερματολογία, για τη διάγνωση δερματικών αλλοιώσεων, η οποία βασίζεται κυρίως σε εικόνες. Σε μελέτη που πραγματοποιήθηκε, χρησιμοποιήθηκε μόνο ένα CNN για την ταξινόμηση διάφορων δερματικών αλλοιώσεων. Έγινε χρήση 129.450 κλινικών δεδομένων και σύγκριναν τα αποτελέσματα, που εξήγαγε το CNN, με τις διαγνώσεις που έκαναν 21 πιστοποιημένοι δερματολόγοι και αφού είχε πραγματοποιηθεί βιοψία. Χρησιμοποιήθηκαν δύο ομάδες εικόνων, οι οποίες εμφάνιζαν τους πιο συνηθισμένους καρκίνους του δέρματος και εκείνους με υψηλότερη θνησιμότητα. Σε αυτή τη μελέτη, τα αποτελέσματα της ΤΝ συμφωνούσαν σε όλες τις περιπτώσεις με τους ειδικούς, αποδεικνύοντας έτσι ότι η ΤΝ είναι εξίσου ικανή με τους δερματολόγους [10].

Η ΤΝ κερδίζει έδαφος σε πολλούς κλάδους λόγω της βελτίωσης της απόδοσης, της αποτελεσματικότητας, της αποδοτικότητας του χρόνου και της μείωσης του κόστους. Οι συνέπειες όλων αυτών στον τομέα της υγείας είναι η βελτιωμένη φροντίδα των ασθενών, η μείωση των ιατρικών σφαλμάτων, του κόστους, της νοσηρότητας και της θνησιμότητας. Σκοπός της ΤΝ δεν είναι η αντικατάσταση των γιατρών, αλλά η αύξηση της ποιότητας της ιατρικής περίθαλψης που προσφέρουν οι γιατροί στους ασθενείς. Φυσικά τίθενται θέματα όπως εγκρίσεις από τον FDA (Food and Drug Administration), προσωπικά δεδομένα και να ξεκαθαριστεί στο κοινό η έννοια και ο σκοπός της ΤΝ στην ιατρική. Η χρήση της θα πρέπει να έχει τη θέση ενός συστήματος υποστήριξης του ιατρικού και νοσηλευτικού προσωπικού, με την τελική απόφαση να ανήκει στον γιατρό [10],[13].

Κεφάλαιο 4 – Μηχανική Μάθηση

Η *Μηχανική Μάθηση* είναι ένας από τους ταχύτερα αναπτυσσόμενους τομείς της επιστήμης των υπολογιστών, με πάρα πολλές εφαρμογές σε διάφορα πεδία, και αναφέρεται στην αυτοματοποιημένη ανίχνευση σημαντικών μοτίβων στα δεδομένα. Τα εργαλεία που χρησιμοποιεί η ΜΜ είναι αλγόριθμοι που έχουν τη δυνατότητα να μάθουν και να προσαρμοστούν ανάλογα με τα δεδομένα που επεξεργάζονται. Το πλήθος των δεδομένων συνεχώς αυξάνεται και, επομένως, είναι επιτακτική η ανάγκη εύρεσης νέων τρόπων διαχείρισης και ανάλυσής τους. Οι εφαρμογές ΜΜ ποικίλουν με σημαντικότερη την εξόρυξη δεδομένων, δηλαδή την εύρεση κρυφών μοτίβων στα δεδομένα ώστε να παραχθεί πληροφορία. Με αυτόν τον τρόπο μειώνεται κατά πολύ η πιθανότητα λάθους από τον άνθρωπο. Οι αλγόριθμοι ΜΜ χωρίζονται σε τέσσερις κατηγορίες (*Εικόνα 7*) αναλόγως το σκοπό τους [10][11] :

- Επιβλεπόμενη Μάθηση
- Μη-Επιβλεπόμενη Μάθηση
- Ημι-Επιβλεπόμενη Μάθηση
- Ενισχυμένη Μάθηση



Εικόνα 7. Κατηγορίες Μηχανικής Μάθησης

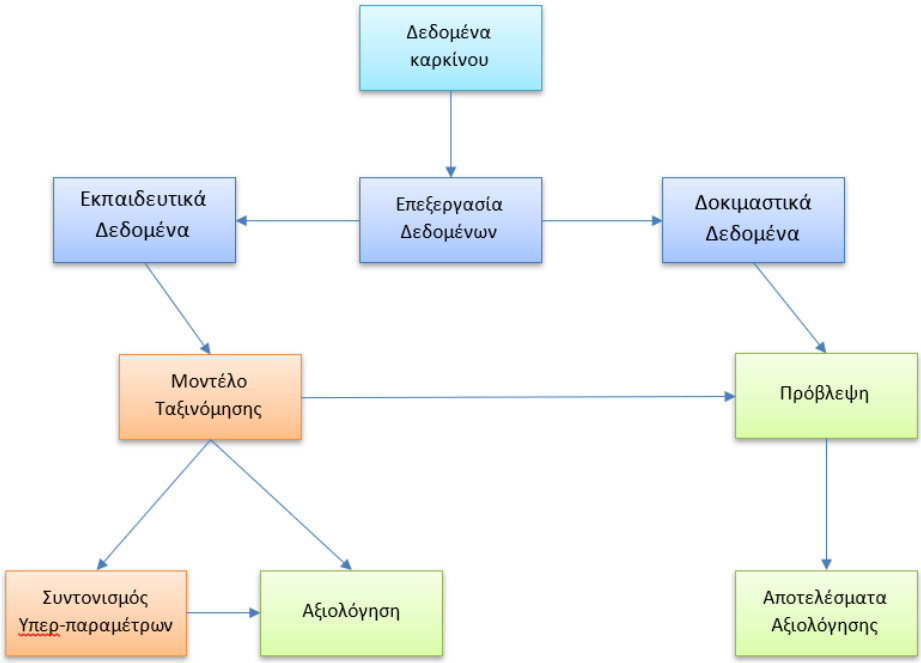
4.1 Επιβλεπόμενη Μηχανική Μάθηση

Η *Επιβλεπόμενη Μάθηση (Supervised Learning – SL)* είναι η πιο ευρέως χρησιμοποιούμενη τεχνική ΜΜ. Η ΜΜ απαιτεί την εκμάθηση μίας συνάρτησης, η οποία προσαρμόζει τα ζεύγη τιμών εισόδου της στην έξοδο. Η συνάρτηση αυτή δέχεται σαν είσοδο επισημασμένα δεδομένα εκπαίδευσης και από αυτά εξάγει αποτέλεσμα. Ουσιαστικά, οι αλγόριθμοι SL ανιχνεύουν το μοτίβο που υπάρχει στα δεδομένα εκπαίδευσης και παράγουν μια συνάρτηση, η οποία έχει τη δυνατότητα να προβλέψει είτε νέα ζεύγη εισόδου, είτε να κάνει παρατηρήσεις οι οποίες, κάτω από άλλες συνθήκες, δεν θα είχαν γίνει ποτέ. Τέλος, οι αλγόριθμοι αυτής της

κατηγορίας έχουν τη δυνατότητα να βελτιώσουν τον κώδικά τους προκειμένου να κάνουν γενικεύσεις με ακρίβεια [16].

Οι αλγόριθμοι SL επιλύουν προβλήματα ακολουθώντας κάποια βήματα (*Εικόνα 8*):

1. Το πρώτο βήμα είναι η απόκτηση ενός συνόλου δεδομένων, τα οποία σχετίζονται με το πρόβλημα που πρέπει να επιλύσει ο αλγόριθμος, και πρέπει να είναι αρκετά σε πλήθος. Επίσης, το μέγεθος του συνόλου των δεδομένων εξαρτάται από το πρόβλημα που επιλύουμε.
2. Η επεξεργασία των δεδομένων είναι το πιο κρίσιμο βήμα στη διαδικασία της ΜΜ, αφού τυχόν προβλήματα στα δεδομένα μπορούν να επηρεάσουν την ακρίβεια της πρόβλεψης που θα παραχθεί από τον αλγόριθμο. Απαιτεί τον καθαρισμό των δεδομένων όπως για παράδειγμα η κατάργηση των περιττών τιμών.
3. Το επόμενο βήμα είναι ο προσδιορισμός του τύπου της μεταβλητής προορισμού, ο οποίος καθορίζει ένα σύνολο αλγορίθμων επιβλεπόμενης μάθησης που μπορούν να εφαρμοστούν. Εάν ο τύπος της μεταβλητής είναι συνεχής, τότε πρόκειται για πρόβλημα παλινδρόμησης. Ενώ, αν ο τύπος των δεδομένων είναι κατηγορικός, τότε είναι πρόβλημα ταξινόμησης.
4. Στη συνέχεια, τα δεδομένα διαχωρίζονται τυχαία σε δύο υποσύνολα, εκπαίδευσης και δοκιμής. Έχει διασφαλιστεί ότι τα δύο υποσύνολα, στα οποία χωρίσαμε τα δεδομένα μας, περιέχουν ισσοροπημένες τιμές διάγνωσης. Επομένως, δεν υπάρχει πρόβλημα υπερπροσάρτησης ή υποτιμήσεως.
5. Το υποσύνολο εκπαίδευσης εφαρμόζεται στον αλγόριθμο ταξινόμησης ΜΜ. Υπάρχουν διάφοροι τέτοιοι αλγόριθμοι και κάθε ένας από αυτούς εκπαιδεύει διαφορετικά.
6. Κάθε αλγόριθμος μπορεί να βελτιωθεί χρησιμοποιώντας ένα σύνολο παραμέτρων. Σε αυτό το σημείο, η εκπαίδευση αλγορίθμων ξεκινάει με τυχαία αρχικοποιημένες παραμέτρους, οι οποίες βελτιώνονται συνεχώς, μέχρις ότου να επιτευχθεί η υψηλότερη ακρίβεια. Όταν επιτευχθεί το βέλτιστο αποτέλεσμα, οι παράμετροι χρησιμοποιούνται ως τελικός αλγόριθμος ΜΜ για την πρόβλεψη των δεδομένων δοκιμών.
7. Τέλος, το μοντέλο αυτό εφαρμόζεται στα δεδομένα εισόδου. Η έξοδος του αλγορίθμου θα είναι μία πρόβλεψη των ετικετών των δεδομένων. Τα αποτελέσματα αυτά αξιολογούνται ανάλογα, βασιζόμενα στα ορθά αποτελέσματα του μοντέλου [16].



Εικόνα 8. Διάγραμμα ροής επιβλεπόμενης μάθησης.

4.1.1. Παλινδρόμηση

Οι αλγόριθμοι SL χωρίζονται σε δύο κατηγορίες. Η πρώτη κατηγορία είναι οι αλγόριθμοι *παλινδρόμησης (Regression)* και η δεύτερη κατηγορία είναι οι αλγόριθμοι *ταξινόμησης (classification).* Οι αλγόριθμοι παλινδρόμησης είναι μια στατιστική διαδικασία, στόχος της οποίας είναι η πρόβλεψη των τιμών μιας δεδομένης μεταβλητής με βάση τις τιμές μίας ή περισσότερων άλλων μεταβλητών. Η δεδομένη μεταβλητή, της οποίας τις τιμές θέλουμε να προβλέψουμε και παίρνει συνεχείς τιμές, ονομάζεται *εξαρτώμενη μεταβλητή, αποτέλεσμα ή μεταβλητή απόκρισης*, ενώ οι δεύτερες μεταβλητές, που αναφέραμε προηγουμένως, ονομάζονται *ανεξάρτητες ή προγνωστικές* [17].

Το αποτέλεσμα μιας παλινδρόμησης είναι συνήθως μια εξίσωση ή ένα μοντέλο, το οποίο συνοψίζει τη σχέση μεταξύ των εξαρτημένων και ανεξάρτητων μεταβλητών. Το μοντέλο αυτό συνοδεύεται από συνοπτικά στατιστικά στοιχεία, τα οποία περιγράφουν πόσο καλά ταιριάζει το μοντέλο στα δεδομένα, το ποσό της διακύμανσης του αποτελέσματος που αντιπροσωπεύει το μοντέλο και μια βάση για τη σύγκριση του υπάρχοντος μοντέλου με άλλα παρόμοια. Με αυτόν τον τρόπο, ο χρήστης είναι σε θέση να καθορίσει ένα συνδυασμό και μια σειρά ανεξάρτητων μεταβλητών που προβλέπουν πιο ικανοποιητικά τις τιμές του αποτελέσματος [17].

Πολυάριθμες μορφές παλινδρόμησης έχουν αναπτυχθεί με στόχο την πρόβλεψη των τιμών μιας μεγάλης ποικιλίας αποτελεσμάτων. Δεδομένου ότι η μοντελοποίηση της παλινδρόμησης εστιάζεται στην εξαρτώμενη μεταβλητή, ο τύπος της παλινδρόμησης που θα χρησιμοποιείται, εξαρτάται από την μεταβλητή που αναλύεται – δηλαδή την εξαρτώμενη – και από τον τελικό στόχο [17].

4.1.1.1. Γραμμική Παλινδρόμηση

Αποτελεί έναν από τους ευκολότερους αλγορίθμους ΜΜ και ανήκει στην κατηγορία της SL. Είναι αλγόριθμος παλινδρόμησης και στατιστική μέθοδος που χρησιμοποιείται προκειμένου να γίνουν προβλέψεις για συνεχείς/πραγματικές ή αριθμητικές μεταβλητές όπως η ηλικία, ο μισθός, τιμές προϊόντων κλπ. Οι γραμμικοί ταξινομητές έχουν αξιολογηθεί ως οι πιο γρήγοροι και, επομένως, χρησιμοποιούνται όταν η ταχύτητα είναι εμπόδιο στο πρόβλημα που θέλουμε να λύσουμε [10][14].

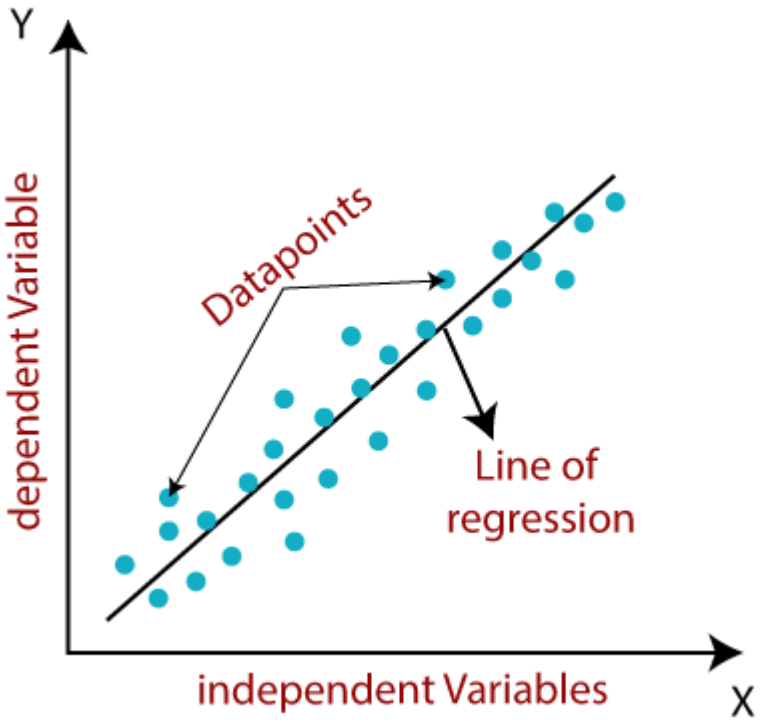
Ο αλγόριθμος *Γραμμικής Παλινδρόμησης - ΓΠ (Linear Regression)* εμφανίζει μια γραμμική σχέση μεταξύ μιας εξαρτημένης μεταβλητής και μίας ή περισσότερων ανεξάρτητων μεταβλητών. Δεδομένου ότι η γραμμική παλινδρόμηση δείχνει τη γραμμική σχέση, αυτό σημαίνει ότι βρίσκει πως αλλάζει η τιμή της εξαρτημένης μεταβλητής σύμφωνα με την τιμή της ανεξάρτητης μεταβλητής. Το συγκεκριμένο μοντέλο παρέχει μία κεκλιμένη ευθεία γραμμή, η οποία αντιπροσωπεύει τη σχέση μεταξύ των μεταβλητών, και βελτιστοποιείται με βάση τη μέθοδο ελαχίστων τετραγώνων [17]. Τα γραμμικά μοντέλα ταξινόμησης διαχωρίζουν τα διανύσματα εισόδου σε κατηγορίες με τη χρήση γραμμικών ορίων απόφασης. Στόχος είναι η ταξινόμηση των στοιχείων με παρόμοιες τιμές χαρακτηριστικών, το οποίο επιτυγχάνεται με τη λήψη μιας απόφασης ταξινόμησης, βασισμένη στην αξία του γραμμικού συνδυασμού των χαρακτηριστικών γνωρισμάτων [18].

Το μοντέλο γραμμικής παλινδρόμησης, όπως ήδη αναφέραμε, χρησιμοποιεί μια κεκλιμένη ευθεία γραμμή η οποία αντιπροσωπεύει τη σχέση που υπάρχει μεταξύ των μεταβλητών (*Εικόνα 9*) και μαθηματικά μπορεί να παρουσιαστεί ως εξής:

$$y = a_0 + a_1 x_1 + \cdots + a_k x_k + \varepsilon \quad (1)$$

Όπου:

- y = εξαρτημένη μεταβλητή
- $x_1 \dots x_k$ = ανεξάρτητες μεταβλητές που περιλαμβάνονται στο μοντέλο
- a_0 = σημείο τομής του y
- $a_1 \dots a_k$= συντελεστές που υποδεικνύουν το βαθμό συσχέτισης μεταξύ κάθε ανεξάρτητης μεταβλητής και του αποτελέσματος.
- ε = σφάλμα που προκύπτει από τη διαφορά μεταξύ των τιμών που προέκυψαν από τα αποτελέσματα και εκείνων που προβλέπονται από το μοντέλο

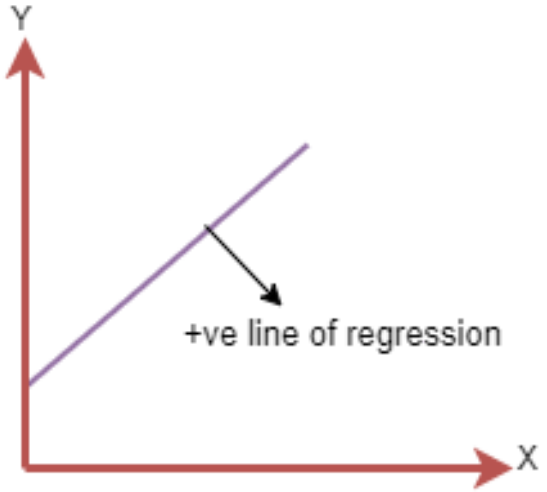


Εικόνα 9. Η γραμμή του μοντέλου Γραμμικής Παλινδρόμησης[18].

Η κεκλιμένη αυτή ευθεία ονομάζεται γραμμή παλινδρόμησης και έχει τη δυνατότητα να δείξει δύο τύπους σχέσεων μεταξύ των μεταβλητών, τη θετική γραμμική σχέση και την αρνητική.

- Η γραμμική σχέση είναι θετική όταν η εξαρτημένη μεταβλητή αυξάνεται στον y-άξονα και η εξαρτημένη στον x-άξονα (*Εικόνα 10*) και η μαθηματική σχέση που την εκφράζει είναι:

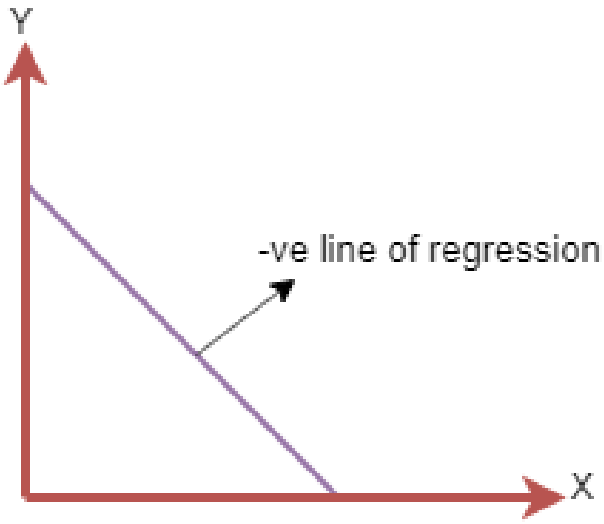
$$y = a_0 + a_1 x$$



Εικόνα 10. Θετική Γραμμική Παλινδρόμηση[18].

- Αρνητική ονομάζεται η γραμμή παλινδρόμησης εάν η εξαρτημένη μεταβλητή μειώνεται στον y-άξονα και η εξαρτημένη αυξάνεται στον x-άξονα (*Εικόνα 11*) και η σχέση που την περιγράφει είναι της μορφής:

$$y = -a_0 + a_1 x$$



Εικόνα 11. Αρνητική Γραμμική Παλινδρόμηση[18].

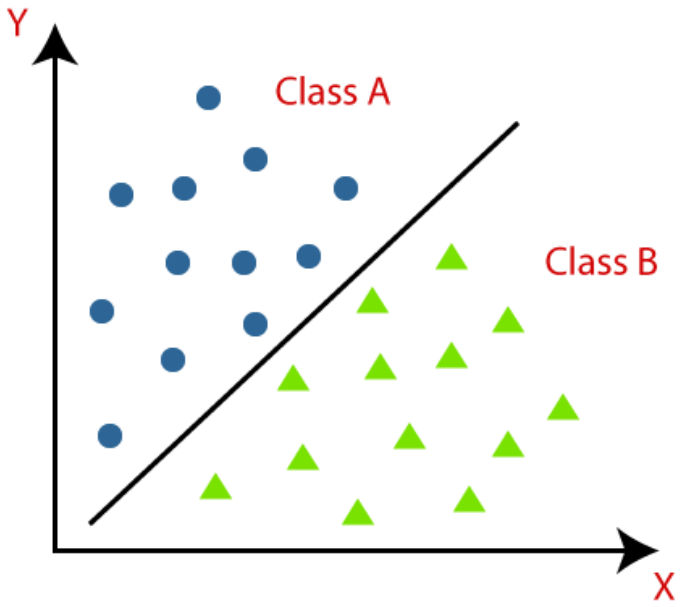
Τέλος, η γραμμική παλινδρόμηση ονομάζεται απλή όταν μία μόνο ανεξάρτητη μεταβλητή χρησιμοποιείται για την πρόβλεψη της τιμής μιας εξαρτημένης αριθμητικής μεταβλητής. Αντίθετα, αν έχουμε περισσότερες από μία ανεξάρτητες μεταβλητές για την πρόβλεψη μιας εξαρτημένης, τότε η γραμμική παλινδρόμηση ονομάζεται πολλαπλή. Όταν εργαζόμαστε με γραμμική παλινδρόμηση, ο κύριος στόχος είναι η εύρεση της καλύτερης γραμμής προσαρμογής, με απώτερο σκοπό την ελαχιστοποίηση του σφάλματος μεταξύ των προβλεπόμενων τιμών και των πραγματικών τιμών. Τελικά, η γραμμή προσαρμογής με το μικρότερο δυνατό σφάλμα θα είναι η βέλτιστη [18].

4.1.2. Ταξινόμηση

Η δεύτερη κατηγορία αλγορίθμων SL είναι, όπως ήδη είπαμε, οι αλγόριθμοι *ταξινόμησης (classification).* Στους αλγορίθμους παλινδρόμησης, όπως προαναφέρθηκε, έχουμε προβλέψει την έξοδο για συνεχείς τιμές. Όταν όμως υπάρχει ανάγκη να προβλέψουμε την έξοδο για κατηγορικές τιμές, χρειαζόμαστε αλγόριθμους ταξινόμησης. Στην ταξινόμηση, ένα πρόγραμμα μαθαίνει από ένα σύνολο δεδομένων και στη συνέχεια ταξινομεί τη νέα παρατήρηση σε διάφορες κλάσεις ή ομάδες, όπως για παράδειγμα NAI/OXI, 1 ή 0, κακοήθεια ή καλοήθεια κλπ., οι οποίες ονομάζονται *ετικέτες ή κατηγορίες*. Αντίθετα από την παλινδρόμηση, στην ταξινόμηση το αποτέλεσμα είναι κάποια κατηγορία και όχι μία τιμή, θα

μπορούσε για παράδειγμα να είναι «Λαχανικό ή Φρούτο», «Μαύρο ή Λευκό» κλπ. Επομένως, αφού ο αλγόριθμος της ταξινόμησης ανήκει στις τεχνικές της SL, αυτό σημαίνει πως τα δεδομένα εισόδου του θα είναι επισημασμένα [18].

Στους αλγορίθμους ταξινόμησης, μια διακριτή συνάρτηση εξόδου y αντιστοιχίζεται στη μεταβλητή εισόδου x, $y = f(x)$, όπου y κατηγορική έξοδος. Ο κύριος στόχος ενός τέτοιου αλγορίθμου είναι η πρόβλεψη της εξόδου για κατηγορικά δεδομένα. Οι αλγόριθμοι ταξινόμησης χωρίζονται σε γραμμικούς και οι μη-γραμμικούς. Γραμμικοί αλγόριθμοι είναι η Λογιστική Παλινδρόμηση και οι Μηχανές Διανυσμάτων Υποστήριξης. Μη-γραμμικοί είναι οι Κ-Κοντινότερος Γείτονας, Kernel Μηχανές Διανυσμάτων Υποστήριξης, Naïve Bayes, Δέντρα Απόφασης και Τυχαία Δάση [18]. Ένας τέτοιος αλγόριθμος, ο οποίος υλοποιεί την μέθοδο της ταξινόμησης, ονομάζεται *ταξινομητής*. Υπάρχουν δύο είδη ταξινομητών, ο *δυαδικός ταξινομητής* και ο *ταξινομητής πολλαπλών κλάσεων*. Δυαδικός ονομάζεται εάν το πρόβλημα έχει μόνο δύο πιθανά αποτελέσματα, όπως για παράδειγμα ΝΑΙ/ΟΧΙ ή ΑΡΣΕΝΙΚΟ/ΘΥΛΗΚΟ. Αντίθετα, αν το πρόβλημα έχει περισσότερα από δύο πιθανά αποτελέσματα, όπως για παράδειγμα η ταξινόμηση τύπων μουσικής, ονομάζεται ταξινομητής πολλαπλών κλάσεων [18].



Εικόνα 12. Διάγραμμα δύο κλάσεων Α και Β[18].

Τα μοντέλα ΜΜ χρησιμοποιούν αλγορίθμους προκειμένου να μάθουν από τα δεδομένα. Ένα τέτοιο πρόγραμμα ονομάζεται *Μάθηση (Learner*) και μπορεί να είναι είτε *Αναβλητική Μάθηση* (*Lazy Learner)* είτε *Έγκαιρη Μάθηση (Eager Learner)*. Η αναβλητική αποθηκεύει πρώτα το σύνολο δεδομένων εκπαίδευσης και περιμένει μέχρι να λάβει το σύνολο δεδομένων δοκιμής. Σε αυτήν την περίπτωση, η ταξινόμηση γίνεται με βάση τα πιο σχετικά δεδομένα, που είναι αποθηκευμένα στο σύνολο δεδομένων εκπαίδευσης. Επίσης, χρειάζεται λιγότερος χρόνος στην εκπαίδευση αλλά περισσότερος χρόνος για τις προβλέψεις. Σε αντίθεση με την

αναβλητική μάθηση, η έγκαιρη μάθηση αναπτύσσει ένα μοντέλο ταξινόμησης βασιζόμενη σε ένα σύνολο δεδομένων εκπαίδευσης και χρειάζεται περισσότερο χρόνο για την εκπαίδευση και λιγότερο χρόνο στην πρόβλεψη [18].

Μόλις το μοντέλο ολοκληρωθεί, είναι απαραίτητο να αξιολογηθεί η απόδοση του και αυτό γίνεται με έναν από τους τρεις ακόλουθους τρόπους:

1. **Log Loss ή Cross-Entropy Loss:** Χρησιμοποιείται σε μοντέλα των οποίων η έξοδος είναι μία τιμή πιθανότητας μεταξύ του 0 και 1. Για ένα καλό μοντέλο δυαδικής ταξινόμησης, η τιμή της απωλειας καταγραφής πρέπει να είναι κοντά στο 0, ενώ αυξάνεται εάν η προβλεπόμενη τιμή αποκλίνει από την πραγματική. Η χαμηλότερη απώλεια καταγραφής αντιπροσωπεύει την υψηλότερη ακρίβεια του μοντέλου.
2. **Πίνακας Σύγχυσης:** Αυτή η μέθοδος αξιολόγησης παρέχει ως έξοδο έναν πίνακα (Πίνακας *1*) ο οποίος περιγράφει την απόδοση του μοντέλου. Αποτελείται από προβλέψεις που οδηγούν σε μια συνοπτική μορφή με συνολικό αριθμό έγκυρων και εσφαλμένων προβλέψεων.

	Πραγματικά Θετικά	**Πραγματικά Αρνητικά**
Θετικές Προβλέψεις	Έγκυρα Θετικά (ΕΘ)	Λανθασμένα Θετικά (ΛΘ)
Αρνητικές Προβλέψεις	Λανθασμένα Αρνητικά (ΛΑ)	Έγκυρα Αρνητικά (ΕΑ)

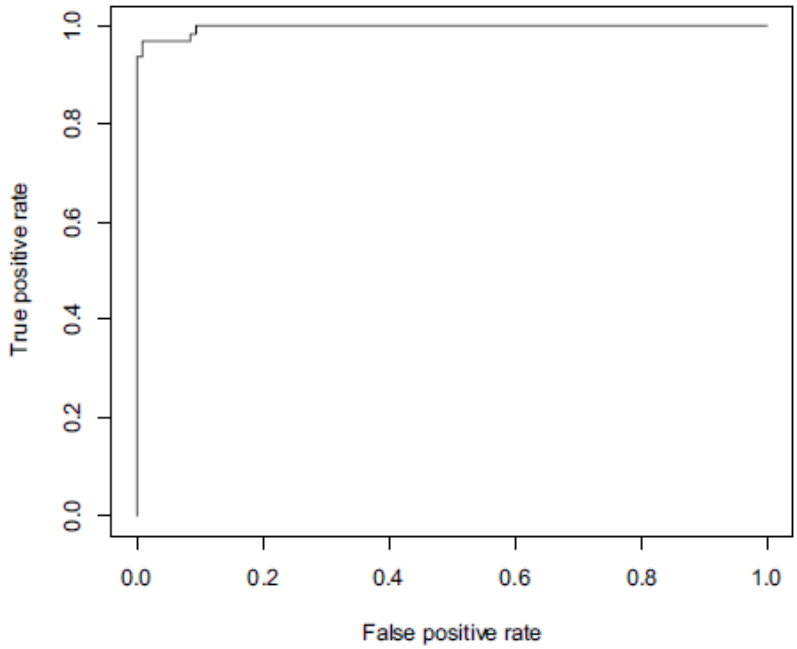
Πίνακας 1. Πίνακας Σύγχυσης για την αξιολόγηση του μοντέλου ταξινόμησης.

Όπου:

- (ΕΘ) : η αναλογία των θετικών προβλέψεων στο σύνολο των πραγματικών θετικών
- (ΛΘ) : η αναλογία των θετικών προβλέψεων στο σύνολο των πραγματικών αρνητικών
- (ΛΑ) : η αναλογία των αρνητικών προβλέψεων στο σύνολο των πραγματικών θετικών
- (ΕΑ) : η αναλογία των αρνητικών προβλέψεων στο σύνολο των πραγματικών αρνητικών

Το Accuracy του μοντέλου δίνεται από τον τύπο : $\frac{(E\Theta)+(EA)}{Συνολικός\ Πληθυσμός}$

3. **Καμπύλη AUC-ROC (Area Under the Curve – Receiver Operating Characteristics Curve):** Η καμπύλη ROC χρησιμοποιείται για την απεικόνιση των επιδόσεων των μοντέλων ταξινόμησης, παρουσιάζοντας την αντιστάθμιση μεταξύ του κόστους και του οφέλους του εν λόγω ταξινομητή, δηλαδή απεικονίζει το σύνολο των δυνατών ζευγαριών (ΛΘ) και (ΕΘ) που μπορεί να ληφθούν. Η περιοχή κάτω από την καμπύλη ROC εμφανίζει την απόδοση του μοντέλου ΜΜ και οι υψηλές επιδόσεις σημειώνονται με τιμές κοντά στο 1 [19].

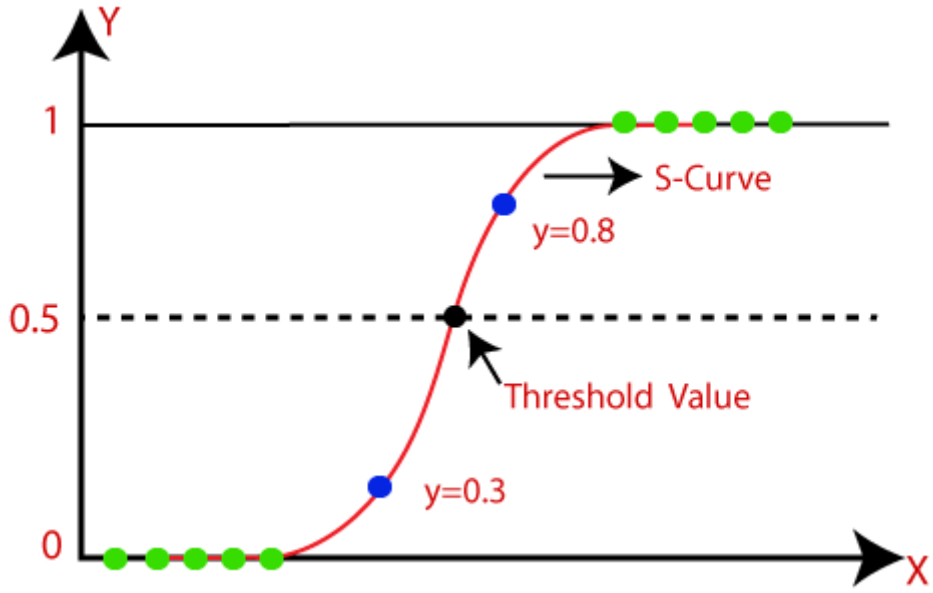


Εικόνα 13. Καμπύλη ROC [19].

4.1.2.1. Λογιστική Παλινδρόμηση

Η *Λογιστική Παλινδρόμηση (Logistic Regression – LR)* είναι ένας από τους πιο δημοφιλής και σημαντικούς αλγορίθμους SL για την επίλυση προβλημάτων ταξινόμησης και βασίζεται στην έννοια της πιθανότητας. Μπορεί να χρησιμοποιηθεί για την ταξινόμηση παρατηρήσεων, χρησιμοποιώντας διαφορετικούς τύπους δεδομένων, και μπορεί εύκολα να προσδιορίσει τις πιο αποτελεσματικές μεταβλητές που χρησιμοποιούνται για την ταξινόμηση. Ο συγκεκριμένος αλγόριθμος προβλέπει μια εξαρτώμενη κατηγορική μεταβλητή, χρησιμοποιώντας ένα σύνολο δεδομένων ανεξάρτητων μεταβλητών. Το αποτέλεσμα είναι πιθανότητες, επομένως παίρνουν τιμές μεταξύ 0 έως 1, οι οποίες στη συνέχεια αντιστοιχίζονται σε 0 ή 1, αληθής ή ψευδής [18].

Προκειμένου να γίνει η πρόβλεψη, κάνει χρήση μιας σύνθετης συνάρτησης κόστους, η οποία χαρτογραφεί οποιαδήποτε πραγματική τιμή σε μια άλλη τιμή, μεταξύ των 0 και 1. Με αυτόν τον τρόπο δείχνει την πιθανότητα για κάτι, όπως για παράδειγμα αν τα κύτταρα είναι καρκινικά ή όχι. Η συνάρτηση αυτή ονομάζεται *«Σιγμοειδής Συνάρτηση»* (Εικόνα *14*) και χρησιμοποιείται για την αντιστοίχιση των προβλεπόμενων τιμών στις πιθανότητες. Δηλαδή, αντιστοιχίζει οποιαδήποτε πραγματική αξία σε μια άλλη τιμή, μέσα σε μια περιοχή 0 και 1, η οποία δε μπορεί να παραβιάσει αυτό το όριο και γι’ αυτό το λόγο η καμπύλη που θα σχηματιστεί θα έχει τη μορφή του S [14], [18].



Εικόνα 14. Σιγμοειδής Καμπύλη[18].

Το μοντέλο Λογιστικής Παλινδρόμησης μπορεί να παρουσιαστεί μέσω του μαθηματικού τύπου :

$$\log\left[\frac{y}{1-y}\right] = a_0 + a_1 x + \cdots + a_k x_k + \varepsilon \quad (2)$$

ο οποίος προκύπτει από τη σχέση (1).
Όπου:

- y = η πιθανότητα ενός γεγονότος
- a_0 = σημείο τομής του y
- $x_1 \ldots x_k$ = ανεξάρτητες μεταβλητές που περιλαμβάνονται στο μοντέλο
- $a_1 \ldots a_k$= συντελεστές που υποδεικνύουν το βαθμό συσχέτισης μεταξύ κάθε ανεξάρτητης μεταβλητής και του αποτελέσματος
- ε = σφάλμα που προκύπτει από τη διαφορά μεταξύ των τιμών που προέκυψαν από τα αποτελέσματα και εκείνων που προβλέπονται από το μοντέλο

Η Λογιστική Παλινδρόμηση μπορεί να χωριστεί σε τρεις κατηγορίες αναλόγως τη φύση της εξαρτημένης κατηγορικής μεταβλητής:

- **Διωνυμική**: υπάρχουν μόνο δύο πιθανοί τύποι εξαρτημένων μεταβλητών, π.χ. 0 ή 1, ΝΑΙ/ΟΧΙ κλπ.
- **Πολυωνυμική:** μπορούν να υπάρχουν τρεις ή περισσότεροι πιθανοί τύποι εξαρτώμενης μεταβλητής χωρίς κάποια φυσική διαβάθμιση, όπως για παράδειγμα ο χαρακτηρισμός του χρώματος ενός αντικειμένου ως κόκκινου, πράσινου ή μπλε κλτ.

- **Τακτική:** η εξαρτημένη μεταβλητή συνίσταται από τρεις ή περισσότερες κατηγορίες μεταξύ των οποίων ισχύει η έννοια της ανισότητας, για παράδειγμα σε μια ερώτηση κλίμακας διαφωνώ καθόλου, λίγο, μέτρια, πολύ, πάρα πολύ [20].

Τελικά, η Λογιστική Παλινδρόμηση είναι παρόμοια με τη Γραμμική Παλινδρόμηση. Σε αυτό το σημείο, είναι αναγκαίο να τονιστεί η διαφορά τους, η οποία βρίσκεται στον τρόπο που χρησιμοποιείται κάθε μία από αυτές. Η πρώτη χρησιμοποιείται για την επίλυση προβλημάτων ταξινόμησης, ενώ η δεύτερη για την επίλυση προβλημάτων παλινδρόμησης [18].

4.1.2.2. Μηχανές Διανυσμάτων Υποστήριξης

Οι *Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines – SVM)* είναι ένας αλγόριθμος SL και μπορεί να χρησιμοποιηθεί για την επίλυση προβλημάτων ταξινόμησης και παλινδρόμησης. Ωστόσο, η χρήση τους είναι κυρίως για προβλήματα ταξινόμησης και για μικρά σύνολα δεδομένων, λόγω του ότι χρειάζονται αρκετό χρόνο για να τα επεξεργαστούν [18],[21].

Ας υποθέσουμε ότι υπάρχει ένα σύνολο δεδομένων δειγμάτων, όπου το κάθε ένα ανήκει σε μία από δύο γνωστές κατηγορίες (τετράγωνο ή αστέρι -

*Εικόνα 15*Α) , και πως τα δεδομένα αυτά είναι γραμμικώς διαχωρίσιμα, δηλαδή υπάρχει μία ευθεία γραμμή που διαχωρίζει τα δείγματα αυτά σωστά σε δύο κατηγορίες. Παρατηρώντας την

Εικόνα *15*Α βλέπουμε πως, συνήθως, υπάρχουν περισσότερες από μία τέτοιες ευθείες γραμμές. Όλες αυτές οι ευθείες γραμμές μπορούν να αποδίδουν εξίσου καλά στα δεδομένα εκπαίδευσης, η απόδοση τους όμως στα δεδομένα δοκιμών μπορεί να διαφέρει [22].

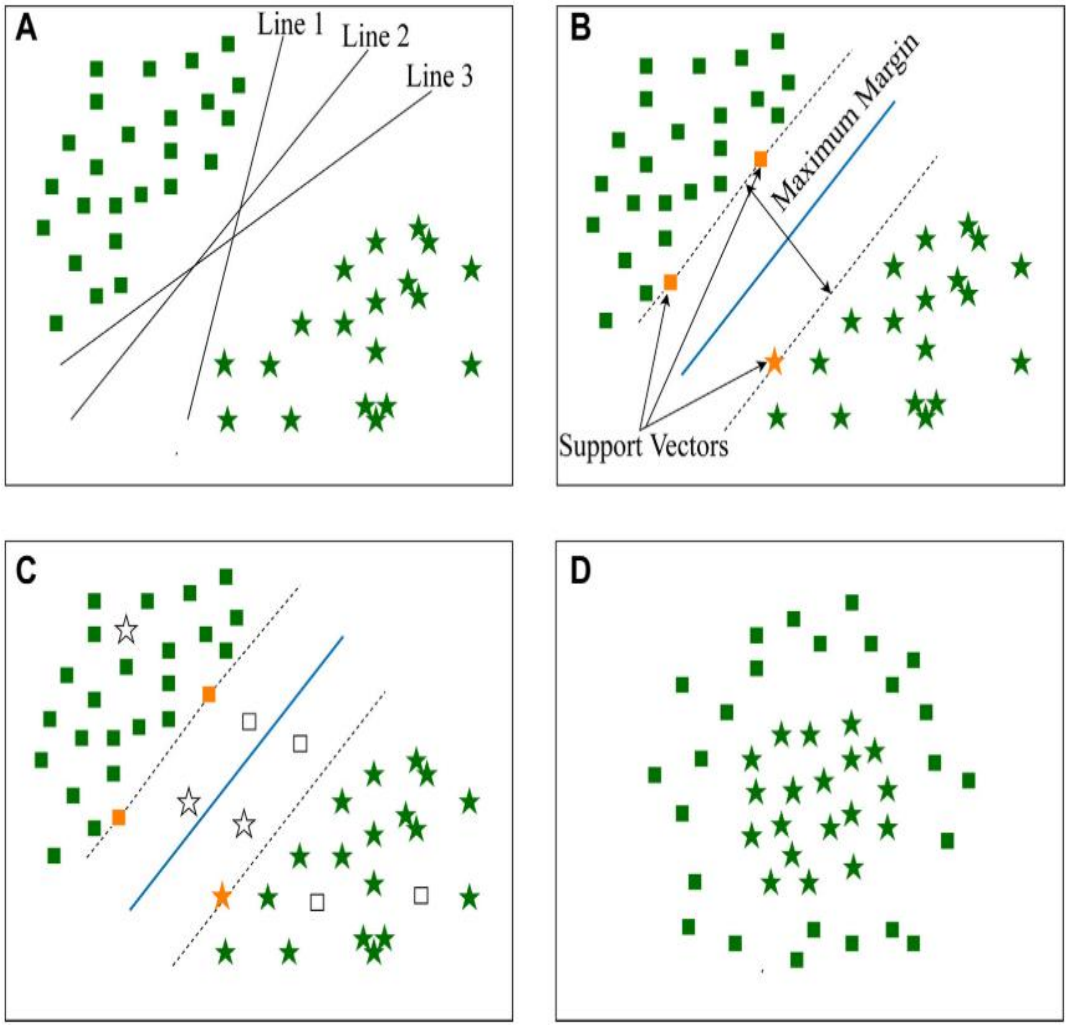
Οι αλγόριθμοι SVM βοηθούν στην εύρεση της καλύτερης γραμμής (ή όριο απόφασης), που μπορούν να διαχωρίσουν έναν χώρο ν-διαστάσεων σε κλάσεις, προκειμένου να υπάρχει η δυνατότητα να τοποθετηθούν τα δεδομένα στη σωστή κατηγορία. Αυτό το καλύτερο όριο απόφασης που θα δημιουργηθεί, ονομάζεται *υπερεπίπεδο (hyperplane)*. Ένα τέτοιο υπερεπίπεδο μπορεί να είναι μία ευθεία γραμμή (σε δισδιάστατο χώρο) ή ένα επίπεδο (σε τρισδιάστατο χώρο). Ο αλγόριθμος βρίσκει το πλησιέστερο σημείο των γραμμών και από τις δύο κλάσεις (

*Εικόνα 15*Β). Τα σημεία ονομάζονται *διανύσματα υποστήριξης*, ενώ η απόσταση μεταξύ τους ονομάζεται *περιθώριο*. Στόχος αυτού του αλγορίθμου είναι η μεγιστοποίηση του περιθωρίου, δηλαδή η εύρεση του *βέλτιστου υπερεπιπέδου* [18].

Λόγω της ύπαρξης θορύβου και ακραίων τιμών, μπορεί να υπάρχει αλληλοεπικάλυψη μεταξύ των δεδομένων και από τις δύο κλάσεις (

*Εικόνα 15*C). Οι SVM μπορούν να αντιμετωπίσουν αυτό το πρόβλημα, χαλαρώνοντας την κατάσταση του βέλτιστου υπερεπιπέδου και, επομένως, επιτρέποντας σε μερικά δεδομένα να παραβιάσουν αυτά τα διαχωριστικά υπερεπίπεδα (

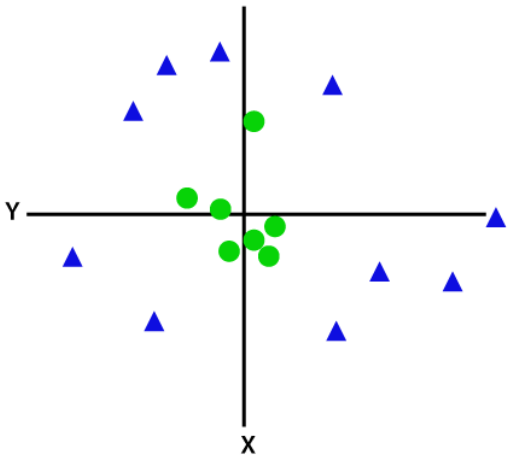
*Εικόνα 15*C) Για να μπορεί να ελέγξει την ποσότητα της χαλάρωσης, χρησιμοποιεί μία μη-αρνητική παράμετρο C. Όσο πιο υψηλή είναι η τιμή του C, τόσο πιο χαλαρή θα είναι η παραβίαση του περιθωρίου. Επομένως, αν αυτή η τιμή της C είναι η μηδενική, τότε οδηγούμαστε σε ένα σκληρό περιθώριο, με αποτέλεσμα να μην είναι δυνατή καμία παραβίαση. Οι μεγαλύτερες τιμές του C καθιστούν τις SVM λιγότερο ευαίσθητες στο θόρυβο στα δεδομένα εκπαίδευσης [22].



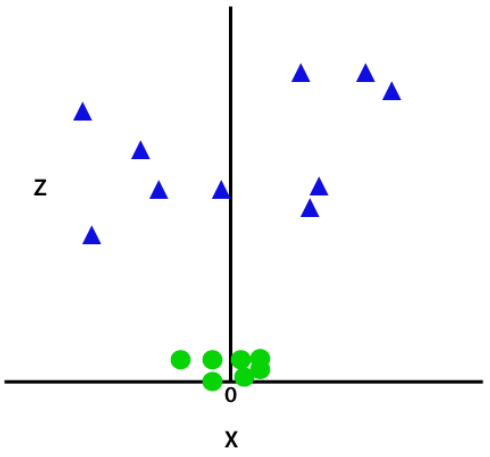
Εικόνα 15. Περιπτώσεις ταξινόμησης με χρήση SVM[22].

Για δεδομένα που είναι αδύνατο να διαχωριστούν γραμμικά (
*Εικόνα 15*D), χρησιμοποιούνται μη-γραμμικές SVM. Για να κατηγοριοποιήσουμε δεδομένα όπως αυτά στις
*Εικόνα 15*D και Εικόνα *16*, θα προσθέσουμε μία ακόμα διάσταση z, στις ήδη υπάρχουσες διαστάσεις x και y, η οποία θα υπολογίζεται από τον τύπο $z = x^2 + y^2$. Με την προσθήκη της τρίτης διάστασης, ο χώρος του δείγματος θα μοιάζει πλέον με το χώρο που απεικονίζεται στην Εικόνα *16*. Χώρος δείγματος για

μη-γραμμικά *δεδομένα[18]*. [22]. Επομένως, ο αλγόριθμος SVM μπορεί πλέον πολύ εύκολα να κάνει τον διαχωρισμό των δύο κλάσεων.



Εικόνα 16. Χώρος δείγματος για μη-γραμμικά δεδομένα[18].

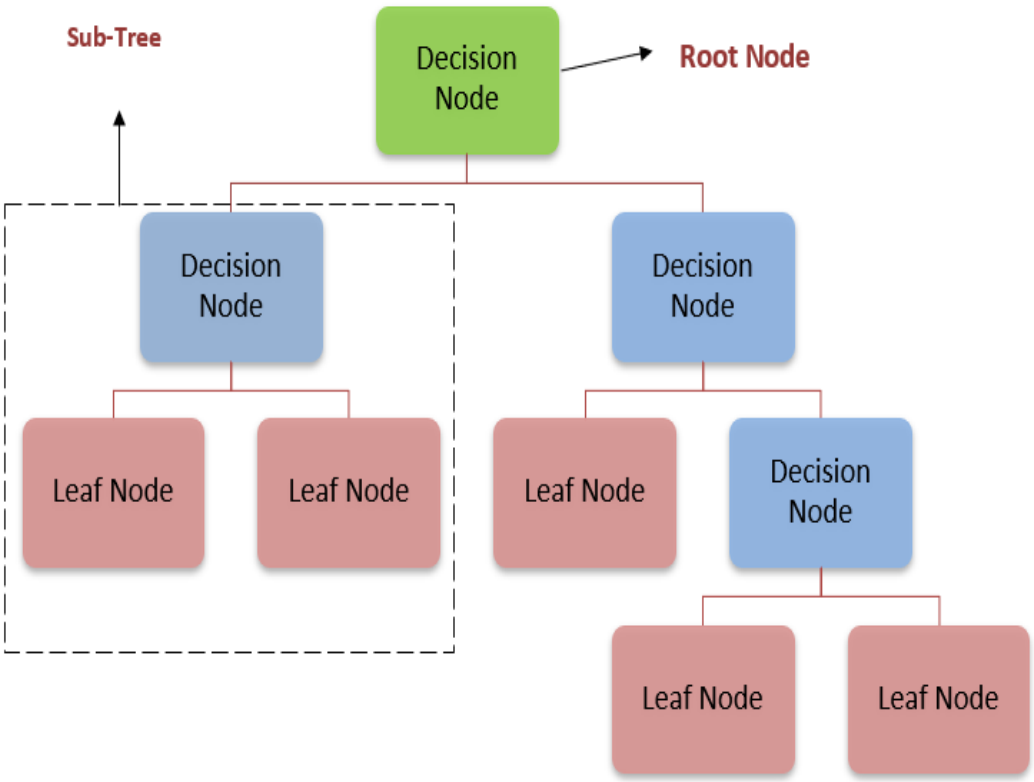


Εικόνα 17.Χώρος δείγματος μετά την προσθήκη της τρίτης διάστασης[18]

4.1.2.3. Δέντρα Απόφασης

Τα *Δέντρα Αποφάσεων (Decision Trees – DT),* επίσης, είναι μια τεχνική SL που χρησιμοποιείται κυρίως για προβλήματα ταξινόμησης, ωστόσο μπορεί να χρησιμοποιηθεί και για επίλυση προβλημάτων παλινδρόμησης. Μιμούνται την ανθρώπινη ικανότητα σκέψης κατά τη διαδικασία λήψης μίας απόφασης. Είναι ένας ταξινομητής η λογική του οποίου είναι δομημένη σε δέντρα. Οι εσωτερικοί κόμβοι αντιπροσωπεύουν τα χαρακτηριστικά ενός συνόλου δεδομένων, οι κλάδοι αντιπροσωπεύουν τους κανόνες απόφασης και κάθε κόμβος φύλλων αντιπροσωπεύει το αποτέλεσμα. Ουσιαστικά, πρόκειται για μία γραφική αναπαράσταση όλων των πιθανών λύσεων σε ένα πρόβλημα ή μια απόφαση που βασίζεται σε δεδομένες συνθήκες [23].

Στα DT υπάρχουν δύο τύποι κόμβων, οι *κόμβοι αποφάσεων (decision node)* και ο *κόμβοι φύλλων (leaf node)*. Οι κόμβοι αποφάσεων είναι υπεύθυνοι για τη λήψη οποιασδήποτε απόφασης και έχουν πολλαπλούς κλάδους. Οι κόμβοι φύλλων είναι το αποτέλεσμα των αποφάσεων, που πάρθηκαν από τους κόμβους αποφάσεων, και δεν περιέχουν περαιτέρω κλάδους. Οι αποφάσεις ή οι δοκιμές εκτελούνται με βάση τα χαρακτηριστικά του συνόλου δεδομένων που έχουμε στη διάθεση μας. Ο λόγος που ονομάζεται Δέντρα Απόφασης είναι επειδή ξεκινά με έναν *ριζικό κόμβο (Root Node)*, ο οποίος επεκτείνεται σε περαιτέρω κλαδιά και κατασκευάζει μια δομή που μοιάζει με δέντρο. Για να χτίσουμε ένα δέντρο απόφασης, χρησιμοποιούμε τον αλγόριθμο CART (Classification and Regression Tree). Τέλος, ένα δέντρο απόφασης μπορεί να περιέχει και κατηγορικά δεδομένα και αριθμητικά [23].



Εικόνα 18. Γενική δομή ενός Δέντρου Αποφάσεων.

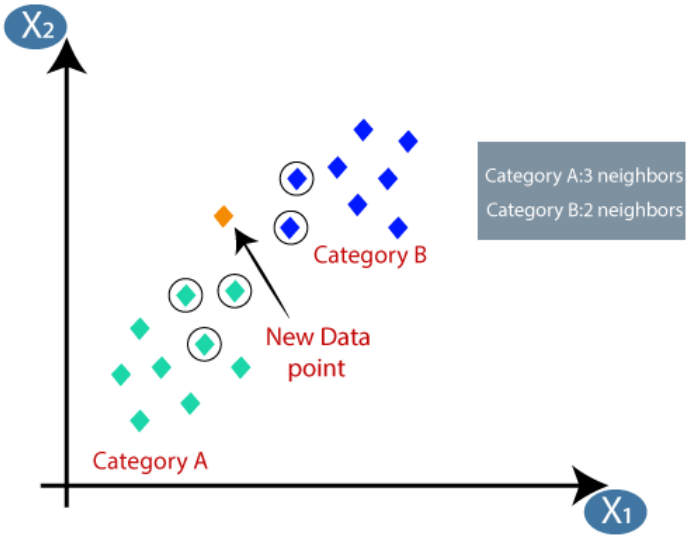
Στην Εικόνα *18* βλέπουμε τη γενική δομή ενός DT. Ο *ριζικός κόμβος (Decision Node)* είναι η αρχή του, αντιπροσωπεύει ολόκληρο το σύνολο δεδομένων και χωρίζεται σε δύο ή και περισσότερα ομοιογενή σύνολα. Οι *κόμβοι φύλλων (Leaf Nodes)* είναι ο τελικός κόμβος εξόδου και το δέντρο δε μπορεί να χωριστεί παραπάνω μετά τη λήψη μίας απόφασης. *Διαχωρισμός (Splitting)* ονομάζεται η διαδικασία διαίρεσης ενός κόμβου σε νέους, σύμφωνα με τις εκάστοτε συνθήκες. *Κλάδος/Υπο-δέντρο (Branch/Sub-Tree)* είναι το νέο δέντρο που σχηματίζεται από το διαχωρισμό ενός κόμβου. *Κλάδεμα (Pruning)* είναι η διαδικασία απομάκρυνσης των ανεπιθύμητων κλαδιών από το δέντρο. Τέλος, ο ριζικός κόμβος του δέντρου ονομάζεται *Γονικός Κόμβος (Parent Node)* ενώ οι κόμβοι που δημιουργούνται από τον διαχωρισμό, ονομάζονται *θυγατρικοί κόμβοι (Child Nodes)* [23].

Για την πρόβλεψη της κλάσης των δεδομένων χρησιμοποιώντας ένα DT, ο αλγόριθμος ξεκινάει από τον ριζικό κόμβο, συγκρίνοντας τις τιμές του χαρακτηριστικού της ρίζας με το πραγματικό σύνολο δεδομένων και με βάση αυτό ακολουθεί τον κλάδο και μεταπηδά στον επόμενο κόμβο. Στον επόμενο κόμβο, ο αλγόριθμος επαναλαμβάνει την ίδια διαδικασία συγκρίνοντας την τιμή του χαρακτηριστικού με τον εκάστοτε κόμβο και η διαδικασία αυτή συνεχίζεται μέχρις ότου να φτάσει σε κόμβο φύλλων. Η επιλογή του καλύτερου χαρακτηριστικού για τον ριζικό κόμβο, αλλά και για τους δευτερεύοντες κόμβους, γίνεται μέσω της

τεχνικής *Μέτρο Επιλογής Χαρακτηριστικών (Attribute Selection Measure – ASM)* [23].

4.1.2.4. Κ-Κοντινότερος Γείτονας

Ο αλγόριθμος *Κ-Κοντινότερος Γείτονας (K-Nearest Neighbor – KNN)* είναι ένας από τους απλούστερους αλγόριθμους μηχανικής μάθησης και χρησιμοποιείται κυρίως για προβλήματα ταξινόμησης, ωστόσο μπορεί να χρησιμοποιηθεί και για προβλήματα παλινδρόμησης. Ο αλγόριθμος αποθηκεύει όλα τα δεδομένα και μόλις εμφανιστούν νέα δεδομένα, υποθέτει την ομοιότητα μεταξύ των νέων δεδομένων με τα ήδη υπάρχοντα. Στη συνέχεια, τοποθετεί τα νέα δεδομένα σε μία ήδη υπάρχουσα κατηγορία, με την οποία μοιάζει περισσότερο. Ο K-NN είναι ένας μη-παραμετρικός αλγόριθμος, το οποίο σημαίνει πως δεν κάνει καμία υπόθεση για τα δεδομένα του. Ο K-NN στη φάση εκπαίδευσης αποθηκεύει μόνο το σύνολο δεδομένων, όταν λάβει τα νέα δεδομένα, τότε τα ταξινομεί σε μια κατηγορία, που είναι πολύ παρόμοια με τα νέα δεδομένα [24].



Εικόνα 19. Άφιξη νέου δεδομένου στον K-NN[24].

Ας υποθέσουμε ότι έχουμε το πορτοκαλί σα νέο δεδομένο (Εικόνα *19*) και θέλουμε να το τοποθετήσουμε στην κατάλληλη κατηγορία Α ή Β. Αρχικά, θα επιλέξουμε έναν αριθμό γειτόνων Κ, πχ. Κ=5. Ουσιαστικά, θα πρέπει να υπολογιστούν οι 5 κοντινότερες Ευκλείδειες αποστάσεις, με εκκίνηση το νέο σημείο δεδομένων (πορτοκαλί) και σημείο άφιξης οποιοδήποτε σημείο της Κατηγορίας Α ή Β. Με αυτόν τον τρόπο, θα βρεθούν οι κοντινότερους γείτονες. Στο συγκεκριμένο παράδειγμα που θέσαμε, έχουμε 3 κοντινούς γείτονες στην κατηγορία Α και 2 στην κατηγορία Β. Όπως βλέπουμε, οι 3 πλησιέστεροι γείτονες είναι από την κατηγορία Α, επομένως το νέο δεδομένο μας πρέπει να ανήκει στην Α κατηγορία [24].

Για την επιλογή του K δεν υπάρχει κάποιος συγκεκριμένος τρόπος, επομένως χρειάζεται να δοκιμάσουμε κάποιες τιμές προκειμένου να εντοπιστεί η βέλτιστη. Μία τιμή που προτιμάτε συνήθως είναι η K=5. Μία πολύ χαμηλή τιμή για το K, όπως η τιμή K=1 ή K=2, θα μπορούμε να προκαλέσει σφάλματα. Μεγαλύτερες τιμές από την K=5 μπορούν να χρησιμοποιηθούν αλλά υπάρχει περίπτωση να προκύψουν δυσκολίες [24].

4.2. Μη-Επιβλεπόμενη Μάθηση

Η *Μη-Επιβλεπόμενη Μάθηση (Unsupervised Learning - USL)* είναι μία τεχνική κατά την οποία τα ίδια τα μοντέλα έχουν τη δυνατότητα να βρίσκουν τα κρυμμένα μοτίβα και τις πληροφορίες στα δεδομένα και δεν επιβλέπουν χρησιμοποιώντας ένα σύνολο δεδομένων εκπαίδευσης. Αντίθετα από την SL, στη USL έχουμε τα δεδομένα εισόδου αλλά όχι τα δεδομένα εξόδου. Στόχος της είναι να βρεθεί η υποκείμενη δομή του συνόλου των δεδομένων, να ομαδοποιηθούν βάση των ομοιοτήτων τους και να αναπαρασταθεί αυτό το σύνολο δεδομένων σε συμπιεσμένη μορφή. Δεν υπάρχουν σωστές απαντήσεις και δεν υπάρχει κάποιος αλγόριθμος να επιβλέπει τη διαδικασία [22].
Τα μοντέλα μη επιβλεπόμενης μάθησης χωρίζονται σε τρείς κατηγορίες:

- Ομαδοποίηση
- Συσχέτιση
- Μείωση διαστάσεων

4.2.1. Ομαδοποίηση

Η *ομαδοποίηση (Clustering)* είναι μία τεχνική εξόρυξης δεδομένων η οποία ομαδοποιεί δεδομένα χωρίς σήμανση με βάση τις ομοιότητες ή τις διαφορές τους. Οι αλγόριθμοι ομαδοποίησης χρησιμοποιούνται για την επεξεργασία μη επεξεργασμένων και μη ταξινομημένων δεδομένων σε ομάδες και χωρίζονται σε τέσσερις κατηγορίες [25].

Η πρώτη κατηγορία ονομάζεται *αποκλειστική (exclusive) ή «σκληρή» (“hard”) ομαδοποίηση* και ορίζει ότι ένα σημείο δεδομένων μπορεί να υπάρχει μόνο σε μία ομάδα. Η δεύτερη κατηγορία ονομάζεται *επικαλυπτόμενη (overlapping) ή «μαλακή» (“soft”) ομαδοποίηση* και, αντίθετα από την αποκλειστική ομαδοποίηση, επιτρέπει στα δεδομένα της να ανήκουν σε πολλές ομάδες [25].

Η τρίτη κατηγορία ονομάζεται *ιεραρχική ομαδοποίηση (hierarchical clustering) ή ιεραρχική ανάλυση ομάδων (hierarchical cluster analysis)* και χωρίζεται σε δύο κατηγορίες, τη *συσσωρευτική* (agglomerative) και τη *διαιρετική (divisive)*. Η συσσωρευτική ομαδοποίηση θεωρείται *«προσέγγιση από κάτω προς τα πάνω» (bottoms-up approach)* και τα δεδομένα της αρχικά απομονώνονται ως ξεχωριστές

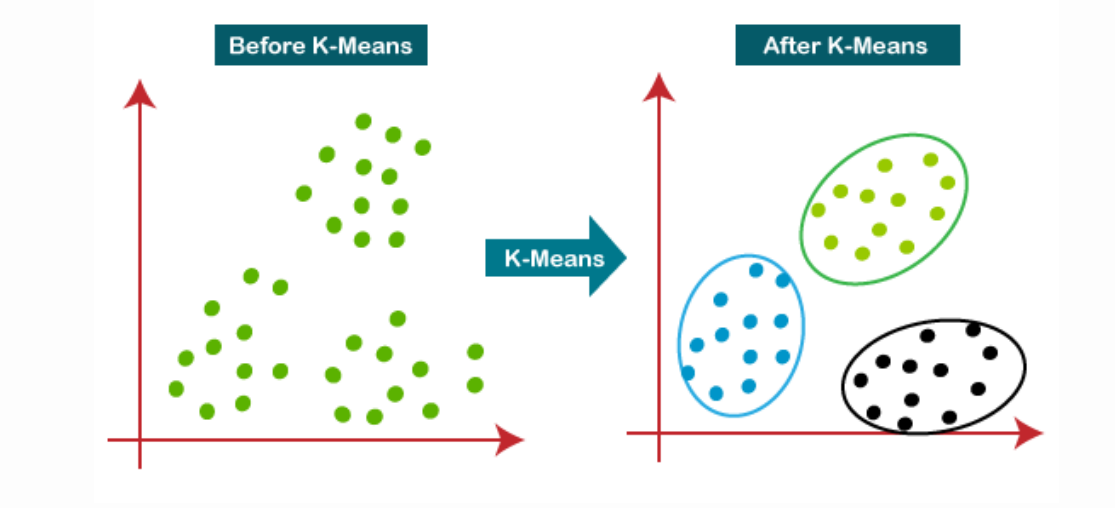
ομάδες και στη συνέχεια συγχωνεύονται επαναληπτικά με βάση τις ομοιότητές τους, μέχρις ότου να επιτευχθεί μία ομάδα. Υπάρχουν τέσσερις διαφορετικές μέθοδοι που χρησιμοποιούνται για να μετρηθεί η ομοιότητα, η μέθοδος Ward, η μέθοδος Μέσης Σύνδεσης, η μέθοδος Ελάχιστης Σύνδεσης και, τέλος, η μέθοδος Μέγιστης Σύνδεσης. Η διαιρετική ομαδοποίηση είναι το αντίθετο της συσσωρευτικής, δηλαδή είναι *«προσέγγιση από πάνω προς τα κάτω» (top-down approach)*, όπου μία ομάδα δεδομένων διαιρείται με βάση τις διαφορές τους. Η τελευταία κατηγορία είναι τα *πιθανοτικά μοντέλα ομαδοποίησης (probabilistic clustering)* κατά τα οποία τα δεδομένα συγκεντρώνονται με βάση την πιθανότητα να ανήκουν σε μία συγκεκριμένη κατανομή [25].

4.2.1.1. Ο αλγόριθμος K-Μέσων

Ο αλγόριθμος *K-Μέσων (K-Means)* είναι ένας αλγόριθμος USL και χρησιμοποιείται για την επίλυση προβλημάτων ομαδοποίησης. Είναι ένας επαναληπτικός αλγόριθμος που διαιρεί το σύνολο δεδομένων χωρίς σήμανση σε K διαφορετικά συμπλέγματα, με τέτοιο τρόπο ώστε κάθε σύνολο δεδομένων να ανήκει σε ένα μόνο σύμπλεγμα που έχει παρόμοιες ιδιότητες. Το K καθορίζει τον αριθμό των προκαθορισμένων συμπλεγμάτων που πρέπει να δημιουργηθούν κατά τη διαδικασία, για K=2 θα υπάρχουν δύο συμπλέγματα, για K=3 θα υπάρχουν τρία συμπλέγματα κ.ο.κ [9].

Ο αλγόριθμος ξεκινάει με K τυχαία σημεία, τα οποία ονομάζονται *κεντροειδή* του συμπλέγματος και δηλώνουν το κέντρο βάρους του. Το K υποδηλώνει σε πόσα συμπλέγματα θέλουμε ο αλγόριθμος να δημιουργήσει. Ο αλγόριθμος ξεκινάει με την αρχικοποίηση του K, ο αριθμός του οποίου δίνεται ως είσοδος από τον χρήστη και η επιλογή του επαφίεται στη δική του γνώση και εμπειρία. Ωστόσο, υπάρχει η *Μέθοδος του Αγκώνα* η οποία μπορεί να βοηθήσει σε κάποιες περιπτώσεις [9].

Το πρώτο βήμα του αλγορίθμου είναι να εξετάσει κάθε δείγμα σε σχέση με τα κεντροειδή των συμπλεγμάτων. Με χρήση κάποιου μέτρου απόστασης, αναθέτει το εξεταζόμενο δείγμα στο σύμπλεγμα, του οποίου το κεντροειδές είναι το πλησιέστερο ως προς το συγκεκριμένο δείγμα. Το δεύτερο βήμα αφορά τον επαναπροσδιορισμό και τη μετατόπιση του κεντροοειδούς κάθε συμπλέγματος. Παίρνοντας τον μέσο όρο των δειγμάτων κάθε συμπλέγματος, υπολογίζονται ξανά τα κεντροειδή τους, ώστε το κεντροειδές να είναι πιο αντιπροσωπευτικό στο πρόσφατα διαμορφωμένο σύμπλεγμα. Ο αλγόριθμος εκτελεί επαναληπτικά αυτά τα δύο βήματα, μέχρις ότου τα κεντροειδή των συμπλεγμάτων να μετατοπίζονται ελάχιστα και σε απόσταση μικρότερη από την τιμή του κατωφλίου που δίνεται. Ως εναλλακτικό κριτήριο τερματισμού του αλγορίθμου μπορεί να χρησιμοποιηθεί και ο αριθμός επαναλήψεων του αλγορίθμου [9].



Εικόνα 20. Αλγόριθμος K-Μέσων[26].

4.2.2. Κανόνες Συσχέτισης – Μείωση Διαστάσεων

Η μέθοδος των *κανόνων συσχέτισης (Association)* βασίζεται σε κανόνες για την εύρεση σχέσεων μεταξύ μεταβλητών σε ένα μεγάλο σύνολο δεδομένων. Αλγόριθμοι που βασίζονται σε τέτοιες μεθόδους χρησιμοποιούνται κυρίως στο marketing. Τέλος, η τεχνική της *Μείωσης Διαστάσεων (Dimensionality Reduction)* χρησιμοποιείται όταν σε ένα σύνολο δεδομένων ο αριθμός των χαρακτηριστικών ή των διαστάσεων είναι πολύ υψηλός. Μειώνει τον αριθμό των δεδομένων εισόδων σε ένα μέγεθος εύκολα διαχειρίσιμο, ενώ παράλληλα διατηρεί την ακεραιότητα του συνόλου δεδομένων, όσο το δυνατόν περισσότερο. Χρησιμοποιείται συνήθως στο στάδιο επεξεργασίας των δεδομένων [25].

4.2.2.1. Ο Αλγόριθμος Apriori

Ο αλγόριθμος Apriori χρησιμοποιεί συχνά στοιχεία για τη δημιουργία κανόνων συσχέτισης και έχει σχεδιαστεί για να λειτουργεί στις βάσεις δεδομένων που περιέχουν συναλλαγές. Με τη βοήθεια αυτών των κανόνων συσχέτισης, καθορίζει πόσο έντονα ή πόσο αδύναμα συνδέονται δύο αντικείμενα. Αυτός ο αλγόριθμος χρησιμοποιεί μια αναζήτηση και ένα δέντρο κατακερματισμού για τον αποτελεσματικό υπολογισμό των αντικειμένων συσχέτισης. Είναι η επαναληπτική διαδικασία για την εύρεση των συχνών στοιχείων από ένα μεγάλο σύνολο δεδομένων. Χρησιμοποιείται κυρίως για την ανάλυση καλαθιών αγοράς, αλλά μπορεί επίσης να χρησιμοποιηθεί στον τομέα της υγειονομικής περίθαλψης [9].

Για να αντιληφθούμε καλύτερα την έννοια του συχνού στοιχείου, θα υποθέσουμε ότι υπάρχουν δύο συναλλαγές A={1,2,3,4,5} και B={2,3,7}. Οι τιμές 2 και 3 επαναλαμβάνονται. Αυτές είναι τα συχνά στοιχεία. Επομένως, *συχνά στοιχεία (frequent itemsets)* ονομάζονται εκείνα τα στοιχεία των οποίων η υποστήριξη είναι

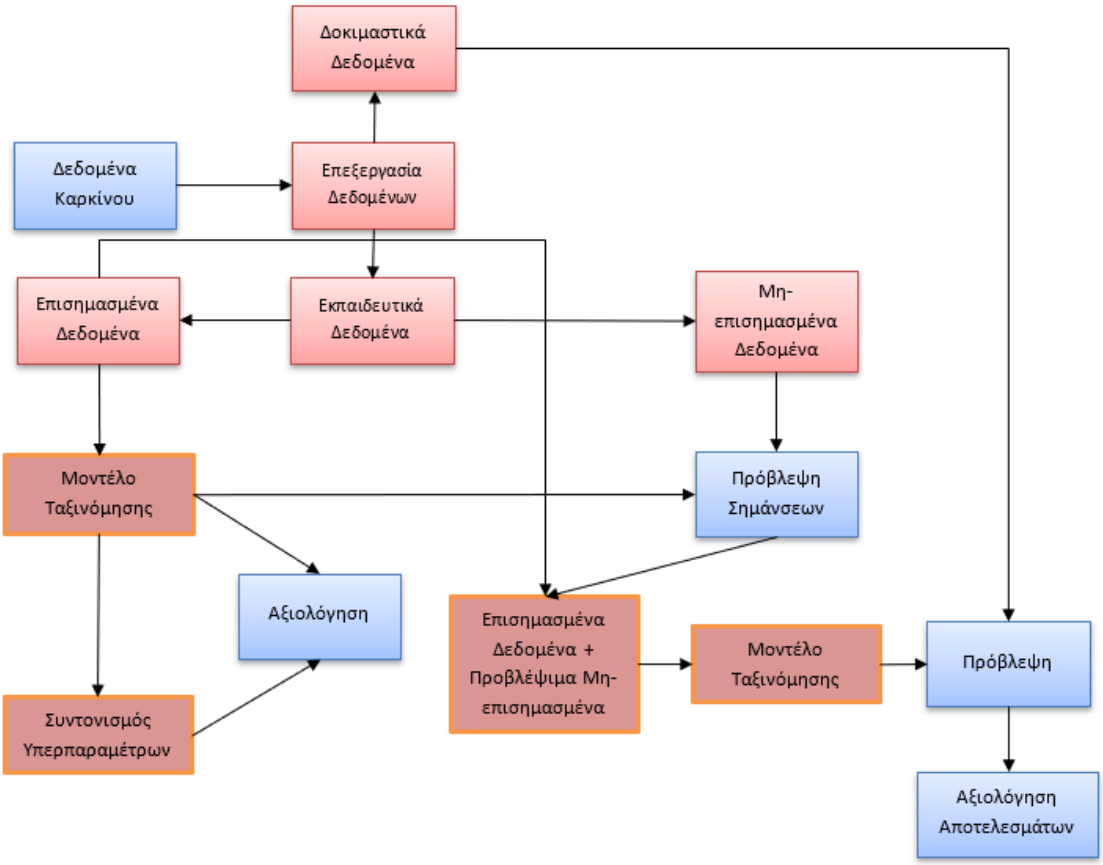
μεγαλύτερη από την οριακή τιμή ή την ελάχιστη υποστήριξη που καθορίζεται από το χρήστη [9].

4.3. Ημι-επιβλεπόμενη μάθηση

Η *Ημι-Επιβλεπόμενη μάθηση (Semi-supervised Learning – SSL)* θεωρείται ένας συνδυασμός της επιβλεπόμενης με την μη-επιβλεπόμενη μάθηση. Τέτοιοι αλγόριθμοι χρησιμοποιούνται όταν ένας τεράστιος όγκος δεδομένων χωρίς σήμανση εμποδίζει το μοντέλο να ενσωματώσει δεδομένα χωρίς σήμανση. Ο αλγόριθμος παρέχεται με δεδομένα χωρίς ετικέτα, μαζί με τις πληροφορίες επίβλεψης, σε μικρή ποσότητα. Η έξοδος που περιέχει η SSL είναι μεταβλητές-στόχοι που χρησιμοποιούνται για την εκπαίδευση και την πρόβλεψη των στόχων για τα δεδομένα χωρίς σήμανση [16].

Για την επίλυση προβλημάτων ημι-επιβλεπόμενης μάθησης, ο αλγόριθμος ακολουθεί τα εξής βήματα (Εικόνα *21*):

1. Σε πρώτη φάση τα δεδομένα κανονικοποιούνται προκειμένου να βρίσκονται όλες οι μεταβλητές στην ίδια κλίμακα και κατανομή. Κάποια από τα εκπαιδευτικά δεδομένα θα χρησιμοποιηθούν ως επισημασμένα δεδομένα, για να προσαρμόσουν τον αλγόριθμο ML, και τα υπόλοιπα ως δεδομένα χωρίς σήμανση.
2. Το κύριο πλεονέκτημα ενός αλγορίθμου ημι-επιβλεπόμενης μάθησης είναι να έχει δεδομένα χωρίς σήμανση και μικρότερο αριθμό δεδομένων με σήμανση. Τα δεδομένα εκπαίδευσης θα διαχωριστούν σε δεδομένα με σήμανση και σε δεδομένα χωρίς σήμανση. Στην πραγματικότητα, υπάρχει ένας τεράστιος αριθμός μη επισημασμένων δεδομένων, δεδομένου ότι τα επισημασμένα έχουν μεγαλύτερο κόστος και είναι χρονοβόρα. Επομένως, η προσέγγιση της ημι-επιβλεπόμενης μάθησης μπορεί να εφαρμοστεί χρησιμοποιώντας τα επισημασμένα δεδομένα, ενώ τα μη επισημασμένα τα εκμεταλλεύτεται για την εκπαίδευση του αλγορίθμου.
3. Το μοντέλο εκπαιδεύεται από το ποσοστό των δεδομένων που ορίσαμε ως δεδομένα εκπαίδευσης (χωρίς σήμανση) με χρήση αλγορίθμων ταξινόμησης.
4. Το μοντέλο αποκτά την υψηλότερη ακρίβεια αρχικοποιώντας τυχαία τις παραμέτρους, τις οποίες αλλάζει μέχρι να επιτευχθεί υψηλότερη ακρίβεια.
5. Οι ετικέτες για τα δεδομένα χωρίς σήμανση προβλέπονται και συνδυάζονται με δεδομένα με σήμανση. Αυτό δημιουργεί ένα μεγάλο σύνολο δεδομένων και στη συνέχεια το μοντέλο εκπαιδεύεται και πάλι με βελτιστοποίηση παραμέτρων.
6. Τέλος, τα δεδομένα δοκιμής προβλέπονται με χρήση του εκπαιδευμένου μοντέλου και τα αποτελέσματα που εξάγονται, αξιολογούνται [16].



Εικόνα 21.Διάγραμμα ροής Ημι-Επιβλεπόμενης Μάθησης.

4.4. Ενισχυμένη Μάθηση

Πριν μιλήσουμε για την *Ενισχυμένη Μάθηση (Reinforcement Learning - RL)* και τον τρόπο λειτουργίας της, είναι αναγκαίο να εξηγήσουμε την έννοια των παρακάτω όρων.

- **Πράκτορας:** Μια οντότητα που μπορεί να αντιληφθεί/εξερευνήσει το περιβάλλον στο οποίο βρίσκεται και να ενεργήσει πάνω σε αυτό.
- **Περιβάλλον:** Μια κατάσταση στην οποία ένας πράκτορας είναι παρών ή περιβάλλεται από αυτήν. Στην Ενισχυμένη Μάθηση, υποθέτουμε το στοχαστικό περιβάλλον, που σημαίνει ότι είναι τυχαίας φύσης.
- **Ενέργεια:** Οι κινήσεις που γίνονται από έναν πράκτορα μέσα στο περιβάλλον.
- **Κατάσταση:** Μια κατάσταση η οποία επιστρέφεται από το περιβάλλον μετά από κάθε ενέργεια που λαμβάνεται από τον πράκτορα.
- **Ανταμοιβή:** Μια ανατροφοδότηση που επέστρεψε στον πράκτορα από το περιβάλλον, προκειμένου να αξιολογήσει την ενέργεια του πράκτορα.
- **Πολιτική:** Στρατηγική που εφαρμόζεται από τον πράκτορα για την επόμενη ενέργεια με βάση την τρέχουσα κατάσταση.
- **Αξία:** Η αναμενόμενη μακροπρόθεσμη απόδοση με έκπτωση, σε αντίθεση με τη βραχυπρόθεσμη ανταμοιβή.

- **Τιμή-Q:** Είναι ως επί το πλείστον παρόμοια με την αξία, αλλά λαμβάνει ως πρόσθετη παράμετρο την τρέχουσα ενέργεια [27].

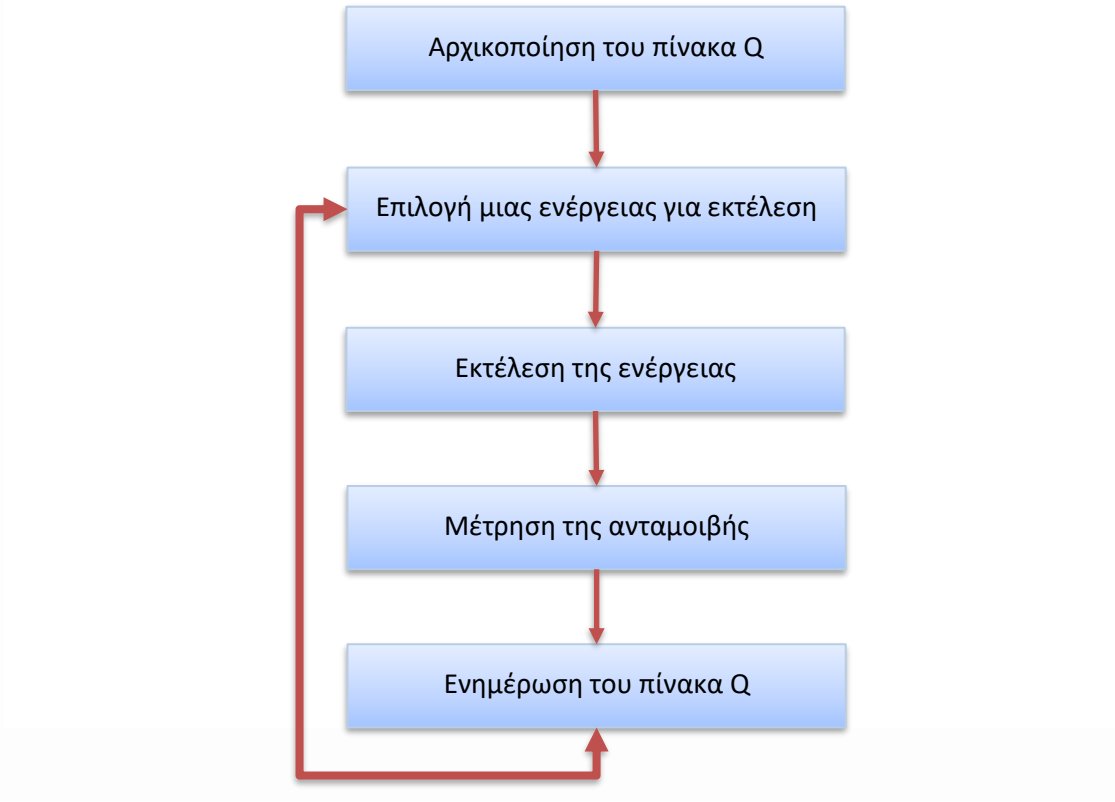
Η Ενισχυμένη Μάθηση είναι μια τεχνική μηχανικής μάθησης η οποία βασίζεται στην ανατροφοδότηση, κατά την οποία ένας πράκτορας μαθαίνει να συμπεριφέρεται σε ένα περιβάλλον εκτελώντας συγκεκριμένες ενέργειες και παρατηρώντας τα αποτελέσματα τους. Ο πράκτορας μαθαίνει αυτόματα, χρησιμοποιώντας σχόλια, χωρίς επισημασμένα δεδομένα. Επομένως, είναι υποχρεωμένος να μάθει μόνο από την εμπειρία του. Η ενισχυμένη μάθηση επιλύει ένα συγκεκριμένο είδος προβλημάτων όπου η λήψη αποφάσεων είναι διαδοχική και ο στόχος είναι μακροπρόθεσμος, όπως το παιχνίδι, η ρομποτική κ.λπ. [27].

Ο πράκτορας αλληλοεπιδρά με το περιβάλλον και το εξερευνά, με πρωταρχικό στόχο τη βελτίωση της απόδοσης, λαμβάνοντας τις μέγιστες θετικές ανταμοιβές. Μαθαίνει, με βάση την εμπειρία του, να εκτελεί την εργασία του με καλύτερο τρόπο. Επομένως, δεν προγραμματίζεται από πριν και δρα χωρίς καμία ανθρώπινη παρέμβαση. Ως εκ τούτου, μπορούμε να πούμε ότι *"Η ενισχυμένη μάθηση είναι ένας τύπος μεθόδου μηχανικής μάθησης όπου ένας ευφυής παράγοντας (πρόγραμμα υπολογιστή) αλληλοεπιδρά με το περιβάλλον και μαθαίνει να ενεργεί μέσα σε αυτό"*.Ένα παράδειγμα ενισχυμένης μάθησης είναι το πώς ένας ρομποτικός σκύλος μαθαίνει την κίνηση των ποδιών του [27].

Για να καταλάβουμε όμως καλύτερα τη μέθοδο της Ενισχυμένης Μάθησης θα υποθέσουμε ότι υπάρχει ένας παράγοντας τεχνητής νοημοσύνης μέσα σε ένα περιβάλλον λαβύρινθου και στόχος του είναι να βρει το διαμάντι. Ο πράκτορας αλληλοεπιδρά με το περιβάλλον, εκτελώντας ορισμένες ενέργειες, και με βάση αυτές, η κατάσταση του πράκτορα αλλάζει και λαμβάνει ανταμοιβή ή ποινή ως ανατροφοδότηση. Ο πράκτορας συνεχίζει να κάνει αυτά τα τρία πράγματα, δηλαδή να αναλαμβάνει δράση, να αλλάζει κατάσταση ή να παραμείνει στην ίδια και να λαμβάνει ανατροφοδότηση. Κάνοντας αυτές τις ενέργειες, μαθαίνει και εξερευνά το περιβάλλον στο οποίο βρίσκεται. Ο πράκτορας μαθαίνει ποιες ενέργειες οδηγούν σε θετικά σχόλια ή ανταμοιβές και ποιες ενέργειες οδηγούν σε αρνητική ποινή ανατροφοδότησης. Ως θετική ανταμοιβή, ο πράκτορας παίρνει ένα θετικό σημείο, και ως ποινή, παίρνει ένα αρνητικό σημείο [27].

4.4.1. Αλγόριθμοι Q-Learning και Βαθύ Νευρωνικό Δίκτυο Q

Ο Q-Learning είναι ένας δημοφιλής αλγόριθμος RL, βασίζεται στην εξίσωση Bellman και στόχος της είναι να μάθει την πολιτική που μπορεί να ενημερώσει τον πράκτορα σχετικά με το ποιες ενέργειες πρέπει να ληφθούν προκειμένου να μεγιστοποιηθεί η ανταμοιβή του. Δηλαδή, προσπαθεί να βρει την καλύτερη δράση που πρέπει να γίνει σε μια τρέχουσα κατάσταση και να μεγιστοποιήσει την αξία του Q, η οποία προκύπτει από την εξίσωση Bellman [27].



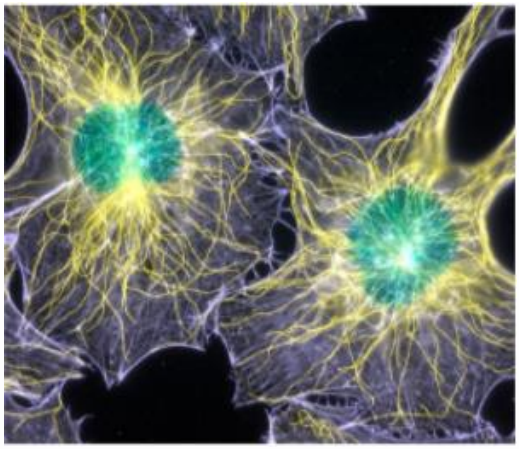
Εικόνα 22. Διάγραμμα Ροής Q-Learning

Ένας άλλος αλγόριθμος RL είναι το *Βαθύ Νευρωνικό Δίκτυο Q (Deep Q Neural Network)*, το οποίο είναι ένας Q-learning αλγόριθμος που χρησιμοποιεί νευρωνικά δίκτυα. Είναι αντιληπτό πως αν ο αλγόριθμος έχει να αντιμετωπίσει ένα περιβάλλον με μεγάλο χώρο κατάστασης, θα είναι μια δύσκολη και σύνθετη διεργασία ο ορισμός και η ενημέρωση του πίνακα Q. Προκειμένου να λυθεί ένα τέτοιο πρόβλημα, γίνεται χρήση του αλγορίθμου αυτού, όπου αντί να ορίσει έναν πίνακα Q, το νευρωνικό δίκτυο προσεγγίζει τις τιμές Q για κάθε ενέργεια και κατάσταση.

4.5. Τεχνητά Νευρωνικά Δίκτυα

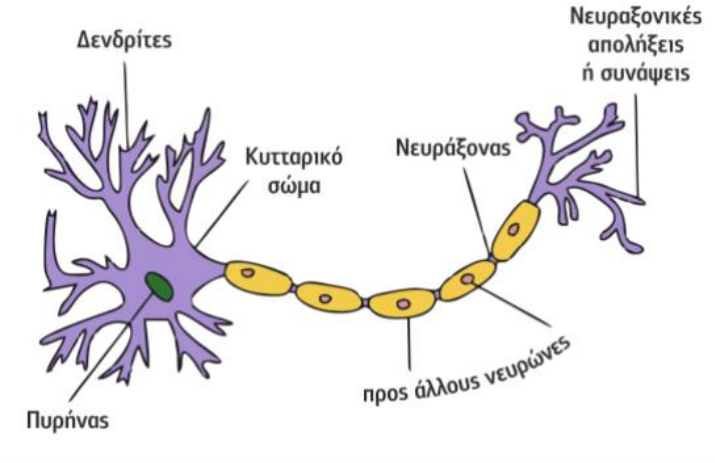
4.5.1. Νευρώνες

Νευρώνες είναι τα διακριτά στοιχεία από τα οποία αποτελείται ο ανθρώπινος εγκέφαλος, επικοινωνούν μεταξύ τους και υπολογίζεται ότι σε έναν εγκέφαλο περιέχονται περίπου 10 δις νευρώνες. Είναι τοποθετημένοι σε ομάδες και καθεμία από αυτές συνιστά ένα φυσικό νευρωνικό δίκτυο. Επομένως, ο ανθρώπινος εγκέφαλος περιέχει εκατοντάδες φυσικά νευρωνικά δίκτυα, καθένα από τα οποία περιέχει χιλιάδες διασυνδεδεμένους νευρώνες [28].



Εικόνα 23. Μικροσκοπική φωτογραφία φυσικών νευρώνων[28].

Ένας νευρώνας διαχωρίζεται από τα υπόλοιπα κύτταρα μέσω μίας μεμβράνης. Ο νευρώνας έχει την ικανότητα να επικοινωνεί με τους υπόλοιπους νευρώνες μέσω ηλεκτρικών σημάτων. Κάθε νευρώνας (Εικόνα *24*) αποτελείται από τους *δενδρίτες (dendrites)*, το *κυρίως κυτταρικό σώμα (cell body)*, τον *άξονα του κυττάρου-νευροάξονα (axon)* και τις *νευροαξονικές απολήξεις ή συνάψεις (synapse)* [28].



Εικόνα 24. Διάγραμμα νευρώνα[28].

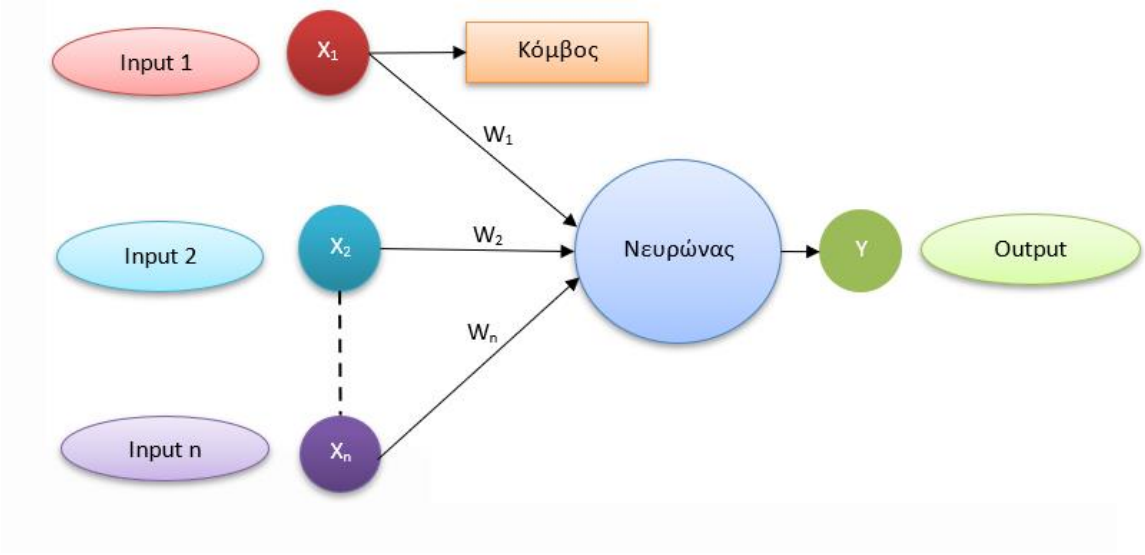
Όπως ήδη αναφέρθηκε, οι νευρώνες επικοινωνούν μεταξύ τους. Οι νευροαξονικές απολήξεις είναι τα σημεία εκείνα μέσω των οποίων ο άξονας του ενός νευρώνα μεταφέρει σήματα στους δενδρίτες των γειτονικών νευρώνων. Επίσης, ένας νευρώνας έχει τη δυνατότητα να λάβει σήματα από πολλαπλούς γειτονικούς νευρώνες από τους δενδρίτες, τα οποία επεξεργάζεται και τροφοδοτεί την έξοδό του μέσω του άξονα προς ένα άλλο σύνολο γειτονικών νευρώνων. Τα σήματα που φτάνουν «ζυγίζονται», τα αποτελέσματα αθροίζονται κι αν το αποτέλεσμα του αθροίσματος ξεπεράσει την τιμή κατωφλίου, ο νευρώνας δημιουργεί μια έξοδο, με τη μορφή ηλεκτρικού σήματος, στον άξονα του, η οποία στη συνέχεια θα μεταφερθεί στους γειτονικούς νευρώνες μέσω των συνάψεων [28].

Υπάρχουν δυο διακριτές καταστάσεις σημάτων, το *δυναμικό ηρεμίας* και το *δυναμικό ενέργειας*. Όταν ένας νευρώνας λαμβάνει σήματα, τα μεταβάλλει από τα

ηλεκτρικά χαρακτηριστικά των επαφών των συνάψεων, προκειμένου μερικά να εμποδίζονται και άλλα να διαδίδονται. Τα ηλεκτρικά χαρακτηριστικά των συνάψεων αποτελούν πληροφορίες για κάθε νευρώνα και, κατ' επέκταση, οι πληροφορίες που κρατούνται από ένα δίκτυο, κατανέμονται στους νευρώνες του. Το δυναμικό ενέργειας είναι εκείνο στο οποίο βασίζεται η μετάβαση πληροφορίας, το οποίο καθορίζεται από την οδό του εγκεφάλου μέσα από διακριτά επικοινωνούντες νευρώνες από τους οποίους περνάει το σήμα [28].

4.5.2. Τεχνητά Νευρωνικά Δίκτυα

Τα *Νευρωνικά Δίκτυα – ΝΔ (Neural Networks)*, γνωστά και ως *Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks – ANN)*, αποτελούν υποσύνολο της μηχανικής μάθησης και βρίσκονται στο επίκεντρο των αλγορίθμων βαθιάς μάθησης. Το όνομα και η δομή τους είναι εμπνευσμένα από τον ανθρώπινο εγκέφαλο, μιμούμενα τον τρόπο που οι βιολογικοί νευρώνες σηματοδοτούν ο ένας στον άλλο [29].



Εικόνα 25. Διάγραμμα απλού τεχνητού νευρώνα.

Στην *Εικόνα 25* βλέπουμε το διάγραμμα ενός τεχνητού νευρώνα, ο οποίος αντιστοιχίζεται πλήρως με έναν βιολογικό νευρώνα σύμφωνα με τον Πίνακας *2*.

Βιολογικό Νευρωνικό Δίκτυο	**Τεχνητό Νευρωνικό Δίκτυο**
Δενδρίτες	Είσοδοι
Κυτταρικός Πυρήνας	Κόμβοι
Σύναψη	Βάρη
Άξονας	Έξοδοι

Πίνακας 2. Αντιστοιχία Βιολογικού Νευρωνικού Δικτύου με Τεχνητό Νευρωνικό Δίκτυο.

Τα ANN επεξεργάζονται πληροφορίες ανταποκρινόμενα δυναμικά στις εισόδους τους. Κάθε τεχνητός νευρώνας αποτελείται από πολλές εισόδους x_i και μόνο μια έξοδο y. Κάθε είσοδος x_i «ζυγίζεται» με ένα βάρος w_i και τα αποτελέσματα αθροίζονται μέσω της συνάρτησης αθροίσματος:

$$F = \sum_{i}^{n} x_i w_i$$

Ο τεχνητός νευρώνας δίνει έξοδο μόνο όταν το ζυγισμένο άθροισμα των εισόδων είναι μεγαλύτερο μιας ορισμένης τιμής κατωφλίου θ, δηλαδή όταν:

$$\sum_{i}^{n} x_i w_i - \theta > 0$$

Ένα από τα απλούστερα ANN είναι ο στοιχειώδης Perceptron, ο οποίος είναι ένα ANN που αποτελείται μόνο από έναν τεχνητό νευρώνα και η έξοδος του, για ένα διάνυσμα εισόδου $x = (x_1, x_2, .., x_n)$ δίνεται μέσω της συνάρτησης μετάβασης:

$$\alpha = g\left(\sum_{i}^{n} x_i w_i\right)$$

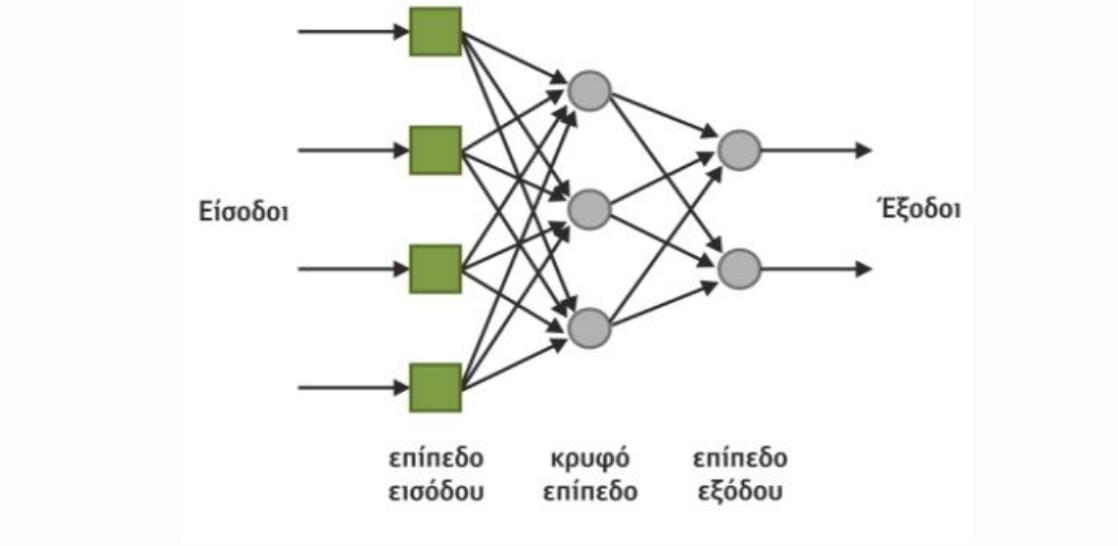
Συνοπτικά, τα ANN συνήθως οργανώνονται σε *επίπεδα (Layers)* τα οποία καλούνται και στρώματα. Τα ενδιάμεσα επίπεδα καλούνται *κρυμμένα επίπεδα (Hidden Layers)* και δεν είναι απαραίτητο να υπάρχουν. Τα επίπεδα αποτελούνται από έναν αριθμό *κόμβων (Nodes)* που είναι συνδεδεμένοι μεταξύ τους, έτσι ώστε ένας κόμβος να είναι συνδεδεμένος με πολλούς άλλους κόμβους του ίδιου ή άλλου επιπέδου. Οι κόμβοι επιδρούν σε άλλους κόμβους με το να τους διεγείρουν ή να αναστέλλουν την ενεργοποίησή τους. Για να επιτευχθεί αυτό, ο κόμβος λαμβάνει το σταθμισμένο άθροισμα όλων των εισόδων μέσω των συνδέσμων που καταλήγουν σε αυτόν. Εάν το άθροισα υπερβαίνει την τιμή κατωφλίου, τότε παράγει μία μοναδική έξοδο, μέσω της συνάρτησης μετάβασης. Τέλος, οι είσοδοι παρουσιάζονται στο δίκτυο μέσω του *επιπέδου εισόδου (Input Layer)* το οποίο επικοινωνεί με ένα ή περισσότερα κρυμμένα επίπεδα, τα οποία συνδέονται με το *επίπεδο εξόδου (Output Layer)* από το οποίο εξάγεται το αποτέλεσμα [28].

Κατά τη δημιουργία ενός ANN υπάρχουν κάποια βασικά στοιχεία της αρχιτεκτονικής του, τα οποία πρέπει να καθοριστούν:

- Το πλήθος των ενδιάμεσων κρυφών επιπέδων
- Το πλήθος των κόμβων ανά επίπεδο
- Ο τρόπος σύνδεσης των κόμβων μεταξύ τους
- Η τιμή κατωφλίου

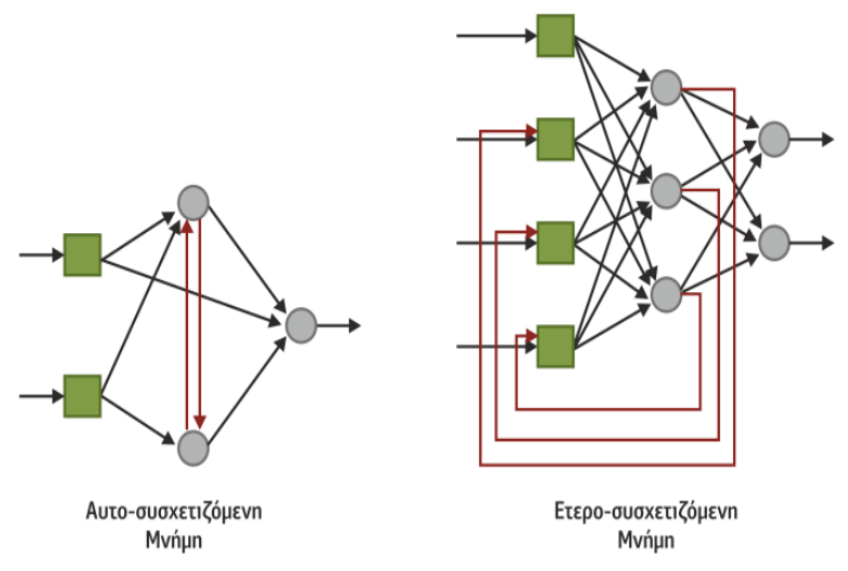
- Η συνάρτηση μετάβασης
- Οι τιμές των αρχικών βαρών μεταξύ των κόμβων
- Οι αλγόριθμοι που χρησιμοποιούνται κατά τη διαδικασία της εκπαίδευσης [28].

Οι κόμβοι μπορούν να συνδέονται μεταξύ τους με δύο τρόπους. Ο πρώτος τρόπος είναι τα *ANN πρόσθιας τροφοδότησης (Feed-Forward Neural Networks)*, όπου οι κόμβοι είναι οργανωμένοι σε διαφορετικά επίπεδα, ώστε οι κόμβοι του ενός επιπέδου να τροφοδοτούν τους κόμβους του επόμενου, μέχρις ότου να τροφοδοτηθούν και οι κόμβοι του τελευταίου επιπέδου (Εικόνα *26*). Δηλαδή, δεν υπάρχει έξοδος ενός επιπέδου που να αποτελεί είσοδο του ίδιου ή προηγούμενων επιπέδων. Τέτοια ΑΝΝ είναι τα δίκτυα *οπισθοδιάδοσης (backpropagation)* [28].



Εικόνα 26. ΑΝΝ πρόσθιας τροφοδότησης[28].

Αντίθετα, στα δίκτυα οπισθοδιάδοσης επιτρέπεται στους κόμβους ενός επιπέδου να τροφοδοτούν τους κόμβους του ιδίου ή προηγούμενων επιπέδων. Αν η ανατροφοδότηση αφορά κόμβους στο ίδιο επίπεδο, τότε τα δίκτυα καλούνται *αυτό-συσχετιζόμενες μνήμες (autoassociated memories)* διαφορετικά, καλούνται *ετερο-συσχετιζόμενες μνήμες (heteroassociated memories).* Στα οπισθίως τροφοδοτούμενα ΑΝΝ συνήθως δεν υπάρχουν πάνω από ένα κρυφά επίπεδα [28].



Εικόνα 27. Ανατροφοδοτούμενα ANN [28].

4.6. Βαθιά Μάθηση

Η *Βαθιά Μάθηση* βασίζεται στον κλάδο της μηχανικής μάθησης και μιμείται την λειτουργία του ανθρώπινου εγκεφάλου, όπως ακριβώς και τα νευρωνικά δίκτυα. Στη βαθιά μάθηση τίποτα δεν προγραμματίζεται ρητά αφού χρησιμοποιεί πολλές, μη γραμμικές, μονάδες επεξεργασίας, προκειμένου να εκτελέσει μετασχηματισμούς και να εξάγει αποτελέσματα. Η έξοδος από κάθε προηγούμενο επίπεδο λαμβάνεται ως είσοδος από κάθε ένα από τα διαδοχικά επίπεδα. Τα μοντέλα βαθιάς μάθησης είναι αρκετά ικανά να επικεντρωθούν στα ακριβή χαρακτηριστικά, απαιτώντας λίγη καθοδήγηση από τον προγραμματιστή, και δεν χρειάζεται να γίνει χρήση αναλυτικών μεθόδων εξαγωγής χαρακτηριστικών. Είναι πολύ βοηθητικά στην επίλυση του προβλήματος των διαστάσεων και χρησιμοποιούνται ειδικά όταν έχουμε έναν τεράστιο αριθμό εισόδων και εξόδων [30]

Η βαθιά μάθηση έχει εξελιχθεί από τη μηχανική μάθηση, η οποία από μόνη της είναι ένα υποσύνολο της τεχνητής νοημοσύνης, που έχει ως στόχο να μιμηθεί την ανθρώπινη συμπεριφορά. Το ίδιο και η βαθιά μάθηση, η οποία εφαρμόζεται με τη βοήθεια των Νευρωνικών Δικτύων. Η ιδέα πίσω από τα Νευρωνικά Δίκτυα είναι οι βιολογικοί νευρώνες, οι οποίοι δεν είναι παρά ένα κύτταρο του εγκεφάλου. Επομένως, η βαθιά μάθηση υλοποιείται με τη βοήθεια βαθιών δικτύων, τα οποία δεν είναι παρά νευρωνικά δίκτυα με πολλαπλά κρυμμένα στρώματα [30].

Υπάρχουν τρεις διαφορετικές αρχιτεκτονικές βαθιάς μάθησης:

- **Βαθιά Νευρωνικά Δίκτυα (Deep Neural Networks – DNN):** Είναι ένα νευρωνικό δίκτυο που ενσωματώνει μεγάλο πλήθος κρυφών στρώσεων μεταξύ των επιπέδων εισόδου και εξόδου και είναι ιδιαίτερα ικανά σε μη γραμμικούς συσχετισμούς μοντέλων και διεργασιών.

- **Δίκτυα Βαθιάς Πίστης (Deep Belief Networks – DBN):** Ένα τέτοιο δίκτυο είναι μια κατηγορία ενός ΒΝΔ που αποτελείται από πολυεπίπεδα δίκτυα πίστης. Ένα ΔΒΠ εκτελείται με τη βοήθεια του αλγορίθμου αντιπαρατιθέμενων αποκλίσεων, όπου ένα επίπεδο χαρακτηριστικών μαθαίνεται από αντιληπτές μονάδες. Στη συνέχεια, τα παλαιότερα εκπαιδευμένα χαρακτηριστικά αντιμετωπίζονται ως ορατές μονάδες, οι οποίες εκτελούν εκμάθηση χαρακτηριστικών. Τέλος, όταν ολοκληρωθεί η εκμάθηση του τελικού κρυμμένου στρώματος, εκπαιδεύεται ολόκληρο το ΔΒΠ.
- **Επαναλαμβανόμενα Νευρωνικά Δίκτυα (Recurrent Neural Networks – RNN):** Επιτρέπει παράλληλο και διαδοχικό υπολογισμό. Είναι ακριβώς παρόμοιο με αυτό του ανθρώπινου εγκεφάλου, δηλαδή ένα μεγάλο δίκτυο ανάδρασης συνδεδεμένων νευρώνων. Δεδομένου ότι είναι αρκετά ικανά να θυμούνται όλα τα δεδομένα που έχουν λάβει από την είσοδο, είναι και πιο ακριβή [30].

4.6.1. Νευρωνικό Δίκτυο Εμπρόσθιας Τροφοδότησης

Σε αυτό το είδος νευρωνικού δικτύου, όλοι οι νευρώνες είναι οργανωμένοι μέσα σε στρώματα, έτσι ώστε το επίπεδο εισόδου να παίρνει την είσοδο και το επίπεδο εξόδου να παράγει την έξοδο. Επειδή τα ενδιάμεσα στρώματα δεν συνδέονται με τον έξω κόσμο, ονομάζονται *κρυμμένα στρώματα*. Κάθε ένας από τους νευρώνες, που περιέχονται σε μία μόνο στρώση, συνδέεται με κάθε κόμβο στην επόμενη στρώση. Επομένως, όλοι οι κόμβοι είναι πλήρως συνδεδεμένοι. Δεν υπάρχει ορατή ή αόρατη σύνδεση μεταξύ των κόμβων στο ίδιο επίπεδο, ούτε βρόχοι επιστροφής προς τα πίσω. Για την ελαχιστοποίηση του σφάλματος πρόβλεψης, ο αλγόριθμος backpropagation μπορεί να χρησιμοποιηθεί προκειμένου να ενημερωθούν οι τιμές βάρους [30].

4.6.2. Επαναλαμβανόμενα Νευρωνικά Δίκτυα

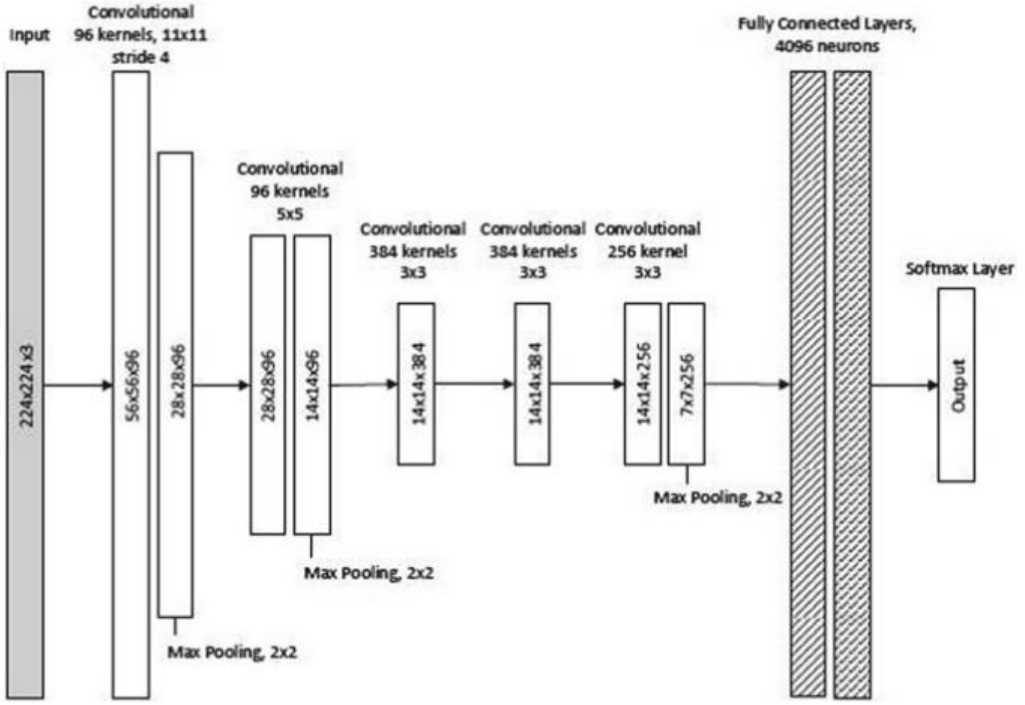
Τα Επαναλαμβανόμενα Νευρωνικά Δίκτυα είναι μια παραλλαγή των νευρωνικών δικτύων εμπρόσθιας τροφοδοσίας. Σε αυτού του τύπου νευρωνικά δίκτυα, κάθε νευρώνας, που υπάρχει στα κρυμμένα στρώματα, λαμβάνει μια είσοδο με μια συγκεκριμένη καθυστέρηση στο χρόνο. Το επαναλαμβανόμενο νευρωνικό δίκτυο αποκτά κυρίως πρόσβαση σε προηγούμενες πληροφορίες των υπαρχουσών επαναλήψεων. Για παράδειγμα, για να μαντέψετε την επόμενη λέξη σε οποιαδήποτε πρόταση, πρέπει να έχετε γνώσεις σχετικά με τις λέξεις που χρησιμοποιήθηκαν προηγουμένως. Όχι μόνο επεξεργάζεται τις εισόδους, αλλά μοιράζεται επίσης το μήκος και τα βάρη. Το μέγεθος του μοντέλου δεν αυξάνεται με την αύξηση του πλήθους των εισόδων. Ωστόσο, το μειονέκτημα ενός επαναλαμβανόμενου νευρωνικού δικτύου είναι ότι έχει αργή υπολογιστική

ταχύτητα και δεν εξετάζει καμία μελλοντική είσοδο για την τρέχουσα κατάσταση [30].

4.6.3. Συνελικτικά Νευρωνικά Δίκτυα

Τα Συνελικτικά Νευρωνικά Δίκτυα (CNN) είναι ένα είδος νευρωνικού δικτύου εμπρόσθιας τροφοδότησης, όπου το μοτίβο σύνδεσης μεταξύ του νευρώνα του εμπνέεται από τον οπτικό φλοιό. Δεν είναι παρά νευρωνικά δίκτυα, που μοιράζονται τις παραμέτρους τους. Χρησιμοποιούνται κυρίως για την ταξινόμηση - ομαδοποίηση εικόνων και την αναγνώριση αντικειμένων και επιτρέπουν τη μη επιβλεπόμενη κατασκευή ιεραρχικών αναπαραστάσεων εικόνων. Για να επιτευχθεί καλύτερη ακρίβεια, τα βαθιά συνελικτικά νευρωνικά δίκτυα προτιμώνται περισσότερο από οποιοδήποτε άλλο νευρωνικό δίκτυο[30].

- **AlexNet:** Είναι ένα από τα πιο περίπλοκα CNNs (Εικόνα *28*) που χρησιμοποιείται για αναγνώριση και ταξινόμηση. Αποτελείται από οκτώ επίπεδα. Τα πέντε πρώτα επίπεδα είναι περίπλοκα και τα τρία τελευταία είναι πλήρως συνδεδεμένα. Το πρώτο επίπεδο χρησιμοποιείται για το φιλτράρισμα της εισερχόμενης εικόνας με 96 πυρήνες με διασκελισμό τεσσάρων εικονοστοιχείων (pixels). Αυτό το επίπεδο λειτουργεί ως είσοδος στο δεύτερο επίπεδο, φιλτράροντάς έτσι την εισερχόμενη εικόνα με 256 πυρήνες. Το τρίτο, το τέταρτο και το πέμπτο επίπεδο συνδέονται χωρίς καμία παρέμβαση ή κανονικοποίηση [31].



Εικόνα 28. AlexNet [31].

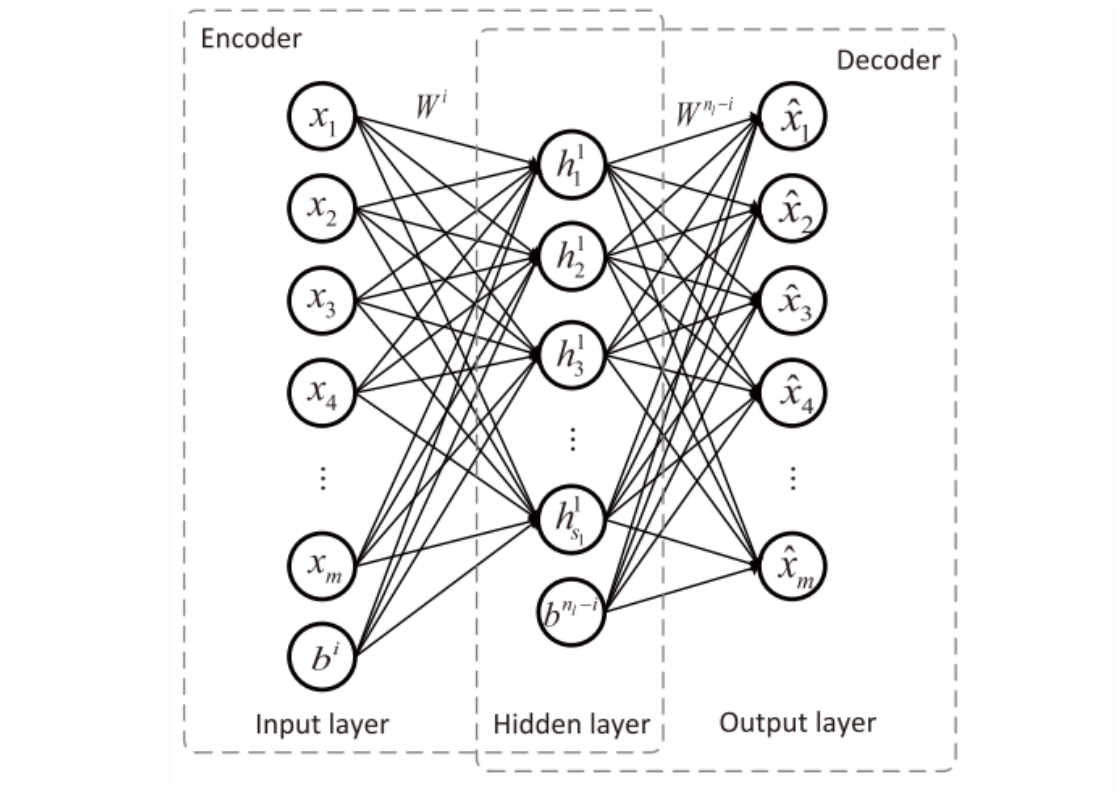
- **GoogLeNet:** Το συγκεκριμένο CNN εκτιμάει την τιμή του accuracy, βρίσκοντας την ιδανική τοπική δομή και κατασκευάζοντας ένα πολύ-επίπεδο δίκτυο. Το επίπεδο ομαδοποίησης τοποθετείται μεταξύ των λειτουργικών μονάδων με σκοπό την εξαγωγή χαρακτηριστικών από το σύνολο δεδομένων. Περιλαμβάνει επίσης τη χρήση ταξινομητών που εφαρμόζουν κανόνες στην εξαγωγή χαρακτηριστικών από τιμές [31].

4.6.4. Περιορισμένη Μηχανή Boltzmann

Στις Περιορισμένες Μηχανές Boltzmann (Restricted Boltzmann Machine – RBM) οι νευρώνες που υπάρχουν στο στρώμα εισόδου και το κρυφό στρώμα, περιλαμβάνουν συμμετρικές συνδέσεις ανάμεσά τους. Ωστόσο, στις RBM δεν υπάρχει εσωτερική συσχέτιση εντός του αντίστοιχου επιπέδου, σε αντίθεση με τις μηχανές Boltzmann οι οποίες περιλαμβάνουν εσωτερικές συνδέσεις μέσα στο κρυφό στρώμα. Αυτοί οι περιορισμοί στις μηχανές Boltzmann βοηθούν το μοντέλο να εκπαιδεύεται αποτελεσματικά [30].

4.6.5. Αυτόματοι Κωδικοποιητές

Ένα νευρωνικό δίκτυο *αυτόματου κωδικοποιητή* (Autoencoder – AE) είναι ένα άλλο είδος αλγόριθμου μη επιβλεπόμενης μηχανικής μάθησης. Εδώ ο αριθμός των κρυφών επιπέδων είναι λιγότερος από των κελιών εισόδου, αλλά ο αριθμός των εισόδων είναι ισοδύναμος με τον αριθμό των εξόδων. Ένα τέτοιο δίκτυο εκπαιδεύεται για να εμφανίσει την έξοδο παρόμοια με την τροφοδοτούμενη είσοδο, προκειμένου να αναγκάσει τους AEs να βρουν κοινά μοτίβα και να γενικεύσουν τα δεδομένα. Ο αλγόριθμος ενός AE αποτελείται από δύο μέρη, μια συνάρτηση κωδικοποιητή που μετατρέπει τις εισόδους σε αναπαραστάσεις χαμηλής διάστασης και μια συνάρτηση αποκωδικοποιητή που παράγει μια ανακατασκευή από τις αναπαραστάσεις. Οι αναπαραστάσεις στο κρυφό επίπεδο θεωρούνται ότι διατηρούν τις μέγιστες πληροφορίες και την κύρια διασπορά των μη επισημασμένων δεδομένων εισόδου. Οι AEs χρησιμοποιούνται κυρίως για τη μικρότερη αναπαράσταση της εισόδου και βοηθούν στην ανακατασκευή των αρχικών δεδομένων από συμπιεσμένα δεδομένα. Αυτός ο αλγόριθμος είναι συγκριτικά απλός, καθώς το μόνο που απαιτεί είναι η είσοδος να είναι ίδια με την έξοδο [30], [32].



Εικόνα 29. Βασική αρχιτεκτονική ενός αυτόματου κωδικοποιητή[32].

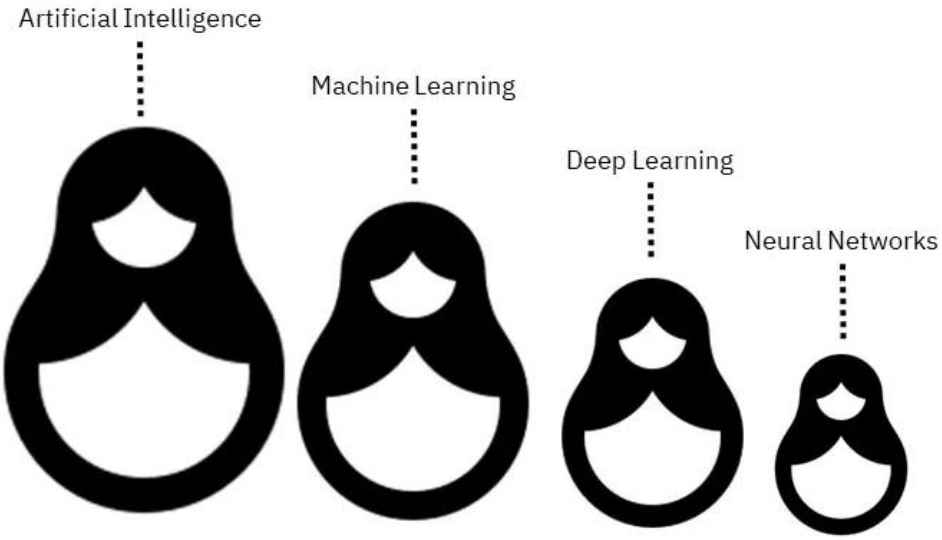
4.7. Διαφορές ΤΝ, ΜΜ, ΒΜ και ΝΔ

Είναι αναγκαίο να γίνει κατανοητό ότι η ΜΜ είναι ένας κλάδος της ΤΝ, η ΒΜ είναι ένας κλάδος της ΜΜ, ενώ, τέλος, τα ΝΔ αποτελούν τη ραχοκοκαλιά των αλγορίθμων ΒΜ (*Εικόνα 30*). Η ΤΝ είναι ο ευρύτερος όρος για να την περιγραφή των μηχανών που μιμούνται την ανθρώπινη νοημοσύνη, όπως η ομιλία, η λήψη αποφάσεων, η αναγνώριση προσώπων και η μετάφραση. Υπάρχουν πολλοί τρόποι επίτευξης της ΤΝ. Η ΜΜ είναι ο συνηθέστερος τρόπος επίτευξης ΤΝ σήμερα, ενώ η ΒΜ είναι ένας ειδικός τύπος ΜΜ [33],[34].

Στην πραγματικότητα, η διαφορά μεταξύ ΜΜ και ΒΜ βρίσκεται στον τρόπο μάθησης των αλγορίθμων που χρησιμοποιεί κάθε μία από αυτές. Η ΜΜ είναι εξαρτώμενη από την ανθρώπινη επέμβαση, ώστε το σύστημα να εκπαιδευτεί, απαιτώντας σύνολα δεδομένων, στα οποία επισημαίνεται η ιδιότητά τους, προκειμένου να γίνει κατανοητή η διαφορά μεταξύ των δεδομένων που εισάγονται. Για παράδειγμα, αν τροφοδοτήσουμε το σύστημα με ένα σύνολο φωτογραφιών με ζώα και ονομάσουμε κάθε φωτογραφία ανάλογα με το ζώο που απεικονίζει, δηλαδή «σκύλος», «γάτα» ή «κουνέλι», το μοντέλο θα εκπαιδευτεί βασιζόμενο στις ετικέτες των φωτογραφιών που το τροφοδοτήσαμε. Αυτή η διαδικασία είναι γνωστή και ως μάθηση με επίβλεψη ή εποπτευόμενη μάθηση. Αντίθετα, στους αλγόριθμους ΒΜ δε χρειάζεται να δηλωθούν τα σημαντικά χαρακτηριστικά των δεδομένων που εισήχθησαν, αφού έχουν τη δυνατότητα να ανακαλύψουν από μόνοι τους τα σημαντικά χαρακτηριστικά με χρήση ΝΔ. Επομένως, τα δεδομένα μας μπορεί να

είναι χωρίς δομή ή χωρίς σήμανση. Η ΒΜ έχει τη δυνατότητα να ενημερώσει τον αλγόριθμό της αξιοποιώντας τα δεδομένα της, είτε έχουν σήμανση είτε όχι. Η διαδικασία αυτή ονομάζεται μάθηση χωρίς επίβλεψη ή μη εποπτευόμενη μάθηση [33][34].

Η διαφορά μεταξύ της ΒΜ με τα ΝΔ είναι ο αριθμός των κρυφών επιπέδων που διακρίνει ένα ΝΔ από ένα αλγόριθμο ΒΜ. Η λέξη «βαθιά» στον όρο, αναφέρεται στο βάθος των στρωμάτων του ΝΔ. Αλγόριθμος ΒΜ μπορεί να θεωρηθεί ένα ΝΔ το οποίο αποτελείται από περισσότερα από τρία κρυφά επίπεδα. Τα περισσότερα ΝΔ ρέουν προς μία κατεύθυνση, συνήθως από την είσοδο προς την έξοδο. Ωστόσο, το μοντέλο μπορεί να εκπαιδευτεί να ρέει από την έξοδο στην είσοδο, το οποίο μας επιτρέπει να υπολογίσουμε το σφάλμα που σχετίζεται με κάθε νευρώνα και, επομένως, να προσαρμόσουμε τον αλγόριθμο κατάλληλα [33].



Εικόνα 30. Σχέση Μεταξύ ΤΝ, ΜΜ, ΒΜ και ΝΔ [33].

Κεφάλαιο 5 – Μηχανική Μάθηση Για Τον Καρκίνο Του Μαστού

Οι συμβατικές μέθοδοι παρακολούθησης και διάγνωσης του καρκίνου του μαστού βασίζονται στην ανίχνευση συγκεκριμένων χαρακτηριστικών από τον άνθρωπο. Λόγω της σοβαρότητας της νόσου, του μεγάλου πλήθους ασθενών και της ανάγκης για συνεχή παρακολούθηση τέτοιων περιστατικών, έχουν αναπτυχθεί αρκετά αυτοματοποιημένα υπολογιστικά συστήματα, τα οποία βοηθούν στον εντοπισμό του καρκίνου του μαστού *(Computer Aided Detection - CAD),* για την προσπάθεια επίλυσης αυτού του προβλήματος [35]. Τα συστήματα CAD έχουν την ικανότητα να ανιχνεύουν τον καρκίνο του μαστού σε πρώιμο στάδιο. Κατά συνέπεια, το ποσοστό επιβίωσης των ανθρώπων, που έχουν προσβληθεί από την ασθένεια, αυξάνεται επειδή η θεραπεία θα εφαρμοστεί σε πρώιμο στάδιο. Σε αυτό το κεφάλαιο θα μελετήσουμε την ικανότητα τέτοιων τεχνικών Μηχανικής Μάθησης για διάγνωση του καρκίνου του μαστού, με βάση πραγματικά δεδομένα ασθενών [19].

5.1. Βήματα Μοντέλου

Προκειμένου να μπορέσουν να εκπαιδευτούν, να δοκιμαστούν και να αξιολογηθούν τα διάφορα μοντέλα Μηχανικής Μάθησης που μελετήθηκαν, οι ερευνητές έκαναν χρήση διάφορων συνόλων δεδομένων, αναλόγως τις ανάγκες και το στόχο του εκάστοτε μοντέλου.

5.1.1. Σύνολα Δεδομένων

5.1.1.1. Σύνολο διαγνωστικών δεδομένων καρκίνου του μαστού της Ουισκόνσιν

Ένα από τα πιο γνωστά και ευρέως διαδεδομένα σύνολα δεδομένων που χρησιμοποιούνται για τις μελέτες των διάφορων μοντέλων για τη διάγνωση του καρκίνου του μαστού, είναι το *Σύνολο Διαγνωστικών Δεδομένων καρκίνου του μαστού της Wisconsin (Wisconsin Diagnostic Breast Cancer Dataset , WDBC)*. Αποτελείται από 569 δείγματα, εκ των οποίων τα 357 είναι καλοήθεις, ενώ τα 212 κακοήθεις [36].

Κάθε δείγμα του συνόλου δεδομένων λαμβάνεται με παρακέντηση μαστού, με χρήση μιας λεπτής βελόνας (Fine Needle Aspirate, FNA), τα χαρακτηριστικά του δείγματος περιγράφουν τα χαρακτηριστικά των πυρήνων των κυττάρων που υπάρχουν στην εικόνα και είναι 10 (Πίνακας *3*). Επίσης, το σύνολο δεδομένων αποτελείται από την ταυτότητα του ασθενούς και τη διάγνωση (Β=καλοήθης , Μ=κακοήθης). Η ταυτότητα είναι ο αριθμός αναγνώρισης του ασθενούς [36].

Χαρακτηριστικά WDBC	Περιγραφή
Ακτίνα (Radius)	Ο μέσος όρος των αποστάσεων από το κέντρο μέχρι τα σημεία της περιμέτρου
Υφή (Texture)	Τυπική απόκλιση των τιμών της κλίμακας του γκρι
Περίμετρος (Perimeter)	
Περιοχή (Area)	
Ομαλότητα (Smoothness)	Η τοπική μεταβολή στα μήκη της ακτίνας
Όγκος (Compactness)	$\frac{περίμετρος^2}{περιοχή} - 1{,}0,$
Κοιλότητα (Concavity)	Η τραχύτητα των περιγραμμάτων των κοίλων τμημάτων
Κοίλα Σημεία (Concave Points)	Το πλήθος των περιγραμμάτων των κοίλων Τμημάτων
Συμμετρία (Symmetry)	
Fractal	«Προσέγγιση ακτών» - 1

Πίνακας 3. Χαρακτηριστικά συνόλου δεδομένων WDBC[36].

5.1.1.2. Σύνολο δεδομένων καρκίνου του μαστού του Ουισκόνσιν (Πρωτότυπο)

Το πρωτότυπο σύνολο δεδομένων καρκίνου του μαστού της Ουισκόνσιν (Wisconsin Breast Cancer (original) Dataset, WBCD) περιέχει 699 περιπτώσεις. Από αυτές, οι 458 είναι καλοήθεις, ενώ οι 241 είναι κακοήθεις. Αποτελείται από 2 κλάσεις, με το 65,5% να είναι κακοήθεις και το 34,5% καλοήθεις, και 11 χαρακτηριστικά. Από αυτά τα 11 χαρακτηριστικά, το πρώτο είναι το αναγνωριστικό τους, τα επόμενα 9 χαρακτηριστικά είναι αυτά που βλέπουμε στον Πίνακας 4, ενώ το τελευταίο χαρακτηριστικό είναι η κλάση τους [37].

Χαρακτηριστικό	Περιγραφή
Clumb Thickness	Προσδιορισμός του πάχους του δείγματος
Uniformity of Cell Size	Αξιολόγηση της ομοιομορφίας του μεγέθους του κυττάρου
Uniformity of Cell Shape	Εκτίμηση της ισότητας των κυτταρικών σχημάτων και προσδιορισμός των οριακών διακυμάνσεων, επειδή τα καρκινικά κύτταρα τείνουν να ποικίλλουν σε σχήμα
Marginal Adhesion	Τα καρκινικά κύτταρα εξαπλώνονται σε όλο το όργανο και τα φυσιολογικά κύτταρα συνδέονται μεταξύ τους
Single Epithelial Cell Size	Μέτρηση της ομοιομορφίας, τα διευρυμένα επιθηλιακά κύτταρα είναι ένα σημάδι της κακοήθειας
Bare Nuclei	Σε καλοήθεις όγκους οι πυρήνες δεν περιβάλλονται από κυτταρόπλασμα
Bland Chromatin	Περιγραφή της υφής του πυρήνα, σε καλοήθη κύτταρα έχει ομοιόμορφο σχήμα. Η χρωματίνη τείνει να είναι πιο χονδροειδής σε όγκους

Normal Nucleoli	Στα φυσιολογικά κύτταρα, ο πυρήνας είναι συνήθως αόρατος και πολύ μικρός. Στα καρκινικά κύτταρα, υπάρχουν περισσότεροι από ένας πυρήνες και γίνεται πολύ πιο εμφανής
Mitoses	Εκτίμηση του αριθμού της μίτωσης. Όσο μεγαλύτερη είναι η τιμή, τόσο μεγαλύτερη είναι η πιθανότητα κακοήθειας

Πίνακας 4. Χαρακτηριστικά του WBCD.

5.1.1.3. Σύνολο διαγνωστικών δεδομένων καρκίνου του μαστού της Κοΐμπρα

Ένα άλλο σύνολο δεδομένων που χρησιμοποιείται στα πλαίσια της διάγνωσης του καρκίνου του μαστού, με χρήση τεχνικών ΜΜ, είναι το *Σύνολο Δεδομένων της Coimbra (Breast Cancer Coimbra Dataset – BCCD)*. Υπάρχουν 10 προγνωστικοί παράγοντες, οι 9 είναι ποσοτικοί, ο 10^{ος} είναι δυαδική εξαρτώμενη μεταβλητή και όλοι δείχνουν την παρουσία ή απουσία καρκίνου του μαστού [38].

Οι προγνωστικοί παράγοντες (Πίνακας *5*) είναι ανθρωπομετρικά δεδομένα και παράμετροι που μπορούν να συγκεντρωθούν από μία απλή ανάλυση αίματος ρουτίνας. Σύμφωνα με αυτά τα χαρακτηριστικά εισόδου, τα δεδομένα μπορούν να ταξινομηθούν σε καρκινικά ή μη. Τα κλινικά αυτά χαρακτηριστικά παρατηρήθηκαν ή μετρήθηκαν για 64 ασθενείς με καρκίνο του μαστού και 52 υγιείς ασθενείς. Τέλος, αυτό το σύνολο δεδομένων διαφέρει από άλλα, όσον αφορά τις δυνατότητες που περιέχει, αφού τα μοντέλα πρόβλεψης που βασίζονται σε αυτούς τους προγνωστικούς παράγοντες, εάν είναι ακριβή, μπορούν δυνητικά να χρησιμοποιηθούν ως βιοδείκτες του καρκίνου του μαστού [38].

	Χαρακτηριστικά BCCD	**Μονάδα Μέτρησης**
Ποσοτικά Χαρακτηριστικά	Ηλικία	Έτη
	Δείκτης Μάζας Σώματος (ΔΜΣ)	Kg/m^2
	Γλυκόζη	Mg/dL
	Ινσουλίνη	Mu/mL
	HOmeostasis Model Assessment (HOMA)	
	Λεπτίνη	ng/mL
	Αδιπονεκτίνη	μg/mL
	Ανθεκτικίνη	ng/mL
	Monocyte Chemoattractant Protein 1 (MCP1)	pg/dL
Δυαδική Μεταβλητή	1 = Υγιής 2 = Ασθενής	

Πίνακας 5. Χαρακτηριστικά Συνόλου Δεδομένων BCCD [38].

5.1.2. Προ-επεξεργασία Δεδομένων

Η προ-επεξεργασία των δεδομένων πραγματοποιείται για τη βελτίωση της ποιότητας του συνόλου δεδομένων, προκειμένου να εξαλειφθεί ο θόρυβος και να μην υπάρχουν ασυνέπειες στα δεδομένα. Υπάρχουν διάφορες διαδικασίες που εμπλέκονται στην προ-επεξεργασία δεδομένων, όπως ο καθαρισμός των δεδομένων, η επιλογή χαρακτηριστικών, η εξαγωγή χαρακτηριστικών κ.λπ. Στο στάδιο της προ-επεξεργασίας, τα δεδομένα χωρίζονται σε δύο κατηγορίες, στα δεδομένα εκπαίδευσης και στα δεδομένα δοκιμής. Στόχος είναι να μειωθούν/επιλεχθούν τα χαρακτηριστικά των δεδομένων, που θα επιτρέψουν στο μοντέλο να τα ταξινομήσει στη σωστή κατηγορία. Το σύνολο δεδομένων εκπαίδευσης χρησιμοποιείται στην εκπαίδευση του μοντέλου μηχανικής μάθησης, ενώ το σύνολο δεδομένων δοκιμής χρησιμοποιείται κατά τη διάρκεια του σταδίου πρόβλεψης [19].

5.1.2.1. Τεχνικές Επιλογής Χαρακτηριστικών

Η επιλογή χαρακτηριστικών περιλαμβάνει την επιλογή συνδυασμού χαρακτηριστικών που είναι σημαντικός για την ταξινόμηση στόχων και αγνοεί τα λιγότερο σημαντικά χαρακτηριστικά [19]. Είναι η διαδικασία αυτόματης ή χειροκίνητης επιλογής των χαρακτηριστικών προκειμένου να [39]:

•**Μειωθεί η υπερβολική τοποθέτηση:** η υπερβολική τοποθέτηση σημαίνει ότι το μοντέλο δεν γενικεύει καλά τα δεδομένα εκπαίδευσής λόγω ύπαρξης θορύβου στα δεδομένα. Το μοντέλο θα είναι καλά γενικευμένο κατά την αφαίρεση τέτοιων δεδομένων.

•**Βελτιωθεί το Accuracy:** η εκπαίδευση του μοντέλου με λιγότερο παραπλανητικά δεδομένα θα βελτιώσει το accuracy.

•**Μειωθεί ο χρόνος εκπαίδευσης:** Όσο μικρότερος είναι ο αριθμός των χαρακτηριστικών, τόσο λιγότερος χρόνος υπολογισμού απαιτείται για την εκπαίδευση.

• **Προσφερθούν πληροφορίες** στους βιολόγους σχετικά με το μηχανισμό μεταξύ των γονιδίων και των ασθενειών [39].

Η επιλογή χαρακτηριστικών μπορεί να ταξινομηθεί με βάση την ενσωμάτωση, μεταξύ του αλγορίθμου που επιλέχθηκε και του υλοποιημένου μοντέλου σε τέσσερις κύριες κατηγορίες. Η πρώτη είναι η προσέγγιση φίλτρου (Filter Approach), η δεύτερη είναι η προσέγγιση περιτυλίγματος (Wrapper Approach), η ενσωματωμένη προσέγγιση (Embedded Approach) είναι η τρίτη και, τέλος, η υβριδική προσέγγιση (Hybrid Approach) [39].

Μέθοδος Correlation based Feature Selection – CFS:

Τεχνική επιλογής χαρακτηριστικών που χρησιμοποιεί την προσέγγιση φίλτρου. Αυτή η τεχνική εξαρτάται από τα βάρη των χαρακτηριστικών γνωρισμάτων, τα οποία αξιολογούνται, εξετάζοντας μόνο τις ιδιότητες των δεδομένων. Συχνά, τα χαρακτηριστικά γνωρίσματα σε ένα σύνολο δεδομένων ίσως συσχετίζονται ιδιαίτερα μεταξύ τους. Αυτά τα χαρακτηριστικά που συσχετίζονται σε μεγάλο βαθμό με άλλα χαρακτηριστικά, παρέχουν περιττές πληροφορίες. Η συγκεκριμένη τεχνική λοιπόν, βρίσκει τη συσχέτιση μεταξύ των χαρακτηριστικών και τα χαρακτηριστικά που σχετίζονται σε μεγάλο βαθμό, εξαιρούνται από το CFS. Ομοίως, χαρακτηριστικά που είναι ιδιαίτερα αλληλένδετα με τη σήμανση της κλάσης, διατηρούνται και επιλέγονται. Σε αυτή τη μελέτη, το φίλτρο συσχέτισης που χρησιμοποιείται είναι 0,7 και τα χαρακτηριστικά με συσχέτιση μεγαλύτερη από αυτή, εξαιρούνται από το σύνολο δεδομένων εκπαίδευσης και επιλέγονται τα χαρακτηριστικά με χαμηλότερο μέσο όρο [19].

Μέθοδος Recursive Feature Elimination - RFE:

Μέθοδος επιλογής χαρακτηριστικών που χρησιμοποιεί την προσέγγιση περιτυλίγματος. Η RFE περιλαμβάνει τη δημιουργία ενός μοντέλου ΜΜ με όλα τα αρχικά χαρακτηριστικά του συνόλου δεδομένων, τα οποία κατατάσσονται ανάλογα με την ποσοτική σημασία τους για τη μείωση του σφάλματος μοντελοποίησης. Σε αυτή τη μελέτη, η RFE χρησιμοποιεί έναν αλγόριθμο Τυχαίου Δάσους, για να δοκιμάσει τους συνδυασμούς χαρακτηριστικών. Κάθε υποσύνολο βαθμολογείται για την ακρίβεια του και, τελικά, επιλέγονται τα υποσύνολα με τη μεγαλύτερη βαθμολογία [19].

5.1.2.2. Μέθοδοι Εξαγωγής Χαρακτηριστικών

Η εξαγωγή χαρακτηριστικών από την άλλη πλευρά, μειώνει τον αριθμό των διαστάσεων μετατρέποντας τα χαρακτηριστικά υψηλότερων διαστάσεων σε χαρακτηριστικά με λιγότερες διαστάσεις. Η μείωση των διαστάσεων είναι πολύ σημαντική στη διαδικασία ταξινόμησης. Ο κύριος στόχος της είναι η βελτίωση της απόδοσης και η εξασφάλιση της ταχύτερης πρόβλεψης. Όταν χρησιμοποιείται η συγκεκριμένη τεχνική, διευκολύνεται η απεικόνιση και η κατανόηση των δεδομένων. Επίσης, η εξαγωγή χαρακτηριστικών μειώνει τον απαιτούμενο χώρο αποθήκευσης και τους χρόνους εκπαίδευσης [19].

Μέθοδος Principal component analysis - PCA:

Είναι μία από τις τεχνικές εξαγωγής χαρακτηριστικών που χρησιμοποιείται στην εν λόγω μελέτη. Μετατρέπει το αρχικό σύνολο δεδομένων σε ένα άλλο σύνολο, με μειωμένο αριθμό παράγωγων μεταβλητών που δεν συσχετίζονται, και ονομάζονται

κύρια συστατικά. Η μέθοδος αυτή χρησιμοποιείται, στα πλαίσια της μελέτης, σε νευρωνικά δίκτυα και εφαρμόζεται στο σύνολο δεδομένων WDBC, προκειμένου να προσδιοριστεί ο συνδυασμός των κύριων στοιχείων που αντιπροσωπεύει τη μεγαλύτερη διακύμανση στα δεδομένα. Κατά την εκτέλεση της, η αθροιστική διακύμανση χρησιμοποιείται, κατά κανόνα, στη μείωση της διάστασης του χαρακτηριστικού του συνόλου δεδομένων [19].

Μέθοδος Linear Discriminant Analysis – LDA:

Πρόκειται για μία άλλη μέθοδο εξαγωγής χαρακτηριστικών που χρησιμοποιείται προκειμένου να μειωθούν, με επίβλεψη, οι υψηλές διαστάσεις των δεδομένων. Η LDA υπολογίζει το μετασχηματισμό, μεγιστοποιώντας το, μεταξύ της διασποράς κλάσης και, ταυτόχρονα, ελαχιστοποιεί τη διασπορά εντός της κλάσης, προκειμένου να επιτευχθούν οι διακρίσεις υψηλότερης κατηγορίας. Η LDA προσδιορίζει χαρακτηριστικά που αντιπροσωπεύουν τη μεγαλύτερη διακύμανση μεταξύ κλάσεων και παράγει καλύτερα αποτελέσματα σε σχέση με την PCA [19].

5.1.3. Cross Validation

Η Cross-Validation είναι μια στατιστική προσέγγιση, η οποία διαχωρίζει το σύνολο δεδομένων σε δεδομένα εκπαίδευσης και δοκιμής, προκειμένου να ελεγχθεί και να αξιολογηθεί η απόδοση ενός μοντέλου ΜΜ. Τα σύνολα εκπαίδευσης και δοκιμών χωρίζονται τυχαία. Η K Fold Cross Validation είναι η βασική μορφή. Στις ερευνητικές εργασίες που διεξάγονται, χρησιμοποιείται συνήθως 10 Fold Cross Validation. Όλα τα δεδομένα χωρίζονται σε 10 ίσα μέρη, όπου 9 μέρη εφαρμόζονται για εκπαίδευση και 1 μέρος για τη δοκιμή του μοντέλου. Η διαδικασία επαναλαμβάνεται 10 φορές και κάθε φορά, κάθε ένα από τα 10 επιμέρους δείγματα χρησιμοποιείται τουλάχιστον μία φορά για τη δοκιμή του μοντέλου [40],[41].

5.1.4. Αξιολόγηση Μοντέλου

Η αξιολόγηση του μοντέλου γίνεται συγκρίνοντας τα αποτελέσματα που εξήχθησαν με τα πραγματικά δεδομένα. Σε αυτό το σημείο είμαστε στη φάση της πρόβλεψης, κατά την οποία το σύνολο δεδομένων δοκιμής χρησιμοποιείται για την αξιολόγηση της απόδοσης των μοντέλων στην ταξινόμηση. Τα δεδομένα χρησιμοποιούνται για τον υπολογισμό των επιδόσεων του ταξινομητή υπολογίζοντας διάφορες τιμές [19].

- *Accuracy* είναι ο λόγος του συνολικού αριθμού ορθών προβλέψεων προς τον συνολικό αριθμό δειγμάτων [32] :

$$Accuracy = \frac{(E\Theta) + (EA)}{Συνολικός\ Πληθυσμός}$$

- Η *Kappa* είναι ακριβώς όπως το accuracy της ταξινόμησης και λαμβάνει υπόψη το αναμενόμενο ποσοστό σφάλματος.
- *Ανάκληση (Recall)* ή *Ευαισθησία (Sensitivity)* είναι το ποσοστό των θετικών περιπτώσεων που προσδιορίστηκαν σωστά. Είναι η ικανότητα ενός μοντέλου να ανιχνεύει μία κατάσταση , όταν αυτή υπάρχει και δίνεται από τον τύπο [32], [42] :

$$Recall = Sensitivity = \frac{(E\Theta)}{(E\Theta) + (\Lambda A)}$$

- Η *Ειδικότητα (Specificity)* αντιπροσωπεύει το πραγματικό αρνητικό ποσοστό. Είναι η ικανότητα ενός μοντέλου να αποκλείει την ύπαρξη μιας κατάστασης, όταν αυτή δεν υπάρχει, ονομάζεται *Ειδικότητα (Specificity)* και δίνεται από τον τύπο [42] :

$$Specificity = \frac{(EA)}{(EA) + (\Lambda\Theta)}$$

- R^2 είναι ο συντελεστής προσδιορισμού και εκφράζει πόσο ακριβή παρατηρούμενα αποτελέσματα αναπαράγονται από το μοντέλο, με τιμές από 0 μέχρι 1.

- Η *Θετική Προγνωστική Τιμή (Positive Predictive Value – PPV)* ή *Precision* δείχνει το ποσοστό των θετικών περιπτώσεων που είχαν προβλεφθεί [32], [42] :

$$Precision = PPV = \frac{(E\Theta)}{(E\Theta) + (\Lambda\Theta)}$$

- Η *Αρνητική Προγνωστική Τιμή (Negative Predictive Value – NPV)* [42]*:*

$$NPV = \frac{(EA)}{(EA) + (\Lambda A)}$$

- Ο *Θετικός Λόγος Πιθανότητας (Positive Likelihood Ratio – LR+)* λαμβάνει τιμές μεγαλύτερες από το μηδέν. Η χειρότερη περίπτωση είναι μια μηδενική τιμή και σχετίζεται με μια δοκιμή με μηδενική ευαισθησία. Οι μεγαλύτερες τιμές του κριτηρίου του θετικού λόγου πιθανότητας δείχνουν ότι σε μια δοκιμή περιλαμβάνονται πιο πολύτιμες πληροφορίες [42] :

$$LR+ = \frac{Sensitivity}{1 - Specificity}$$

- Ο *Αρνητικός Λόγος Πιθανότητας (Negative Likelihood Ratio – LR-)* λαμβάνει τιμές μεγαλύτερες από το μηδέν. Όσο μικρότερη είναι η αναλογία αρνητικής

πιθανότητας, τόσο μεγαλύτερες είναι οι πληροφορίες που μπορούν να παρασχεθούν από μια δοκιμή [42] :

$$LR- = \frac{1 - Sensitivity}{Specificity}$$

- Η *Βαθμολογία F1* είναι γνωστή ως ο αρμονικός μέσος όρος της ανάκλησης και του Precision και δίνεται από τον τύπο:

$$F1\ Score = \frac{2 \times (Presicion \times Recall)}{Presicion + Recall}$$

5.2. Μέθοδος Μηχανών Διανυσμάτων Υποστήριξης (SVM)

Η απόδοση τεσσάρων ταξινομητών, των SVM, NB, C4.5 και K-NN, συγκρίθηκε στα πλαίσια μιας έρευνας που πραγματοποιήθηκε από τον Hiba Asri και την ερευνητική του ομάδα στο Μορόκκο το 2016. Στόχος ήταν να αξιολογηθεί η αποδοτικότητα και η αποτελεσματικότητα αυτών των αλγορίθμων όσον αφορά το accuracy, την ευαισθησία, την ειδικότητα και το precision. Το πρωτότυπο σύνολο διαγνωστικών δεδομένων καρκίνου του μαστού του Ουισκόνσιν (WBCD) χρησιμοποιήθηκε για την απόδοση των τεσσάρων αυτών ταξινομητών [43].

Για την εφαρμογή και αξιολόγηση των ταξινομητών, εφαρμόστηκε η μέθοδος 10-Fold Cross Validation. Τα μοντέλα αξιολογήθηκαν για την αποτελεσματικότητα τους από το χρόνο που χρειάστηκε για την κατασκευή του μοντέλου (ΧΚΜ), τις σωστά ταξινομημένες περιπτώσεις (ΣΤΠ), τις εσφαλμένα ταξινομημένες περιπτώσεις (ΕΤΠ) και το accuracy. Τα αποτελέσματα παρουσιάζονται στον *Πίνακας 6*.

Κριτήρια Αξιολόγησης	**Ταξινομητές**			
	C4.5	SVM	NBC	K-NN
ΧΚΜ	0.06	0.07	0.05	0.01
ΣΤΠ	665	678	671	666
ΛΤΠ	34	21	28	33
Accuracy (%)	95.13	97.13	95.99	95.27

Πίνακας 6. Απόδοση ταξινομητών της μελέτης του H. Asri [43].

Προκειμένου να μετρηθεί καλύτερα η απόδοση των ταξινομητών, εξετάστηκε και το σφάλμα προσομοίωσης αξιολογώντας την αποτελεσματικότητα του κάθε ταξινομητή με βάση :

- Τα Στατιστικά στοιχεία kappa (Kappa Statistic, KS) ως πιθανό διορθωμένο μέτρο συμφωνίας μεταξύ των ταξινομήσεων και των πραγματικών κλάσεων,
- Το Μέσο Απόλυτο Σφάλμα (Mean Absolute Error, MAE) ως προς το πόσο κοντά είναι οι προβλέψεις στα τελικά αποτελέσματα,
- Ριζα του μέσου τετραγωνικού σφάλματος (Root Mean Squared Error, RMSE),
- Σχετικό απόλυτο σφάλμα (Relative Absolute Error, RAE),
- Ρίζα σχετικού τετραγωνικού σφάλαμτος (Root Relative Squared Error, RRSE).

Τα αποτελέσματα παρουσιάζονται στον (Πίνακας 7) .

Κριτήρια Αξιολόγησης	**Ταξινομητές**			
	C4.5	**SVM**	**NBC**	**K-NN**
KS	0.89	0.93	0.91	0.89
MAE	0.06	0.02	0.03	0.04
RMSE	0.21	0.16	0.19	0.21
RAE (%)	14	6.33	8.59	10.46
RRSE (%)	45	35.58	40.95	44.77

Πίνακας 7. Σφάλμα εκπαίδευσης και προσομοίωσης[43].

Για να συγκριθεί η αποτελεσματικότητα των μοντέλων, συγκρίθηκαν πρώτα οι τιμές του precision, του recall, το TPR και το FPR. Τα αποτελέσματα φαίνονται στον πίνακα *Πίνακας 8*.

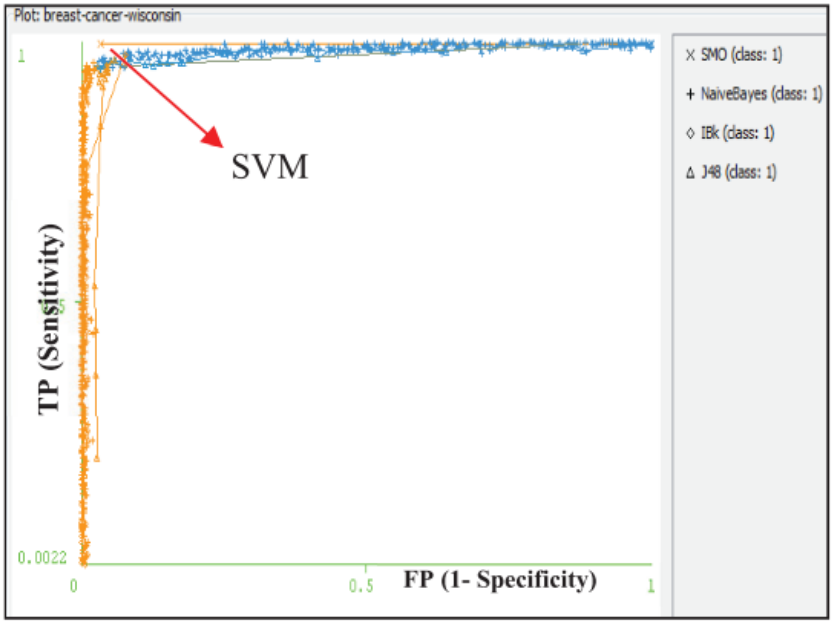
	TPR	**FPR**	**Precision**	**Recall**	**F-Measure**	**Class**
C4.5	0.95	0.05	0.96	0.95	0.96	Καλοήθης
	0.94	0.04	0.94	0.94	0.93	Κακοήθης
SVM	0.97	0.03	0.98	0.97	0.97	Καλοήθης
	0.96	0.02	0.95	0.96	0.95	Κακοήθης
NBC	0.95	0.02	0.98	0.95	0.96	Καλοήθης
	0.97	0.04	0.91	0.97	0.94	Κακοήθης
K-NN	0.97	0.08	0.95	0.97	0.96	Καλοήθης
	0.91	0.02	0.94	0.91	0.93	Κακοήθης

Πίνακας 8. Συγκριτικός πίνακας των τιμών Accuracy[43].

Η καμπύλη ROC στην *Εικόνα 31* δίνει το γράφημα της απόδοσης των τεσσάρων διαφορετικών ταξινομητών που μελετήθηκαν στη συγκεκριμένη μελέτη. Τέλος, ο *Πίνακας 9* είναι η Μήτρα Σύγχυσης που προέκυψε.

	TPR	FPR	Class
C4.5	438	20	Καλοήθης
	14	227	Κακοήθης
SVM	446	12	Καλοήθης
	9	232	Κακοήθης
NBC	436	22	Καλοήθης
	6	235	Κακοήθης
K-NN	445	13	Καλοήθης
	20	221	Κακοήθης

Πίνακας 9. Μήτρα Σύγχυσης μελέτης H. Asri [43].



Εικόνα 31
. Καμπύλη ROC της μελέτης του H. Asri [43].

Από τον Πίνακας *6* φαίνεται πως ο SVM χρειάζεται περίπου 0,07s για να κατασκευάσει το μοντέλο σε αντίθεση με τον K-NN που διαρκεί μόλις 0,01s. Αυτό μπορεί να εξηγηθεί από το γεγονός ότι ο K-NN είναι ένας "lazy learner" και δεν κάνει πολλά κατά τη διάρκεια της διαδικασίας εκπαίδευσης, σε αντίθεση με άλλους ταξινομητές. Από την άλλη πλευρά, ο SVM λαμβάνει το καλύτερο accuracy (97.13%) από όλους τους υπόλοιπους ταξινομητές. Επίσης, ο SVM έχει την υψηλότερη τιμή σωστά ταξινομημένων περιπτώσεων και τη χαμηλότερη τιμή στις λανθασμένες ταξινομημένες περιπτώσεις. Επίσης, από τον Πίνακας *7*, φαίνεται και πάλι πως ο SVM έχει τα καλύτερα αποτελέσματα, όσον αφορά την πιθανότητα καλύτερης ταξινόμησης (0.93 %), το μικρότερο Μέσο Απόλυτο Σφάλμα (0.02) και την καλύτερη συμβατότητα μεταξύ της αξιοπιστίας των δεδομένων που συλλέγονται και της εγκυρότητάς τους. Ενώ οι C4.5 και K-NN έχουν την υψηλότερη τιμή ποσοστού σφάλματος, το οποίο εξηγεί τον αυξημένο αριθμό εσφαλμένων ταξινομημένων

περιπτώσεων για κάθε αλγόριθμο (34 για τον C4.5 και 33 για τον K-NN, βλ. Πίνακας *6*). Από τα αποτελέσματα του Πίνακας *8*, της καμπύλης ROC και της μήτρας σύγχυσης και πάλι φαίνεται πως η ακρίβεια του SVM είναι καλύτερη από τις υπόλοιπες τεχνικές ταξινόμησης που χρησιμοποιούνται. Επομένως, μπορούμε να καταλάβουμε γιατί ο SVM έχει ξεπεράσει τους άλλους ταξινομητές [43].

5.3. Μέθοδος SVM-Linear Discriminant Analysis (SVM-LDA)

5.3.1. Διάγνωση

Σε μία έρευνα που πραγματοποιήθηκε το 2019, οι Omondiagbe, Veeramani και Sidhu προσπάθησαν να αποδείξουν ότι τα μοντέλα των SVMs, ΝΔ και ο ταξινομητής Naive Bayes (Naïve Bayes Classifier – NBC), μπορούν να επιλύσουν αποτελεσματικά προβλήματα ταξινόμησης προτύπων και να επιλέξουν έναν κατάλληλο αλγόριθμο ΜΜ για την κατασκευή ενός ολοκληρωμένου ευφυούς μοντέλου για την διάγνωση του καρκίνου του μαστού [19]. Η μέθοδος που πρότειναν ακολουθεί την έννοια των τεχνικών ταξινόμησης. Επομένως, το πρώτο βήμα της μεθοδολογίας αφορά την κατασκευή του μοντέλου ταξινόμησης, μαθαίνοντας από ένα σύνολο επισημασμένων δεδομένων εκπαίδευσης [19]. Το δεύτερο βήμα είναι το στάδιο της πρόβλεψης, όπου το σύνολο δεδομένων δοκιμής χρησιμοποιείται για την αξιολόγηση της ακρίβειας ταξινόμησης του μοντέλου που έχει κατασκευαστεί προηγουμένως [19].

Τα δεδομένα που χρησιμοποιήθηκαν σε αυτή τη μελέτη είναι το σύνολο δεδομένων WDBC. Στο στάδιο της προ-επεξεργασίας, τα δεδομένα χωρίστηκαν σε δύο κατηγορίες, στα δεδομένα εκπαίδευσης και στα δεδομένα δοκιμής. Το σύνολο δεδομένων εκπαίδευσης (399 παρατηρήσεις ~ 70%) χρησιμοποιείται στην εκπαίδευση του μοντέλου, ενώ το σύνολο δεδομένων δοκιμής (170 παρατηρήσεις ~ 30%) χρησιμοποιείται στο στάδιο της πρόβλεψης. Οι τεχνικές επιλογής χαρακτηριστικών που χρησιμοποιήθηκαν είναι οι μέθοδοι CFS και RFE και οι μέθοδοι εξαγωγής χαρακτηριστικών που χρησιμοποιήθηκαν είναι PCA και LDA [19].

Μετά την προ-επεξεργασία των δεδομένων, είναι το στάδιο της εφαρμογής τεχνικών ταξινόμησης ΜΜ στα επεξεργασμένα δεδομένα. Κατά τη διάρκεια αυτού του σταδίου, τα επεξεργασμένα δεδομένα θα χρησιμοποιηθούν για την εκπαίδευση και την κατασκευή του μοντέλου ΜΜ. Τα δεδομένα εκπαίδευσης περιέχουν όλα τα χαρακτηριστικά γνωρίσματα έτσι ώστε να μπορούν να ταξινομήσουν τα δεδομένα σε καλοήθεις και κακοήθεις όγκους. Επίσης, οι τεχνικές επιλογής χαρακτηριστικών και εξαγωγής χαρακτηριστικών που χρησιμοποιήθηκαν μειώνουν τις διαστάσεις των δεδομένων. Τα δεδομένα με μειωμένα χαρακτηριστικά χρησιμοποιούνται, επίσης, για την εκπαίδευση των μοντέλων [19].

Ο υπολογισμός των επιδόσεων του ταξινομητή έγινε υπολογίζοντας τις τιμές του precision, την περιοχή κάτω από την καμπύλη ROC, την ανάκλαση, την ευαισθησία,

την ειδικότητα, το kappa και το accuracy. Στον Πίνακας 10 εμφανίζονται τα μοντέλα που κατασκευάστηκαν και τα αντίστοιχα αποτελέσματά τους στις διάφορες μετρήσεις απόδοσης. [19].

Μοντέλα ΜΜ	Accuracy	Περιοχή κάτω από την καμπύλη ROC	Precision	Ανάκληση	Ευαισθησία	Ειδικότητα	Kappa
SVM	0.9647	0.9964	0.9385	0.9682	0.9682	0.9626	0.9248
SVM-CFS	0.9647	0.9954	0.9524	0.9524	0.9524	0.972	0.9243
SVM-RFE	0.9647	0.9976	0.9524	0.9524	0.9524	0.972	0.9243
SVM-LDA	0.9882	0.9994	0.9841	0.9841	0.9841	0.9907	0.9748
ΝΔ	0.9706	0.9985	0.9833	0.9365	0.9365	0.9907	0.9363
ΝΔ-CFS	0.9706	0.9973	0.9531	0.9683	0.9683	0.9719	0.9372
ΝΔ-RFE	0.9824	0.9989	0.9839	0.9683	0.9683	0.9907	0.962
ΝΔ-PCA	0.9765	0.9937	0.9836	0.9524	0.9524	0.9907	0.9492
ΝΔ-LDA	0.9882	0.9994	0.9841	0.9841	0.9841	0.9907	0.9748
NBC	0.9118	0.9860	0.8750	0.8889	0.8889	0.9252	0.8115
NBC-CFS	0.9176	0.9799	0.9152	0.8571	0.8571	0.9533	0.8211
NBC-RFE	0.9118	0.9860	0.8750	0.8889	0.8889	0.9352	0.8115
NBC-LDA	0.9824	0.9994	0.9839	0.9683	0.9683	0.9907	0.962

Πίνακας 10. Αποτελέσματα Μοντέλων Μηχανικής Μάθησης[19].

Από τα αποτελέσματα που εμφανίζονται στον Πίνακας 10, παρατηρείται ότι οι συνδυασμοί SVM-LDA, ΝΔ-LDA, ΝΔ-RFE, ΝΔ-PCA, NBC-LDA και το ΝΔ έλαβαν την καλύτερη απόδοση στην ειδικότητα (99.07%). Ωστόσο, οι συνδυασμοί SVM-LDA και ΝΔ-LDA απέκτησαν την καλύτερη απόδοση στο accuracy (98.82%), στο precision (98.41%), την ανάκληση (98.41%) και την ευαισθησία (98,41%). Το μοντέλο ΜΜ με τις χαμηλότερες επιδόσεις στην ευαισθησία (85,71%) και την ανάκληση (85,71%) ήταν ο συνδυασμός NBC-CFS, ενώ οι χαμηλότερες επιδόσεις στο precision (87,50%), την ειδικότητα (92,52% και 93,52%) και το accuracy (91,18%), παρατηρήθηκαν στον NBC και στον NBC-RFE [19].

Η μέθοδος μείωσης των διαστάσεων φαίνεται να επηρεάζει την απόδοση του αλγορίθμου ΜΜ. Τα αποτελέσματα της προσομοίωσης δείχνουν ότι οι μέθοδοι CFS και RFE αυξάνουν το precision και την ειδικότητα της SVM, ενώ το LDA αυξάνει τόσο το accuracy όσο και την ευαισθησία του (ανίχνευση κακοήθων περιπτώσεων). Στην περίπτωση των ΝΔ, το CFS αυξάνει την ευαισθησία του, ενώ το RFE, το PCA και το LDA αυξάνουν την ευαισθησία και το accuracy της ταξινόμησης. Το accuracy του NBC βελτιώνεται από το CFS και το LDA, ενώ η ευαισθησία του βελτιώνεται από το LDA. Οι αλγόριθμοι ταξινόμησης ΜΜ συγκρίνονται παρατηρώντας και την περιοχή κάτω από την καμπύλη της ROC και τις τιμές kappa. Οι SVM-LDA, ΝΔ-LDA και NBC-LDA είχαν την καλύτερη περιοχή κάτω από τις καμπύλες ROC τους (0,9994) και εμφανίζει την υψηλή απόδοσή τους. Επιπλέον, η καλύτερη τιμή kappa που παρατηρήθηκε (97,48%) αποκτήθηκε από τη SVM-LDA και το ΝΔ-LDA [19].

Η συγκεκριμένη μελέτη έδειξε ότι οι επιδόσεις ταξινόμησης εξαρτώνται από τον αλγόριθμο ΜΜ που θα επιλεχθεί. Τα αποτελέσματα έδειξαν ότι η SVM-LDA και το

ΝΔ-LDA ξεπερνούν τα άλλα μοντέλα ταξινομητών ΜΜ. Παρ 'όλα αυτά, η SVM-LDA επιλέγεται έναντι του ΝΔ-LDA επειδή το δεύτερο μοντέλο απαιτεί περισσότερο χρόνο. Η μελέτη προτείνει μια έξυπνη προσέγγιση ταξινόμησης που ενσωματώνει τη μέθοδο LDA με τη SVM για τη διάγνωση του καρκίνου του μαστού. Αυτή η προσέγγιση έδειξε καλά και ελπιδοφόρα αποτελέσματα, αφού έλαβε accuracy ταξινόμησης 98,82%, ευαισθησία 98,41%, ειδικότητα 99,07% και η περιοχή κάτω από τη χαρακτηριστική καμπύλη λειτουργίας ήταν 0,9994 [19].

5.3.2. Πρόβλεψη επιβίωσης και μετάστασης

Το 2019 ο L. Tapak και η ερευνητική του ομάδα μελέτησαν, σύγκριναν και αξιολόγησαν την απόδοση έξι τεχνικών ΜΜ και δύο κλασσικών τεχνικών πρόβλεψης για την επιβίωση και την εμφάνιση μετάστασης σε ασθενείς με καρκίνο του μαστού. Επομένως, υπήρχε μία μεταβλητή-στόχος για την επιβίωση και περιελάμβανε δύο κατηγορίες, ζωντανός ή νεκρός, και άλλη μία για τη μετάσταση, επίσης με δύο κατηγορίες ναι ή όχι. Οι τεχνικές ΜΜ που χρησιμοποιήθηκαν ήταν ο Naïve Bayes (NB), τα τυχαία δάση (Random Forest - RF), ο AdaBoost, οι Μηχανές Διανυσμάτων Υποστήριξης (SVM), οι SVMs ελαχίστου τετραγώνου (LSSVM), ο Adabag, η Λογιστική Παλινδρόμηση (LR) και η LDA [42].

Το σύνολο δεδομένων που χρησιμοποιήθηκε περιλαμβάνει 550 αρχεία ασθενών με καρκίνο του μαστού, από τα οποία οι 463 (83,4%) ήταν ζωντανοί ασθενείς και οι 92 (16,6%) ασθενείς ήταν αποβιώσαντες. Επιπλέον, περίπου το 85% των ασθενών δεν εμφάνισε μετάσταση. Το σύνολο αυτό προέρχεται από μια μελέτη που διεξήχθη το 2014 στην Τεχεράνη. Όλοι οι ασθενείς που είχαν διάγνωση και οι ασθενείς με άγνωστη παθολογία αποκλείστηκαν από την ανάλυση και η μελέτη επικεντρώθηκε στις πληροφορίες σχετικά με την επιβίωση (νεκρός/ζωντανός) και τη μετάσταση (ναι/όχι) σαν εξόδους [42].

Επιλέχθηκαν 9 παράγοντες κινδύνου που πιστεύεται ότι σχετίζονται με την επιβίωση των ασθενών με καρκίνο του μαστού, συμπεριλαμβανομένης της ηλικίας, την επιθετικότητα (καλά, μέτρια και άσχημα - καρκινικά κύτταρα «καλής» επιθετικότητας μοιάζουν και συμπεριφέρονται περισσότερο σαν φυσιολογικά κύτταρα. Όγκοι καλής επιθετικότητας τείνουν να είναι λιγότερο επιθετικοί), στάδιο, υποδοχέα οιστρογόνων (Estrogen Receptor - ER, ως αρνητικό ή θετικό), υποδοχέα προγεστερόνη (Progesterone Reseptor - PR, ως αρνητικό ή θετικό), υποδοχέα ανθρώπινου επιδερμικού αυξητικού παράγοντα 2 (Human epidermal growth factor receptor 2 - HER2 ως αρνητικός ή θετικός), Παθολογικός τύπος (Πορογενές/λοβιακό καρκίνωμα in situ, Επεμβατικό λοβιακό καρκίνωμα και επεμβατικό πορογενές καρκίνωμα) και χειρουργική προσέγγιση (Τροποποιημένη Ριζική Μαστεκτομή και Χειρουργική Επέμβαση Διατήρησης μαστού) για τη σύγκριση της απόδοσης των επιλεγμένων μοντέλων [42].

Για την αξιολόγηση των μοντέλων υπολογίστηκαν η ευαισθησία, η ειδικότητα, η PPV, η NPV, ο LR+, ο LR– και το accuracy, όπου το (ΛΘ) αντιπροσωπεύει ζωντανά άτομα με καρκίνο του μαστού που αναγνωρίστηκαν λανθασμένα ως νεκρά, το (ΕΘ) αντιπροσωπεύει νεκρούς ανθρώπους με καρκίνο του μαστού που διαγνώστηκαν σωστά ως νεκροί, το (ΕΑ) αντιπροσωπεύει ζωντανά άτομα με καρκίνο του μαστού που αναγνωρίστηκαν σωστά ως νεκρά και το (ΛΑ) αντιπροσωπεύει νεκρούς ανθρώπους με καρκίνο του μαστού που εσφαλμένα αναγνωρίστηκαν ως ζωντανός [42].

Τα δεδομένα χωρίστηκαν σε σύνολα εκπαίδευσης και δοκιμών με δύο διαφορετικούς τρόπους. Η πρώτη περίπτωση που εξετάστηκε ήταν 70% δεδομένα εκπαίδευσης και 30% δεδομένα δοκιμής, ενώ η δεύτερη ήταν 50% και 50%. Έγινε επανάληψη της διαδικασίας 100 φορές για κάθε σενάριο και αναφέρθηκαν οι μέσοι όροι των κριτηρίων αξιολόγησης [42].

Τα χαρακτηριστικά των ασθενών με καρκίνο του μαστού παρουσιάστηκαν στον Πίνακας 11. Οι ασθενείς με καρκίνο του μαστού με διάγνωση στην ηλικία των 47.86 ± 11.79 (μέσος όρος ± τυπική απόκλιση) ετών κατά μέσο όρο, με ελάχιστο και μέγιστο τα 17 και 84 έτη αντίστοιχα. Στους περισσότερους ασθενείς παρουσιάστηκε Επιθετικότητα II (52.36%) και βρίσκονταν στο στάδιο II (41.46%), με ER+ 71.27%, PR+ 68.36%, HER2- 76.36%, το 90.19% διαγνώστηκαν με Επεμβατικό Πορογενές Καρκίνωμα, ενώ το 65.09% του συνολικού δείγματος υποβλήθηκε σε χειρουργική επέμβαση διατήρησης του μαστού. Οι κατανομές των χαρακτηριστικών των ασθενών που χωρίστηκαν τυχαία σε δύο σύνολα (εκπαίδευση και δοκιμή) παραχωρούνται επίσης στον Πίνακας 11. Όπως φαίνεται στις περισσότερες από 100 επαναλήψεις, δεν υπήρχαν σημαντικές διαφορές μεταξύ των συνόλων εκπαίδευσης και δοκιμών ($P > 0,05$) [42].

Ο Πίνακας 12 δείχνει την απόδοση των οκτώ ταξινομητών για την πρόβλεψη της επιβίωσης των ασθενών με καρκίνο του μαστού, όσον αφορά την ευαισθησία, την ειδικότητα, την PPV, την NPV, την LR+, την LR και το accuracy, που λαμβάνονται από 100 φορές επανάληψης της τεχνικής cross validation στο πλαίσιο δύο διαφορετικών περιπτώσεων. Όπως φαίνεται στον Πίνακας 12, όλοι οι χρησιμοποιούμενοι αλγόριθμοι είχαν υψηλή ειδικότητα (≥94%) για τα δύο διαφορετικά σετ δοκιμής. Ωστόσο, οι τιμές της ευαισθησίας για τα μοντέλα ήταν μέτριες και κυμάνθηκαν μεταξύ 0.61 και 0.73 κατά μέσο όρο για τα σύνολα δοκιμής, με το χαμηλότερο και υψηλότερο να ανήκει στο AdaBoost και το SVM αντίστοιχα. Η ευαισθησία του LDA ήταν παρόμοια με εκείνης της SVM. Ο μέσος PPV των μεθόδων κυμαινόταν μεταξύ 0.69 και 0.82, για την τα σύνολα δοκιμών, με το χαμηλότερο και το υψηλότερο να ανήκουν στο AdaBoost και το RF αντίστοιχα και η μέση απόδοση NPV όλων των μεθόδων ήταν μεγαλύτερη από 0.92. Επιπλέον, το accuracy κυμαινόταν μεταξύ 0.89 (για το AdaBoost) και 0.93 (SVM και LDA). Ο μέσος όρος LR+ των μεθόδων κυμαινό -

ταν μεταξύ 10.17 (AdaBoost) και 24.33 (για το SVM και το LDA) για τα σύνολα δοκιμής. Οι τιμές του LR- για τα μοντέλα κυμάνθηκαν μεταξύ 0.27 και 0.41 κατά μέσο όρο για τα σύνολα δοκιμών, με το χαμηλότερο και το υψηλότερο να ανήκει στο LDA και το LSSVM και το AdaBoost, αντίστοιχα [42].

Ο Πίνακας 13 δείχνει την απόδοση των οκτώ ταξινομητών για την πρόβλεψη της μετάστασης των ασθενών με καρκίνο του μαστού όσον αφορά την ευαισθησία, την ειδικότητα, την PPV, την ΚΠΑ, την LR+, την LR και την ακρίβεια που λαμβάνονται από 100 φορές επανάληψης της τεχνικής cross validation στο πλαίσιο δύο διαφορετικών περιπτώσεων. Όπως φαίνεται, όλοι οι χρησιμοποιούμενοι αλγόριθμοι είχαν υψηλή ειδικότητα (≥90%) κατά μέσο όρο για τα σύνολα δοκιμής με το ελάχιστο για το AdaBoost και το μέγιστο για το RF (0.98). Από την άλλη, όλες οι μέθοδοι είχαν χαμηλή ευαισθησία (κατά μέσο όρο), η οποία κυμαινόταν μεταξύ 0.07 (για το RF) και 0.36 (για τον Naive Bayes) κατά μέσο όρο. Ο μέσος PPV των μεθόδων κυμαινόταν μεταξύ 0.32 (AdaBoost και SVM) και 0.61 (για το LDA), ενώ ο μέσος όρος NPV όλων των μεθόδων ήταν μεγαλύτερη από 0.82 (το ελάχιστο ανήκε στο Adabag και το μέγιστο ανήκε στον Naive Bayes). Επιπλέον, το συνολικό accuracy κυμαινόταν μεταξύ 0.80 (για το AdaBoost) και 0.86 (για το LR και το LDA). Ο μέσος όρος LR+ των μεθόδων κυμαινόταν μεταξύ 2.81 (για το AdaBoost) και 9.05 (για το LDA). Οι τιμές του LR- για τα μοντέλα κυμαίνονταν μεταξύ 0.71 και 0.94, κατά μέσο όρο για τα σύνολα δοκιμών, με το χαμηλότερο και το υψηλότερο να ανήκει στον NB και το RF, αντίστοιχα [42].

Διάφορες μέθοδοι μηχανικής μάθησης υλοποιήθηκαν και συγκρίθηκαν για την πρόβλεψη της επιβίωσης και της μετάστασης σε ασθενείς με καρκίνο του μαστού. Με βάση το accuracy, παρατηρείται πως όλες οι μέθοδοι ταξινόμησης ήταν αρκετά αποτελεσματικές στην πρόβλεψη για την κατάσταση επιβίωσης του καρκίνου του μαστού και τη μετάσταση (κυμάνθηκαν μεταξύ 0.80 και 0.93). Όσον αφορά την ευαισθησία, η SVM και η LDA προέβλεψαν καλύτερα την επιβίωση των ασθενών με 0.73. Η ελάχιστη ευαισθησία παρατηρήθηκε στον AdaBoost με 0.61 και η μέγιστη τιμή ήταν 0.73 (SVM και LDA). Επιπλέον, παρά την καλή επίδοση των μεθόδων όσον αφορά την ειδικότητα και το accuracy, η ευαισθησία για την πρόβλεψη της μετάστασης των ασθενών με καρκίνο του μαστού ήταν σχετικά κακή, αφού κυμαινόταν μεταξύ 0.07 και 0.36, και, επομένως, καμία από τις μεθόδους δεν έχει ευαισθησία μεγαλύτερη από 0.5. Τα ευρήματα αυτής της μελέτης έδειξαν ότι το SVM και το LDA ήταν τα καλύτερα μοντέλα για την πρόβλεψη της επιβίωσης για τα διάφορα κριτήρια, ενώ το LDA ήταν η καλύτερη τεχνική για την πρόβλεψη της μετάστασης μεταξύ των ασθενών με καρκίνο του μαστού [42].

Μεταβλητή	Πλήθος (%)	ΣΕ		ΣΔ		P-Value
		ΜΟ	ΤΑ	ΜΟ	ΤΑ	
Στάδιο						0.833
I	110 (220.00)	77.80	4.88	32.20	4.88	
II	228 (41.46)	162.49	4.96	68.51	4.96	
III	188 (34.18)	131.72	5.03	57.28	5.03	
IV	24 (4.36)	16.99	2.23	7.01	2.23	
Επιθετικότητα						0.835
1	66 (12.00)	46.51	3.36	19.49	3.36	
2	288 (52.36)	204.92	5.35	87.08	5.35	
3	196 (35.64)	137.57	5.63	59.43	5.63	
Μετάσταση						0.903
Όχι	467 (84.91)	331.1	3.74	140.9	3.74	
Ναι	83 (15.09)	87.9	3.74	25.1	3.74	
ER						0.841
Αρνητικό	158 (28.73)	112.21	5.04	47.79	5.04	
Θετικό	392 (71.27)	276.79	5.04	118.21	5.04	
PR						0.916
Αρνητικό	174 (31.64)	124.09	4.99	51.91	4.99	
Θετικό	376 (68.36)	264.91	4.99	114.09	4.99	
HER2						0.871
Αρνητικό	420 (76.36)	296.36	4.53	127.36	4.53	
Θετικό	130 (23.64)	92.36	4.53	38.64	4.53	
Παθολογικός Τύπος						0.931
Πορογενές/Λοβιακό καρκίνωμα in situ	29 (5.27)	20.74	2.43	8.26	2.43	
Επεμβατικό Λοβιακό Καρκίνωμα	25 (4.54)	17.44	2.22	7.56	2.22	
Επεμβατικό Πορογενές Καρκίνωμα	496 (90.19)	350.82	3.25	150.18	3.25	
Χειρουργική Προσέγγιση						0.780
Τροποποιημένη Ριζική Μαστεκτομή	192 (34.91)	252.74	5.85	108.26	5.85	
Χειρουργική Επέμβαση Διατήρησης μαστού	358 (65.09)	136.26	5.85	57.74	5.85	
Ηλικία (ΜΟ(ΤΑ))	47.86 (11.79)	52.61	0.38	52.50	0.80	0.181

Πίνακας 11. Χαρακτηριστικά ασθενών με καρκίνο του μαστού, όπου ΣΕ=Σύνολο Εκπαίδευσης, ΣΔ= Σύνολο Δοκιμής, ΜΟ= Μέσος Όρος, ΤΑ=Τυπική Απόκλιση [42].

Περίπτωση	Μέθοδος	Σύνολο	Sensitivity	Specificity	PPV	NPV	LR+	LR-	Accuracy
70% Ε,	NB	Ε	0.74±0.04	0.96±0.01	0.80±0.03	0.95±0.01	18.50±7.68	0.27±0.08	0.93±0.01
		Δ	0.69±0.07	0.96±0.02	0.78±0.07	0.94±0.02	17.25±6.27	0.32±0.07	0.92±0.02
30% Δ	LS-SVM	Ε	0.79±0.05	0.98±0.01	0.90±0.03	0.96±0.01	39.50±6.23	0.21±0.04	0.95±0.01
		Δ	0.62±0.09	0.97±0.01	0.78±0.08	0.93±0.02	20.67±11.73	0.39±0.07	0.91±0.02
	Adabag	Ε	0.73±0.03	0.97±0.01	0.82±0.02	0.95±0.01	24.33±6.23	0.28±0.04	0.93±0.01
		Δ	0.70±0.08	0.97±0.01	0.81±0.07	0.94±0.02	23.33±11.73	0.31±0.07	0.92±0.02
	Adaboost	Ε	0.99±0.01	0.99±0.01	0.99±0.01	0.99±0.01	99.00±4.27	0.01±0.04	0.99±0.01
		Δ	0.61±0.10	0.94±0.02	0.69±0.08	0.92±0.02	10.17±11.73	0.41±0.07	0.89±0.02
	RF	Ε	0.72±0.04	0.97±0.01	0.85±0.03	0.95±0.01	24.00±4.30	0.29±0.04	0.93±0.01
		Δ	0.68±0.08	0.97±0.01	0.82±0.07	0.94±0.02	22.67±11.73	0.33±0.07	0.92±0.02
	SVM	Ε	0.73±0.03	0.97±0.01	0.81±0.03	0.95±0.01	24.33±4.27	0.28±0.04	0.93±0.01
		Δ	0.73±0.07	0.97±0.01	0.81±0.07	0.95±0.01	24.33±11.73	0.28±0.07	0.92±0.02
	Logit	Ε	0.72±0.03	0.97±0.01	0.84±0.02	0.95±0.01	24.00±4.30	0.29±0.04	0.93±0.01
		Δ	0.68±0.07	0.97±0.02	0.81±0.08	0.94±0.02	22.67±11.73	0.33±0.07	0.92±0.02
	LDA	Ε	0.73±0.03	0.96±0.01	0.80±0.03	0.95±0.01	18.25±4.27	0.28±0.04	0.93±0.01
		Δ	0.73±0.07	0.97±0.01	0.81±0.07	0.95±0.01	24.33±11.72	0.27±0.07	0.93±0.02
50% Ε,	NB	Ε	0.74±0.05	0.96±0.01	0.80±0.05	0.95±0.01	18.5±8.74	0.27±0.05	0.93±0.01
		Δ	0.69±0.07	0.96±0.01	0.78±0.06	0.94±0.02	17.25±9.61	0.32±0.07	0.91±0.01
50% Δ	LS-SVM	Ε	0.83±0.06	0.99±0.006	0.95±0.04	0.97±0.01	83.00±8.74	0.17±0.05	0.97±0.01
		Δ	0.62±0.07	0.96±0.01	0.76±0.06	0.93±0.02	15.25±9.61	0.41±0.07	0.90±0.02
	Adabag	Ε	0.71±0.06	0.97±0.009	0.83±0.04	0.94±0.01	23.67±8.74	0.30±0.05	0.93±0.01
		Δ	0.70±0.08	0.97±0.01	0.81±0.05	0.94±0.02	23.33±9.61	0.31±0.07	0.92±0.01
	Adaboost	Ε	0.99±0.02	0.99±0.002	0.99±0.01	0.99±0.01	99.00±8.75	0.01±0.05	0.99±0.002
		Δ	0.61±0.07	0.94±0.02	0.69±0.06	0.92±0.02	10.33±9.61	0.40±0.07	0.89±0.01
	RF	Ε	0.71±0.06	0.98±0.01	0.87±0.04	0.94±0.01	35.5±8.74	0.30±0.05	0.93±0.01
		Δ	0.66±0.07	0.97±0.01	0.82±0.04	0.93±0.02	22.00±9.61	0.35±0.07	0.92±0.01
	SVM	Ε	0.72±0.06	0.97±0.01	0.82±0.04	0.95±0.01	24.00±8.75	0.29±0.05	0.93±0.01
		Δ	0.72±0.06	0.97±0.01	0.81±0.04	0.95±0.01	24.00±8.75	0.29±0.05	0.93±0.01
	Logit	Ε	0.72±0.06	0.97±0.01	0.85±0.04	0.95±0.01	24.00±8.75	0.29±0.05	0.93±0.01
		Δ	0.68±0.07	0.95±0.01	0.80±0.06	0.94±0.01	22.67±9.61	0.33±0.07	0.92±0.01
	LDA	Ε	0.73±0.04	0.97±0.01	0.81±0.04	0.95±0.01	24.33±8.74	0.28±0.05	0.93±0.01
		Δ	0.73±0.05	0.97±0.01	0.81±0.04	0.95±0.01	24.33±9.61	0.28±0.07	0.93±0.02

Πίνακας 12. Τιμές απόδοσης αλγορίθμων για πρόβλεψη επιβίωσης των ασθενών με καρκίνο του μαστού [42]

Ε= Εκπαίδευσης, Δ=Δοκιμή

Περίπτωση	Μέθοδος	Σύνολο	Sensitivity	Specificity	PPV	NPV	LR+	LR-	Accuracy
70% Ε	NB	Ε	0.36±0.04	0.94±0.01	0.51±0.05	0.89±0.01	5.93±0.98	0.68±0.05	0.85±0.01
30% Δ		Δ	0.33±0.08	0.94±0.02	0.49±0.14	0.89±0.02	5.50±3.54	0.72±0.09	0.85±0.02
	LS-SVM	Ε	0.46±0.10	0.98±0.01	0.81±0.01	0.91±0.01	30.20±17.72	0.55±0.10	0.90±0.02
		Δ	0.21±0.07	0.95±0.02	0.44±0.14	0.87±0.02	5.07±3.07	0.84±0.08	0.84±0.02
	Adabag	Ε	0.33±0.06	0.98±0.01	0.74±0.05	0.89±0.01	18.44±10.24	0.69±0.06	0.88±0.01
		Δ	0.20±0.07	0.97±0.02	0.53±0.16	0.82±0.02	6.67±3.84	0.83±0.07	0.85±0.02
	Adaboost	Ε	0.92±0.03	0.94±0.01	0.97±0.02	0.99±0.01	189.84±56.78	0.08±0.03	0.98±0.01
		Δ	0.25±0.08	0.91±0.02	0.32±0.09	0.87±0.02	2.81±1.01	0.83±0.08	0.81±0.02
	RF	Ε	0.20±0.07	0.99±0.01	0.86±0.06	0.88±0.01	35.24±15.14	0.81±0.07	0.87±0.01
		Δ	0.08±0.05	0.98±0.01	0.49±0.20	0.86±0.02	5.47±3.82	0.94±0.04	0.85±0.02
	SVM	Ε	0.58±0.14	0.99±0.01	0.94±0.07	0.93±0.04	59.75±42.27	0.43±0.26	0.93±0.04
		Δ	0.14±0.11	0.95±0.03	0.32±0.17	0.87±0.02	3.33±2.47	0.89±0.10	0.83±0.02
	Logit	Ε	0.26±0.06	0.97±0.01	0.63±0.07	0.88±0.02	10.27±3.65	0.76±0.06	0.86±0.02
		Δ	0.21±0.07	0.97±0.02	0.59±0.15	0.87±0.02	7.61±3.73	0.81±0.07	0.86±0.02
	LDA	Ε	0.28±0.05	0.97±0.01	0.66±0.06	0.89±0.01	11.88±6.03	0.74±0.05	0.87±0.01
		Δ	0.26±0.09	0.97±0.01	0.61±0.13	0.88±0.02	9.05±3.90	0.77±0.09	0.86±0.02
50% Ε	NB	Ε	0.39±0.07	0.94±0.01	0.53±0.07	0.90±0.01	6.86±2.07	0.65±0.07	0.86±0.02
50% *Δ*		Δ	0.34±0.07	0.93±0.02	0.46±0.08	0.89±0.02	5.20±1.69	0.71±0.07	0.84±0.02
	LS-SVM	Ε	0.55±0.11	0.98±0.01	0.84±0.06	0.93±0.02	30.20±17.72	0.46±0.11	0.92±0.02
		Δ	0.24±0.07	0.93±0.02	0.39±0.09	0.88±0.02	3.95±1.61	0.81±0.07	0.83±0.02
	Adabag	Ε	0.33±0.09	0.98±0.01	0.76±0.06	0.89±0.01	19.97±8.19	0.68±0.09	0.88±0.01
		Δ	0.21±0.07	0.96±0.02	0.71±0.11	0.87±0.02	6.59±3.84	0.82±0.07	0.85±0.02
	Adaboost	Ε	0.94±0.04	0.99±0.01	0.98±0.02	0.99±0.01	189.84±56.78	0.06±0.04	0.99±0.01
		Δ	0.27±0.08	0.90±0.02	0.33±0.06	0.88±0.02	2.83±0.77	0.82±0.07	0.81±0.02
	RF	Ε	0.18±0.10	0.99±0.01	0.86±0.06	0.87±0.01	35.24±15.14	0.82±0.10	0.87±0.01
		Δ	0.07±0.05	0.98±0.01	0.49±0.20	0.86±0.02	5.47±3.82	0.94±0.05	0.85±0.01
	SVM	Ε	0.49±0.22	0.99±0.01	0.91±0.06	0.92±0.03	65.67±57.65	0.51±0.21	0.92±0.03
		Δ	0.14±0.09	0.96±0.03	0.37±0.16	0.87±0.02	4.85±6.58	0.89±0.08	0.84±0.02
	Logit	Ε	0.28±0.08	0.97±0.01	0.64±0.09	0.89±0.01	11.47±4.68	0.74±0.08	0.86±0.02
		Δ	0.24±0.09	0.96±0.02	0.55±0.11	0.87±0.02	7.61±3.73	0.79±0.08	0.86±0.02
	LDA	Ε	0.30±0.07	0.97±0.01	0.67±0.09	0.89±0.01	13.65±9.21	0.72±0.07	0.87±0.01
		Δ	0.27±0.09	0.97±0.02	0.59±0.09	0.88±0.02	9.05±3.90	0.75±0.09	0.86±0.01

Πίνακας 13. Τιμές απόδοσης αλγορίθμων για πρόβλεψη μετάστασης των ασθενών με καρκίνο του μαστού [42].

Ε= Εκπαίδευσης, Δ=Δοκιμή

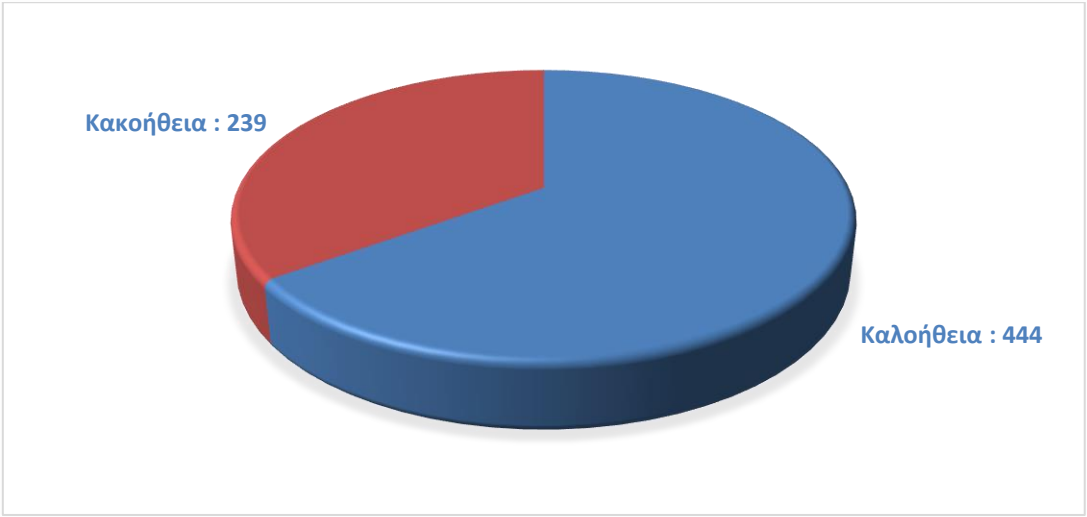
5.4. Μέθοδος Κ-Κοντινότερων Γειτόνων (Κ-ΝΝ)

Στην μελέτη που πραγματοποιήθηκε από τους M. Amrane και I. Gagaoua, παρουσιάστηκαν δύο ταξινομητές, ο ταξινομητής Naive Bayes (NBC) και ο Κ-Κοντινότερος Γείτονας (Κ-ΝΝ), για τη διάγνωση του καρκίνου του μαστού. Οι δύο υλοποιήσεις συγκρίθηκαν και αξιολογήθηκαν χρησιμοποιώντας την τεχνική Cross Validation και με χρήση του συνόλου δεδομένων WBCD. Η απόδοση κάθε ταξινομητή αξιολογήθηκε από το accuracy, τη διαδικασία εκπαίδευσης και τη διαδικασία δοκιμής [41].

Για να γίνει σωστή πρόγνωση, η ταξινόμηση του χρειάζεται τα εννέα χαρακτηριστικά της βάσης δεδομένων WBCD του Πίνακας 4. Προκειμένου να

ταξινομηθεί ο καρκίνος του μαστού, οι παθολόγοι αποδίδουν σε καθένα από αυτά τα χαρακτηριστικά έναν αριθμό από το 1 έως το 10. Η πιθανότητα κακοήθειας χρειάζεται τα εννέα κριτήρια, ακόμη και αν ένα από αυτά είναι πολύ μεγάλο [41].

Στο σύνολο των δεδομένων που χρησιμοποιήθηκε υπάρχουν 11 χαρακτηριστικά. Το πρώτο είναι το αναγνωριστικό τους, το οποίο θα αφαιρέσουμε αφού δεν είναι ένα χαρακτηριστικό που θέλουμε να τροφοδοτήσουμε στην ταξινόμησή μας. Τα εννέα κριτήρια του Πίνακας 4 προορίζονται να καθορίσουν εάν ένας όγκος είναι καλοήθης ή κακοήθης. Το τελευταίο χαρακτηριστικό είναι η κλάση (Class) και περιέχει δυαδική τιμή (2 για καλοήθη όγκο και 4 για κακοήθη όγκο). Το σύνολο αποτελείται από 699 κλινικές περιπτώσεις. Στην αρχική ταξινόμηση για τον καρκίνο του μαστού λείπουν δεδομένα για 16 παρατηρήσεις, οι οποίες περιόρισαν τα δεδομένα μας σε 683 δείγματα. Ο Πίνακας *14* δείχνει πόσοι όγκοι από το δείγμα αντιστοιχούν σε κακοήθεια ή καλοήθεια.



Πίνακας 14. Σύνολο Διαγνωστικών Δεδομένων καρκίνου του μαστού του Wisconsin.

Από τον Πίνακας 15 μπορούμε να δούμε τα αποτελέσματα της μελέτης. Οι δύο αλγόριθμοι έδειξαν υψηλό επίπεδο ακρίβειας παρά το μικρό σύνολο δεδομένων. Σε αυτή τη μελέτη, φαίνεται πως ο K-NN είναι ο πιο αποτελεσματικός ταξινομητής, με accuracy 97,51%, ενώ ο NBC έχει εξίσου καλό accuracy στο 96,19% . Ωστόσο, εάν αυξηθεί το μέγεθος του συνόλου δεδομένων, ο K-NN θα έρθει δεύτερος λόγω της χρονικής πολυπλοκότητας του υπολογισμού [41].

Μέθοδος	Accuracy	Διαδικασία Εκπαίδευσης	Διαδικασία Δοκιμής	Συνολική Διαδικασία
K-NN	0.975109	0.000735	0.001744	0.002479
NBC	0.961932	0.000759	0.000422	0.001182

Πίνακας 15. Σύγκριση αποτελεσμάτων μεταξύ K-NN και NBC [41].

5.5. Μέθοδος Ενισχυμένης Μάθησης στη βάση του Deep Q-Network (DQ-N)

Η μαστογραφία αποτελεί τη συνηθέστερη απεικονιστική μέθοδο για την αξιολόγηση των ασθενών με καρκίνο του μαστού, λόγω του μειωμένου κόστους αλλά και της μη επεμβατικότητάς της. Ωστόσο, σε περιπτώσεις κατά τις οποίες ο ασθενής έχει πυκνό μαστό, τα αποτελέσματα μιας μαστογραφίας δεν είναι και τόσο ικανοποιητικά, λόγω του μεγάλου ποσοστού λανθασμένων θετικών ανιχνεύσεων. Σε αυτούς τους ασθενείς συνιστάται ο προληπτικός τους έλεγχος να περιλαμβάνει Δυναμικές Εικόνες Μαγνητικού Συντονισμού Ενισχυμένες με Αντίθεση (Dynamically Contrast Enhanced Magnetic Resonance Images, DCE-MRI), οι οποίες αυξάνουν την ευαισθησία και την ειδικότητα της ανίχνευσης του καρκίνου του μαστού [44].

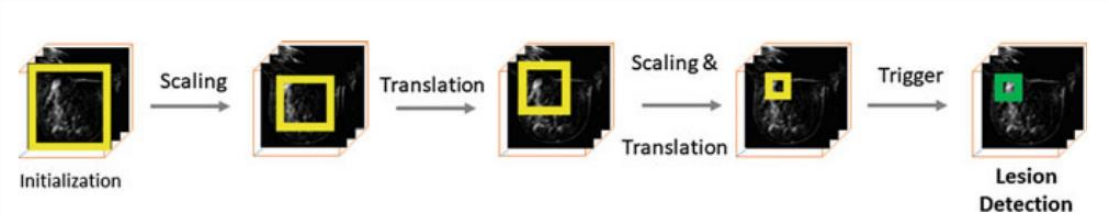
Η ερμηνεία των εικόνων DCE-MRI είναι μια χρονοβόρα και επίπονη διαδικασία λόγω του πλήθους των τομών και των πολλών διαστάσεών τους, με αποτέλεσμα μεγάλη ποικιλία σφαλμάτων. Με στόχο τη μείωση του χρόνου ερμηνείας των τομών και τα σφάλματα, αναπτύσσονται συστήματα διάγνωσης με τη βοήθεια υπολογιστή (CAD), τα οποία αποδεικνύεται ότι βελτιώνουν την ευαισθησία και την ειδικότητα, προκειμένου να παρέχεται στους ακτινολόγους μια δεύτερη γνώμη [44].

Τα συστήματα CAD μπορούν να ακολουθήσουν ένα pre-hoc ή ένα post-hoc σύστημα. Σε ένα σύστημα pre-hoc, ύποπτες περιοχές ενδιαφέροντος (Regions Of Interest, ROIs) εντοπίζονται για πρώτη φορά στην τομή εισόδου. Στη συνέχεια, ένας ταξινομητής διακρίνει τις αλλοιώσεις μεταξύ κακοήθων και μη-κακοηθών (όπου οι μη-κακοήθεις αλλοιώσεις μπορούν να ταξινομηθούν σε καλοήθεις ή ψευδώς θετικές). Τελικά, η διάγνωση παράγεται συνδυάζοντας τέτοια μεμονωμένα αποτελέσματα ταξινόμησης αλλοιώσεων. Από την άλλη, τα post-hoc συστήματα εκτελούν διάγνωση ταξινομώντας πρώτα ολόκληρη την τομή εισόδου σε αρνητικά (φυσιολογικά ευρήματα) ή θετικά (κακοήθη ευρήματα). Μόνο για τις θετικές περιπτώσεις, η μέθοδος στη συνέχεια εντοπίζει τις κακοήθεις αλλοιώσεις [44].

Οι pre-hoc και post-hoc διαφέρουν και ως προς το είδος των παρατηρήσεων που απαιτούνται στο στάδιο της εκπαίδευσής. Ενώ τα pre-hoc συστήματα απαιτούν ένα έντονα επισημασμένο σύνολο εκπαίδευσης, οι post-hoc προσεγγίσεις απαιτούν μόνο αδύναμα επισημασμένα σύνολα εκπαίδευσης. Αυτό οδηγεί σε μειωμένη ακρίβεια (accuracy) απόδοσης εντοπισμού των αλλοιώσεων συγκριτικά με τις pre-hoc προσεγγίσεις. Αυτή η έλλειψη ακρίβειας για την επισήμανση των κακοηθών αλλοιώσεων είναι ένας σημαντικός περιορισμός, καθώς μπορεί να αποτρέψει τη χρήση των συστημάτων CAD στην κλινική δοκιμή. Η συντριπτική πλειοψηφία των συστημάτων CAD βασίζεται στις pre-hoc προσεγγίσεις, δεδομένου ότι οι υψηλότερες πραγματικές θετικές και πραγματικές αρνητικές ανιχνεύσεις τους είναι πιο αποτελεσματικές στο να βοηθήσουν τους ακτινολόγους να βελτιώσουν την απόδοση της διάγνωσής τους [44].

Ο G. Maicas και η ερευνητική του ομάδα εστίασαν στη μείωση του χρόνου συμπεράσματος που απαιτείται από τις μεθόδους αναζήτησης που

χρησιμοποιούνται για το στάδιο ανίχνευσης αλλοιώσεων των pre-hoc συστημάτων. Για την αντιμετώπιση αυτών των περιορισμών, λόγω της μεγάλης χρονικής διάρκειας συμπερασμάτων και της ανάγκης για μεγάλα σύνολα εκπαίδευσης, πρότειναν μια μέθοδο Ενισχυμένης Μάθησης (EM) βασισμένη στο βαθύ δίκτυο Q (Deep Q-Network, DQN) για την ανίχνευση μαζών σε μαστούς από DCE-MRI. Το DQN διαμορφώνει μια πολιτική που δείχνει πώς να μετασχηματίσει διαδοχικά μία μεγάλη τομή οριοθέτησης, έτσι ώστε μετά από κάθε μετασχηματισμό η βλάβη να είναι καλύτερα εστιασμένη και τελικά να ανιχνευθεί. Στην Εικόνα 32 το σύστημα ξεκινά αναλύοντας ένα μεγάλο ποσοστό της τομής εισόδου και εφαρμόζει διαδοχικά αρκετούς μετασχηματισμούς οριοθέτησης, μέχρις ότου οι αλλοιώσεις να είναι στενά στοχευμένες. Σε σύγκριση με άλλες προτεινόμενες μεθόδους βαθιάς μάθησης, που μπορούν να ανιχνεύσουν αλλοιώσεις και να εκτελέσουν διάγνωση, η προσέγγισή αυτή θεωρείται ταχύτερη και μπορεί να εκπαιδευτεί με μικρά σύνολα εκπαίδευσης [44].



Εικόνα 32. Μοντέλο βασισμένο στην Ενισχυμένη Μάθηση για την ανίχνευση αλλοιώσεων του μαστού με DCE-MRI [44].

Η μέθοδος αυτή αξιολογήθηκε με χρήση ενός συνόλου δεδομένων μαγνητικής τομογραφίας μαστού, το οποίο αποτελείται από 117 ασθενείς. Στον Πίνακας *16* βλέπουμε το πλήθος των ασθενών και των αλλοιώσεων που υπάρχουν σε κάθε σύνολο δεδομένων, από αυτά που χωρίστηκε το συνολικό δείγμα. Όλες οι αλλοιώσεις επιβεβαιώθηκαν με βιοψία και παρόλο που οι ασθενείς πάσχουν από τουλάχιστον μία αλλοίωση, δεν περιέχουν όλοι οι μαστοί αλλοιώσεις [44].

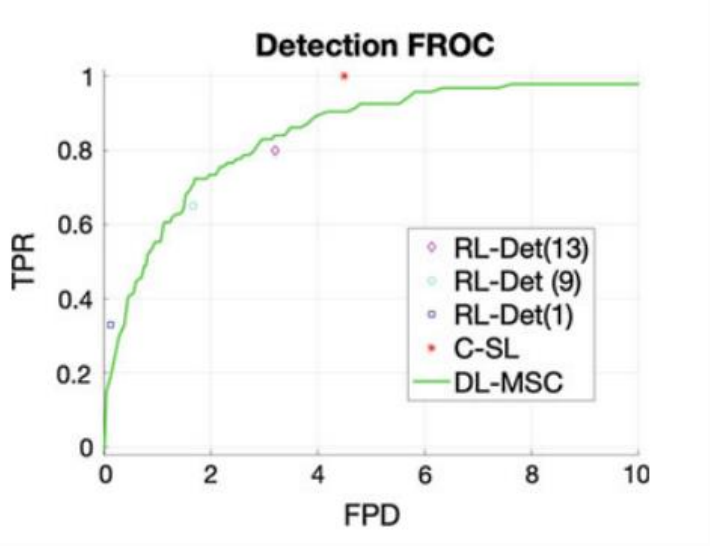
	Σύνολο Εκπαίδευσης	**Σύνολο Επαλήθευσης**	**Σύνολο Δοκιμής**
Ασθενείς	45	13	59
Κακοήθεις Αλλοιώσεις	38	11	46
Καλοήθεις Αλλοιώσεις	19	4	23
Συνολικό Πλήθος Αλλοιώσεων	57	15	69

Πίνακας 16. Διαχωρισμός του συνόλου δεδομένων μαγνητικής τομογραφίας μαστού[44].

Κατά τη διαδικασία της προ-επεξεργασίας των δεδομένων αφαιρέθηκαν οι δεξιές και αριστερές περιοχές του μαστού και ο θωρακικός μυς, προκειμένου να τονιστεί αυτόματα η περιοχή του μαστού από το θωρακικό τοίχωμα. Ο θωρακικός μυς αφαιρείται προκειμένου να μειωθεί ο αριθμός των ψευδών θετικών ανιχνεύσεων. Τέλος, η περιοχή του μαστού χωρίζεται στο αριστερό και το δεξί στήθος και αλλάζει μέγεθος σε τομή μεγέθους 100 × 100 × 50 [44].

Η απόδοση της συγκεκριμένης μελέτης, για την προτεινόμενη μέθοδο διάγνωσης του καρκίνου του μαστού, αξιολογήθηκε σύμφωνα με την καμπύλη λειτουργίας ελεύθερης απόκρισης (Free Response Operating Curve, FROC), η οποία συγκρίνει τον πραγματικό θετικό ρυθμό (True Positive Rate, TPR) όσον αφορά τον αριθμό των ψευδώς θετικών ανιχνεύσεων και το χρόνο που χρειάστηκε ο υπολογιστής ώστε να βγάλει συμπέρασμα ανά ασθενή. Ο υπολογιστής που χρησιμοποιήθηκε ήταν Intel Core i7, με 12 GB μνήμη RAM και GPU Nvidia Titan X 12 GB. Τέλος, μια εντοπισμένη περιοχή θεωρείται πραγματική θετική ανίχνευση μόνο εάν ο συντελεστής Dice σε σχέση με μια αλλοίωση είναι ίσος ή μεγαλύτερος από 0,2 [44].

Τα αποτελέσματα που επιτεύχθηκαν από την προτεινόμενη μέθοδος EM, παρουσιάζονται στην Εικόνα *33* και στον Πίνακας 17. Στην Εικόνα *33* η ένδειξη RL-Det (Reinforcement Learning Detection) αναφέρεται στο μοντέλο που προτάθηκε από τον G. Maicas, το οποίο το έτρεξαν με διαφορετικό αριθμό αρχικοποιήσεων (1, 9 και 13) για να αποκτηθεί η καμπύλη FROC. Η πράσινη καμπύλη FROC αφορά τη μέθοδο BM, η οποία στο γράφημα χαρακτηρίζεται ως DL-MSC (Deep Learning Multi-Scale Cascade), ενώ το μοντέλο μη-επιβλεπόμενης μάθησης επισημαίνεται ως C-SL (Clustering and Structure Learning) και αντιστοιχεί στο κόκκινο σημείο του γραφήματος [44].



Εικόνα 33. *Καμπύλη FROC* [44].

Στον Πίνακας 17 συγκρίνεται η RL-Det με τις DL-MSC και C-SL, όσον αφορά τον χρόνο συμπεράσματος, το TPR (True Positive Rate) και το FPR (False Positive Rate). Και τα τρία μοντέλα έκαναν χρήση του ίδιου συνόλου δεδομένων [44].

	Χρόνος Συμπεράσματος	**TPR**	**FPR**
RL-Det	**92±21 s**	**0.8**	**3.2**
DL-MSC	164±137 s	0.8	2.8
C-SL	≈3600 s	1.0	4.5

Πίνακας 17. Αποτελέσματα μελέτης RL-Det [44].

Η σημασία της διαφοράς για το χρόνο συμπεράσματος ανά ασθενή, μεταξύ του προτεινόμενου μοντέλου με τα άλλα δύο, πραγματοποιήθηκε με χρήση της μεθόδου paired t-test, λαμβάνοντας $p \leq 9 \times 10^{-5}$. Δεδομένης μιας τόσο μικρής ποσότητας στην p-value, συμπεραίνουμε πως η διαφορά της προτεινόμενης μεθόδου με τις άλλες δύο είναι σημαντική [44].

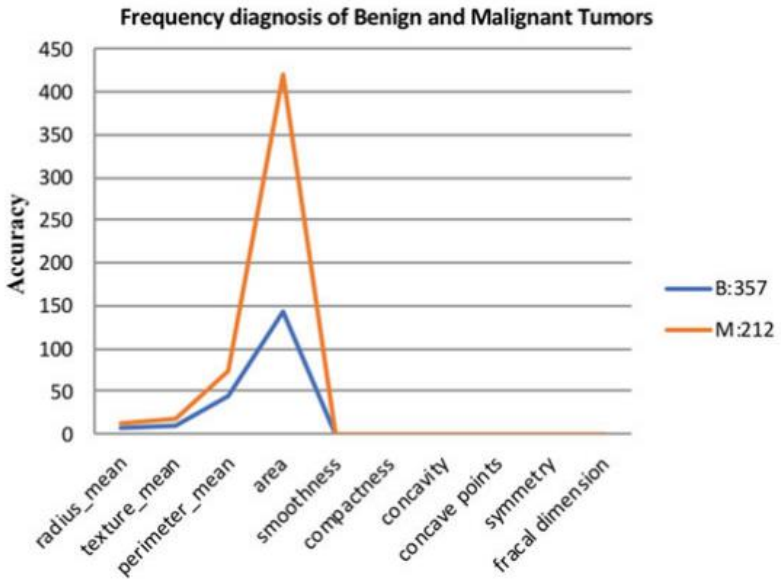
Τελικά, η μέθοδός που προτείνεται είναι η γρηγορότερη μέθοδος από τις τρεις, με το 90% του χρόνου συμπεράσματος να δαπανάται για την αλλαγή μεγέθους της τρέχουσας οριοθετημένης τομής στο μέγεθος εισόδου (100 × 100 × 50). Θεωρείται ότι αυτή η μειωμένη διακύμανση οφείλεται στην ικανότητα του μοντέλου να επικεντρώνεται στις πιο σημαντικές περιοχές, ενώ η προσέγγιση DL-MSC δεν είναι τόσο αποτελεσματική στην παράβλεψη των θορύβων, αυξάνοντας το χρόνο ανάλυσης που απαιτεί. Ως εκ τούτου, το μοντέλο φαίνεται να είναι πιο ισχυρό σε θορυβώδεις εισόδους. Τέλος, ο λιγότερος χρόνος συμπεράσματος καθιστά το συγκεκριμένο μοντέλο καταλληλότερο για ανάπτυξη σε κλινικές πρακτικές [44].

5.6. Μέθοδος GoogLeNet και AlexNet

Ο A. Sharma και η ερευνητική του ομάδα απέδειξαν πως οι τεχνικές GoogLeNet και AlexNet, οι οποίες αποτελούν CNN άρα και αλγορίθμους BM, απέδωσαν πολύ καλύτερα συγκριτικά με τις τεχνικές LR και SVM για την ανίχνευση του καρκίνου του μαστού. Στη συνέχεια, τα αποτελέσματα που προέκυψαν από τις τεχνικές BM, συγκρίθηκαν με τεχνικές LR και SVM [31].

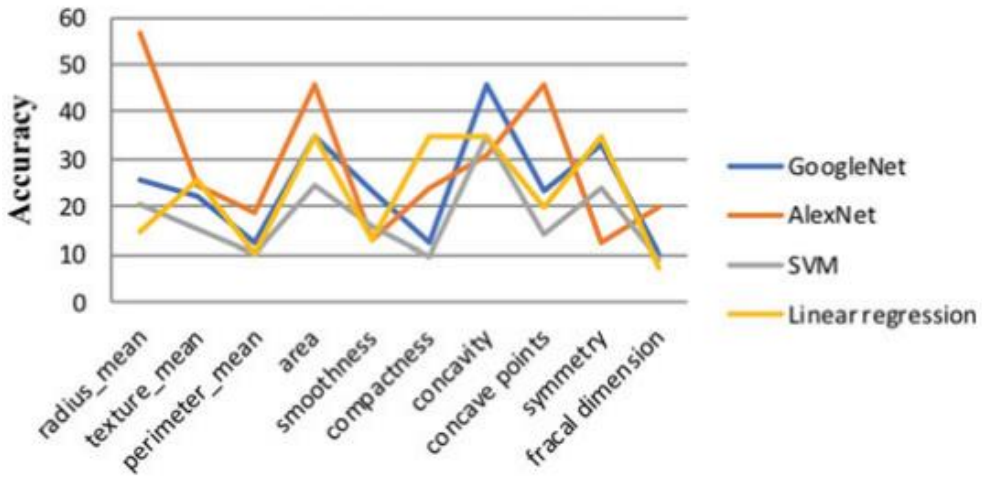
Το σύνολο δεδομένων που χρησιμοποιήθηκε ήταν το WDBC. Τα δείγματα εικόνων από το WDBC αναδιαμορφώθηκαν, προκειμένου να μπορέσουν να τροφοδοτήσουν τα μοντέλα. Τα αποτελέσματα των εικόνων ορίστηκαν σε διαφορετικούς μεγεθυντικούς παράγοντες όπως 40x, 100x και 200x. Η ανάλυση εικόνας έγινε από τεχνικές BM αλλά και τεχνικές MM. Το μοντέλο γραμμικής

παλινδρόμησης εφαρμόζεται για να προβλέψει το accuracy, αφού καθορίσει τα τμήματα του δείγματος που έχουν επηρεαστεί από τον όγκο. Προκειμένου να βελτιωθεί τυχόν υψηλό ποσοστό εσφαλμένης ταξινόμησης που δημιουργείται από ένα μόνο χαρακτηριστικό, δηλαδή η ένταση, γίνεται προσθήκη περισσότερων χαρακτηριστικών, όπως ο μέσος όρος του παραθύρου που εφαρμόζεται και η τυπική απόκλιση. Με αυτόν τον τρόπο, η τμηματοποίηση που δημιουργείται από διάφορους κανόνες, έχει επίσης βελτιωθεί [31].



Εικόνα 34. Συχνότητα διάγνωσης καλοηθών και κακοηθών όγκων [31].

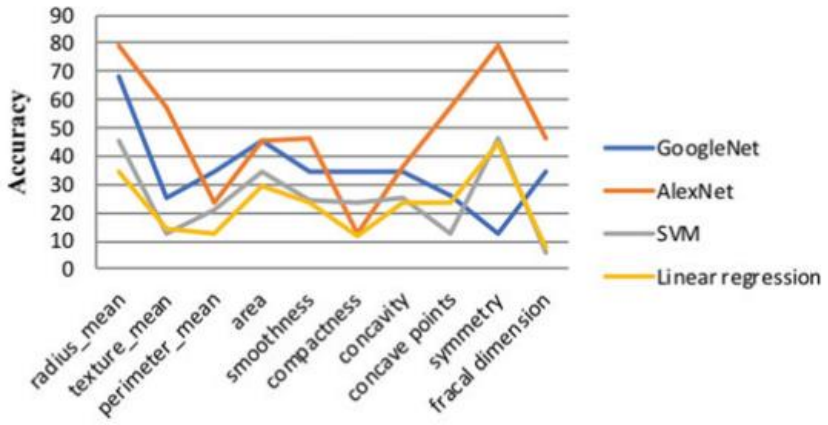
Ο Πίνακας 18, ο Πίνακας 19 και ο Πίνακας 20 παρουσιάζουν οπτικοποιημένα αποτελέσματα της εικόνας, που ορίζεται σε διαφορετικούς μεγεθυντικούς παράγοντες. Επίσης, στην Εικόνα 35, Εικόνα 36, Εικόνα 37 συγκρίνεται το accuracy μεταξύ των τεχνικών CNN και ΜΜ. Σε όλους τους πίνακες, φαίνεται ότι η βαθιά μάθηση αποδίδει πολύ καλύτερα, συγκριτικά με τα αποτελέσματα της ΜΜ στη διάγνωση του καρκίνου του μαστού [31].



Εικόνα 35. CNN και τεχνικές ΜΜ για 40x μεγεθυντικό παράγοντα εικόνας[31].

Μέθοδος	GoogLeNet	AlexNet	SVM	Logistic Regression
Radius_mean	25.7	56.34	20.34	14.87
Texture_mean	22.3	24.65	15.23	25.9
Perimeter_mean	12.5	18.98	10.1	10.12
Area	34.7	45.98	24.8	34.98
Smoothness	23.5	12.87	15.76	12.98
Compactness	12.6	23.8	9.8	34.98
Concavity	45.76	30.98	34.87	34.8
Concave points	23.6	45.98	14.3	20
Symmetry	33.4	12.43	23.76	34.8
Fractal_Dimension	10.22	19.87	8.9	7.6

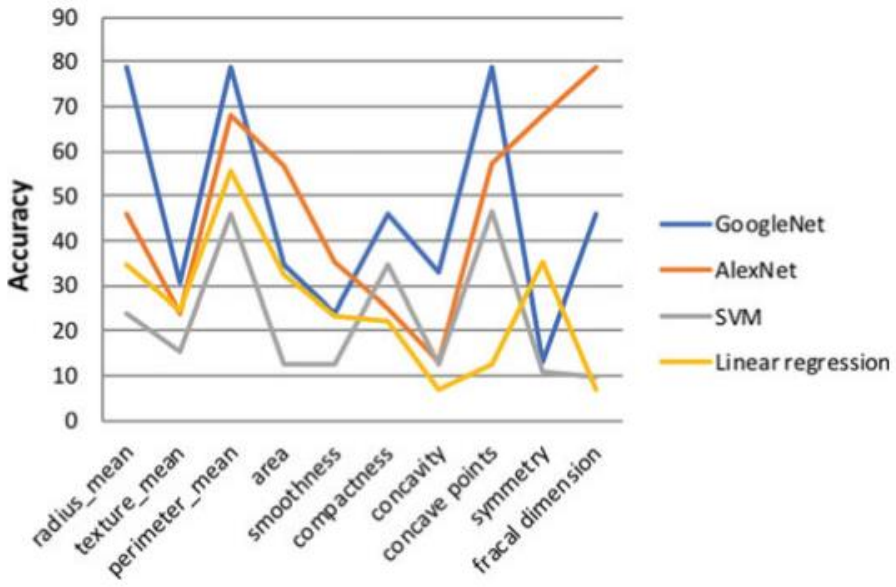
Πίνακας 18. Ταξινόμηση με βάση 40x μεγεθυντικό παράγοντα εικόνας [31].



Εικόνα 36. CNN και τεχνικές ΜΜ για 100x μεγεθυντικό παράγοντα εικόνας[31].

Μέθοδος	GoogLeNet	AlexNet	SVM	Logistic Regression
Radius_mean	67.9	78.9	45.6	34.67
Texture_mean	25.6	56.89	12.4	13.9
Perimeter_mean	34.5	23.6	21.4	12.5
Area	45.8	45.7	34.8	29.8
Smoothness	34.6	45.89	24.6	23.9
Compactness	34.8	12.9	23.9	11.4
Concavity	34.9	35.9	24.9	23.5
Concave points	25.9	56.9	12.9	23.9
Symmetry	12.9	78.9	45.9	45
Fractal_Dimension	34.9	45.9	5.9	7.9

Πίνακας 19. Ταξινόμηση με βάση 100x μεγεθυντικό παράγοντα εικόνας [31].



Εικόνα 37. CNN και τεχνικές ΜΜ για 200x μεγεθυντικό παράγοντα εικόνας[31].

Μέθοδος	GoogLeNet	AlexNet	SVM	Logistic Regression
Radius_mean	78.9	45.7	23.7	34.6
Texture_mean	30.8	23.7	15.6	24.6
Perimeter_mean	78.9	67.9	45.8	55.6
Area	34.8	56.5	12.5	32.4
Smoothness	23.9	34.9	12.4	23.5
Compactness	45.9	24.9	34.5	22.4
Concavity	32.9	12.9	12.8	6.7
Concave points	78.9	56.97	46.5	12.3
Symmetry	12.9	67.9	10.6	34.89
Fractal_Dimension	45.8	78.9	9.7	6.9

Πίνακας 20. Ταξινόμηση με βάση 200x μεγεθυντικό παράγοντα εικόνας [31].

5.7. Μέθοδος Multi-layer Perceptron με 10-Fold Cross Validation

Το 2018 πραγματοποιήθηκε μία έρευνα στο τμήμα Μηχανικών Πληροφορικής και Τεχνολογίας Πληροφορικής, στο Ινστιτούτο Πληροφορικής Jaypee στην Νόιντα της Ινδίας. Οι Madhuri και Bharat Gupta παρουσίασαν μια επισκόπηση εφαρμογής των αλγορίθμων ΜΜ. Σκοπός ήταν και εδώ η διάγνωση του καρκίνου του μαστού, χωρίζοντας τα δεδομένα σε καρκινικά και μη-καρκινικά. Οι ταξινομητές ΜΜ που χρησιμοποιήθηκαν ήταν K-NN, SVM, Multi-layer Perceptron (MP) και τα DTs. Η αξιολόγηση των μοντέλων πραγματοποιήθηκε υπολογίζοντας το Precision, το accuracy, την ανάκληση και το R^2 [40].

Το σύνολο δεδομένων που χρησιμοποιήθηκε ήταν και εδώ το WBCD. Κάθε δείγμα του συνόλου αποδίδεται σε ένα διάνυσμα 9 διαστάσεων, ενώ κάθε διάνυσμα παίρνει τιμές μεταξύ των τιμών 1-10, με την τιμή 1 να δείχνει την

κανονική κατάσταση και την τιμή 10 να δείχνει την πιο ανώμαλη κατάσταση. Τα χαρακτηριστικά των δειγμάτων (Πίνακας *3*) υπολογίζονται για κάθε εικόνα. Η μέση τιμή, το τυπικό σφάλμα και το "χειρότερο" (μέση τιμή των τριών μεγαλύτερων τιμών) υπολογίζονται για όλα αυτά τα 10 χαρακτηριστικά, γεγονός που έχει ως αποτέλεσμα 30 τιμές. Τα δεδομένα διαχωρίζονται σε δεδομένα εκπαίδευσης και δοκιμών κατά 70% και 30%, αντίστοιχα. Κατά τη διαδικασία της προ-επεξεργασίας δεδομένων καταργήθηκαν οι στήλες των σημάνσεων και αναγνωριστικών από το σύνολο δεδομένων και υπολογίστηκε η μέση τιμή και η τυπική απόκλιση κάθε χαρακτηριστικού [40].

Κατά την εξαγωγή χαρακτηριστικών εξετάστηκε η μέση τιμή κάθε χαρακτηριστικού και, στη συνέχεια, εκτελέστηκε ανάλυση διασποράς με έναν παράγοντα (one-way ANalysis Of VAriance, ANOVA). Τα χαρακτηριστικά εγκαθίστανται με αυξανόμενη σειρά, βάση τη διαφορά μεταξύ των μέσων τιμών απόδοσης. Εδώ, το accuracy αποδίδεται σε όλα τα χαρακτηριστικά και στους πέντε καλύτερους προγνωστικούς παράγοντες [40].

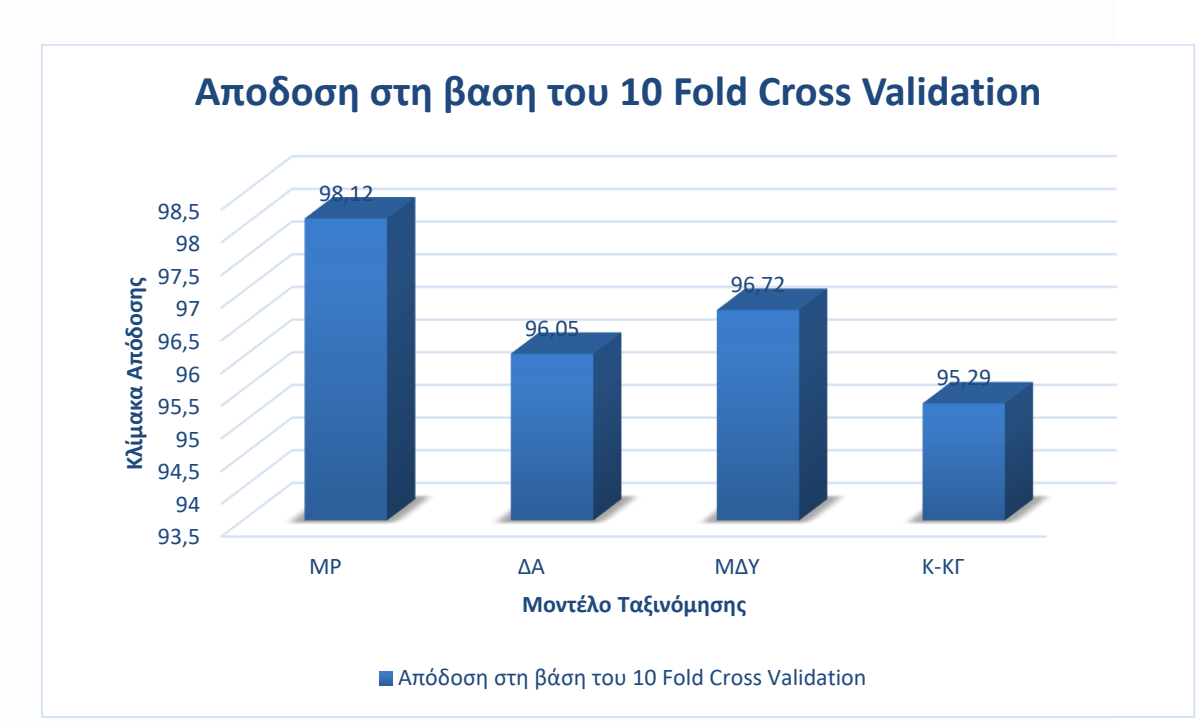
Τα αποτελέσματα του πειράματος φαίνονται στον Πίνακας 21, σύμφωνα με τον οποίο ο MP λειτουργεί καλύτερα από τα υπόλοιπα μοντέλα, όταν εφαρμόζονται όλοι οι προγνωστικοί παράγοντες και χρησιμοποιούνται 7,2 κρυφά στρώματα. Το DT έχει κορεστεί πιθανότατα από το μεγάλο μέγεθος των προγνωστικών παραγόντων (*). Γι' αυτό το λόγο, εφαρμόζονται οι 5 καλύτεροι παράγοντες για την εκπαίδευση του μοντέλου παρουσιάζοντας πολύ καλά αποτελέσματα στο accuracy. Το accuracy του MP, με 7,2 κρυφά επίπεδα, είναι στο 97,2, δηλαδή υψηλότερη από των SVM (93,9), του K-NN (94,4) και του DT (κορεσμένο) για όλους τους προγνωστικούς παράγοντες. Το Precision, η ανάκληση και R^2 του MP είναι καλύτερα και ακολουθούν τα αποτελέσματα των SVM, K-NN και DT. Συγκριτικά, ο MP αποδίδει καλύτερα από τις SVM, το DT και τον K-NN. Για την εκτίμηση της απόδοσης του μοντέλου MM μέσω του συνόλου δεδομένων επικύρωσης, πραγματοποιείται μια δοκιμή απόδοσης στη βάση του "10 Fold Cross Validation" [40].

Αλγόριθμος Ταξινόμησης	K-NN		SVM		DT		MP	
Αριθμός Προγνωστικών Παραγόντων	**Όλα**	**Καλύτερα 5**	**Όλα**	**Καλύτερα 5**	**Όλα**	**Καλύτερα 5**	**5,2**	**7,2**
Accuracy	94.4	90.6	93.9	92.7	100*	96.9	90.9	**97.7**
Precision	100%		97.21		100%		**99.2%**	
Ανάκληση	95.11%		96.67%		97.22%		**97.85%**	
R^2	0.6		0.7		1.0		**0.8**	

Πίνακας 21. Απόδοση των τεχνικών MM[40].

Ο Πίνακας 22 παρουσιάζει τη συγκριτική απόδοση των τεσσάρων τεχνικών ξεχωριστά, με βάση τη "10 Fold Cross Validation". Η απόδοση του MP είναι 98,12%, το DT είναι 96,05%, οι SVMs είναι 96,19% και ο K-NN 94,29%. Σύμφωνα με το

γράφημα, το μοντέλο MP αποδίδει καλύτερα μεταξύ όλων των άλλων μοντέλων με βάση τη "10 Fold Cross Validation". Τελικά, η απόδοση, όσον αφορά το accuracy, είναι καλύτερη με τη χρήση του MP συγκριτικά με άλλες τεχνικές και αποδίδει επίσης καλύτερα από άλλες τεχνικές, όταν χρησιμοποιούνται μετρήσεις cross validation στην πρόβλεψη για τον καρκίνου του μαστού [40].



Πίνακας 22. Απόδοση των μοντέλων ΜΜ στη βάση του 10-Fold Cross Validation.

5.8. Μέθοδος Extreme Learning Machine

Οι M. Aslan, Y. Celik, K. Sabanci, A. Durdu στην Τουρκία το 2018 πραγματοποίησαν μια έρευνα, στόχος της οποίας ήταν να επεξεργαστεί τα αποτελέσματα μίας απλής ανάλυσης αίματος ρουτίνας. Η μελέτη αυτή πραγματοποιήθηκε με χρήση τεσσάρων διαφορετικών μεθόδων ΜΜ, τα ANN, Μηχανές Ακραίας Μάθησης (Extreme Learning Machines - ELM), SVM και K-NN. Το σύνολο δεδομένων που χρησιμοποιήθηκε ήταν το WCCD [38].

Λαμβάνοντας υπόψη τις τιμές εισόδου, οι μέγιστες και ελάχιστες από αυτές τις τιμές είναι αρκετά διαφορετικές μεταξύ τους. Επομένως, η κανονικοποίηση, με χρήση της μεθόδου Feature Scaling, είναι απαραίτητη προκειμένου να αυξηθεί το ποσοστό επιτυχίας. Αφού γίνει κανονικοποίηση, τα δεδομένα χωρίζονται τυχαία σε δεδομένα εκπαίδευσης και δοκιμής κατά 80% και 20%, αντίστοιχα, και στη συνέχεια λαμβάνονται τα διάφορα αποτελέσματα που προέκυψαν από τις διάφορες τεχνικές ΜΜ που χρησιμοποιήθηκαν [38].

Στη μέθοδο ANN, υπάρχουν διάφοροι υπερ-παράμετροι που επηρεάζουν την ακρίβεια του συστήματος. Οι σημαντικές παράμετροι μπορούν να καταχωρηθούν

ως Αριθμός Κρυφού Νευρώνα Στρώματος, Αριθμός Epoch, Ποσοστό Μάθησης και Συντελεστής Ορμής, οι οποίες πρέπει να καθορίζονται βασιζόμενες σε δοκιμές και σφάλματα, προκειμένου να επιτευχθεί το βέλτιστο αποτέλεσμα. Για το λόγο αυτό, ένα συγκεκριμένο εύρος αυτών των παραμέτρων μπορεί να προσαρμοστεί από το χρήστη. Τα γραφήματα δίνουν τιμές Μέσης Τετραγωνικής Ρίζας Σφάλματος (Root Mean Square Error - RMSE) σύμφωνα με τις μεταβαλλόμενες τιμές παραμέτρων. Τα αποτελέσματα των τιμών RMSE σχεδιάζονται σύμφωνα με την τροποποιημένη παράμετρο και καταγράφηκαν οι παράμετροι με το ελάχιστο σφάλμα. Μετά από αυτό, η διαδικασία εκπαίδευσης και δοκιμής διαχειρίστηκε χρησιμοποιώντας τις καλύτερες παραμέτρους. Τα αποτελέσματα που πήραν για το μέσο ποσοστό του Accuracy της δοκιμής ήταν 79,4304%, ο μέσος χρόνος εκπαίδευσης 0,4282 δευτερόλεπτα και η μέση τιμή RMSE υπολογίστηκε στο 0,3954 [38].

Στην μέθοδο ELM, η υπέρ-παράμετρος που επηρεάζει την ακρίβεια του συστήματος είναι ο αριθμός των κρυφών νευρώνων στρώματος, ο οποίος αλλάζει μέσα σε ένα ορισμένο εύρος, το οποίο προσδιορίζεται από το χρήστη, για να επιτευχθεί το βέλτιστο αποτέλεσμα με την ELM. Οι τιμές RMSE σχεδιάζονται σύμφωνα με την τροποποιημένη παράμετρο. Ο καλύτερος αριθμός κρυφών στρώσεων νευρώνα που υπολογίστηκε ήταν 1800. Το μέσο ποσοστό ακρίβειας της δοκιμής που υπολογίστηκε ήταν 80%, ο μέσος χρόνος εκπαίδευσης είναι 0,0075 δεύτερος και η μέση τιμή RMSE επιτεύχθηκε ως 0,4755. Ο Πίνακας 23 δείχνει ότι οι τιμές του Accuracy της ANN και της ELM βρίσκονται κοντά η μία στην άλλη, αλλά η ELM είναι πολύ πιο γρήγορη από την ANN. Όταν ο αριθμός των δειγμάτων εκπαίδευσης είναι πολύ υψηλός, η χρήση ELM συμφέρει πολύ περισσότερο από άποψη χρόνου [38].

Η μέθοδος υπερ-παραμετρικής βελτιστοποίησης χρησιμοποιείται επίσης για ταξινόμηση με χρήση του K-NN. Αυτές οι παράμετροι μπορούν να θεωρηθούν ως ο αριθμός των γειτόνων και η απόσταση για τον K-NN. Το μέσο ποσοστό του Accuracy υπολογίστηκε στο 77,5%, χρησιμοποιώντας τις καλύτερες παραμέτρους, τα δεδομένα εκπαίδευσης ταξινομήθηκαν στα 0,15781 δευτερόλεπτα [38].

Η ίδια μέθοδος χρησιμοποιείται επίσης για ταξινόμηση με χρήση του αλγορίθμου SVM. Οι υπερ-παράμετροι του SVM μπορούν να θεωρηθούν ως σταθερά κανονικοποίησης (C) και κλίμακα kernel. Η βέλτιστη τιμή κλίμακας kernel βρέθηκε 0,0287, η βέλτιστη τιμή C 0,4869, το μέσο ποσοστό Accuracy επιτεύχθηκε στο 73,5% και, με χρήση των καλύτερων παραμέτρων, τα δεδομένα εκπαίδευσης ταξινομήθηκαν σε 0,1866 δευτερόλεπτα [38].

Αριθμός Δεδομέ-νου	ELM				ANN			
	Εκπαίδευση		Δοκιμή		Εκπαίδευση		Δοκιμή	
	Accuracy (%)	Χρόνος Εκπαί-δευσης	Accuracy (%)	RMSE	Accuracy (%)	Χρόνος Εκπαί-δευσης	Accuracy (%)	RMSE
1	83.8710	0.0073	78.2609	0.4802	76.3441	0.4620	78.2610	0.3618
2	83.8710	0.0136	82.6087	0.4706	70.9677	0.3797	78.2610	0.3952
3	83.8710	0.0060	82.6087	0.4631	80.6452	0.3864	73.9130	0.3858
4	83.8710	0.0079	73.9130	0.4863	75.2688	0.5253	69.5652	0.4961
5	83.8710	0.0060	78.2610	0.4817	74.1935	0.3666	82.6087	0.3820
6	83.8710	0.0063	82.6087	0.4814	83.8710	0.4406	82.9087	0.3932
7	84.9462	0.0069	78.2610	0.4634	82.7957	0.4107	89.9565	0.3527
8	84.9462	0.0072	82.6087	0.4820	80.6452	0.4761	82.6087	0.3733
9	83.8710	0.0068	82.6087	0.4701	74.1935	0.4105	73.913	0.4108
10	81.7204	0.0073	78.2610	0.4759	78.4946	0.4244	82.6087	0.4034
MO	83.8710	0.0075	80.00	0.4755	77.7419	0.4282	79.4304	0.3954

Πίνακας 23. Σύγκριση μεθόδων ANN-ELM για 10 δεδομένα[38]

Τα ποσοστά Accuracy και οι χρόνοι εκπαίδευσης που επιτεύχθηκαν παρουσιάζονται στον Πίνακας 24. Η μέθοδος K-NN δεν περιέχει στην πραγματικότητα τη φάση εκπαίδευσης, επομένως η τιμή που βλέπουμε στο αντίστοιχο πεδίο του Πίνακας 24 αντιπροσωπεύει την περίοδο υπολογισμού των δεδομένων εκπαίδευσης. Όταν εξετάζονται οι τιμές του Πίνακας 24, το υψηλότερο ποσοστό Accuracy και η χαμηλότερη περίοδος εκπαίδευσης παρέχονται από τον ELM. Σύμφωνα με αυτά τα αποτελέσματα, η χρήση των ELM είναι αυτή που συμφέρει περισσότερο όταν υπάρχει μεγάλος αριθμός δειγμάτων από άποψη χρόνου [38].

Αλγόριθμος	ANN	ELM	K-NN	SVM
Ποσοστό Accuracy (%)	79.4304	80	77.5	73.5
Χρόνος Εκπαίδευσης (sec)	0.4282	0.0075	0.15781	0.1866

Πίνακας 24. Σύγκριση τεχνικών ANN, ELM, K-NN και SVM με WCCD [38].

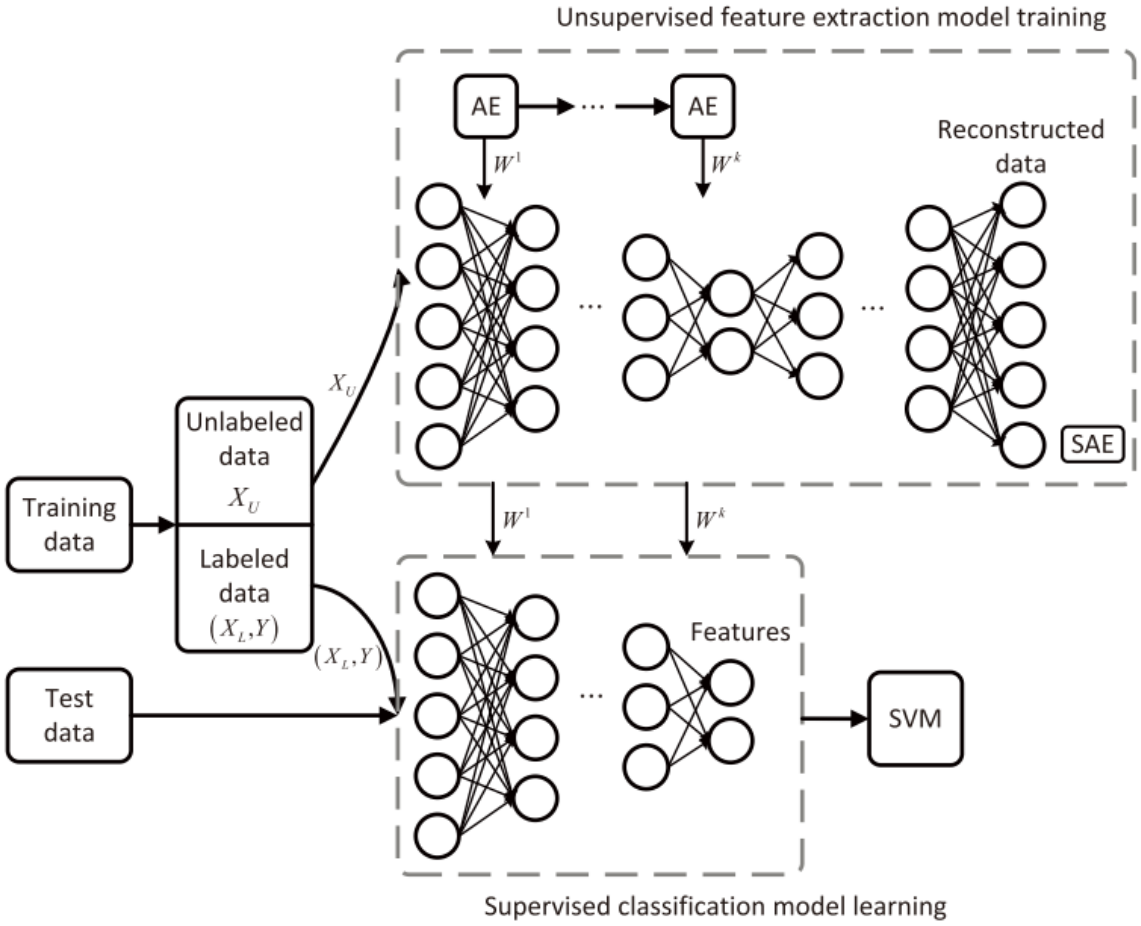
5.9. Μέθοδος Stacked Auto-Encoders με Μηχανές Διανυσμάτων Υποστήριξης

Το 2018 στην Γουχάν της Κίνας πραγματοποιήθηκε μια έρευνα από τους Y. Xiao, J. Wu, Z. Lin, X. Zhao και προτάθηκε ένα σύστημα εξαγωγής χαρακτηριστικών που βασίζεται στη βαθιά μάθηση χωρίς επίβλεψη. Το σύστημα αυτό ενσωματώνει

Στοιβαγμένους Αυτόματους Κωδικοποιητές με Μηχανές Διανυσμάτων Υποστήριξης (Stacked Autoencoders, SAE-SVM) για τη διάγνωση του καρκίνου του μαστού. Το σύνολο δεδομένων που χρησιμοποιήθηκε για την επικύρωση της μεθόδους, ήταν το WDBC. Επιπλέον, έγινε χρήση 10 Fold Cross Validation οι καμπύλες precision-recall (PR) και ROC χρησιμοποιήθηκαν για τη μέτρηση των επιδόσεων του προτεινόμενου SAE-SVM [32].

Η διάγνωση του καρκίνου του μαστού, με βάση το SAE-SVM, στοχεύει στην εξαγωγή των βασικών πληροφοριών των αρχικών δεδομένων και στην αντιμετώπιση των δεδομένων χωρίς σήμανση, που λαμβάνονται από την ανίχνευση καρκίνου του μαστού. Η βασική δομή του μοντέλου (*Εικόνα 38*) αποτελείται από δύο στάδια, το μη-επιβλεπόμενο στάδιο προ-εκπαίδευσης βασιζόμενο σε αυτόματους κωδικοποιητές (ΑΚ) και το επιβλεπόμενο στάδιο βασιζόμενο σε Μηχανή Διανυσμάτων Υποστήριξης. Στο πρώτο στάδιο, η παράμετρος $W^i = 1, 2, \ldots, k$, εκπαιδεύεται από τον i ΑΚ μονού επιπέδου και δηλώνει τα βάρη του i-επιπέδου του SAE. Όλα τα επίπεδα και οι παράμετροι, που εκπαιδεύονται από τους k ΑΚ, συνδυάζονται για να σχηματίσουν το βαθύ SAE, το οποίο εξάγει χαρακτηριστικά για την τελική ταξινόμηση στο SVM. Στο δεύτερο στάδιο, τα δεδομένα εκπαίδευσης για το μοντέλο SAE-SVM περιλαμβάνουν δεδομένα X_U χωρίς σήμανση και δεδομένα με σήμανση (X_L, Y). Τα μη επισημασμένα δεδομένα X_U χρησιμοποιούνται για την εκπαίδευση του SAE και το επισημασμένο ζεύγος (X_L, Y) χρησιμοποιείται για την εφαρμογή της ταξινόμησης στο SVM. Τα δεδομένα δοκιμής χρησιμοποιούνται για την αξιολόγηση της απόδοσης του μοντέλου[32].

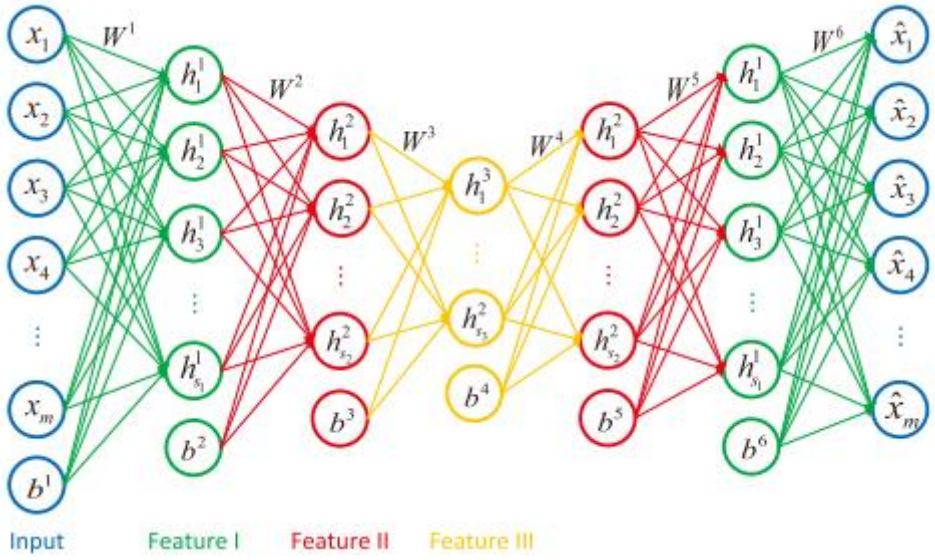
Στόχος είναι η ταξινόμηση ενός νέου όγκου σε καλοήθη ή κακοήθη. Προκειμένου να διευκολυνθεί η ταξινόμηση, χρησιμοποιείται η στρατηγική των ΑΕ. Στη διαδικασία της προ-εκπαίδευσης, μαθαίνονται αρκετοί ΑΕ και πολλά επίπεδα στοιβάζονται. Ένας SAE αποτελείται από πολλαπλά επίπεδα ΑΕ, όπου η έξοδος κάθε επιπέδου χρησιμοποιείται ως είσοδος του επόμενου επιπέδου. Για την προ-εκπαίδευση ενός βαθύ SAE, χρησιμοποιείται η προσέγγιση greedy layer-wise. Ένα SAE με n_l επίπεδα, στην greedy layer-wise προσέγγιση, χρησιμοποιούνται $(n_l - 1)/2$ ΑΕ για την προ-εκπαίδευση του SAE. Η *Εικόνα 29* απεικονίζει την αρχιτεκτονική του ΑΕ i με ένα δείγμα x που αποτελείται από m χαρακτηριστικά. Η παράμετρος W_i δηλώνει το βάρος μεταξύ της σύνδεσης των μονάδων στο επίπεδο εισόδου και των μονάδων στο κρυφό επίπεδο του i ΑΕ και το βάρος του i επιπέδου του SAE. Η $W^{n_l - i}$ δηλώνει το βάρος μεταξύ της σύνδεσης των μονάδων του κρυφού επιπέδου και των μονάδων του επιπέδου εξόδου του i ΑΕ και το βάρος του επιπέδου $n_l - i$ του SAE.



Εικόνα 38. Διάγραμμα μοντέλου SAE-SVM[32].

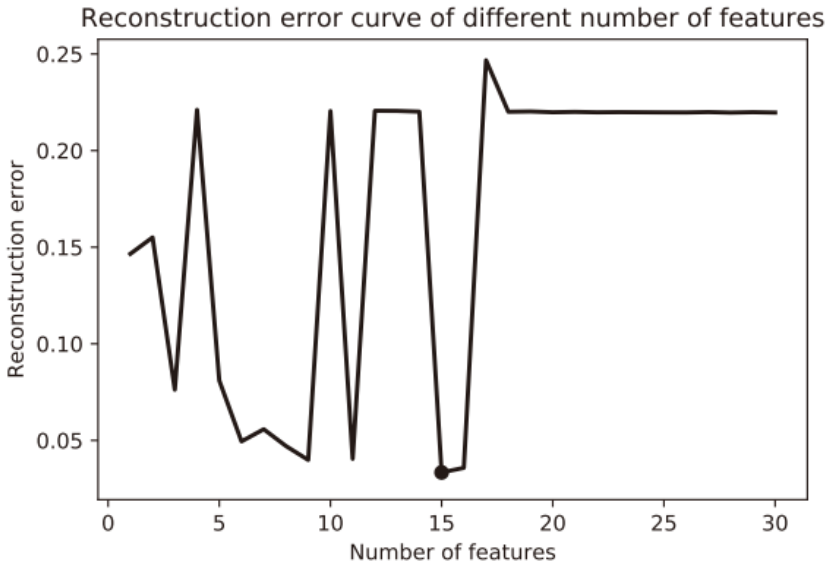
Η παράμετρος b^i υποδηλώνει την απόκλιση των μονάδων στο κρυφό επίπεδο του i AE και την απόκλιση του $(i+1)$ επιπέδου του SAE, και b^{n_l-i} δηλώνει την απόκλιση των μονάδων στο επίπεδο εξόδου του i AE και την απόκλιση του (n_l-i+1) επιπέδου του SAE. Η σταθερά $n_l = 2k+1$ είναι ο αριθμός των επιπέδων του SAE. Η αρχιτεκτονική ενός SAE εμφανίζεται στην *Εικόνα 39*, για την περίπτωση όπου στοιβάζονται τρεις AE. Στην εκπαίδευση greedy layer-wise, ο πρώτος AE εκπαιδεύεται στην είσοδο, προκειμένου να αποκτήσει τις παραμέτρους $W^1, W^{n_l-1}, b^1, b^{n_l-1}$ και τα κύρια χαρακτηριστικά $h_i^1, i = 1, 2, \ldots, s_1$ όπου s_1 είναι ο αριθμός των μονάδων του κρυμμένου επιπέδου. Στη συνέχεια, η κύρια αναπαράσταση χαρακτηριστικών τροφοδοτείται στον δεύτερο AE ως είσοδος για να μάθουν τα δευτερεύοντα χαρακτηριστικά $h_i^2, i = 1, 2, \ldots, s_1$ και τις αντίστοιχες παραμέτρους $W^2, W^{n_l-2}, b^2, b^{n_l-2}$. Στη συνέχεια, η διαδικασία επαναλαμβάνεται μέχρι τον τελευταίο AE. Η στοίβα του κάθε AE συνθέτει το βαθύ SAE με μια σειρά παραμέτρων που έχουν προ-εκπαιδευτεί από την προσέγγιση greedy layer-wise. Προκειμένου να επιτευχθούν καλύτερα αποτελέσματα, μετά τη διαδικασία προ-εκπαίδευσης, εφαρμόζεται η τεχνική back-propagation για να τελειοποιηθούν όλοι

οι παράμετροι του SAE. Το σφάλμα μεταξύ της εισόδου και της ανακατασκευασμένης εξόδου χρησιμοποιείται για την ενημέρωση των παραμέτρων. Μετά την εξαγωγή χαρακτηριστικών, τα νέα σύνολα δεδομένων χρησιμοποιήθηκαν ως είσοδοι στον αλγόριθμο ταξινόμησης SVM, για να βρεθεί το μέγιστο υπερ-επίπεδο με το μέγιστο περιθώριο [32].



Εικόνα 39. Αρχιτεκτονική ενός SAE[32].

Για να βρούμε τον κατάλληλο αριθμό χαρακτηριστικών που μειώθηκαν από το ΣΑΕ, υπολογίσαμε το σφάλμα ανακατασκευής, το οποίο ήταν το σφάλμα μεταξύ των ανακατασκευασμένων δεδομένων και των αρχικών δεδομένων του SAE. Από την καμπύλη σφάλματος ανακατασκευής διαφορετικών αριθμών χαρακτηριστικών, που απεικονίζονται στην Εικόνα *40*, το 15 επιλέχθηκε για τον αριθμό των χαρακτηριστικών, καθώς αντιστοιχούσε στο τοπικό ελάχιστο στην περιοχή από 2 έως 30. Η έξοδος του μεσαίου επιπέδου του ΣΑΕ, το οποίο αποτελούνταν από 15 αναπαραστάσεις, συμβόλιζε τα 15 νέα χαρακτηριστικά. Στην εποπτευόμενη ταξινόμηση, τα δείγματα πρώτα μετατράπηκαν σε νέα σύνολα δεδομένων με 15 χαρακτηριστικά και στη συνέχεια ταξινομήθηκαν σε κακοήθεις και καλοήθεις όγκους με υψηλότερη ακρίβεια [32].



Εικόνα 40. Καμπύλη σφάλματος ανακατασκευής διαφορετικών αριθμών χαρακτηριστικών [32].

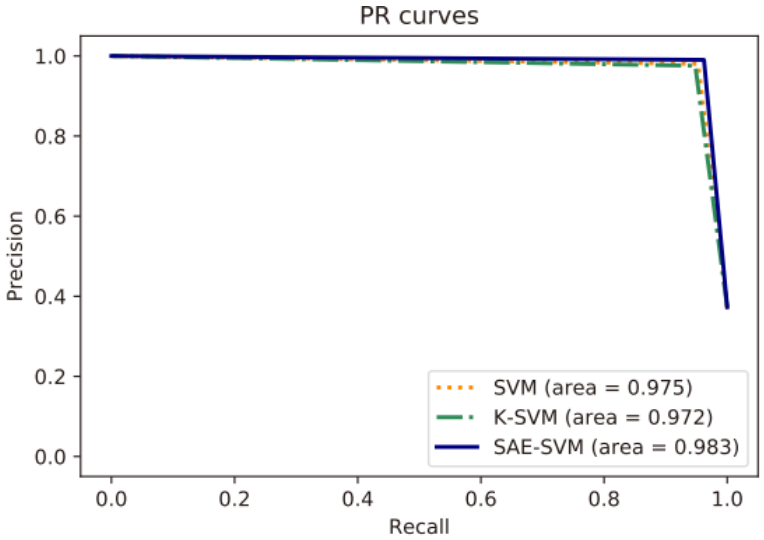
Για τη φάση ταξινόμησης, εφαρμόστηκε το μοντέλο SVM με συνάρτηση Linear Kernel και συγκρίθηκε με προηγούμενα πειραματικά αποτελέσματα που είχαν μελετηθεί από τους Prasad, Zheng, Ramadevi και τις ερευνητικές τους ομάδες, τα διαγνωστικά αποτελέσματα των οποίων παρουσιάζονται στον Πίνακας *25*.

	Accuracy (%)	**Διάσταση Χαρακτηριστικών**
ACO-SVM	95.96	15
GA-SVM	97.19	18
PSO-SVM	97.37	17
K-SVM	97.38	6
PCA-SVM	96.84	-
SAE-SVM	98.25	15

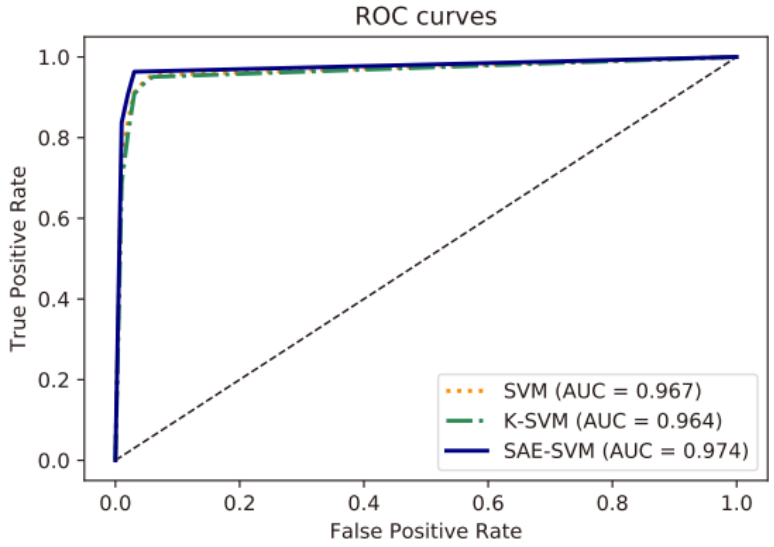
Πίνακας 25. Σύγκριση διάφορων μεθόδων με βάση το Accuracy [32].

Παρατηρούμε ότι η προτεινόμενη μέθοδος SAE-SVM είχε ως αποτέλεσμα το υψηλότερο Accuracy με 98,25%. Οι προτεινόμενες μέθοδοι του Prasad απαιτούν την εκτέλεση ολόκληρης της διαδικασίας, συμπεριλαμβανομένης της εκπαίδευσης και της δοκιμής της ταξινόμησης, για την επιλογή του καλύτερου συνδυασμού χαρακτηριστικών και, άρα, τα μοντέλα ACO-SVM, GA-SVM και PSO-SVM κατανάλωσαν μεγάλο υπολογιστικό χρόνο. Συγκρίνοντας τα K-SVM, PCA-SVM και SAE-SVM έσωσαν σημαντικό χρόνο εκπαίδευσης, με τη διατήρηση των μειωμένων πληροφοριών χαρακτηριστικών γνωρισμάτων. Επιπλέον, συγκρίνοντας τα αποτελέσματα του Zheng και της ομάδας του, η υψηλότερη ακρίβεια που προκύπτει από το SAE-SVM δείχνει ότι έχουν διατηρηθεί ορισμένα απαραίτητα χαρακτηριστικά. Επιπλέον, συγκρίνοντας τον PCA με το SAE-SVM , ο δεύτερος είχε μεγαλύτερο Accuracy στα αποτελέσματά του λόγω του περιορισμού του PCA που πραγματοποίησε μόνο γραμμικό μετασχηματισμό στα εισερχόμενα δεδομένα [32].

Στις καμπύλες PR (Εικόνα 41. Οι PR καμπύλες για τους SVM, K-SVM και SAE-SVM [32].) παρατηρούμε ότι η μέθοδος SAE-SVM καλύπτει μεγαλύτερη περιοχή από την K-SVM και την κλασσική SVM, γεγονός που δείχνει ότι το SAE-SVM είχε καλύτερο αποτέλεσμα ταξινόμησης των δειγμάτων. Οι καμπύλες ROC (Εικόνα *42.* Οι καμπύλες ROC για τους SVM, K-SVM και SAE-SVM [32].) δείχνουν ότι το SAE-SVM ήταν καλύτερο από το K-SVM και το SVM, κρίνοντας από την υψηλότερη βαθμολογία AUC.



Εικόνα 41. Οι PR καμπύλες για τους SVM, K-SVM και SAE-SVM [32].



Εικόνα 42. Οι καμπύλες ROC για τους SVM, K-SVM και SAE-SVM [32].

Με βάση τα αποτελέσματα, η προτεινόμενη μέθοδος SAE-SVM αποδίδει καλύτερα από άλλες μεθόδους για την πρόβλεψη του καρκίνου του μαστού. Μέσω της διαδικασίας εξαγωγής χαρακτηριστικών γνωρισμάτων του SAE-SVM, τα περιττά και θορυβώδη χαρακτηριστικά απορρίπτονται και διατηρούνται οι απαραίτητες πληροφορίες. Επομένως, εκτός από τη διατήρηση της ακρίβειας, μειώνεται σημαντικά και ο χρόνος υπολογισμού. Επίσης, τα δείγματα χωρίς σήμανση επαρκούν για τη μείωση των χαρακτηριστικών γνωρισμάτων στο SAE-SVM. Με τη

δυνατότητα χειρισμού δεδομένων χωρίς σήμανση, το SAE μπορεί να εφαρμοστεί στις μαζικές πληροφορίες που συλλέγονται χωρίς σήμανση. Εκτός αυτού, στο SAE η προσέγγιση greedy layer-wise έχει εφαρμοστεί για την προ-εκπαίδευση του βαθιού νευρικού δικτύου. Αρκετοί βασικοί AEs έχουν εκπαιδευτεί διαδοχικά και τελικά στοιβάζονται σε ένα βαθύ αυτόματο κωδικοποιητή. Στη συνέχεια, μια βελτιωμένη αρχή ενημέρωσης δυναμικής, η οποία αποδίδει καλύτερα σε παραμέτρους υψηλών διαστάσεων και έχει καλύτερο ρυθμό σύγκλισης στα βαθιά δίκτυα, έχει χρησιμοποιηθεί στον αλγόριθμο back-propagation για να τελειοποιήσει όλες τις παραμέτρους του SAE. Γι' αυτό το λόγο, η προτεινόμενη μέθοδος έχει μεγαλύτερη ικανότητα να μαθαίνει αυτόματα κρυμμένες περίπλοκες δομές πίσω από τα δεδομένα καλύτερα, βαθύτερα και ταχύτερα, και έχει μεγάλη χρησιμότητα για την εκμετάλλευση των πλεονεκτημάτων της αύξησης του όγκου των πληροφοριών και των μη επισημασμένων δεδομένων [32].

5.10. Πρόγνωστικά SVM μοντέλα για την ταξινόμηση υποτύπων του καρκίνου του μαστού με χρήση διαφορετικών τύπων δεδομένων

Το 2020, οι M. Ozer, P. Sarica και K. Arga πραγματοποίησαν μια έρευνα κατά την οποία παρουσιάστηκε μια σύντομη ανασκόπηση της ταξινόμησης των διάφορων υποτύπων καρκίνου του μαστού (Breast Cancer Subtypes – BCS) μέσω μοντέλων SVM, συζητήθηκε η δυνατότητα εφαρμογής διαφόρων τύπων δεδομένων σε αυτά τα μοντέλα και συγκρίθηκε το accuracy των σημερινών μοντέλων πρόβλεψης [45].

<u>Προγνωστικά μοντέλα βασιζόμενα σε Omics</u>

Υπάρχουν μοντέλα SVM που αναπτύχθηκαν χρησιμοποιώντας σύνολα δεδομένων x-omics, τα οποία περιλαμβάνουν δεδομένα υψηλής απόδοσης από διαφορετικά βιολογικά επίπεδα (*Πίνακας 26*). Ανάλογα με την ποιότητα των δεδομένων, αυτά τα μοντέλα παρουσιάζουν επαρκείς απαιτήσεις πρόβλεψης [45].

Τα επίπεδα έκφρασης της κωδικοποίησης και των μη κωδικοποιημένων RNAs (noncoding RNAs – ncRNAs) διαφέρουν σε φυσιολογικές συνθήκες και σε συγκεκριμένα στάδια ανάπτυξης και, ως εκ τούτου, χρησιμοποιούνται συχνά για την κατασκευή προγνωστικών μοντέλων που βασίζονται στην αλληλεπίδραση των διαφορών έκφρασης των αγγελιαφόρων RNAs (mRNAs) και ncRNAs. Ομοίως, τα σύνολα δεδομένων mRNA και microRNA (miRNA) που παράγονται με τη χρήση διαφορετικών τεχνολογιών, όπως η αλληλουχία RNA (RNA sequencing – RNA-seq), συγκαταλέγονται μεταξύ των πιο συχνά χρησιμοποιούμενων omics συνόλων δεδομένων, που χρησιμοποιούνται με τεχνικές ΜΜ, συμπεριλαμβανομένου του SVM, για τη διαφοροποίηση των υποτύπων καρκίνων του μαστού (Πίνακας *26*). Συνήθως, οι μέθοδοι SVM εφαρμόζονται χρησιμοποιώντας δύο ή περισσότερες κλάσεις [45].

Προσέγγιση	Τεχνολογία/Δεδομένα	Τύπος SVM	Αναφορά
Transcriptome	RNA-seq	v-SVM SVDD	Sokolov et al. (2016)
	miRNA	SVM with 5-fold CV	Hsu et al. (2012)
	Microarray	SVM	De Ronde et al. (2014)
Methylome	DNA methylation	SVM with 5-fold CV	Flanagan et al. (2010)
Proteome	SILAC	SVM One-vs-Rest	Tyanova et al. (2016)
	2D-DIGE	SVM-LOOCV	Waldemarson et al. (2016)
Pathway	Pathway enrichment analysis	SVM SVM with 20-fold CV	Wu et al. (2017) Graudenzi et al. (2017)
Radiome	MRI	SVM with LOOCV	Sutton et al. (2016)
	Ultrasound	SVM tih 3-fold CV	Guo et al. (2018)
	DCE-MRI	SVM	Agner et al. (2014)
	DWI	SVM	Vidić et al. (2018)
Clinical-pathological	Patient clinical data and tumor characteristics	SVM	Wu et al. (2014)
Biochemical	Raman spectra of lipids, Nucleic acids, and proteins	SVM	Becker-Putsche et al. (2013)

Πίνακας 26. Επιλεγμένες μελέτες με σημαντική συνάφεια με τη διαστρωμάτωση των BCSs με χρήση SVM [45].

2D-DIGE – 2-Dimensional-Differential In Gel-Electrophoresis, LOOCV – Leave-One-Out Cross-Validation, miRNA – microRNA, DCE-MRI – Dynamic Contrast-Enhanced Magnetic Resonance Imaging, RNA-seq – RNA sequencing, SILAC – Stable Isotope Labeling with Amino Acids in Cell Culture, SVDD – Support Vector Data Description, v-SVM – variant of Support Vector Machine

Διάφορες μέθοδοι ΜΜ που χρησιμοποιούν σύνολα δεδομένων μεταγραφής (transcriptome), χρησιμοποιούνται στη διαστρωμάτωση των υποτύπων του καρκίνων του μαστού. Ωστόσο, οι ακρίβειες των μεθόδων ποικίλλουν ανάλογα με το εφαρμοζόμενο σύνολο δεδομένων ή τη δειγματοληψία που χρησιμοποιείται [45].

Οι βέλτιστοι προγνωστικοί παράγοντες, για διαφορετικούς υποτύπους, εντοπίστηκαν μέσω του πλησιέστερου μέσου, των Naive Bayes και του 3-Κοντινότερου Γείτονα μαζί με λογιστική παλινδρόμηση, ενώ το μοντέλο πρόβλεψης SVM απέτυχε στην απόδοση. Επιπλέον, στην περίπτωση της χρήσης δεδομένων miRNA, βάσει RNA-seq για τη διάκριση των υποτύπων, η SVM με 5-fold CV έδειξε υψηλότερες επιδόσεις όσον αφορά τη μέτρηση AUC σε σύγκριση με τις μεθόδους που αφορούν τη βαθμολογία Fisher και την απόσταση Hellinger. Το υψηλό accuracy του SVM αποδείχθηκε και στον προσδιορισμό δυνητικών βιοδεικτών miRNA, χρησιμοποιώντας ειδικά δεδομένα μικροσυστοιχιών για τον υποτύπο [45].

Εκτός από τα δεδομένα μεταγραφής, τα δεδομένα προφίλ μεθυλίωσης του DNA, σε όλο το γονιδίωμα, εξετάστηκαν επίσης σε καρκίνο του μαστού για τον προσδιορισμό διακριτών προφίλ που ορίζονται από την κατάσταση μετάλλαξης. Το 2010 ο Flanagan και η ερευνητική του ομάδα, ανέλυσαν τα προφίλ μεταγραφής και μεθυλίωσης για να συγκρίνουν την κατάσταση μετάλλαξης (BRCA1, BRCA2 και

BRCAx) και τα χαρακτηριστικά των υποτύπων. Το SVM με 5-fold CV, με χρήση δεδομένων γονιδιακής έκφρασης, απέδωσε 100% accuracy στις προβλέψεις των εγγενών υποτύπων και 90% accuracy στην πρόβλεψη της μετάλλαξης BRCA1, ενώ οι προβλέψεις για BRCA2 και BRCAx απέτυχαν. Στην περίπτωση των προφίλ μεθυλίωσης, τα εγγενή BCS απέτυχαν και βελτιώθηκαν οι μελέτες μετάλλαξης. Αρκετές άλλες μελέτες, που αξιολόγησαν επιγενετικά δεδομένα μέσω μοντέλων πρόβλεψης ΜΜ, απέδωσαν υψηλό δυναμικό σε ορισμένες από τις περιπτωσιολογικές μελέτες ταξινόμησης [45].

Λαμβάνοντας υπόψη ότι η πρωτεωμική αντικατοπτρίζει τις κυτταρικές λειτουργίες περισσότερο από τις πληροφορίες στη γονιδιωματική, την επιγενετική και τη μεταγραφική, το υποτυπικό του καρκίνου του μαστού μελετήθηκε, επίσης, σε πρωτεωμικό επίπεδο. Ποσοτικοποιώντας τις πρωτεΐνες μέσω σταθερής επισήμανσης ισοτόπων με αμινοξέα στην τεχνολογία κυτταρικής καλλιέργειας και SVM με την προσέγγιση one-vs-rest, το 2016 επιτεύχθηκε από τους Tyanova και την υπόλοιπη ερευνητική ομάδα, διαστρωμάτωση των υποτύπων με υψηλή ακρίβεια. Επιπλέον, το 2016 ο Waldemarson, χρησιμοποίησε δισδιάστατες, διαφορετικές τεχνολογίες ηλεκτροφόρησης και μικροσυγκέντρωσης για την απόκτηση πρωτεωμικών και μεταγραφικών συνόλων δεδομένων, αντίστοιχα. Πραγματοποίησαν συγκρίσεις κατά ζεύγη, μεταξύ των υποτύπων, και παρουσίασαν μια σαφή διάκριση μεταξύ βασικών και φωτεινών Α όγκων χρησιμοποιώντας SVM με την προσέγγιση της LOOCV [45].

Μοντέλα πρόβλεψης που βασίζονται σε μονοπάτια

Η αξιοποίηση των πληροφοριών και των σχέσεων μεταξύ των κυτταρικών μορίων μπορεί να βοηθήσει στη μείωση ή την εξάλειψη του θορύβου στα δεδομένα omics. Η ιδέα αυτή έχει ανοίξει το δρόμο για υψηλότερης απόδοσης μοντέλα πρόβλεψης που βασίζονται σε μονοπάτια [45].

Η ερευνητική ομάδα του Wu το 2017, έδειξε πως η ταξινόμηση των βιοδεικτών που βασίζονται σε SVM με τη χρήση βιοδεικτών που βασίζονται σε μονοπάτια ήταν ακριβής (>90%) στην πραγματική πρόβλεψη των ασθενών, αλλά ανακριβής στην πρόβλεψη άλλων καρκίνων του μαστού. Επίσης, μελετήθηκε η μεταστατική συμπεριφορά των όγκων του μαστού, χρησιμοποιώντας μοντέλα πρόβλεψης που βασίζονται σε μονοπάτια. Μια συνολική ακρίβεια ελαφρώς χαμηλότερη από 90% επιτεύχθηκε με τη χρήση SVM με 20-fold CV[45].

Μοντέλα Πρόβλεψης με βάση την απεικόνιση

Επίσης, συχνά χρησιμοποιούνται μοντέλα πρόβλεψης που βασίζονται στην απεικόνιση. Τα δεδομένα αποκτημένης απεικόνισης μπορούν να εφαρμοστούν για την ανάπτυξη μοντέλων πρόβλεψης. Με αυτή την έννοια, η απεικόνιση μαγνητικού

συντονισμού (MRI) και η υπερηχογραφία είναι τεχνικές ευρείας εξάπλωσης για τη διάγνωση του καρκίνου του μαστού και αρκετές μελέτες χρησιμοποίησαν τη διαδικασία SVM για να βελτιώσουν την ερμηνεία των εικόνων και να παρατηρήσουν αν τα χαρακτηριστικά τους θα μπορούσαν να παρέχουν διαφοροποίηση μεταξύ των υποτύπων [45].

Το 2016 ο Sutton ανέλυσε τα δεδομένα μαγνητικής τομογραφίας μέσω της διαδικασίας SVM με LOOCV, με την απόδοση των προβλέψεων να είναι αμφισβητήσιμη και το accuracy περιορισμένο σε κάθε υποτύπο. Ομοίως, η ανάλυση των χαρακτηριστικών υπερήχων υψηλής απόδοσης και οι πληροφορίες των βιοδεικτών καρκίνου μέσω SVM με 3-fold CV, είχαν ως αποτέλεσμα περιορισμένη ακρίβεια όσον αφορά τις τιμές AUC. Από την άλλη, μία μελέτη που πραγματοποιήθηκε το 2014, από τον Agner και την ερευνητική του ομάδα, έδειξε ανώτερη απόδοση της χρήσης μεθόδων διάγνωσης με τη βοήθεια υπολογιστή (CAD) με δυναμική μαγνητική τομογραφία ενισχυμένη με αντίθεση (DCE-MRI) στη διαφοροποίηση των υποτύπων. Αν και η DCE-MRI αναφέρεται ως μια ευαίσθητη τεχνική στην ανίχνευση του TNBC (Triple Negative Breast Cancer) και τον έλεγχο φορέων μετάλλαξης BRCA, διαπιστώθηκε ότι είναι προβληματική λόγω της υψηλής ομοιότητας στην απεικονιστική κατάρτιση προφίλ των τριπλά αρνητικών αλλοιώσεων και καλοηθών ινομυωμάτων [45].

Ωστόσο, η απασχόληση των μεθόδων CAD έχει προταθεί για την αύξηση της διαγνωστικής ιδιαιτερότητας της DCE-MRI, δεδομένου ότι η υψηλή ακρίβεια (97%) θα μπορούσε να επιτευχθεί μέσω της ταξινόμησης SVM αυτών των υποομάδων. Επιπλέον, η ενσωμάτωση των παραμέτρων που προέρχονται από διάχυση (μέση, τυπική απόκλιση, κλίση και κύρτωση του συντελεστή φαινομενικής διάχυσης, της εκ νέου ενισχυμένης διάχυσης και της ενδοοξειδικής ασυνάρτητης κίνησης) με μαγνητική τομογραφία ενίσχυσε σημαντικά την απόδοση του SVM στην ταξινόμηση καλοήθων και κακοήθων όγκων του μαστού[45].

<u>Μοντέλα πρόβλεψης που χρησιμοποιούν κλινικά, παθολογικά και βιοχημικά δεδομένα</u>

Μια πλούσια πηγή πληροφοριών για τους όγκους μπορεί να συγκεντρωθεί μέσω κλινικών, παθολογικών και βιοχημικών αναλύσεων, οι οποίες θα πρέπει να θεωρηθούν απαραίτητο μέρος της έρευνας για τον καρκίνο, ειδικά για τη διεξαγωγή προβλέψεων με χρήση ΜΜ σε κρίσιμες και αβέβαιες καταστάσεις. Σε μια μελέτη, που παρουσιάστηκε το 2014 από τον Wu και την ερευνητική του ομάδα, κατασκεύασαν το μοντέλο SVM για να ταξινομήσουν καρκίνους του μαστού καθώς και θετικές και αρνητικές ομάδες μετάστασης ALN (Axillary Lymph Node, ομάδα λεμφαδένων στην περιοχή της μασχάλης) με βάση παθολογικές πληροφορίες του πρωτογενούς όγκου και κλινικά χαρακτηριστικά (όπως ηλικία κατά τη διάγνωση, μέγεθος όγκου, κατάσταση ER, κατάσταση PR, κατάσταση HER2 και άλλα). Το SVM

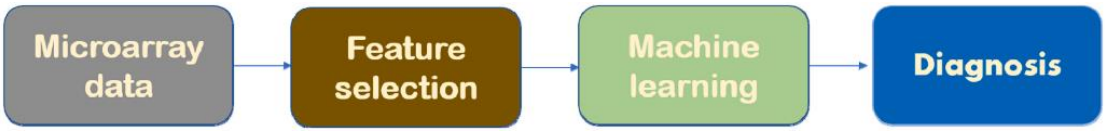
προέβλεψε σωστά τις μεταστάσεις ALN στο 75% των ασθενών που χρησιμοποιούσαν παθολογικές και κλινικές πληροφορίες. Η προγνωστική ικανότητα των υποτύπων που χρησιμοποιούν ανάλυση υποομάδας δεν έδειξε καμία διαφορά και αυτή η προγνωστική απόδοση ήταν κατώτερη, με ακρίβεια μόνο 60%. Από τη βιοχημική άποψη, ο Becker-Putsche το 2013 κατασκεύασε το SVM χρησιμοποιώντας δεδομένα πρωτεϊνών, λιπιδίων και νουκλεϊνικού οξέος που αποκτήθηκαν από φασματοσκοπία Raman. Η έρευνα εφαρμόστηκε σε έξι σειρές κυττάρων καρκίνου του μαστού που αντιπροσωπεύουν διαφορετικούς υποτύπους σε επίπεδο ενός κυττάρου. Οι επιδόσεις ταξινόμησης , σε κυτταρικό επίπεδο, παρατηρήθηκαν μέχρι ένα ποσοστό του 97% [45].

Τελικά, μέσω αυτής της ανασκόπησης εξετάστηκε η αποτελεσματικότητα μοντέλων πρόβλεψης SVM στην ταξινόμηση των υποτύπων του καρκίνου του μαστού, χρησιμοποιώντας διαφορετικούς τύπους δεδομένων. Η προγνωστική απόδοση των μεθόδων SVM, που περιλαμβάνουν ακτινολογικά δεδομένα, ήταν σημαντικά υψηλότερη και σχεδόν χωρίς αποτυχία στη διάκριση των υποτύπων. Στην περίπτωση ζητημάτων συνέπειας και σχέσης κόστους-αποτελεσματικότητας, θα πρέπει να προτιμώνται τα μοντέλα πρόβλεψης που βασίζονται σε omics, παρόλο που θα μπορούσαν να βελτιωθούν με τη βελτίωση της ποιότητας των δεδομένων, αλλά και μοντέλα που βασίζονται σε μονοπάτια, και παρουσιάζουν αποδεκτή ακρίβεια [45].

5.11. Μοντέλα Πρόγνωσης με χρήση δεδομένων μικροσυστοιχιών DNA

Ένα αποτελεσματικό εργαλείο για τη διάγνωση του καρκίνου του μαστού είναι η τεχνολογία μικροσυστοιχιών (Microarray) DNA. Τα δεδομένα γονιδιακής έκφρασης της μικροσυστοιχίας του DNA αντιπροσωπεύουν την κατάσταση ενός κυττάρου σε μοριακό επίπεδο και γι’ αυτό το λόγο έχουν πολλές προοπτικές στην ιατρική διάγνωση. Χρησιμοποιούνται για την επίλυση προβλημάτων δύο ή περισσότερων κλάσεων, μπορούν να προβλέψουν την αντίδραση ενός φαρμάκου ή να προσδιορίσουν τους όγκους με την εύρεση ομάδων με όμοια εκφρασμένα γονίδια. Η μέθοδος αυτή έχει αναλυθεί αποτελεσματικά με χρήση MM [39].

Η ανάλυση των microarray δεδομένων γίνεται σε δύο βήματα (*Εικόνα 43*). Αρχικά, τα δεδομένα προ-επεξεργάζονται χρησιμοποιώντας τεχνικές επιλογής χαρακτηριστικών, προκειμένου να αφαιρεθούν θορυβώδη και περιττά χαρακτηριστικά. Στη συνέχεια, το προκύπτων υποσύνολο χρησιμοποιείται για την εκπαίδευση του μοντέλου MM για τη διάγνωση υποτύπων καρκίνου[39].



Εικόνα 43. Ανάλυση Microarray δεδομένων[39].

Στον Πίνακας *27* παρουσιάζονται διάφορες μέθοδοι ταξινόμησης του καρκίνου του μαστού, οι πληροφορίες του οποίου αντλήθηκαν από [39].

Προσέγγιση	**Αναφορά**	**Μέθοδος Επιλογής Χαρακτηριστικού**	**Ταξινομητής**	**Accuracy Ταξινομητή**	**Πλήθος Γονιδίων**
Φίλτρου	Purbolaksono et al.	MI	BN	84%	NA
	Cilia et al.	GR	KNN	91.96%	50
	V. Bolón et al.	ReliefF	NB	89%	50
	Al-Batah et al.	CFS	JRip	88.7%	138
	Gao et al.	FCBF	PA-SVM	88.66%	92
	Baliarsingh et al.	Wilcoxon rank sum test	JELM	90.91%	505
	Su et al.	K-S test + CFS	SVM	87.4%	11.7
	Ahmad, F. K.	IG	SVM	80%	200
Υβριδική	Medjaheda et al.	SVM-RFE + BDF	SVM	89.47%	7237
	Jain et al.	CFS-iBPSO	NB	92.75%	32.7
	Shahbeig et al.	TLBO-PSO	SVM	91.88%	195
	Lu et al.	MIMAGA	ELM	87.12%	59
	Alomari et al.	MRMR-FPA		85.88%	16.8
	Turgut et al.	RFE + RLR	SVM	87.87%	50
	Mufassirin and Ragel	GR+Wrapper	NB	89.69%	NA
	Hameed et al.	PCC-BPSO	SVM	90.72%	41
	Utami and Rustama	ABC	SVM	88%	NA
	Sardana et al.	ClusterQGA	SVM	86.6%	21
	Singh and Sivabalakrishnan	mRMRAGA	ELM	86.73%	140
	Nagpala and Singhb	QMI	IB1	90.72%	98
	Loev et al.	IG + GWO	SVM	94.87%	240
	Hamim et al.	FC5	C5.0	93.28%	5
Άλλη	Jinthanasatian et al.	A neuro-fuzzy algorithm	Rule set generation	82.37%	7
	Li et al.		SVM-RFE	86.09%	NA

Πίνακας 27. Μέθοδοι Ταξινόμησης του καρκίνου του μαστού.

Αν και τα δεδομένα μικροσυστοιχιών αποδεικνύονται αποτελεσματικά για τη διάγνωση του καρκίνου, ο τεράστιος αριθμός χαρακτηριστικών τους σε σχέση με το μικρό μέγεθος του δείγματος, προκαλούν το λεγόμενο πρόβλημα διαστάσεων. Για να αποφευχθεί αυτό, οι τεχνικές επιλογής που χρησιμοποιούνται είναι, συνήθως, οι υβριδικές προσεγγίσεις ή οι προσεγγίσεις φίλτρων [39].

Στην μελέτη των Su και της ερευνητικής του ομάδας, εφαρμόστηκε προσέγγιση φίλτρου και η μέθοδος επιλογής K-S test+CFS δημιούργησε ένα μικρό υποσύνολο σε σύγκριση με το υποσύνολο που παράγεται από άλλη τεχνική φίλτρου, αλλά με χαμηλότερο accuracy περίπου στο 87.4%. Αντίθετα, παρατηρείται πως η μέθοδος φίλτρου Wilcoxon rank sum test πέτυχε accuracy περίπου στο 90.91%, αλλά με ένα μεγάλο υποσύνολο γονιδίων περίπου στα 505 γονίδια. Αντίθετα, η εφαρμογή GR πέτυχε υψηλότερο accuracy (91,96%) με μόνο 50 γονίδια. Ενώ η προσέγγιση του φίλτρου μπορεί να πετύχει υψηλό αποτέλεσμα, με επιλογή μεγαλύτερου αριθμού χαρακτηριστικών, η ικανότητα της προσέγγισης περιτυλίγματος να βρει το βέλτιστο, ή σχεδόν βέλτιστο, υποσύνολο βοηθάει την υβριδική προσέγγιση να πετύχει υψηλότερο accuracy με χρήση ενός μικρού υποσύνολου. Στα υβριδικά μοντέλα, με χρήση FC5 επιτεύχθηκε 93.28% accuracy με χρήση 5 γονιδίων. Το υψηλότερο αποτέλεσμα στο accuracy (94.87%) στην υβριδική προσέγγιση, επιτεύχθηκε από τους Loev και την ερευνητική του ομάδα, χρησιμοποιώντας την τεχνική IG-GWO, τον αλγόριθμο SVM και ένα μεγάλο υποσύνολο γονιδίων, περίπου στα 240 γονιδίων [39].

Επομένως, παρατηρήθηκε ότι η υβριδική προσέγγιση μπορεί να πετύχει καλύτερη απόδοση σε σχέση με την απόδοση που θα πετύχει η προσέγγιση του φίλτρου για την ταξινόμηση του καρκίνου του μαστού. Η χειρότερη απόδοση, όσον αφορά τον αριθμό των επιλεγμένων γονιδίων (7237), παρατηρήθηκε στην έρευνα που χρησιμοποίησαν SVM-RFE+BDF με το accuracy να είναι στο 89.47%, το οποίο θεωρείται αποδεκτό [39].

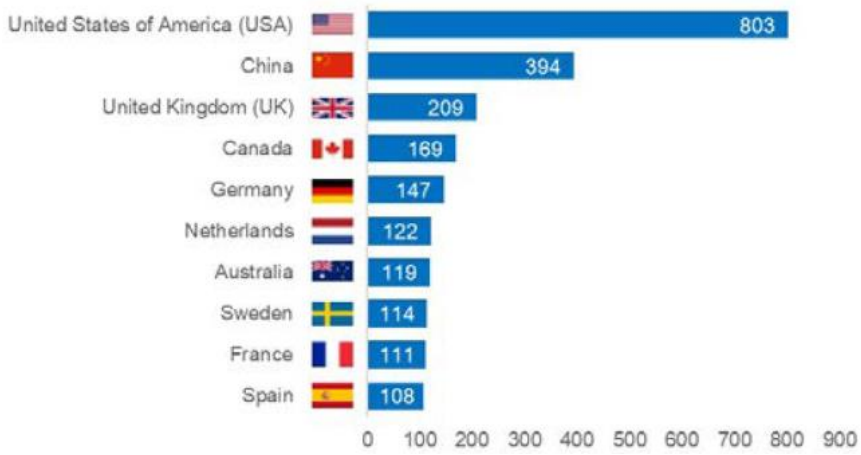
5.12. Ανασκόπηση προγωνστικών μοντέλων Μηχανικής Μάθησης για τον καρκίνο του μαστού με βάση τη χώρα

Το 2020 το τμήμα τηλε-Ιατρικής του πανεπιστήμιου της Κουαζούλου - Νατάλ, προσπάθησε να κάνει μία πενταετή (2015 μέχρι 2019) ανάλυση των μελετών, που έχουν επίκεντρο την πρόγνωση του καρκίνου του μαστού με χρήση τεχνικών μηχανικής μάθησης. Ο πρώτος στόχος αυτής της μελέτης ήταν να διερευνηθούν οι τάσεις στην πρόβλεψη του καρκίνου του μαστού, με χρήση ΜΜ, αναλύοντας τη χώρα, τον συγγραφέα, το περιοδικό, τις συνεργασίες με άλλα ιδρύματα και τις λέξεις-κλειδιά. Ο δεύτερος στόχος ήταν να παρασχεθεί μια ανασκόπηση τέτοιων μελετών σχετικά με το σύνολο δεδομένων ανάλυσης αίματος BCCD. Τέλος, ο τρίτος στόχος ήταν να παρασχεθεί μια σύντομη ανασκόπηση και πάλι τέτοιων μελετών βασιζόμενες στο σύνολο WBCD (Original) [46].

<u>Πρώτο αντικείμενο μελέτης</u>

Οι βιβλιογραφικές αναζητήσεις επικεντρώθηκαν γύρω από την pubMed, αφού τα δεδομένα της είναι πιο ελεύθερα προσβάσιμα. Αναζητήθηκαν όλες οι δημοσιεύσεις ταξινομημένες με βάση τις πιο πρόσφατες ημερομηνίες και τις ημερομηνίες δημοσίευσης που ορίζονται σε πέντε έτη για το ανθρώπινο είδος και όλους τους τύπους άρθρων.

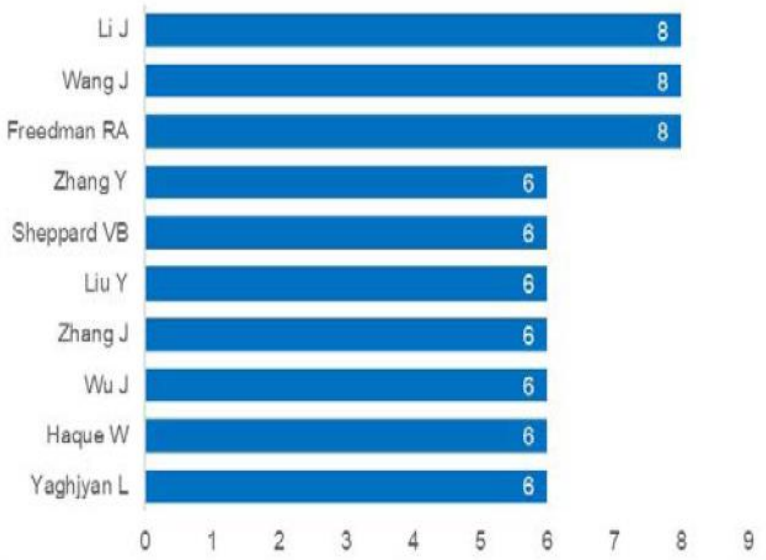
Για ανάλυση με βάση τη χώρα, βρέθηκαν αποτελέσματα για 86 από τις 195 χώρες. Τα αποτελέσματα ταξινομήθηκαν από το υψηλότερο στο χαμηλότερο ανάλογα με τον συνολικό αριθμό δημοσιεύσεων. Οι δέκα πρώτες χώρες εντοπίστηκαν και επιλέχθηκαν με βάση τον υψηλότερο συνολικό αριθμό δημοσιεύσεων (*Εικόνα 44*) [46].



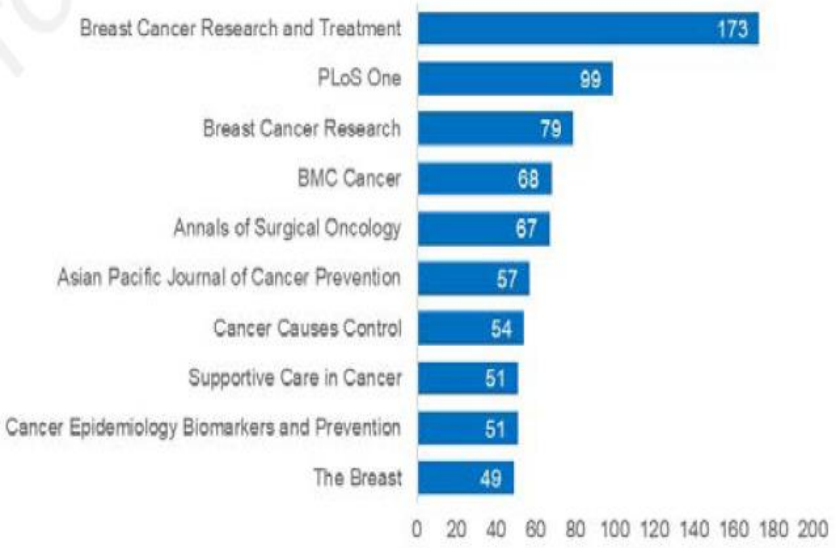
Εικόνα 44. Οι δέκα χώρες με τις περισσότερες δημοσιεύσεις για την πρόβλεψη του καρκίνου του μαστού με χρήση ΜΜ από το 2015-2019 [46].

Για ανάλυση με βάση τον πρώτο συγγραφέα και το περιοδικό, επιστράφηκαν 2928 αποτελέσματα από την αναζήτηση, από τα οποία υπήρχαν περίπου 2419 ξεχωριστοί πρώτοι συγγραφείς και 670 ξεχωριστά περιοδικά. Οι δέκα πρώτοι συγγραφείς και περιοδικά επιλέχθηκαν μετά από υπολογισμούς του συνολικού αριθμού δημοσιεύσεων και ταξινομώντας τα από την υψηλότερη στη χαμηλότερη. Τα αποτελέσματα των δέκα πρώτων συγγραφέων και των δέκα πρώτων περιοδικών απεικονίζονται στην Εικόνα *45* και στην Εικόνα *46*, αντίστοιχα [46].

Για την εμφάνιση των θεσμικών συνεργασιών και των λέξεων-κλειδιών του συγγραφέα, εξήχθη το αρχείο MEDLINE, από την PubMed, με τα αποτελέσματα από την αναζήτηση της βάσης δεδομένων. Το αρχείο MEDLINE χρησιμοποιήθηκε ως είσοδος σε ένα πρόγραμμα υπολογιστών που ονομάζεται VOSviewer για την οπτική ανάλυση της βιβλιογραφίας. Το πρόγραμμα υπολογιστών VOSviewer βρήκε 10100 ξεχωριστά ιδρύματα από το αρχείο MEDLINE. Από αυτά τα 10100 ιδρύματα, βρέθη -



Εικόνα 45. Οι δέκα πρώτοι συγγραφείς με τις περισσότερες δημοσιεύσεις για την πρόγνωση του καρκίνου του μαστού με χρήση ΜΜ από το 2015 εως 2019 [46].



Εικόνα 46. Τα δέκα πρώρα περιοδικά με τις περισσότερες δημοσιεύσεις για την πρόγνωση του καρκίνου του μαστού με χρήση ΜΜ από το 2015 εώς 2019 [46].

καν θεσμικές συνεργασίες για εννέα και αυτές συμπεριλήφθηκαν στην ανάλυση αυτής της κατηγορίας. Το πρόγραμμα υπολογιστών VOSviewer αναγνώρισε 4755 διαφορετικές λέξεις-κλειδιά στο αρχείο MEDLINE, εκ των οποίων 123 βρέθηκαν να έχουν συνυπάρχοντας με λέξεις-κλειδιά συγγραφέα και συμπεριλήφθηκαν σε αυτήν την ανάλυση [46].

Όσον αφορά την ανάλυση με βάση τη χώρα, από την *Εικόνα 44* φαίνεται πως οι Ηνωμένες Πολιτείες της Αμερικής παρήγαγαν τις περισσότερες δημοσιεύσεις, 803 από τις 2928 συνολικά (27% των συνολικών παγκόσμιων εκδόσεων για εκείνη την περίοδο). Η Κίνα κατέλαβε τη δεύτερη θέση με 394 δημοσιεύσεις (περίπου το 13% των εκδόσεων του 2928). Επίσης, στην *Εικόνα 44* παρατηρείται πως, με εξαίρεση

την Κίνα, οι υπόλοιπες εννέα από τις δέκα χώρες είναι ανεπτυγμένες. Αυτό συμβαίνει παρά το γεγονός ότι ο καρκίνος του μαστού αποτελεί παγκόσμιο πρόβλημα. Ωστόσο, το εύρημα αυτό είναι αναμενόμενο δεδομένου ότι οι αναπτυσσόμενες χώρες δαπανούν συγκριτικά λιγότερα χρήματα για έρευνα και ανάπτυξη λόγω περιορισμένων πόρων [46].

Τέλος, ένα σημαντικό εύρημα, που αφορούσε το πρώτο αντικείμενο μελέτης, ήταν πως στην ανάλυση με βάση τις θεσμικές συνεργασίες, παρατηρήθηκε πως καμία αναπτυσσόμενη χώρα δεν συμπεριλήφθηκε σε συλλογική έρευνα με ιδρύματα από τις ανεπτυγμένες χώρες [46].

Δεύτερο και τρίτο αντικείμενο μελέτης

Για το δεύτερο και τρίτο αντικείμενο της μελέτης, βρέθηκαν όλες οι επιστημονικές μελέτες, με χρήση του Google Scholar και με αναζήτηση που πραγματοποιήθηκε τον Σεπτέμβριο του 2019, που έκαναν χρήση των BCCD και του WBCD, αντίστοιχα. Αφού τα μοντέλα ΜΜ εκπαιδευτούν, δοκιμάζονται με χρήση του συνόλου δοκιμών. Στη συνέχεια, κατασκευάζεται ο πίνακας σύγχυσης, με βάση τον οποίο υπολογίζονται το Accuracy, το Precision, η Ευαισθησία, η Ειδικότητα, η Βαθμολογία F1, η καμπύλη ROC και η AUC-ROC. Με βάση τα αποτελέσματα αυτών των μετρήσεων υπολογίζεται και η τελική απόδοση του μοντέλου [46].

Για το δεύτερο αντικείμενο μελέτης, οι αναζητήσεις του Google και του Google Scholar απέφεραν 16 σχετικά αποτελέσματα, από τα οποία 5 αποκλείστηκαν για διάφορους λόγους. Επομένως, συμπεριλήφθηκαν τελικά 11 επιλέξιμες μελέτες για το δεύτερο αντικείμενο της συγκεκριμένης μελέτης. Για το τρίτο αντικείμενο μελέτης, το Google Scholar απέδωσε 6130 πιθανά έγγραφα μόνο για τις χρονιές 2016 έως 2017. Εξετάστηκαν οι πρώτες 11 μελέτες, πέντε από αυτές αποκλείστηκαν και, τελικά, στη μελέτη αυτή συμπεριλήφθηκαν έξι εργασίες, τρεις για το 2016 και άλλες τρεις για το 2017. Οι έρευνες του 2018 και του 2019 κατέληξαν, με αντίστοιχο τρόπο, σε δύο μελέτες η καθεμία [46].

Προκειμένου να εξακριβωθεί η απόδοση του κάθε μοντέλου, παρατηρήθηκε πως το Accuracy χρησιμοποιήθηκε συχνότερα (σε n=15 μελέτες). Οι πρώτες 3 μελέτες με το υψηλότερο Accuracy με χρήση του BCCD ήταν Hernández-Julio (95.90%), Singh (92.11 %) και Polat και Senturk (91.37%). Ενώ, για το WBCD ήταν Abdar και Makarenkov (100%), Elgedawy (99.42 %) και Hernández-Julio (99.40%). Η AUC ήταν το δεύτερο πιο συχνό μέσο αξιολόγησης μοντέλου, με την καλύτερη μέτρηση της AUC να παρατηρείται στο 87.00, 91.00 και 95%. Ενώ, για το BCCD το υψηλότερο AUC που παρατηρήθηκε ήταν 99.90%. Τέλος, μόνο η μελέτη του Hung. ανέφερε τη βαθμολογία F1 (82%) που ήταν για το BCCD.

Αναφορά	Σύνολο Δεδομένων	Χώρα*	Στρατηγική Δείγματος	Αλγόριθμος ΜΜ	Μέτρηση (%)
Hernandez-Julio et al.	BCCD	Κολομβία	10-fold CV	Clusters + pivot table	**95.90 (Accuracy)**
Singh	BCCD	Ινδία	67-33 Ε-Δ	K-NN	92.11 (Accuracy)
Polat και Senturk	BCCD	Τουρκία	10-fold CV	AdaBoost	91.37 (Accuracy)
Akben	BCCD	Τουρκία	10-fold CV	DT	90.52 (Accuracy)
Islam και Poly	BCCD	Ταϊβάν (Κίνα)	10-fold CV	K-NN	86.00 (Accuracy)
Araujo et al.	BCCD	Βραζιλία	70-30 Ε-Δ 10-fold CV	ΝΔ	80.67 (Accuracy)
Aslam et al.	BCCD	Τουρκία	80-20 Ε-Δ	ELM	80.00 (Accuracy)
Livieris	BCCD	Ελλάδα	10-fold CV	K-NN	62.00 (Accuracy)
Patricio et al.	BCCD	Πορτογαλία	MCCV	SVM	**87.00, 91.00 (95% CI for AUC)**
Li και Chen	BCCD	Ηνωμένη Βασίλειο	70-30 Ε-Δ	RF	78.50 (AUC)
Hung et al.	BCCD	Βιετνάμ	80-20 Ε-Δ	DT	**82.00 (F1 Score)**
Abdar και Makarenkov	WBCD	Καναδάς	50-50 Ε-Δ	CWV-BANN-SVM	**100.00 (Accuracy)**
Elgedawy	WBCD	Σαουδική Αραβία	75-25 Ε-Δ	RF	99.42 (Accuracy)
Hernandez-Julio et al.	WBCD	Κολομβία	10-fold CV	Clusters + pivot table	99.40 (Accuracy)
Chaurasia et al.	WBCD	Ινδία	Stratified 10-fold CV	NB	97.36 (Accuracy)
Asri et al.	WBCD	Μορόκκο	10-fold CV	SVM	97.13 (Accuracy)
Alzubaldi et al.	WBCD	Ηνωμένο Βασίλειο	LOOCV	SVM (quadric-linear kernel)	97.00 (Accuracy)
				K-NN (Minkowsky και Ευκλείδεια απόσταση)	97.00 (Accuracy)
Islam et al.	WBCD	Μπανγκλαντές	10-fold CV	SVM	97.00 (Accuracy)
Chaurasia και Pal	WBCD	Ινδία	10-fold CV	SMO (SVM)	96.20 (Accuracy)
Bazazeh και Shubair	WBCD	Ηνωμένα Αραβικά Εμιράτα	10-fold CV	RF	**99.90 (AUC)**
Li και Chen	WBCD	Ηνωμένο Βασίλειο	70-30 Ε-Δ	RF	98.90(AUC)

Πίνακας 28. Συνοπτικός κατάλογος μελετών για την πρόβλεψη του καρκίνου του μαστού με χρήση ΜΜ για τα σύνολα BCCD και WBCD [46].

Για το δεύτερο αντικείμενο μελέτης, το οποίο αφορά το σύνολο BCCD, ο Hernández-Julio και η ερευνητική ομάδα πρότειναν ένα νέο περιβάλλον βασισμένο σε clusters και Pivot tables, χρησιμοποιώντας το λογισμικό MATLAB. Το Accuracy που επιτεύχθηκε ήταν υψηλό, στο 95.90%, με τη 10-fold CV. Το μοντέλο του Singh με χρήση K-NN έφτασε στη δεύτερη θέση, με accuracy στο 92.11%. Ενώ το τρίτο

καλύτερο Accuracy παρατηρήθηκε στην μελέτη των Polat και Shenturk, με χρήση του υβριδικού μοντέλου Adaptive Boosting (AdaBoost) μαζί με 10-fold CV και πέτυχε ακρίβεια 91.37% [46].

Για το τρίτο αντικείμενο μελέτης, το μοντέλο των Abdar και Makarenkov πέτυχε το υψηλότερο Accuracy στο 100% με χρήση του WBCD. Αν και αυτό το μοντέλο φαίνεται να έχει επιτύχει εξαιρετικά αποτελέσματα, αυτό μπορεί να είναι αναξιόπιστο, καθώς δεν χρησιμοποιήθηκε προσέγγιση CV για να χειριστεί οποιαδήποτε πρόβλημα στο σύνολο δεδομένων. Το μοντέλο της Elgedawy ήταν το αμέσως επόμενο καλύτερο, με accuracy στο 99,42% για το WBCD. Όπως και με τη μελέτη των Abdar και του Makarenkov, και αυτό το μοντέλο θα έπρεπε να είχε χρησιμοποιήσει κάποια μορφή CV για να χειριστεί το σύνολο δεδομένων του. Το μοντέλο των Hernández-Julio ήταν το αμέσως επόμενο με accuracy στο 99,40%, με 10-fold CV και δημιουργήθηκε με το λογισμικό MATLAB [46].

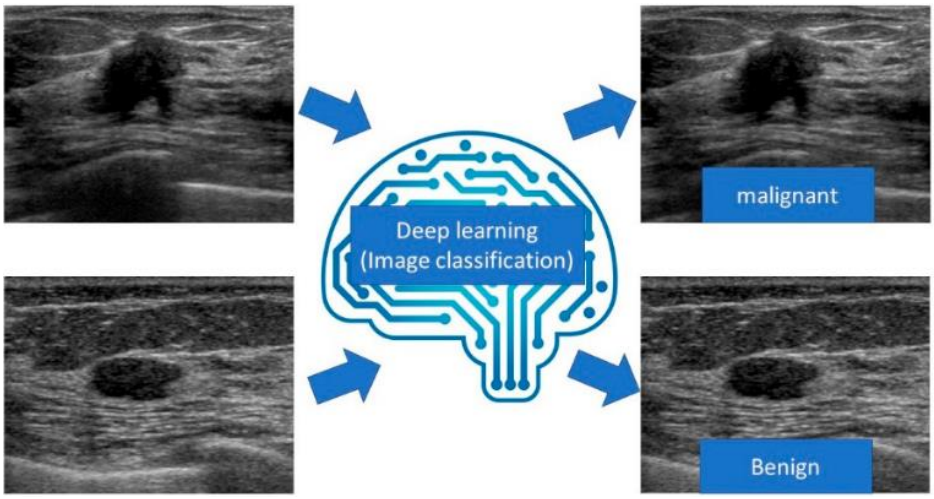
5.13. Η εφαρμογή Βαθιάς Μάθησης σε υπερήχους για την απεικόνιση του μαστού

Ο υπέρηχος μαστού είναι ένας από τους πιο συχνά χρησιμοποιούμενους τρόπους διάγνωσης και ανίχνευσης του καρκίνου του μαστού, λόγω της ασφάλειας αλλά και του χαμηλού του κόστους. Επίσης, η τεχνολογία βαθιάς μάθησης έχει σημειώσει σημαντική πρόοδο στην εξαγωγή και ανάλυση δεδομένων για ιατρικές εικόνες τα τελευταία χρόνια. Επομένως, η χρήση βαθιάς μάθησης στους υπέρηχους για απεικόνιση του μαστού είναι εξαιρετικά σημαντική, καθώς εξοικονομεί χρόνο και αντισταθμίζει την έλλειψη εμπειρίας και δεξιοτήτων σε ορισμένες περιπτώσεις. Η έρευνα που πραγματοποιήθηκε το 2020, μετά από συνεργασία των πανεπιστημίων της Ιαπωνίας, εξετάζει τις βασικές τεχνικές γνώσεις και αλγορίθμους της βαθιάς μάθησης για υπερήχους μαστού και την εφαρμογή τεχνολογίας βαθιάς μάθησης στην ταξινόμηση εικόνας, την ανίχνευση αντικειμένων, την τμηματοποίηση και τη σύνθεση εικόνας [47].

Η ανάπτυξη της έρευνας ΤΝ για τον υπέρηχο μαστού έχει οδηγήσει σε αύξηση των δημοσιεύσεων σε αυτόν τον τομέα. Η PubMed ερευνήθηκε έως τις 31 Οκτωβρίου 2019 για να εξάγει άρθρα σχετικά με την απεικόνιση και την τεχνητή νοημοσύνη που χρησιμοποιούνται για υπέρηχο μαστού και επιλέχθηκαν μόνο μελέτες σχετικά με την τεχνητή τομογραφία για την απεικόνιση του μαστού. Ο ετήσιος αριθμός δημοσιεύσεων φαίνεται στον *Πίνακας 29*. Παρατηρείται ότι δεν είχαμε ιδιαίτερες μεταβολές από το 2010 έως το 2017, αλλά αυξήθηκε σημαντικά από το 2018 έως το 2020, με περισσότερες από 20 δημοσιεύσεις το 2018 και περισσότερες από 40 δημοσιεύσεις το 2019, με τις 50 εργασίες να έχουν δημοσιευθεί το πρώτο 10μηνο του 2020. Πολλές από αυτές τις μελέτες περιέγραψαν την απεικονιστική ταξινόμηση, την ανίχνευση αντικειμένων, την τμηματοποίηση και τη συνθετική απεικόνιση των βλαβών του μαστού [47].

Πίνακας 29. Πλήθος δημοσιεύσεων Βαθιάς Μάθησης σε υπέρηχους μαστού ανά έτος [47].

Η *ταξινόμηση εικόνας (Image Classification)* είναι μια μέθοδος για τον προσδιορισμό και την πρόβλεψη του τι απεικονίζει μια εικόνα (*Εικόνα 47*). Η ανάπτυξη της βαθιάς μάθησης με CNN, έχει διευκολύνει τη δημιουργία μοντέλων υψηλής ακρίβειας και έχει αποκτήσει ηγετικό ρόλο στην ταξινόμηση εικόνων, αφού ικανά δίκτυα όπως το VGGNet, το GoogLeNet, το ResNet και το DenseNet έχουν αναπτυχθεί με βαθύτερα στρώματα [47].



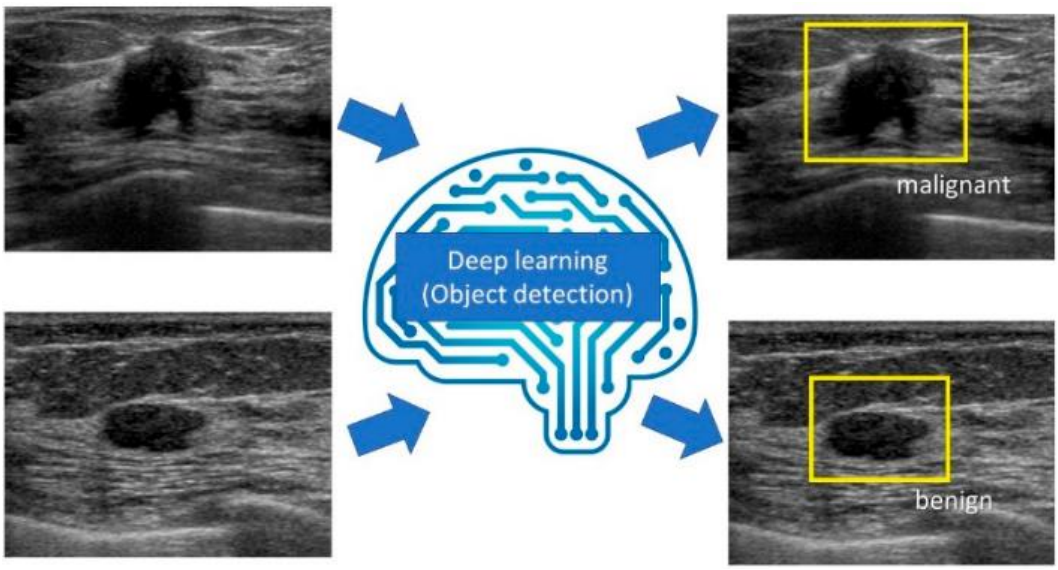
Εικόνα 47. Διάγραμμα ταξινόμησης εικόνας για υπέρηχο μαστού [47].

Στην ταξινόμηση εικόνας του υπέρηχου μαστού, σε πολλές αναφορές συζητήθηκε η διάκριση μεταξύ καλοηθών και κακοήθων αλλοιώσεων σε εικόνες B-mode (*Πίνακας 30*).

Σκοπός (Τύπος Εικόνας)	Μοντέλο	Πλήθος Εικόνων Εκπαίδευσης	Πλήθος Εικόνων Δοκιμής	Αποτέλεσμα	Αναφορά
Αλλοιώσεις Στήθους (B-mode εικόνες)	GoogLeNet	6579	829	Ευαισθησία: 86% Ειδικότητα: 96% Accuracy: 90% AUC: 0.9	Han et al.
Αλλοιώσεις Στήθους (B-mode εικόνες)	GoogLeNet Inception v2	937	120	Ευαισθησία: 95.8% Ειδικότητα: 87.5% Accuracy: 92.5% AUC: 0.913	Fujioka et al.
Αλλοιώσεις Στήθους με CAD (B-mode εικόνες)	Koios DS	Περισσότερα από 400,000	900	AUC χωρίς CAD: 0.83 AUC με CAD: 0.87	Mango et al.
Αλλοιώσεις Στήθους (SWE εικόνες)	PGBM και RBM	227	5-fold CV	Ευαισθησία: 88.6% Ειδικότητα: 97.1% Accuracy: 93.4% AUC: 0.947	Zhang et al.
Αλλοιώσεις Στήθους (SWE εικόνες)	DenseNet 169	304	73	Ευαισθησία: 85.7% Ειδικότητα: 78.9% AUC: 0.898	Fujioka et al.
Μασχαλιαίοι Λεμφαδένες (B-mode εικόνες)	VGG-M model	118	5-fold CV	Ευαισθησία: 84.9% Ειδικότητα: 87.7% Accuracy: 86.4% AUC: 0.937	Coronado-Gutierrez et al.

Πίνακας 30. Μοντέλα Βαθιάς Μάθησης για την ταξινόμηση εικόνας [47].
PGBM - point-wise gated Boltzmann machine, RBM - restricted Boltzmann machine

Η *ανίχνευση αντικειμένων (Object Detection)* αναφέρεται στην ανίχνευση της θέσης και της κατηγορίας (για παράδειγμα κακοήθης ή καλοήθης) ενός καθορισμένου αντικειμένου σε μια εικόνα, όπως η *Εικόνα 48*. Και σε αυτή την περίπτωση, η ανάπτυξη της βαθιάς μάθησης με CNN, έχει δημιουργήσει ταχύτερα και ακριβέστερα μοντέλα ανίχνευσης αντικειμένων, όπως το Spatial Pyramid Pooling (SPP)-net, το Fast R-CNN, το Faster R-CNN, το You Look Only Once (YOLO), ο ανιχνευτής πολλαπλών κιβωτίων μίας βολής (Single Shot multibox Detector - SSD), τα δίκτυα πυραμίδας χαρακτηριστικών γνωρισμάτων και το RetinaNet [47].



Εικόνα 48. Διάγραμμα ανίχνευσης αντικειμένου για υπέρηχο μαστού [47].

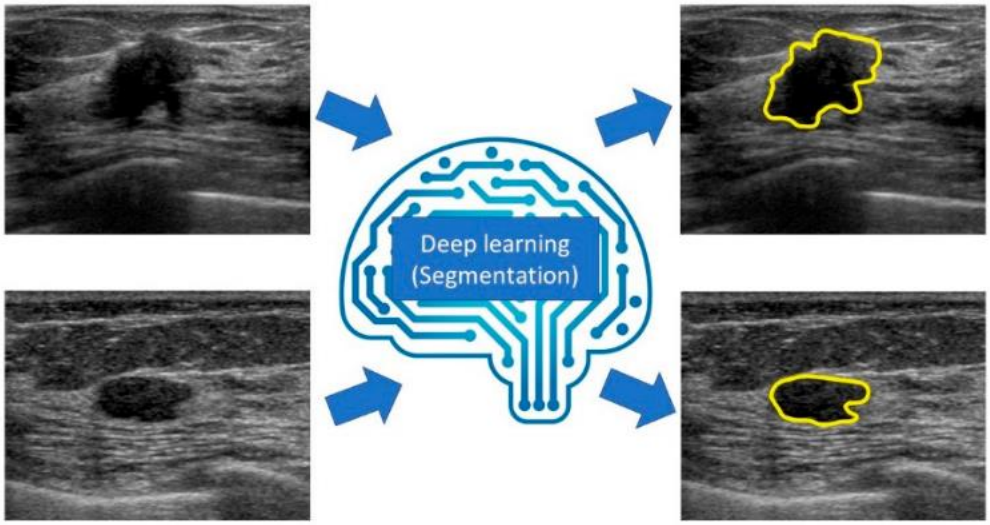
Στον Πίνακας 31 φαίνονται μελέτες σχετικές με μοντέλα βαθιάς μάθησης για την ανίχνευση αντικειμένων σε απεικονίσεις υπερήχων μαστού.

Σκοπός (Τύπος Εικόνας)	**Μοντέλο**	**Πλήθος Εικόνων Εκπαίδευσης**	**Πλήθος Εικόνων Δοκιμής**	**Αποτέλεσμα**	**Αναφορά**
AAAM (B-mode)	SSD300	860	183	Precision: 96.89% Ανάκληση: 67.23% F1: 79.38%	Cao et al.
AAAM με CAD (εικόνα ABUS)	QVCAD	Περισσότερα από 20,000	185	AUC χωρίς CAD: 0.828 AUC με CAD: 0.848	Jiang et al.
AAAM με CAD (εικόνα ABUS)	QVCAD	Περισσότερα από 20,000	1485	AUC χωρίς CAD: 0.88 AUC με CAD: 0.91 Ευαισθησία χωρίς CAD: 67% Ευαισθησία με CAD: 88%	Yang et al.
AAAM με CAD (εικόνα ABUS)	QVCAD	Περισσότερα από 20,000	1000	AUC χωρίς CAD: 0.747 AUC με CAD: 0.784	Xu et al.

Πίνακας 31. Μοντέλα βαθιάς μάθησης για ανίχνευση αντικειμένου σε υπέρηχο μαστού [47].
AAAM – Ανίχνευση Αντικειμένου Αλλοιώσεων Μαστού, ABUS – Automated Breast Ultrasound

Η *σημασιολογική τμηματοποίηση (Semantic Segmentation)* είναι μια μέθοδος που μπορεί να συσχετίσει ετικέτες και κατηγορίες με όλα τα εικονοστοιχεία (pixels) σε μία εικόνα και διαιρεί το αντικείμενο σε πολλές περιοχές σε επίπεδο εικονοστοιχείου (*Εικόνα 49*). Έχουν αναπτυχθεί διάφορα μοντέλα βασισμένα στην αρχιτεκτονική του CNN, τα οποία επιτρέπουν τη σημασιολογική τμηματοποίηση υψηλής ακρίβειας και υψηλής ταχύτητας, όπως το Πλήρως Συζευκτικό Δίκτυο (Fully Convolutional Network – FCN), το Segnet και το U-net. Χρησιμοποιούνται ευρέως σε

διάφορες βιομηχανίες που απαιτούν χαρτογράφηση εικόνων υψηλής ακρίβειας, όπως η ιατρική απεικόνιση, η αυτόνομη οδήγηση, η βιομηχανική επιθεώρηση και οι δορυφορικές εικόνες [47].



Εικόνα 49. Διάγραμμα τμηματοποίησης για υπέρηχο μαστού [47].

Στον *Πίνακας 32* βλέπουμε την απόδοση μοντέλων σημασιολογικής τμηματοποίησης για υπερήχους μαστού.

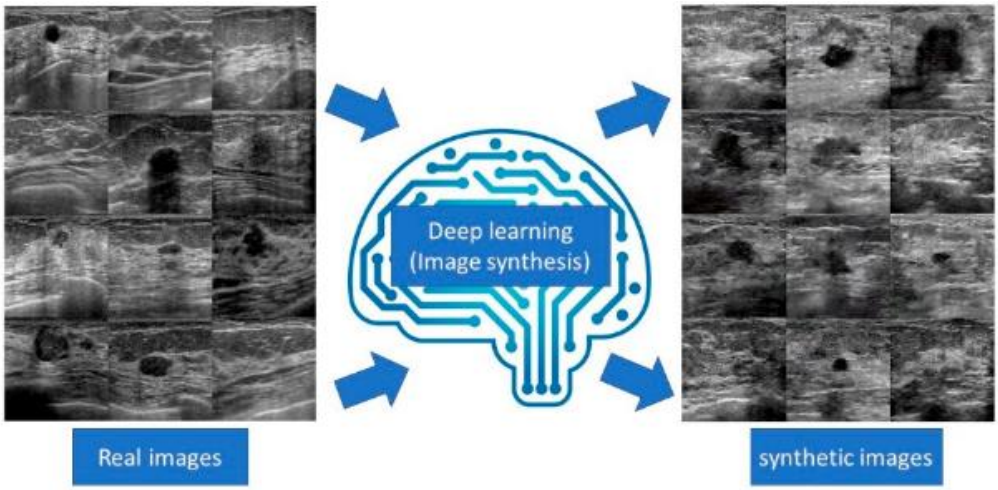
Σκοπός (Τύπος Εικόνας)	**Μοντέλο**	**Πλήθος Εικόνων Εκπαίδευσης**	**Πλήθος Εικόνων Δοκιμής**	**Αποτέλεσμα**	**Αναφορά**
TAM (Εικόνα B-mode)	RDAU-NET	857	205	Precision: 88.58% Ανάκληση:83.19% F1: 84.78	Zhang et al.
TAM (Εικόνα B-mode)	Συνδυασμός DFCN με ένα μοντέλο PBAC	400	170	Συντελεστής ομοιότητας Dice: 88.97% Απόσταση Hausdorff: 35.54 pixels Μέση απόλυτη απόκλιση: 7.67 pixels	Hu et al.
TAM (Εικόνα B-mode)	Αλγόριθμος Multi U-net	372	61	Μέσος συντελεστής Dice: 0.82 Πραγματικό θετικό κλάσμα: 0.84 Λανθασμένο θετικό κλάσμα: 0.01	Kumar et al.

Πίνακας 32. Μοντέλα βαθιάς μάθησης για τμηματοποίηση [47].
TAM – Τμηματοποίηση Αλλοιώσεων Μαστού

Η *σύνθεση εικόνας (Image Synthesis)* περιλαμβάνει τη δημιουργία ρεαλιστικών εικόνων χρησιμοποιώντας έναν αλγόριθμο. Η μέθοδος αυτή βελτιώνει την ποιότητα της εικόνας με την ανακατασκευή και τη δημιουργία δεδομένων εκπαίδευσης, δημιουργώντας εικονικές εικόνες (*Εικόνα 50*) [47].

Μία από τις πιο ενδιαφέρουσες ανακαλύψεις στον τομέα, ήταν τα Παραγωγικά Αντιπαραθετικά Δίκτυα (Generative Adversarial Network – GAN). Τα GANs είναι ένας

ειδικός τύπος μοντέλου νευρωνικού δικτύου, στο οποίο δύο δίκτυα εκπαιδεύονται ταυτόχρονα, με το ένα να επικεντρώνεται στη δημιουργία εικόνων και το άλλο να επικεντρώνεται στις διευκρινήσεις. Ένα άλλο αποτελεσματικό μοντέλο παραγωγής είναι το Variational AutoEncoder (VAE). Το VAE περιέχει δύο μέρη. Το ένα είναι ένας κωδικοποιητής που μαθαίνει αποτελεσματική κωδικοποίηση δεδομένων από το σύνολο δεδομένων, και το μεταβιβάζει σε μια αρχιτεκτονική συμφόρησης, και το άλλο είναι ένας αποκωδικοποιητής που χρησιμοποιεί λανθάνον χώρο στο επίπεδο συμφόρησης για να αναγεννήσει εικόνες παρόμοιες με αυτές στο σύνολο δεδομένων. Το VAE διαφέρει από ένα από AE επειδή παρέχει έναν στατιστικό τρόπο για να περιγράψει τα δείγματα του συνόλου δεδομένων στον λανθάνον χώρο [47].



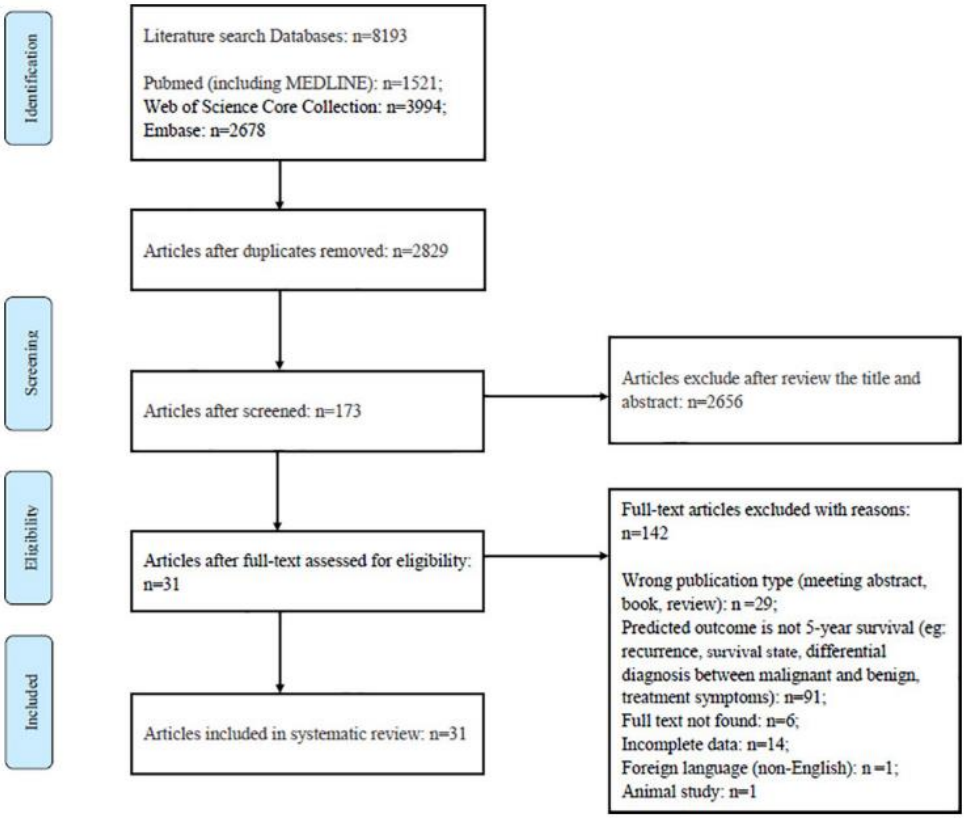
Εικόνα 50. Διάγραμμα σύνθεσης εικόνας για υπέρηχο μαστού [47].

Η ανακατασκευή εικόνας καθιστά δυνατή τη δημιουργία εικόνων με λίγα αντικείμενα. Επιπλέον, μια εικόνα υψηλού ρυθμού καρέ μπορεί να δημιουργηθεί με την εκτέλεση παρεμβολής εικόνας. Επειδή η εικονική εικόνα, που δημιουργείται από το παραγωγικό μοντέλο, δεν περιέχει προσωπικές πληροφορίες, θεωρείται χρήσιμη για την έρευνα και την εκπαίδευση. Κατορθώθηκε λοιπόν, να δημιουργηθούν ρεαλιστικές εικονικές εικόνες όγκων του μαστού και εικονικές εικόνες παρεμβολής της ανάπτυξης όγκων, χρησιμοποιώντας βαθιά περίπλοκα GANs. Επιπλέον, χρησιμοποιήθηκε ένα μοντέλο ανίχνευσης ανωμαλιών, με βάση το GAN, για να διακριθούν οι φυσιολογικοί ιστοί από τις καλοήθεις και κακοήθεις μάζες. Το μοντέλο είχε ευαισθησία 89.2%, ειδικότητα 90.2%, και AUC 0.936. Ο Han και η ερευνητική του ομάδα πρότειναν ένα ημι-επιβλεπόμενο δίκτυο τμηματοποίησης , με βάση τα GANs, και διαπίστωσαν ότι πέτυχε υψηλότερο accuracy τμηματοποίησης σε σχέση με τις υπερσύγχρονες ημι-επιβλεπόμενες μεθόδους τμηματοποίησης [47].

5.14. Ανασκόπηση μοντέλων μηχανικής μάθησης για την επιβίωση από τον καρκίνο του μαστού

Οι Li και Dong, το 2021 στόχευσαν να επανεξετάσουν τη δημοσιευμένη βιβλιογραφία σχετικά με την ανάπτυξη μοντέλων ΜΜ για την επιβίωση των ασθενών από τον καρκίνου του μαστού. Η έρευνα αυτή διεξήχθη σύμφωνα με τις κατευθυντήριες γραμμές για τα Προτιμώμενα Στοιχεία Αναφοράς για Συστηματικές Αναθεωρήσεις και Μετα-Αναλύσεις (Preferred Reporting Items for Systematic Reviews and Meta-Analysis – PRISMA) (*Εικόνα 51*). Οι δύο ερευνητές έψαξαν τις βάσεις δεδομένων PubMed (συμπεριλαμβανομένης της MEDLINE) (1966), Embase (1980) και Web of Science Core Collection (1900), από την έναρξη τους έως τις 30 Νοεμβρίου 2020 [48].

Με την αναζήτηση αυτών των τριών ιατρικών βάσεων δεδομένων, εντοπίστηκαν συνολικά 8193 μελέτες. Αφού αφαιρέθηκαν οι διπλότυπες, με χρήση του EndNote X9, και καταργήθηκαν μελέτες με βάση τον τίτλο τους και τις περιλήψεις τους, έμειναν 173 μελέτες. Από αυτές αποκλείστηκαν 142 για διάφορους λόγους. Συνολικά 31 μελέτες πληρούσαν τα κριτήρια ένταξης των ερευνητών. Η διαδικασία διαλογής της βιβλιογραφίας παρουσιάζεται στην *Εικόνα 51* [48].



Εικόνα 51. Διάγραμμα Ροής PRISMA [48].

Οι ερευνητές χρησιμοποίησαν το εργαλείο αξιολόγησης μοντέλου πρόβλεψης PROBAST (prediction model risk of bias assessment tool), το οποίο χρησιμοποιείται κυρίως στην επικύρωση έρευνας και ανάπτυξης ή στην ενημέρωση πολυποίκιλων μοντέλων διάγνωσης προγνωστικών ή πρόβλεψης πρόγνωσης. Το εργαλείο περιλαμβάνει 20 ερωτήσεις σηματοδότησης σε 4 τομείς (συμμετέχοντες,

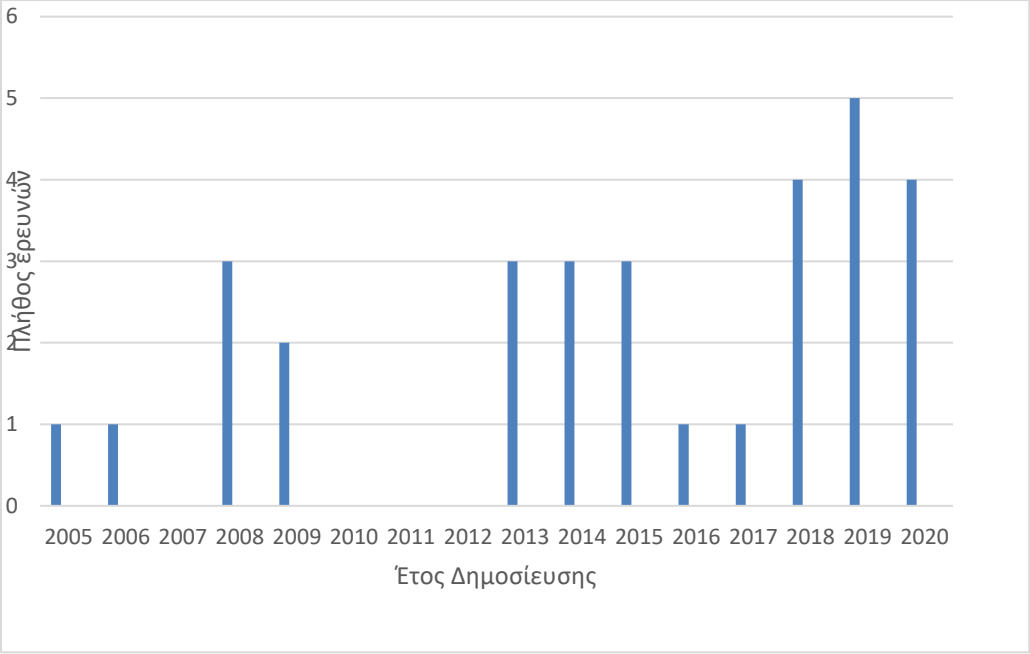
προγνωστικά, αποτελέσματα και ανάλυση) και κάθε ερώτηση απαντάται ως χαμηλός κίνδυνος μεροληψίας (bias), υψηλός κίνδυνος μεροληψίας ή ασαφής. Στον *Πίνακας 33* φαίνεται ποιες από τις 31 έρευνες που μελετήθηκαν είχαν χαμηλό, υψηλό ή ασαφή κίνδυνο. Οι 9 από αυτές είχαν υψηλό κίνδυνο μεροληψίας, οι 17 είχαν μέτριο και οι υπόλοιπες 5 είχαν χαμηλό. [48].

Αναφορά	**Bias συμμετεχόντων**	**Bias Πρόβλεψης**	**Bias Αποτελεσμάτων**	**Bias Ανάλυσης**	**Συνολικό Ποσοστό Bias**	**Συνολική Βαθμολογία Εφαρμογής**
Delen, 2005	Χαμηλό	Χαμηλό	Χαμηλό	Υψηλό	Υψηλό	Χαμηλό
Bellaachia, 2006	Χαμηλό	Χαμηλό	Χαμηλό	Χαμηλό	Χαμηλό	Χαμηλό
Endo, 2007	Χαμηλό	Χαμηλό	Χαμηλό	Μέτριο	Μέτριο	Χαμηλό
Khan, 2008	Χαμηλό	Χαμηλό	Χαμηλό	Υψηλό	Υψηλό	Χαμηλό
Thongkam, 2008	Χαμηλό	Χαμηλό	Χαμηλό	Μέτριο	Μέτριο	Χαμηλό
Choi, 2009	Χαμηλό	Χαμηλό	Χαμηλό	Υψηλό	Υψηλό	Χαμηλό
Liu, 2009	Χαμηλό	Χαμηλό	Χαμηλό	Υψηλό	Υψηλό	Χαμηλό
Wang, 2013	Χαμηλό	Χαμηλό	Χαμηλό	Μέτριο	Μέτριο	Χαμηλό
Kim, 2013	Χαμηλό	Χαμηλό	Χαμηλό	Μέτριο	Μέτριο	Χαμηλό
Park, 2013	Χαμηλό	Χαμηλό	Χαμηλό	Μέτριο	Μέτριο	Χαμηλό
Shin, 2014	Χαμηλό	Χαμηλό	Χαμηλό	Μέτριο	Μέτριο	Χαμηλό
Wang, 2015	Χαμηλό	Χαμηλό	Χαμηλό	Μέτριο	Μέτριο	Χαμηλό
Wang, 2014	Χαμηλό	Χαμηλό	Χαμηλό	Μέτριο	Μέτριο	Χαμηλό
Chao, 2014	Χαμηλό	Χαμηλό	Χαμηλό	Μέτριο	Μέτριο	Χαμηλό
Garcia-Laencina, 2015	Χαμηλό	Χαμηλό	Χαμηλό	Χαμηλό	Χαμηλό	Χαμηλό
Lotfnezhad Afshar, 2015	Χαμηλό	Χαμηλό	Χαμηλό	Χαμηλό	Χαμηλό	Χαμηλό
Khalkhali, 2016	Χαμηλό	Χαμηλό	Χαμηλό	Χαμηλό	Χαμηλό	Χαμηλό
Shawky, 2016	Χαμηλό	Χαμηλό	Χαμηλό	Μέτριο	Μέτριο	Χαμηλό
Sun, 2018	Χαμηλό	Χαμηλό	Χαμηλό	Χαμηλό	Χαμηλό	Χαμηλό
Sun, 2018	Χαμηλό	Χαμηλό	Χαμηλό	Μέτριο	Μέτριο	Χαμηλό
Zhao, 2018	Χαμηλό	Χαμηλό	Χαμηλό	Μέτριο	Μέτριο	Χαμηλό
Fu, 2018	Χαμηλό	Χαμηλό	Υψηλό	Υψηλό	Υψηλό	Υψηλό
Lu, 2019	Χαμηλό	Χαμηλό	Χαμηλό	Υψηλό	Υψηλό	Χαμηλό
Abdikenov, 2019	Χαμηλό	Χαμηλό	Χαμηλό	Μέτριο	Μέτριο	Χαμηλό
Kalafi, 2019	Χαμηλό	Χαμηλό	Χαμηλό	Υψηλό	Υψηλό	Χαμηλό
Shouket, 2019	Χαμηλό	Χαμηλό	Χαμηλό	Μέτριο	Μέτριο	Χαμηλό
Ganggayah, 2019	Χαμηλό	Χαμηλό	Χαμηλό	Υψηλό	Υψηλό	Χαμηλό
Simsek, 2020	Χαμηλό	Χαμηλό	Χαμηλό	Μέτριο	Μέτριο	Χαμηλό
Salehi, 2020	Χαμηλό	Χαμηλό	Χαμηλό	Υψηλό	Υψηλό	Χαμηλό
Tang, 2020	Χαμηλό	Μέτριο	Χαμηλό	Μέτριο	Μέτριο	Μέτριο
Hussain, 2020	Χαμηλό	Χαμηλό	Χαμηλό	Μέτριο	Μέτριο	Χαμηλό

Πίνακας 33. Βαθμολογία κινδύνου μεροληψίας και αξιολόγησης υλοποίησης των 31 μελετών σύμφωνα με τα κριτήρια PROBAST [48].

Στον Πίνακας 35 παρουσιάζονται τα κύρια χαρακτηριστικά των 31 μελετών, οι περισσότερες από τις οποίες δημοσιεύθηκαν από το 2013 έως το 2020. Τα στατιστικά στοιχεία σχετικά με το πόσες από αυτές τις μελέτες δημοσιεύτηκαν ανά έτος, παρουσιάζονται στον *Πίνακας 34*. Μεταξύ αυτών, 22 μελέτες εντοπίστηκαν στην Ασία, 5 στη Βόρεια Αμερική, 2 στην Ωκεανία, 1 στην Ευρώπη και 1 στην

Αφρική. Το κύριο αποτέλεσμα πρόβλεψης ήταν η 5ετής επιβίωση των ασθενών με καρκίνο του μαστού, ενώ οι προβλεπόμενοι τύποι ασθενειών ήταν όλοι καρκίνος του μαστού [48].



Πίνακας 34.. Πλήθος μελετών που δημοσιεύτηκαν κάθε χρόνο[48].

Χαρακτηριστικά	Κατηγορία	Πλήθος (n)	Ποσοστό (%)
Περιοχή έρευνας	Ασία	22	71.0
	Βόρεια Αμερική	5	16.1
	Ωκεανία	2	6.5
	Ευρώπη	1	3.2
	Αφρική	1	3.2
Τύπος δημοσίευσης	Επιστημονικό Άρθρο	27	87.1
	Έγγραφο Συνεδρίου	3	9.6
	Ενημερωτικό Έγγραφο	1	3.2
Πηγή δεδομένων	SEER	18	58.1
	Molecular Taxonomy of Breast Cancer International Consortium	2	6.5
	The Cancer Genome Atlas	1	3.2
	Haberman's Cancer Survival Dataset	1	3.2
	Hospital Registration Data	9	29.0
Τύπος δεδομένων	Δημόσια	22	71.0
	Ιδιωτικά	9	29.0
Πλήθος κέντρων	Μονό	22	71.0
	Πολλαπλό	9	29.0
Μέγεθος δείγματος	<1000	7	22.6
	1000~10000	7	22.6
	>10000	17	54.8
Ελλιπή δεδομένα και περιγραφή επεξεργασίας	Ναι	20	64.5
	Όχι	11	35.3
Περιγραφή προ-επεξεργασίας	Ναι	31	100.0
	Όχι	0	0.0
Περιγραφή	Ναι	8	25.8

επιλογής χαρακτηριστικών	Όχι	23	74.2
Επεξεργασίας ανισορροπίας κλάσης	Ναι	24	77.4
	Όχι	2	6.5
	Άγνωστο	5	16.1
Πλήθος προβλέψεων	<10	4	12.9
	10~100	25	80.6
	>100	2	6.5
Πλήθος αλγορίθμων ΜΜ	1	5	16.1
	>1	26	83.9
Τύπος αλγορίθμου ΜΜ	DT	19	61.3
	ANN	18	58.1
	SVM	16	51.6
	LR	12	38.7
	Bayesian classification algorithms	6	19.4
	K-NN	3	9.7
	Semi-supervised learning	3	9.7
	Ensemble learning	10	32.3
	DNN	3	9.7
Παρουσίαση μοντέλου	Τύπος	6	19.4
	Γράφημα	5	16.1
	Τύπος και γράφημα	16	51.6
	Χωρίς παρουσίαση	4	12.9
Βαθμονόμηση	Ναι	1	3.2
	Όχι	30	96.8
Εσωτερική επικύρωση	Ναι	31	100.0
	Όχι	0	0.0
Εξωτερική επικύρωση	Ναι	1	3.2
	Όχι	30	96.8
Επιλογή Υπέρ-παραμέτρου	Ναι	9	29.0
	Όχι	22	71,0
Αξιολόγηση μοντέλου	Accuracy	29	93.5
	Ευαισθησία/Ανάκληση	25	80.6
	Ειδικότητα	24	77.4
	AUC	20	64.5
	Precision/PPV	6	19.4
	F1 Score	5	16.1
	Mcc	5	16.1
	NPV	2	6.5
	G-mean	2	6.5
	C-index	1	3.2
	Cutoff	1	3.2
	Youden index	1	3.2
	Retaining time	1	3.2
	FPR	1	3.2
	FDR	1	3.2
	FNR	1	3.2
Τύπος προβλέψεων	Κλινικά δεδομένα	29	93.5
	Κλινικά + Μοριακά δεδομένα	1	3.2
	Κλινικά + Μοριακά δεδομένα + Παθολογικές εικόνες	1	3.2
Κατάταξη προβλέψεων	Ναι	15	28.4
	Όχι	16	51.6

Πίνακας 35. Κύρια χαρακτηριστικά και κατηγορίες των 31 μελετών [48].

Η συγκεκριμένη μελέτη αποσκοπούσε στο να εντοπίσει αν υπάρχει βελτίωση στις επιδόσεις των μοντέλων που αφορούν την πρόβλεψη επιβίωσης από τον καρκίνο του μαστού, παρόλο που είχε κάποιους περιορισμούς όπως το ότι μελετήθηκαν μόνο αγγλικές μελέτες. Οι έρευνες που μελετήθηκαν στο πλαίσιο αυτής της έρευνας ήταν 31, με τις περισσότερες από αυτές να έχουν δημοσιευτεί μετά το 2013. Οι πιο συχνές χρησιμοποιούμενες μέθοδοι ΜΜ ήταν τα DTs (19 μελέτες, 61,3%), τα ANNs (18 μελέτες, 58,1%), οι SVMs (16 μελέτες, 51,6%), και η Ensemble Learning (10 μελέτες, 32,3%). Ο μέσος όρος του δείγματος ήταν 37256 ασθενείς, με εύρος από 200 έως 659820, ενώ ο μέσος προγνωστικός ήταν 16 (εύρος 3 έως 625). Το Accuracy 29 μελετών κυμαινόταν από 0,510 έως 0,971, η ευαισθησία 25 μελετών κυμαινόταν από 0,037 έως 1, η ειδικότητα 24 μελετών κυμαινόταν από 0,008 έως 0,993, η AUC από 20 μελέτες κυμάνθηκε από 0,500 έως 0,972 και το Precision 6 μελετών κυμαινόταν από 0,549 έως 1. Όλα τα μοντέλα επικυρώθηκαν εσωτερικά και μόνο ένα επικυρώθηκε εξωτερικά [48].

5.15. Συνοπτικοί Πίνακες Αποτελεσμάτων

Στον *Πίνακας 36* συνοψίζονται όλες οι σημαντικές πληροφορίες των μοντέλων μηχανικής μάθησης που μελετήθηκαν. Παρατηρείται πως πολλές από τις έρευνες, δεν ανέφεραν πληροφορία σχετικά με το αν έγινε χρήση κάποιου αλγορίθμου προ-επεξεργασίας δεδομένων. Επίσης, παρατηρείται πως 6 από τις 10 έρευνες κάνουν ξεκάθαρη στο κοινό την ακριβή στρατηγική δείγματος που χρησιμοποιήθηκε, ενώ ο Abd-Elnaby δεν αναφέρει στην έρευνα του τη στρατηγική που χρησιμοποίησε για να διαχειριστεί το δείγμα, παρόλο που φαίνεται να υπήρχε. Επίσης, η μέτρηση που φαίνεται να είναι κοινή σε κάθε έρευνα, είναι το Accuracy, με εξαίρεση την έρευνα του Maicas G, η οποία βασίστηκε στο TP, το FP και το χρόνο διάγνωσης. Μία ακόμη παρατήρηση που έγινε είναι πως δείγμα WBCD επαναλαμβάνεται σε πολλές μελέτες. Όλα τα μοντέλα αφορούσαν τη διάγνωση του καρκίνου του μαστού, με μόνη εξαίρεση να αποτελεί η έρευνα του Tapak το 2019, ο οποίος εστίασε στην πρόβλεψη επιβίωσης και μετάστασης, η οποία απέδωσε ελπιδοφόρα αποτελέσματα. Τέλος, η χρήση των SVM μοντέλων φαίνεται να χρησιμοποιείται πιο συχνά.

Στον Πίνακας *37* συνοψίζονται όλες οι σημαντικές πληροφορίες των μοντέλων βαθιάς μάθησης, που μελετήθηκαν κατά την εκπόνηση της εργασίας. Όλα τα μοντέλα στόχευαν στη διάγνωση του καρκίνου του μαστού. Το μοντέλο ELM φαίνεται να λειτουργεί αρκετά ικανοποιητικά, αν αναλογιστεί κανείς τη διαφορετικότητα του δείγματος δεδομένων που χρησιμοποίησε αυτή η έρευνα σε σύγκριση με άλλες. Η στήλη της προ-επεξεργασίας δεδομένων αφαιρέθηκε, αφού δεν υπήρχε πληροφορία για κανένα από τα μοντέλα, με εξαίρεση τα μοντέλα ELM και SAE-SVM των Aslan και Xiao , αντίστοιχα. Τέλος, στο κομμάτι της ταξινόμησης εικόνας και της τμηματοποίησης, φαίνεται να έχουν δοκιμαστεί διάφορα μοντέλα

με ικανοποιητικά αποτελέσματα, ενώ στο κομμάτι της ανίχνευσης φαίνεται να χρησιμοποιείται περισσότερο το μοντέλο QVCAD.

Πίνακας 36. Συνοπτικός πίνακας μοντέλων μηχανικής μάθησης.

Πρώτος Συγγραφέας	Έτος	Δεδομένα	Μοντέλο	Στρατηγική Δείγματος	Προ-επεξεργασία Δεδομένων	Αξιολόγηση
Asri H. [43]	2016	WBCD	SVM	10-fold CV		Accuracy: 97.13% Χρόνος: 0.07 s
Amrane M. [41]	2018	WBCD	K-NN	CV		Accuracy: 97.51%
Gupta M. [40]	2018	WBCD	MP	10-fold CV 70-30 Training - Test	NAI	Accuracy : 97.7% Precision: 99.2% Recall: 97.85%
Omondiagbe D. [19]	2019	WDBC	SVM-LDA	70-30 Training-Test	NAI	Accuracy: 98.82% Sensitivity: 98.41% Specificity: 99.07% AUC: 0.9994
Tapak. L. [42]	2019	Tehran' s Dataset	SVM, LDA	100-fold CV 70-30 Training - Test		Accuracy: 92 % , 93% Sensitivity: 73 % , 73% Specificity: 97 %, 97 %
			LDA			Accuracy: 86 % Sensitivity: 26 % Specificity: 97 %
Maicas G. [44]	2019	Breast MRI Dataset	RL-Det		NAI	TPR: 0.8 FPR: 3.2 Χρόνος: 92 s
Ozer M. [45]	2020	X-omics Paths	SVM	Omics: 5-fold CV Paths: 20-fold CV		Accuracy omics: >90% Accuracy paths: >90%
Salod Z. [46]	2020	BCCD	Clusters+pivot table	10-fold CV		Accuracy: 95.90%
		WBCD	Clusters+pivot table	10-fold CV		Accuracy: 99.40%
Adb-Elnaby M. [39]	2021	Microarray DNA	SVM	Αόριστο	NAI	Accuracy: 94.87%

	Πρώτος Συγγραφέας	Έτος	Δεδομένα	Μοντέλο	Στρατηγική Δείγματος	Αξιολόγηση
Ταξινόμηση εικόνας	Zhang Q. [47]	2016	SWE images (n=227)	PGBM και RBM	5-fold CV	Ευαισθησία: 88.6% Ειδικότητα: 97.1% Accuracy: 93.4% AUC: 0.947
	Han S. [47]	2017	B-mode images (n=7408)	GoogLeNet	6579-829 Training-Test	Ευαισθησία: 86% Ειδικότητα: 96% Accuracy: 90% AUC: 0.9
	Coronado-Gutierrez D. [47]	2019	B-mode images (n=118)	VGG-M model	5-fold CV	Ευαισθησία: 84.9% Ειδικότητα: 87.7% Accuracy: 86.4% AUC: 0.937
	Fujioka T. [47]	2020	B-mode images (n=1057)	GoogLeNet Inception v2	937-120 Training-Test	Ευαισθησία: 95.8% Ειδικότητα: 87.5% Accuracy: 92.5% AUC: 0.913
	Mango V.L. [47]	2020	B-mode Images (n>400,000)	Koios DS	400,000-900 Training-Test	AUC χωρίς CAD: 0.83 AUC με CAD: 0.87
	Fujioka T. [47]	2020	SWE images (n=377)	DenseNet 169	304-73 Training-Test	Ευαισθησία: 85.7% Ειδικότητα: 78.9% AUC: 0.898
Ανίχνευση αντικειμένου	Jiang Y.[47]	2018	ABUS images (n>20,000)	QVCAD	20,000-185 Training-Test	AUC χωρίς CAD: 0.828 AUC με CAD: 0.848
	Xu X.[47]	2018	ABUS images(n>20,000)	QVCAD	20,000-1000 Training-Test	AUC χωρίς CAD: 0.747 AUC με CAD: 0.784

Πίνακας 37. Συνοπτικός πίνακας μοντέλων βαθιάς μάθησης.

Πρώτος Συγγραφέας	Έτος	Δεδομένα	Μοντέλο	Στρατηγική Δείγματος	Αξιολόγηση
Cao Z. [47]	2019	B-mode images (n=1043)	SSD300	860-183 Training-Test	Precision: 96.89% Ανάκληση: 67.23% F1: 79.38%
Yang S. [47]	2019	ABUS images (n>20,000)	QVCAD	20,000-1485 Training-Test	AUC χωρίς CAD: 0.88 AUC με CAD: 0.91 Ευαισθησία χωρίς CAD: 67% Ευαισθησία με CAD: 88%
Τμηματοποίηση					
Kumar V.[47]	2018	B-mode images (n=433)	Multi U-net	372-61 Training-Test	Μέσος συντελεστής Dice: 0.82 Πραγματικά θετικά: 0.84 Λανθασμένα θετικά: 0.01
Zhuang Z. [47]	2019	B-mode images (n=1062)	RDAU-NET	857-205 Training-Test	Precision: 88.58% Ανάκληση:83.19% F1: 84.78 %
Hu Y.[47]	2019	B-mode images (n=570)	DFCN + PBAC	400-170 Training-Test	Συντελεστής Dice: 88.97% Απόσταση Hausdorff: 35.54 pixels Μέση απόλυτη απόκλιση: 7.67 pixels
Λοιπές έρευνες					
Aslan M. [38]	2018	WCCD	ELM	80-20 Training-Test	Accuracy: 80% Χρόνος: 0.0075 s
Xiao Y. [32]	2018	WDBC	SAE-SVM	10-fold CV	Accuracy: 98.25%
Sharma A. [31]	2021	WDBC	GoogLeNet AlexNet	Αόριστο	Accuracy όλων των χαρακτηριστικών του δείγματος

Πίνακας 38. Συνέχεια πίνακα 37

Κεφάλαιο 6 – Συμπεράσματα

Από το 2015 μέχρι το τέλος του 2020, υπήρχαν 7,8 εκατομμύρια ζωντανές γυναίκες που διαγνώστηκαν με καρκίνο του μαστού. Η ανάπτυξη της έρευνας στον τομέα της τεχνητής νοημοσύνης για την αντιμετώπιση του καρκίνου του μαστού, έχει οδηγήσει σε αύξηση των δημοσιεύσεων σε αυτόν τον τομέα. Στον *Πίνακας 29* και στον *Πίνακας 34*, φαίνεται μία σημαντική αύξηση, από το 2018 και μετά, στις δημοσιεύσεις σχετικά με την αντιμετώπιση του καρκίνου του μαστού. Επομένως, η ανάγκη να αντιμετωπιστεί και να καταπολεμηθεί η νόσος με την εύρεση νέων τεχνολογικών εργαλείων, είναι τεράστια.

Από τον *Πίνακας 36*, τον *Πίνακας 37* και την ανασκόπηση που πραγματοποιήθηκε συνολικά, παρατηρήθηκε πως οι επιδόσεις των μοντέλων ΜΜ δεν παρουσιάζουν κατ' ανάγκη κάποια βελτίωση. Ωστόσο, σε κάποια μοντέλα τα βήματα προ-επεξεργασίας δεδομένων είτε δεν υπήρχαν, είτε δεν διευκρινιζόταν ποια μέθοδος ακολουθήθηκε, είτε ο συγγραφέας παρέλειπε να το διευκρινίσει. Από [19] είδαμε πως οι τεχνικές ΜΜ σε συνδυασμό με τις μεθόδους εξαγωγής χαρακτηριστικών, μπορούν να κάνουν καλύτερες προσεγγίσεις προκειμένου να διαγνωσθεί ο καρκίνος του μαστού. Επιπροσθέτως, οι υπερβολικές διαφορές στην επιλογή χαρακτηριστικών του δείγματος και ζητήματα που σχετίζονται με την επικύρωση αποτελούν και αυτά ένα εμπόδιο. Τελικά, παρατηρήθηκε πως κάθε μεθοδολογία ήταν διαφορετική, οι διαφορές μεταξύ των δημοσιεύσεων που μελετήθηκαν ήταν υπερβολικές, περιορίζοντας με αυτόν τον τρόπο τη σύγκριση μεταξύ τους.

Επιπλέον, οι πληροφορίες σχετικά με τις προγνωστικές επιδόσεις (accuracy, AUC, Ευαισθησία, Ειδικότητα κλπ.) ήταν, επίσης, ανεπαρκείς και οι περισσότερες από τις μελέτες περιέγραψαν μόνο μία διάσταση της προγνωστικής απόδοσης. Επομένως, συνιστάται να αναφέρονται λεπτομερώς ολοκληρωμένες μεθοδολογικές πληροφορίες, γεγονός που θα επιτρέψει στο ευρύτερο κοινό να υλοποιήσει και να μελετήσει τα διάφορα μοντέλα, προκειμένου να δοκιμαστούν και να αναβαθμιστούν. Δεν είναι τυχαίο που όλες οι έρευνες που μελετήθηκαν, επικεντρώθηκαν στην ανάπτυξη μοντέλων πρόβλεψης χρησιμοποιώντας αλγορίθμους ΜΜ, αντί να επικυρώνουν ή να αναβαθμίζουν τα ήδη υπάρχοντα μοντέλα.

Επίσης, τα περισσότερα μοντέλα αφορούσαν το κομμάτι της διάγνωσης. Οι μελέτες που ασχολούνται με την πρόβλεψη επιβίωσης και την πιθανότητα μετάστασης είναι, συγκριτικά, λίγες. Προκειμένου να αποφασισθεί η κατάλληλη θεραπεία, δεν αρκεί μόνο να διαγνωσθεί η νόσος. Η πιθανότητα επιβίωσης και μετάστασης είναι εξίσου σημαντικές πληροφορίες, που θα παίξουν καθοριστικό ρόλο στην επιλογή της κατάλληλης θεραπείας. Γι΄ αυτό το λόγο, απαιτείται περαιτέρω διερεύνηση με τη χρήση μεγάλων συνόλων δεδομένων για τη σύσταση ενός χρήσιμου εργαλείου για την επιβίωση από τον καρκίνο του μαστού και την πρόβλεψη μετάστασης. Επομένως, δεν μπορεί να προταθεί με ασφάλεια κάποιο

μοντέλο για την αντιμετώπιση της νόσου, ωστόσο, φαίνεται πως τα SVM μοντέλα και οι διάφορες τεχνικές βαθιάς μάθησης προτιμώνται και δείχνουν να λειτουργούν αποδοτικότερα.

Είναι αναμενόμενο πως η εισαγωγή μίας τέτοιας καινοτομίας στο σύστημα περίθαλψης, θα βελτιώσει τη διαγνωστική απόδοση, θα διευκολύνει την εργασία των ιατρών και θα μειώσει το κόστος υγειονομικής περίθαλψης. Οι ιατροί θα είναι σε θέση να περνούν περισσότερο χρόνο επικοινωνώντας με ασθενείς και άλλους ιατρούς, γεγονός που θα βελτιώσει την ποιότητα της περίθαλψης. Αν και έχουν αναπτυχθεί ορισμένα αρμόδια συστήματα βαθιάς μάθησης, τα συστήματα αυτά εφαρμόζονται μόνο για περιορισμένο σκοπό ή με συγκεκριμένο τρόπο. Στο μέλλον που θα έχουν ξεπεραστεί όλα τα εμπόδια που προαναφέρθηκαν, θα είναι απαραίτητο να αναπτυχθεί ένα εξαιρετικά ευέλικτο σύστημα που θα συνδυάζει πολλαπλά σύνολα δεδομένων (εικόνες από μαστογραφίες, δεδομένα εξετάσεων αίματος και κλινικές πληροφορίες, όπως τα συμπτώματα των ασθενών) για την ολοκληρωμένη διάγνωση και τη διαχείριση της θεραπείας.

Επίσης, από την Εικόνα *44* μπορεί κανείς να αντιληφθεί πως η μεγαλύτερη έρευνα για την επίλυση του προβλήματος, πραγματοποιείται στις αναπτυγμένες χώρες. Επομένως, στο μέλλον, θα πρέπει να τεθούν σε εφαρμογή κανονισμοί για την ανάπτυξη διεθνών συνεργασιών μεταξύ ανεπτυγμένων και αναπτυσσόμενων χωρών, με σκοπό να επιτραπεί η ανταλλαγή πολύτιμης γνώσης και πόρων, μεταξύ αναπτυγμένων και αναπτυσσόμενων χωρών, προκειμένου να αυξηθούν οι παραγωγικοί πόροι αλλά και να εξασφαλιστεί πως όλοι οι άνθρωποι θα έχουν πρόσβαση σε αυτά τα τεχνολογικά μέσα για τη διασφάλιση της υγείας τους.

Η μηχανική μάθηση είναι μια νέα μεθοδολογία για την αντιμετώπιση του καρκίνου του μαστού και εξακολουθούν να υπάρχουν πολλά περιθώρια βελτίωσης, αφού προηγούμενες μελέτες αποκάλυψαν ότι ακόμη και αποτελεσματικά μοντέλα διάγνωσης του καρκίνου του μαστού, μπορούν να διαγνώσουν λανθασμένα αλλοιώσεις, ακριβώς όπως και ο άνθρωπος. Όλα αυτά τα συστήματα παραμένουν μια συμπληρωματική-βοηθητική μέθοδος και η τελική απόφαση ανήκει στον ιατρό. Τα υπάρχοντα μοντέλα πρόβλεψης είναι πολλά σε πλήθος και εξακολουθούν να αντιμετωπίζουν περιορισμούς. Οι ερευνητές και οι ιατροί θα πρέπει να επιλέξουν προσεκτικά ένα μοντέλο, να το σχεδιάσουν με αυστηρές μεθόδους σχεδιασμού και επικύρωσης, να το εκπαιδεύσουν και να το δοκιμάσουν με χρήση ενός μεγάλου δείγματος ερευνητικών δεδομένων υψηλής ποιότητας, να το αξιολογήσουν αυστηρά και, στο τέλος, να το ελέγξουν στην πράξη.

Βιβλιογραφία

[1] World Health Organization, “Breast Cancer,” 2021. https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

[2] P. N. Suganthan, “Jou rna lP,” *Expert Syst. Appl.*, p. 114161, 2020, doi: 10.1016/j.eswa.2020.114161.

[3] Γ. Α.-Ο. Ν. Αθηνών, “Καρκίνος μαστού.” http://www.agsavvas-hosp.gr/Μάθεγιατονκαρκίνο/Πρόληψη/Πρωτογενήςπρόληψη/Καρκίνοςμαστού.aspx.

[4] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, and R. Mcbride, “Deep Learning to Improve Breast Cancer Detection on Screening Mammography,” no. May, pp. 1–12, 2019, doi: 10.1038/s41598-019-48995-4.

[5] Γ. Α.-Ο. Ν. Αθηνών, “Τί είναι ο καρκίνος;” http://www.agsavvas-hosp.gr/Μάθεγιατονκαρκίνο/Πληροφορίες/Τίείναιοκαρκίνος;.aspx.

[6] “Δεδομένα για τον καρκίνο στην Ελλάδα από τον Παγκόσμιο Οργανισμό Υγείας.” https://www.almazois.gr/2018/09/greece-cancer-facts/.

[7] G. C. Map, “No Title.” http://globalcancermap.com/.

[8] R. Anyoha, “Can Machines Think?,” 2017. https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/.

[9] Σ. Βλαχάβας, Κεφάλας, Βασιλειάδης, Κοκκόρας, *Τεχνητή Νοημοσύνη*, Γ’ έκδοση. 2006.

[10] Y. Mintz and R. Brodie, “Introduction to artificial intelligence in medicine,” *Minim. Invasive Ther. Allied Technol.*, vol. 28, no. 2, pp. 73–81, 2019, doi: 10.1080/13645706.2019.1575882.

[11] K. S. Hughes, J. Zhou, Y. Bao, M. S. Preeti, J. Wang, and K. Yin, “Natural language processing to facilitate breast cancer research and management,” no. October, pp. 1–8, 2019, doi: 10.1111/tbj.13718.

[12] S. Kulkarni, N. Seneviratne, M. S. Baig, and A. H. A. Khan, “Artificial Intelligence in Medicine: Where Are We Now?,” *Acad. Radiol.*, vol. 27, no. 1, pp. 62–70, 2020, doi: 10.1016/j.acra.2019.10.001.

[13] W. L. Bi *et al.*, “Artificial intelligence in cancer imaging: Clinical challenges and applications,” *CA. Cancer J. Clin.*, vol. 0, no. 0, pp. 1–31, 2019, doi: 10.3322/caac.21552.

[14] F. Y. Osisanwo, J. E. T. Akinsola, O. Awodele, J. O. Hinmikaiye, O. Olakanmi, and J. Akinjobi, “Supervised Machine Learning Algorithms : Classification and Comparison,” no. July, 2017, doi: 10.14445/22312803/IJCTT-V48P126.

[15] D. Fum, “Types of Maching Learning Algorithms You Should Know.” https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861.

[16] N. Al-azzam, I. Shatnawi, and D. Ph, “Comparing supervised and semi-supervised Machine Learning Models on Diagnosing Breast Cancer,” *Ann. Med. Surg.*, vol. 62, no. November 2020, pp. 53–64, 2021, doi: 10.1016/j.amsu.2020.12.043.

[17] J. J. Guido, P. C. Winters, A. B. Rains, and R. Medical, “Logistic Regression Basics,” pp. 1–7, 2006.

[18] “Machine Learning.” https://www.javatpoint.com/machine-learning.

[19] M. Mathur, “Machine Learning Classification Techniques for Breast Cancer Diagnosis Machine Learning Classification Techniques for Breast Cancer Diagnosis,” 2019, doi: 10.1088/1757-899X/495/1/012033.
[20] “Κεφάλαιο 3 : Λογιστική Παλινδρόμηση,” pp. 89–125.
[21] A. Yadav, “Support Vector Machines (SVM).” https://towardsdatascience.com/support-vector-machines-svm-c9ef22815589.
[22] F. Maleki, K. Ovens, and K. Najafian, “Overview of Machine Learning Part 1 Fundamentals and Classic Approaches,” *Neuroimaging Clin. NA*, vol. 30, no. 4, pp. e17–e32, doi: 10.1016/j.nic.2020.08.007.
[23] S. D. Brown, A. J. Myles, and U. States, *Decision Tree Modeling ☆*, 2nd ed. Elsevier Inc., 2019.
[24] “K-Nearest Neighbor(KNN) Algorithm for Machine Learning.” https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning.
[25] “Unsupervised Learning,” 2020. https://www.ibm.com/cloud/learn/unsupervised-learning.
[26] “K-Means Clustering Algorithm.” https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning.
[27] M. Wiering, *ALO 12 - Reinforcement Learning*. .
[28] Κ. Γεωργούλη, *Τεχνητή Νοημοσύνη Μια Εισαγωγική Προσέγγιση*. 2015.
[29] “Artificial Neural Network Tutorial.” https://www.javatpoint.com/artificial-neural-network.
[30] I. Goodfellow, “Deep Learning,” 2015.
[31] D. Gupta, A. Khanna, A. Ella, and H. Sameer, *International Conference on Innovative Computing and Communications*, vol. 1. 2020.
[32] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, “Breast Cancer Diagnosis Using an Unsupervised Feature Extraction Algorithm Based on Deep Learning,” *2018 37th Chinese Control Conf.*, pp. 9428–9433, 2018.
[33] E. Kavlakoglu, “AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What’s the Difference?,” 2020. https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks.
[34] M. Robins, “The Difference Between Artificial Intelligence, Machine Learning and Deep Learning,” 2020. https://www.intel.com/content/www/us/en/artificial-intelligence/posts/difference-between-ai-machine-learning-deep-learning.html.
[35] A. Osareh, “Machine Learning Techniques to Diagnose Breast Cancer,” 2008.
[36] R. Sarmento, E. Text, and M. Visualization, “Breast Cancer Wisconsin (Diagnostic) Data Set,” no. November, 2019, doi: 10.13140/RG.2.2.24243.99364.
[37] “Breast Cancer Wisconsin (Original) Data Set.” https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28original%29.
[38] M. F. Aslan, Y. Celik, K. Sabanci, and A. Durdu, “Breast Cancer Diagnosis by Different Machine Learning Methods Using Blood Intelligent Systems and

Applications in Engineering Breast Cancer Diagnosis by Different Machine Learning Methods Using Blood Analysis Data," no. January 2019, 2018, doi: 10.1039/b000000x.
[39] M. Abd-elnaby, M. Alfonse, and M. Roushdy, "Classification of breast cancer using microarray gene expression data : A survey," *J. Biomed. Inform.*, vol. 117, no. April, p. 103764, 2021, doi: 10.1016/j.jbi.2021.103764.
[40] M. Gupta, "A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques," no. Iccmc, pp. 997–1002, 2018.
[41] M. Amrane, "Breast cancer classification using machine learning," *2018 Electr. Electron. Comput. Sci. Biomed. Eng. Meet.*, pp. 1–4.
[42] L. Tapak, N. Shirmohammadi-khorram, P. Amini, and B. Alafchi, "Prediction of survival and metastasis in breast cancer patients using machine learning classifiers," *Clin. Epidemiol. Glob. Heal.*, vol. 7, no. 3, pp. 293–299, 2019, doi: 10.1016/j.cegh.2018.10.003.
[43] H. Asri, H. Mousannif, H. Al, and T. Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," *Procedia - Procedia Comput. Sci.*, vol. 83, no. Fams, pp. 1064–1069, 2016, doi: 10.1016/j.procs.2016.04.224.
[44] L. Lu, *Advances in Computer Vision and Pattern Recognition Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics*. .
[45] M. E. Ozer, P. O. Sarica, and K. Y. Arga, "New Machine Learning Applications to Accelerate Personalized Medicine in Breast Cancer: Rise of the Support Vector Machines," *Omi. A J. Integr. Biol.*, vol. 24, no. 5, pp. 241–246, 2020, doi: 10.1089/omi.2020.0001.
[46] S. Reviews, Z. Salod, and Y. Singh, "A five-year (2015 to 2019) analysis of studies focused on breast cancer prediction using machine learning : A systematic review and bibliometric analysis on m er ci al us e on on m er ci al us e on," vol. 9, 2020.
[47] T. Fujioka *et al.*, "The Utility of Deep Learning in Breast Ultrasonic Imaging : A Review."
[48] J. L. Id *et al.*, "Predicting breast cancer 5-year survival using machine learning : A systematic review," vol. 16, pp. 1–23, 2021, doi: 10.1371/journal.pone.0250370.