ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΒΙΟΪΑΤΡΙΚΗΣ

Εφαρμογές της Μηχανικής Μάθησης σε επιλεγμένα προβλήματα λειτουργικής ταξινόμησης βιομορίων ή/και φαρμάκων

Γεωργίου Γρηγόρης

Αριθμός Μητρώου: 15007

Επιβλέπων Καθηγητής

Διονύσης Κάβουρας, Ομότιμος Καθηγητής

Λευκωσία 01/06/2021

Η Τριμελής Εξεταστική Επιτροπή

Ο Επιβλέπων Καθηγητής		
Διονύσης Κάβουρας	Γιώργος Σπύρου	Παντελής Ασβεστάς
Ομότιμος Καθηγητής	Καθηγητής	Αναπλ. Καθηγητής
Dionysios Kavouras Digitally signed by Dionysios Kavouras Date: 2021.10.25 15:00:48 +03'00'		 Digitally signed by Panteleimon Asvestas

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Ο υπογράφων Γρηγόρης Γεωργίου του Γεώργιου, με αριθμό μητρώου 15007 φοιτητής του Τμήματος Μηχανικών Βιοϊατρικής του Πανεπιστημίου Δυτικής Αττικής, δηλώνω υπεύθυνα ότι:

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

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

Ημερομηνία

25/07/2021

Ο/Η Δηλών/ούσα

Περίληψη

Σκοπός
Στην σύγχρονη Βιοϊατρική έρευνα προκύπτουν συνεχώς ερωτήματα ταξινόμησης σε διάφορα πεδία. Καθώς η έρευνα έχει προχωρήσει με γοργούς ρυθμούς σε μοριακό επίπεδο, υπάρχουν διαθέσιμα μεγάλα σύνολα δεδομένων για τα οποία είναι αναγκαίο να προβούμε σε λειτουργικές ταξινομήσεις. Η διπλωματική στοχεύει στην ανασκόπηση του πεδίου εφαρμογών της Μηχανικής Μάθησης, όπου θα αναδειχθεί πόσο ικανές είναι αυτοί οι μέθοδοι να διαχειρίζονται δεδομένα από διάφορα πεδία Βιοπληροφορικής αλλά να προβλέπουν μελλοντικά γεγονότα και να ανακαλύπτουν κρυφές πληροφορίες.

Υλικά & Μεθοδολογία
Οι εφαρμογές Μηχανικής Μάθησης βασίζονται από υπολογιστικούς αλγορίθμους και σύνολα δεδομένων. Ουσιαστικά, συλλέχθηκαν τέσσερα σύνολα δεδομένων από διαφορετικά πεδία Βιοπληροφορικής όπου το πρώτο αφορά την Λευχαιμία, το δεύτερο το καρκίνο του μαστού, το τρίτο την νόσο του Πάρκινσον και το τέταρτο την διαπερατότητα του αιματοεγκεφαλικού φραγμού από τα φάρμακα. Με την χρήση Η/Υ και της γλώσσας προγραμματισμού R έγινε προεπεξεργασία για κάθε σύνολο δεδομένων ξεχωριστά και στην συνέχεια έγινε εκπαίδευση, επικύρωση και αξιολόγηση για ευρέως γνωστούς αλγορίθμους που χρησιμοποιούνται στην Μηχανική Μάθηση. Επίσης, για την αξιοπιστία των εφαρμογών αυτών χρησιμοποιήθηκαν μετρητικές μέθοδοι όπως ακρίβεια, ευαισθησία, ειδικότητα κ.τ.λ.

Αποτελέσματα
Η Μηχανική Μάθηση έδειξε για κάθε διαφορετικό πεδίο πόσο ικανή είναι χειριστεί σύνολα δεδομένων. Εντόπισε τα κύρια χαρακτηριστικά που διακρίνουν καλύτερα το κάθε σύνολο δεδομένων αλλά και τον αλγόριθμο που έχει την καλύτερη επίδραση στην ταξινόμηση τους. Βάσει λοιπόν των επιλεγμένων χαρακτηριστικών έγινε βιβλιογραφική ανασκόπηση για την επαλήθευση των αποτελεσμάτων.

Συμπεράσματα
Από τις επιδόσεις της είναι εμφανές ότι είναι μια τεχνική χαμηλού κόστους που έχει προοπτικές να γίνει αρκετά χρήσιμη στην Βιοϊατρική έρευνα και να αντικαταστήσει χρονοβόρες και ελαττωματικές μεθόδους αλλά και επεμβατικές τεχνικές που είναι επιβλαβείς για τους ασθενείς. Σε μελλοντικές έρευνες καλύτερο θα ήταν να γίνουν αναλύσεις με περισσότερα δείγματα ούτως ώστε να υπάρχουν πιο αξιόπιστα αποτελέσματα.

Λέξεις κλειδιά: Μηχανική Μάθηση, Βιοπληροφορική, Βιοϊατρική έρευνα, Βαθιά Μάθηση, Λευχαιμία, Καρκίνος, Πάρκινσον, Αιματοεγκεφαλικός φραγμός,

Abstract

Aim
Modern biomedical research constantly raises classification questions in various fields. Due to rapid molecular research progression, it is necessary to perform functional classifications for the large available data sets. This thesis aims to review the applications of Machine Learning in Bioinformatics and collect certain data sets that are available in public scientific community, which will be used to train, validate and test selected Machine Learning algorithms.

Materials & Methods
Machine Learning applications are based on computational algorithms and datasets. Four datasets, from different fields of Bioinformatics were collected. The first concerns Leukemia, the second breast cancer, the third Parkinson's disease and the fourth is about drug permeability of blood- brain barrier. PC and R- programming language were used to individually pre-process each data set. Afterwards, training, validation and evaluation were performed for well-known algorithms of Machine Learning.

Results
Machine Learning has shown, for each field, how capable it is in handling datasets. It has identified the best features that discriminate each data set and the algorithm with the best effect on their classification. Based on feature selection, a literature review was carried out to verify the results.

Conclusion
Considering its achievements is obvious that is a low cost technique which can potentially become quite useful in biomedical research and replace time consuming, defective and harmful invasive techniques.

Key words: Machine Learning, Bioinformatics, Biomedical research, Deep Learning, Leukaemia, Cancer, Parkinson, Blood-Brain Barrier

Ευχαριστίες

Επιθυμώ να εκφράσω τις θερμές μου ευχαριστίες στον επιβλέποντα Καθηγητή μου κ. Διονύση Κάβουρα για τις βασικές αρχές που με δίδαξε επί του θέματος αλλά και για την εμπιστοσύνη που μου έδειξε. Στη συνέχεια θα ήθελα να ευχαριστήσω ιδιαίτερα τον Καθηγητή κ. Γιώργο Σπύρου επικεφαλής του τμήματος Βιοπληροφορικής στο Ινστιτούτο Νευρολογίας και Γενετικής Κύπρου για το θέμα που μου ανέθεσε αλλά και την άψογη συνεργασία που είχα με τον ίδιο και την ερευνητική του ομάδα. Επίσης, έχω την υποχρέωση να ευχαριστήσω το φίλο και συνάδελφο Σωτήρη Ουζούνη για την πολύτιμη στήριξη και βοήθεια που μου παρείχε καθ' όλη την διάρκεια εκπόνησης της διπλωματικής μου εργασίας. Τέλος, ευχαριστώ πολύ την οικογένεια μου και τους φίλους μου για την στήριξη και την κατανόηση που μου έδειξαν όλο αυτό το διάστημα.

ΠΕΡΙΕΧΟΜΕΝΑ

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Κατάλογος συντομογραφιών	
MM	Μηχανική Μάθηση
AML	Acute Myeloid Leukemia (Οξεία μυελογενή λευχαιμία)
ALL	Acute Lymphoblastic Leukemia (Οξεία λεμφοβλαστική λευχαιμία)
BBB	Blood-Brain Barrier (Αιματοεγκεφαλικός φραγμός)
TSS	Transcription Start Sites (Σημεία έναρξης μεταγραφής)
UTR	Untranslated region (Μη μεταφρασμένη περιοχή μορίου)
GO	Gene Ontology (Οντολογία γονιδίων)
MS	Mass Spectrometry (Φασματογράφος Μάζας)
GA	Genetic Algorithm (Εξελικτικός αλγόριθμος)
SCMF	Self-Consistent mean field (Αυτόνομο μέσο πεδίο)
K-NN	K-Nearest Neighbors (Κ-πλησιέστερων γειτόνων)
NER	Named Entity Recognition (Αναγνώριση ονομάτων οντοτήτων)
KDT	Knowledge Discovery in Text (Ανακάλυψη γνώσεων σε κείμενα)
SVM	Support Vector Machine (Μηχανή διανυσμάτων υποστήριξης)
BN	Bayesian Network (Μπεϋζιανό δίκτυο)
ANN	Artificial Neural Networks (Τεχνητά νευρωνικά δίκτυα)
BM	Βαθιά Μάθηση
DNN	Deep Neural Network (Βαθιά νευρωνικά δίκτυα)
BP	Backpropagation (Οπισθοδιάδοση)
SDG	Stochastic Gradient Descent (Στοχαστική Κατάβαση Δυναμικού)
MLP	Multilayer Perceptron (Πολλαπλές κατηγορίες Perceptron)
DBN	Deep Belief Networks (Δίκτυα βαθιάς πεποίθησης)
SAE	System Architecture Evolution (Εξέλιξη αρχιτεκτονικής συστήματος)
RBM	Restricted Boltzmann Machine (Περιορισμένες μηχανές Boltzmann)
AE	Auto Encoders (Αυτόματος κωδικοποιητής)
ΠΟΥ	Παγκόσμιος Οργανισμός Υγείας
FNA	Fine Needle Aspiration (Αναρρόφηση δια λεπτής βελόνης)
ΜΕΘ	Μονάδα Εντατικής Θεραπείας
CNS	Central Nervous System (Κεντρικό Νευρικό Σύστημα)
QSAR	Quantitative Structure – Activity Relationship (Ποσοτική σχέση – δομής δραστικότητας)
IDE	Integrated Development Environment (Ολοκληρωμένο περιβάλλον ανάπτυξης)
API	Application Programming Interface (Διεπαφή προγραμματισμού εφαρμογών)
ROC	Receiver Operating Characteristic (Λειτουργικού χαρακτηριστικού δέκτη
AUC	Area Under the Curve (Περιοχή κάτω από την καμπύλη
MFCC	Mel Frequency Cepstral Coefficients (Συντελεστές συχνότητας Cepstral Mel)
TWQT	Tunable Q-factor Wavelet Transform (Συντονισμένοι μετασχηματισμοί παράγοντα Q)
RFE	Recursive Feature Elimination (Αναδρομική εξάλειψη χαρακτηριστικών)
LDA	Linear Discriminant Analysis (Γραμμική διαχωριστική ανάλυση)
RF	Random Forest (Τυχαίο δάσος)
GBM	Gradient Boosting Machine (Μηχανή ενίσχυσης κλίσης
QDA	Quadratic Discriminant Analysis (Ανάλυση τετραγωνικών διακρίσεων)
RFE	Recursive Feature Elimination (Αναδρομική εξάλειψη χαρακτηριστικών)

ΕΙΣΑΓΩΓΗ

Η μηχανική μάθηση (Machine Learning - ΜΜ) είναι ένα πεδίο της τεχνητής νοημοσύνης (Artificial Intelligence - AI), που εφαρμόζεται με υπολογιστικούς αλγόριθμους συνθέτοντας σχέσεις μεταξύ δεδομένων και πληροφοριών. Θεωρείται μία από τις καινοτόμες τεχνολογίες της δεκαετίας μας λόγω του ότι επιτρέπει σε μηχανές και υπολογιστές να προσομοιώνουν ανθρωπιστικές γνωστικές ικανότητες, προβλέποντας μελλοντικά γεγονότα με γνώμονα τις παρελθοντικές εμπειρίες. Αυτή η τεχνολογία έχει εφαρμοστεί σε ένα ευρύ φάσμα τομέων όπως η αναγνώριση προτύπων, η όραση υπολογιστών, η μηχανική διαστημικών σκαφών ,τα χρηματοοικονομικά , η ψυχαγωγία, η οικολογία, οι ιατρικές εφαρμογές και στην βιοϊατρική έρευνα. Ένας άλλος τομέας που έχει επωφεληθεί σημαντικά από τη χρήση αλγορίθμων ΜΜ είναι οι φαρμακευτικές εταιρείες. Συγκεκριμένα έχουν χρησιμοποιηθεί για την ανάπτυξη διαφόρων μοντέλων για την πρόβλεψη χημικών, βιολογικών και φυσικών χαρακτηριστικών των ενώσεων στην ανακάλυψη φαρμάκων.

Έχουν αναπτυχθεί διάφορα υπολογιστικά εργαλεία από την κοινότητα της Βιοπληροφορικής για ανάλυση δεδομένων, χρησιμοποιώντας συμβατικούς αλγόριθμους μέσω των επιστημών πληροφορικής. Ωστόσο, πολλά από τα εργαλεία που χρησιμοποιήθηκαν καθίστανται ανίκανα να αντιμετωπίσουν προβλήματα του πραγματικού κόσμου και αυτό οφείλεται κυρίως στην πολυπλοκότητα των βιολογικών συστημάτων και στην έλλειψη θεμελιώδους θεωρίας σε μοριακό επίπεδο. Ένας άλλος κύριος λόγος είναι ότι τα συμβατικά υπολογιστικά εργαλεία δεν είναι σε θέση να χειριστούν τη μεγάλη και ταχέως αυξανόμενη ποσότητα δεδομένων. Επομένως, οι μεθόδοι ΜΜ έχουν γίνει μια από τις πιο αγαπημένες τεχνικές στην βιοϊατρική έρευνα για την ικανότητα που έχουν να μαθαίνουν αυτόματα από τα διαθέσιμα δεδομένα και να παράγουν χρήσιμες πληροφορίες.

Οι βιομοριακές βάσεις δεδομένων αναπτύσσονται ραγδαία. Το να συνοψίσουμε το τεράστιο ποσό των βιολογικών δεδομένων που έχουμε στην διάθεση μας σε ουσιαστικά μοντέλα, για να κατανοηθεί ο πλήρης μηχανισμός των ασθενειών φαίνεται όλο και πιο δύσκολο. Επομένως υπάρχει ανάγκη να αυξηθεί η αλλήλεπίδραση μεταξύ ΜΜ και Βιοπληροφορικής για την εξερεύνηση, την ανακάλυψη και την επεξήγηση στις βάσεις δεδομένων, που έχει αναδειχθεί ως η μεγαλύτερη πρόκληση για τον κλάδο της Βιοπληροφορικής. Η ανάλυση μεγάλων συνόλων βιολογικών δεδομένων απαιτεί την κατανόηση των δεδομένων. Παραδείγματα αυτού του τύπου ανάλυσης περιλαμβάνουν πρόβλεψη δομής πρωτεΐνης, ταξινόμηση γονιδίων, ταξινόμηση καρκίνου με βάση δεδομένα μικροσυστοιχιών, στατιστική μοντελοποίηση αλληλεπίδρασης πρωτεϊνών κ.λπ. Κάθε μία από αυτές τις εργασίες μπορεί να θεωρηθεί ως πρόβλημα στην ΜΜ.

Σε αυτή τη μελέτη με βάση την βιβλιογραφική έρευνα που πραγματοποιήθηκε, έγινε συλλογή δεδομένων τα οποία είναι διαθέσιμα στην επιστημονική κοινότητα και με την βοήθεια βάσεων δεδομένων συμπληρώθηκαν οι ελλιπείς πληροφορίες. Πιο συγκεκριμένα, εφαρμόστηκε ΜΜ σε συνολικά 4 σύνολα δεδομένων: 1) Με βάση τις τιμές έκφρασης γονιδίων ασθενών που έπασχαν είτε από οξεία μυελογενή λευχαιμία (Acute Myeloid Leukemia - AML) είτε από οξεία λεμφοβλαστική λευχαιμία (Acute

Lymphoblastic Leukemia - ALL) υλοποιήθηκε ένα μοντέλο διάκρισης του τύπου λευχαιμίας, 2) με δείγματα από ασθενείς που είχαν διαγνωστεί με καλοήθη ή κακοήθη καρκίνο του μαστού, κατασκευάστηκε μοντέλο για την διάκριση του τύπου καρκίνου, 3) μέσω δεδομένων ηχογραφημένων φωνών από υγιείς ασθενείς και από νοσούντες της ασθένειας Πάρκινσον, εντοπίστηκαν οι συχνότητες που διαχωρίζουν αν κάποιος πάσχει από Πάρκινσον ή όχι, 4) με μοριακά χαρακτηριστικά φαρμάκων που είτε διαπερνούν τον εγκεφαλικό φραγμό (Blood-Brain Barrier - BBB) είτε όχι, υλοποιήθηκε μοντέλο για την πρόβλεψη της διαπερατότητας των φαρμάκων. Με την διαδικασία αυτή έγινε προσπάθεια ανάπτυξης προγραμμάτων για την ανάδειξη της ΜΜ στα πεδία της Βιοπληροφορικής, Έτσι διαφαίνεται πόσο χρήσιμη είναι λόγω χαμηλού κόστους, εξοικονόμησης χρόνου και αποτελεί ένα εναλλακτικό τρόπο από την χρήση συμβατικών υπολογιστικών μεθόδων.

Στην σύγχρονη βιοιατρική έρευνα, προκύπτουν συνεχώς ερωτήματα ταξινόμησης σε διάφορα επίπεδα. Κατέχοντας μεγάλα σύνολα μοριακών δεδομένων είναι απαραίτητο να προβούμε σε λειτουργικές ταξινομήσεις. Στην διπλωματική αυτή έγινε: (I) Ανασκόπηση του πεδίου εφαρμογών της ΜΜ στην Βιοπληροφορική, (II) Συλλογή ορισμένων συνόλων δεδομένων τα οποία είναι δημοσίως διαθέσιμα στην επιστημονική κοινότητα, (III) εκπαίδευση, έλεγχος και επαλήθευση εγκυρότητας επιλεγμένων αλγορίθμων ΜΜ (IV) Τέλος, εξετάστηκε η αξιοπιστία των τεχνικών αυτών για το πόσο ικανές είναι να βοηθούν σε διάφορους τομείς της Βιοπληροφορικής,

Στο 1ο κεφάλαιο θα παρουσιαστούν έννοιες, ορισμοί αλλά και ορολογίες που συνάδουν με την έρευνα αυτή. Αναλυτικά, με βιβλιογραφική ανασκόπηση θα αναδειχθεί η αλληλεπίδραση της ΜΜ στην Βιοϊατρική έρευνα και διάφορες προκλήσεις που υπάρχουν, θα εξηγηθούν οι κατηγορίες της ΜΜ και θα αναφερθούν διάφορες τεχνικές λειτουργικής ταξινόμησης που είναι ευρέως γνωστές. Τέλος, θα προβληθεί το θεωρητικό υπόβαθρο για κάθε σύνολο δεδομένων που υπήρχε στην διάθεση μας ξεχωριστά.

Στο 2ο κεφάλαιο θα παρατεθούν τα υλικά που χρησιμοποιήθηκαν για την ανάλυση των δεδομένων, όπως ο τύπος του υπολογιστή, η γλώσσα προγραμματισμού αλλά και τα διάφορα πακέτα που έλαβαν χρήση. Έπειτα, θα αναφερθεί από που έγινε η συλλογή των δεδομένων και θα επεξηγηθούν. Στην πορεία, θα γίνει αναφορά για την μεθοδολογία έρευνας, δηλαδή η επεξεργασία των δεδομένων πριν και κατά την διάρκεια εφαρμογής ΜΜ και τέλος, θα αναδειχθούν οι αλγόριθμοι που χρησιμοποιήθηκαν σε κάθε σύνολο δεδομένων

Στο 3ο κεφάλαιο θα προβληθούν τα αποτελέσματα των αναλύσεων αυτών. Πιο συγκεκριμένα, θα αναδειχθούν τα αποτελέσματα της μεθοδολογίας με διάφορες μετρήσεις σε πίνακες και γραφικές παραστάσεις που χρησιμοποιούνται στην ΜΜ για την αξιολόγηση της αξιοπιστίας της. Τέλος, θα παρουσιαστεί ο καλύτερος διαχωριστικός συνδυασμός για τα χαρακτηριστικά κάθε συνόλου δεδομένων.

Στο 4ο κεφάλαιο θα σχολιαστούν τα αποτελέσματα των αναλύσεων και θα ελεγχθεί αν συμβαδίζουν με την υπάρχουσα βιβλιογραφία. Επιπρόσθετα, θα συζητηθεί σύμφωνα με τα συμπεράσματα που προκύπτουν από τα αποτελέσματα, κατά πόσο είναι αξιόπιστες οι μέθοδοι ΜΜ και θα αναδειχθούν περιορισμοί και προβλήματα που προέκυψαν κατά την

διαδικασία. Πιο συγκεκριμένα, εδώ θα αξιολογηθεί η ΜΜ για κάθε πεδίο που εφαρμόστηκε ξεχωριστά.

1. Θεωρητικό υπόβαθρο

1.1 Μηχανική Μάθηση και Βιοϊατρική έρευνα

Στις αρχές της δεκαετίας του 1970, ο Ben Hesper και ο Paulien Hogeweg ξεκίνησαν να χρησιμοποιούν τον όρο «βιοπληροφορική» για μία ερεύνα ορίζοντάς την ως «τη μελέτη των πληροφοριακών διεργασιών στα βιοτικά συστήματα». Η μοντελοποίηση και η ανάλυση προτύπων χρειάζεται να συνδυαστούν έτσι ώστε: πρώτον, να αναλυθούν τα πρότυπα διακύμανσης σε πολλαπλά επίπεδα στους οργανισμούς, δεύτερον, για την ανίχνευση φαινομένων σε μοντέλα, τρίτον, να συγκριθεί το αποτέλεσμα τέτοιων μοντέλων με «πραγματικά» δεδομένα, και τέλος, η σχέση μεταξύ του γονότυπου, του φαινοτύπου, της λειτουργικότητας και του περιβάλλοντος μπορεί να θεωρηθεί ως ένας τύπος αναγνώρισης ή μετασχηματισμού μοτίβου. Η κατανόηση αυτών των διαδικασιών ήταν ο πυρήνας της βιοπληροφορικής έρευνας. Ο πρώτος αλγόριθμος για σύγκριση αλληλουχιών πρωτεϊνών ή DNA δημοσιεύθηκε από τους Needleman και Wunsch το 1970. Έπειτα, στα τέλη της δεκαετίας του 1980 έως και σήμερα, ο όρος «βιοπληροφορική» αναφέρεται κυρίως σε υπολογιστικές μεθόδους για συγκριτική ανάλυση και ταξινόμηση, για την πρόβλεψη διαφόρων ασθενειών και γονιδιακών δεδομένων [1], [2]. Η Βιοπληροφορική είναι ένας διεπιστημονικός κλάδος στον οποίο γίνεται σύμπραξη με εκπαιδευμένους επιστήμονες σε διάφορες ειδικότητες (βιολόγοι, μαθηματικοί, επιστήμονες Η/Υ, μηχανικοί). Ουσιαστικά, περιλαμβάνει την αλληλεπίδραση της βιολογίας, της πληροφορικής και της στατιστικής. Η προσέγγιση της Μηχανικής Μάθησης (ΜΜ) είναι κατάλληλη για εφαρμογή στον κλάδο αυτό επειδή τα θέματα της έρευνας αφορούν περίπλοκα βιολογικά προβλήματα [3]. Η βελτιστοποίηση στην ακρίβεια και την αποτελεσματικότητα των τεχνικών ΜΜ έχουν σημειώσει αύξηση τα τελευταία χρόνια όπου την καθιστούν βασικό εργαλείο για την επίλυση προβλημάτων στην ιατρική και την βιολογία.

Ο όρος Μηχανική Μάθηση επινοήθηκε το 1959 από τον Arthur Samuel, έναν Αμερικανό πρωτοπόρο στον τομέα των ηλεκτρονικών παιχνιδιών και της τεχνητής νοημοσύνης. Αργότερα, ο Hutchinson (1995) προτείνει ότι η χρήση αλγορίθμων για την εκμάθηση υφισταμένων δεδομένων είναι η ουσία του όρου μηχανική μάθηση. Επιπροσθέτως, η ΜΜ έχει οριστεί από τον Mitchell (1997) και ως «πρόγραμμα υπολογιστή που μπορεί να μάθει από την εμπειρία βασιζόμενο σε παρελθοντικές αποστολές και αποδόσεις». Είναι μία αποτελεσματική και φθηνή προσέγγιση για την επίλυση προβλημάτων στην μοριακή βιολογία. Επίσης, είναι από τις πιο αγαπημένες τεχνικές στην Βιοπληροφορική, λόγω του ότι μπορεί να χειριστεί μεγάλα σύνολα δεδομένων και βασίζεται σε προσεγγίσεις που μπορούν να μάθουν αυτόματα από τα ήδη υπάρχοντα δεδομένα ούτως ώστε να παράγουν χρήσιμες πληροφορίες [4]. Ως επί το πλείστων οι περισσότεροι χρησιμοποιούν τον όρο "μηχανή" εννοώντας τον υπολογιστή. Ωστόσο, ο αναγνώστης καλό θα ήταν να γνωρίζει ότι η ορολογία "μηχανή" έχει μια πιο ευρύτερη έννοια. Σχεδόν κάθε άψυχος μηχανισμός, ο οποίος μπορεί να ολοκληρώσει μια εργασία, χαρακτηρίζεται ως μηχανή. Την δεκαετία του 1960, μόλις τέθηκαν σε χρήση οι ηλεκτρονικοί υπολογιστές (Η/Υ), αναπτύχθηκαν αλγόριθμοι που επέτρεπαν την μοντελοποίηση και την ανάλυση μεγάλων συνόλων δεδομένων. Γενικά, η ΜΜ είναι μια διαδικασία που κάνει τα συστήματα να βελτιώνονται με την εμπειρία. Η βασική ιδέα είναι

ο σχεδιασμός μίας μηχανής όπου μαθαίνει σαν τον άνθρωπο, δηλαδή να μαθαίνει από την εμπειρία και να ανακαλύπτει πληροφορίες από τα διαθέσιμα δεδομένα. Αυτή η προσέγγιση είναι κατάλληλη για εφαρμογή στη Βιοπληροφορική επειδή τα θέματα της έρευνας αφορούν περίπλοκα βιολογικά προβλήματα [5]. Επιπλέον η έρευνα μοριακής βιολογίας έχει περιορισμένες θεωρίες και τις περισσότερες φορές εξαρτάται από πειραματικά δεδομένα.

1.1.1 Πεδία εφαρμογής στην Βιοπληροφορική

Η εκθετική αύξηση του αριθμού των διαθέσιμων βιολογικών δεδομένων δημιουργεί ανάγκη για αποτελεσματική αποθήκευση, διαχείριση πληροφοριών και την εξαγωγή χρήσιμων πληροφοριών από τα δεδομένα. Υπάρχουν διάφορα βιολογικά πεδία που εφαρμόζονται τεχνικές ΜΜ και μπορούν να χωριστούν: 1) γονιδιωματική (genomics), 2) μεταγραφωμική (transcriptomics) 3) πρωτεομική (proteomics), 4) βιολογία συστημάτων (systems biology), 5) αναγνώριση χαρακτήρων κειμένου (text mining), 6) εξέλιξη (evolution) και 7) άλλα προβλήματα/εφαρμογές.

1) Η μηχανική μάθηση είναι ίσως η πιο χρήσιμη για την ερμηνεία μεγάλων συνόλων δεδομένων των γονιδίων και έχει χρησιμοποιηθεί για να σχολιάσει μια μεγάλη ποικιλία γονιδιωματικών αλληλουχιών. Για παράδειγμα, οι μέθοδοι ΜΜ μπορούν να χρησιμοποιηθούν για να «μάθουν» πώς να αναγνωρίζουν τα σημεία έναρξης μεταγραφής (transcription start sites- TSS) σε μια ακολουθία γονιδιώματος. Μοντέλα τα οποία αναγνωρίζουν έναν μεμονωμένο τύπο γονιδιωματικού στοιχείου, μπορούν να συνδυαστούν για τη δημιουργία συστημάτων ΜΜ, που είναι ικανά να επεξηγήσουν και να κατανοήσουν από τα γονίδια την μη μεταφρασμένη περιοχή μορίου (untranslated region - UTR), ιντρόνια και τα εξόνια σε ολόκληρα τα χρωμοσώματα σε ευκαριωτικά κύτταρα ή οργανισμούς. Εκτός από την εκμάθηση αναγνώρισης σε ακολουθίες DNA, οι αλγόριθμοι μηχανικής μάθησης μπορούν να χρησιμοποιήσουν δεδομένα εισόδου προερχόμενα από γονιδιακή έκφραση π.χ. αλληλουχίες RNA (RNA-seq). Τα δεδομένα έκφρασης γονιδίων μπορούν να χρησιμοποιηθούν για την διάκριση μεταξύ διάφορων φαινοτύπων ασθενειών όπως και στην διαδικασία εντοπισμού πολύτιμων βιοδεικτών μίας νόσου. Οι εφαρμογές ΜΜ έχουν επίσης χρησιμοποιηθεί για την αποτελεσματική γονιδιακή ανάλυση (π.χ. οντολογία γονιδίων (Gene Ontology - GO)). Συνοπτικά, αυτές οι μέθοδοι, είναι αρκετά αποτελεσματικές στην ανάλυση μεγάλων και πολύπλοκων συνόλων δεδομένων, έτσι είναι πιθανόν να γίνονται ολοένα και πιο σημαντικές στην γονιδιωματική καθώς μεγάλα σύνολα δεδομένων διατίθενται μέσω διεθνών συνεργασιών. Από την άλλη όμως, στην πράξη η επίτευξη καλής απόδοσης αυτής της μεθόδου απαιτεί συνήθως θεωρητική και πρακτική γνώση τόσο της μεθοδολογίας όσο και της ερευνητικής περιοχής. Νέες μέθοδοι ΜΜ αλλά και ειδικοί που είναι ικανοί να ανταπεξέλθουν σε μεγάλα σύνολα δεδομένων, είναι απαραίτητοι στο πεδίο αυτό [6].

2) Μεταξύ των διάφορων τεχνολογιών που αφορούν τη μεταγραφωμική, οι δύο που ξεχωρίζουν είναι οι μικροσυστοιχίες και οι αλληλουχίες RNA (RNA–seq) [7]. Η τεχνολογία των μικροσυστοιχιών ανακαλύφθηκε το 1995, και έχει αναπτυχθεί την τελευταία εικοσαετία. Η ανάπτυξη αυτής της τεχνολογίας έδωσε νέες, ενδιαφέρουσες πληροφορίες και αύξησε εκθετικά τα διαθέσιμα δεδομένα για την κατανόηση των

βιολογικών συστημάτων. Ένα άλλο σημαντικό είναι ότι δίνει τη δυνατότητα επαρκούς αποκρυπτογράφησης της γονιδιακής έκφρασης αλλά δίνει και απαντήσεις σε ερωτήματα γονιδιακού προφίλ με ταυτόχρονη ανάλυση πολλών γονίδιων, ακόμα και ολοκλήρου του γονιδιώματος [8]. Σύμφωνα με τον Μαλατρά (2010) [8] "Μπορεί κάνεις να εξάγει χρήσιμες πληροφορίες για τη βιολογική λειτουργία ενός οργανισμού, βρίσκοντας ποια γονίδια επάγονται ή καταστέλλονται σε κάποια φάση του κυτταρικού κύκλου, σε κάποια αναπτυξιακή στιγμή ή σε απόκριση σε ερεθίσματα του περιβάλλοντος, όπως για παράδειγμα η απόκριση σε ορμόνες ή σε υψηλή θερμοκρασία". Επιπρόσθετα, χρησιμοποιείται σε συγκριτικές μελέτες γονιδιωμάτων αλλά και για την ομαδοποίηση γονιδιακών εκφράσεων [5]. Το RNA-Seq έχει σημαντικά πλεονεκτήματα για την μελέτη της δομής του μεταγραφώματος, όπως η ανίχνευση νέων μεταγραφών, αλληλόμορφικές συχνότητες και η θέση ματίσματος. Επίσης, το RNA-Seq δεν εξαρτάται από τον σχολιασμό του γονιδιώματος για επιλογή ανιχνευτών και αποφεύγει τις σχετικές στρεβλώσεις που παρουσιάζονται κατά τον υβριδισμό των μικροσυστοιχιών [9]. Όπως αναφέρεται στην έρευνα του Almas Jabeen et al. σε περιπτώσεις ταξινόμησης μεγάλου όγκου δεδομένων RNA – Seq αρκετά ικανοί φαίνονται να είναι οι μέθοδοι βαθιάς μάθησης [10]. Τέλος, αλγόριθμοι εκπαίδευσης όπως ο κ- πλησιέστερων γειτόνων (K- Nearest Neighbors - K-NN) και ο εξελικτικός αλγόριθμος (GA) έχουν εφαρμοστεί για την ανάλυση δεδομένων γονιδιακής έκφρασης [5].

3)Ο κύριος σκοπός της πρωτεομικής επιστήμης είναι ο εντοπισμός και ο χαρακτηρισμός της πρωτεομικής έκφρασης σε βιολογικά συστήματα. Η μηχανική μάθηση με γοργούς ρυθμούς γίνεται ένα πολύ δημοφιλές εργαλείο στον τομέα της πρωτεομικής. Ο Φασματογράφος Μάζας (Mass Spectrometry - MS) είναι μια βασική συσκευή σε αυτό το βιολογικό πεδίο [11], [12]. Οι στόχοι τέτοιων ερευνών είναι να εντοπίσουν βιοδείκτες και να βοηθήσουν στη διάγνωση, την πρόγνωση και τη θεραπεία συγκεκριμένων ασθενειών (π.χ. Καρκίνος). Έπειτα, να κατανοηθεί πως οι πρωτεΐνες αλληλοεπηρεάζονται μεταξύ τους και τον ρόλο που έχουν στην εξέλιξη μιας νόσου. Η Φασματομετρία Μάζας μπορεί ακόμα να εφαρμοστεί για τη δημιουργία δικτύων που υποδεικνύουν αλληλεπιδράσεις μεταξύ πρωτεϊνών [11]. Επίσης, αρκετά απλοποιημένα μοντέλα όπως η αναζήτηση tabu, οι μέθοδοι Monte Carlo και οι εξελικτικοί αλγόριθμοι (Genetic Algorithms - GAs) έχουν χρησιμοποιηθεί για την αναδίπλωση πρωτεϊνών. Η πρόβλεψη της πλευρικής αλυσίδας πρωτεϊνών, είναι ένα σημαντικό πρόβλημα για την πρόβλεψη και τον σχεδιασμό δομής των πρωτεϊνών. Αυτό μπορεί να επιτευχθεί χρησιμοποιώντας αλγόριθμους όπως GAs. Τέλος, αλγόριθμοι απόπτυσης, βασιζόμενοι σε συμπεράσματα από γραφικά μοντέλα και η μέθοδος self-consistent mean field (Self-Consistent Mean Field - SCMF) επίσης χρησιμοποιούνται για την επίλυση του συγκεκριμένου προβλήματος [5].

4) Η βιολογία συστημάτων είναι ένα διεπιστημονικό πεδίο μελέτης στο οποίο τα βιολογικά συστήματα διερευνώνται με ολιστικές ποσοτικές προσεγγίσεις σε μια προσπάθεια να ανακαλυφθεί πώς οι συλλογικές συμπεριφορές αυτών των συστημάτων προκύπτουν από τις πολύπλοκες αλληλεπιδράσεις μεταξύ των συστατικών τους. Μεταξύ των ποσοτικών προσεγγίσεων που χρησιμοποιούνται στη βιολογία των μοριακών συστημάτων είναι και η μηχανική μάθηση. Οι τεχνικές ΜΜ είναι ικανές να εξάγουν μοτίβα κρυμμένα σε ένα σύνολο δεδομένων και να χρησιμοποιήσουν αυτά τα μοτίβα για να κάνουν ακριβείς προβλέψεις για μελλοντικά δεδομένα. Έτσι, λαμβάνοντας υπόψη ένα

βιολογικό σύστημα, η εφαρμογή τεχνικών ΜΜ μπορεί να αποκαλύψει πώς οι σχέσεις μεταξύ των βιολογικών συστατικών μπορούν να δημιουργήσουν μια συλλογική συμπεριφορά ενδιαφέροντος του βιολογικού συστήματος που μελετάται [13]. Στα συστήματα βιολογίας οι πιο κοινοί βελτιστοποιημένοι αλγόριθμοι που χρησιμοποιούνται είναι ο Markov chain, όπως και ο εξελικτικός αλγόριθμος (GA) (μοντέλα γενετικών δικτύων, επιλογή ρυθμιστικών δομών και εκτίμηση της παραμέτρου των βιολογικών διεργασιών) [5].

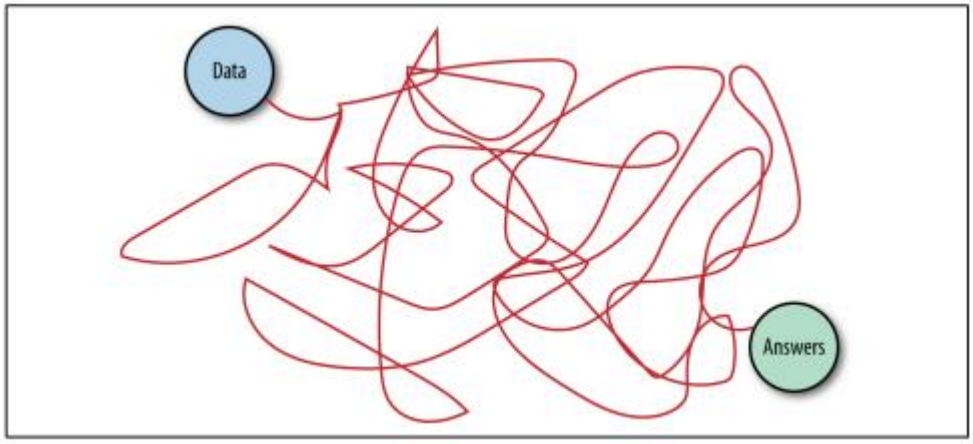
5) Η ταχεία εξέλιξη των μέσων επικοινωνίας και ανταλλαγής πληροφοριών οδήγησε στην δημιουργία μαζικών συνόλων δεδομένων, τα οποία μπορούν να χρησιμοποιηθούν προκειμένου να αξιοποιηθούν πληροφορίες διάφορων κλάδων. Η ανάγκη για μια πιο εύκολη σύλλεξη και ανάλυση δεδομένων οδήγησε τους ερευνητές στη δημιουργία ενός νέου πεδίου ανάλυσης, το οποίο ονομάζεται Εξόρυξη Γνώσης από Κείμενο (Text Mining) [10]. Υπάρχουν πολλές προσεγγίσεις για την επίτευξη εξόρυξης γνώσης από κείμενο σε Βιοϊατρικές έρευνες. Ορισμένες από αυτές ονομάζονται αναγνώριση ονομάτων οντοτήτων (Named Entity Recognition - NER), ταξινόμηση και ομαδοποίηση εγγράφων, ανακάλυψη σχέσεων, διασύνδεση εννοιών , εξαγωγή πληροφοριών και απεικόνιση πληροφοριών. Κύριος στόχος όλων των τεχνικών είναι η κατανόηση μεγάλου μεγέθους δεδομένων στην Βιοϊατρική έρευνα μέσα σε σύντομο χρονικό διάστημα. Αυτά τα σύνολα δεδομένων αποτελούνται συνήθως από διανύσματα λέξεων ή ενσωματωμένων λέξεων. Στην Βιοϊατρική έρευνα χρησιμοποιείται κυρίως για την ανακάλυψη γνώσεων σε κείμενα (Knowledge Discovery in Text - KDT), για να εξερευνηθούν και να συνοψισθούν κρυφές πληροφορίες. Οι μέθοδοι ΜΜ μελετούν τα σύνολα δεδομένων για την εξαγωγή χαρακτηριστικών από το κείμενο, ακολουθώντας την προσέγγιση του NER στον τομέα των ειδήσεων, και γίνεται μια παράλληλη στον βιοϊατρικό τομέα π.χ. σε γονίδια, πρωτεΐνες και ιούς [14].

6) Η μελέτη της εξέλιξης και ειδικά η αναδημιουργία φυλογενετικών δέντρων, εκμεταλλεύονται ωφελούνται επίσης από τις τεχνικές ΜΜ. Τα φυλογενετικά δέντρα είναι σχηματικές αναπαραστάσεις της εξέλιξης των οργανισμών. Οι αλγόριθμοι παραδοσιακά κατασκευάστηκαν σύμφωνα με διαφορετικά χαρακτηριστικά (μορφολογικά χαρακτηριστικά, μεταβολικά χαρακτηριστικά κ.λπ.) αλλά σήμερα με τον μεγάλο αριθμό διαθέσιμων αλληλουχιών γονιδιώματος που υπάρχει, βασίζονται στη σύγκριση μεταξύ διαφορετικών γονιδιωμάτων πραγματοποιώντας την με ευθυγράμμιση πολλαπλών αλληλουχιών [5]. Επιπλέον, οι αλγόριθμοι ΜΜ επιτρέπουν την αυτοματοποιημένη διαπίστωση μεταβολικών φαινοτύπων για χιλιάδες μικροβιακά γονιδιώματα. Τα δεδομένα αυτά παρέχουν άνευ προηγουμένου ευκαιρίες για μεγάλης κλίμακας εξελικτική ανάλυση, όπως ανοικοδόμηση προηγούμενης μεταβολικής ποικιλομορφίας και φαινοτυπικές προβλέψεις για κακώς χαρακτηρισμένα υπάρχοντα τμήματα [15].

7) Η κατηγορία που ονομάζεται «άλλες εφαρμογές ή υπόλοιπα προβλήματα» ομαδοποιεί τα υπόλοιπα προβλήματα [5] όπως την ανάλυση δεδομένων από φασματογράφο μάζας, ανάλυση βιοσήματος και ανάλυση βιοϊατρικών εικόνων.

1.1.2 Προκλήσεις στην Βιοπληροφορική

Η ταχεία ανάπτυξη βιολογικών δεδομένων θέτει ζητήματα για ανάπτυξη υπολογιστικών εργαλείων και τεχνικών που θα μπορούν να μετατρέψουν τα δεδομένα σε χρήσιμες βιολογικές γνώσεις. Αυτό που ονομάζουμε δεδομένα είναι παρατηρήσεις πραγματικών φαινομένων. Για παράδειγμα, τα δεδομένα της χρηματιστηριακής αγοράς ενδέχεται να περιλαμβάνουν παρατηρήσεις των ημερήσιων τιμών των μετοχών, ανακοινώσεις κερδών από μεμονωμένες εταιρείες, ακόμη και άρθρα γνωμοδότησης από ειδικούς. Κάθε κομμάτι δεδομένων παρέχει ένα μικρό παράθυρο σε μια περιορισμένη πτυχή της πραγματικότητας. Η συλλογή όλων αυτών των παρατηρήσεων μας δίνει μια εικόνα του συνόλου αλλά η εικόνα είναι ακατάστατη επειδή αποτελείται από χίλια μικρά κομμάτια και υπάρχει πάντα θόρυβος μέτρησης και ελλιπή κομμάτια. Η διαδρομή από τα δεδομένα προς τις απαντήσεις είναι γεμάτη από λανθασμένες κινήσεις και αδιέξοδα. Αυτό που ξεκινά ως μια πολύ υποσχόμενη προσέγγιση μπορεί τελικά να διαψευσθεί, ενώ αυτό που έδειχνε λάθος να καταλήξει να οδηγεί στην καλύτερη λύση [16].



Εικόνα 1.1: Η περίπλοκη διαδρομή από τα δεδομένα προς τις απαντήσεις. [16]

Οι απαντήσεις που παίρνουμε από τα δεδομένα της Βιοπληροφορικής συνήθως αφορούν αν κάποια από τα χαρακτηριστικά (π.χ. γονίδια, πρωτεΐνες ή και μεταβολίτες) είναι καλοί βιοδείκτες για κάποια ασθένεια ή όχι. Οι βιοδείκτες είναι χαρακτηριστικά που υποδηλώνουν μια φυσιολογική ή παθογόνο βιολογική διαδικασία, ή μια φαρμακολογική απόκριση σε μια θεραπευτική παρέμβαση ή εμβολιασμό. Έτσι μπορούν να παρέχουν την κατάσταση της νόσου, τον κίνδυνο εξέλιξης, την πιθανότητα απόκρισης στην θεραπεία ή την τοξικότητα του φαρμάκου [17].

1.1.3 Κατηγορίες Μηχανικής Μάθησης

Η ΜΜ δεν είναι απλά η επιλογή και η εφαρμογή αλγορίθμων. Χρειάζεται αρχικά να κατανοηθεί το πρόβλημα στα διαθέσιμα δεδομένα, να απορριφθούν οι τεχνικές που είναι ακατάλληλες και να εντοπιστούν οι στόχοι για την λύση του προβλήματος. Συνήθως οι αλγόριθμοι ΜΜ χωρίζονται σε τέσσερις κύριες κατηγορίες: 1) Εποπτευόμενη (Supervised) , 2) Μη – εποπτευόμενη (Unsupervised), 3) Ημί – εποπτευόμενη (Semi – supervised) και 4) Ενισχυτική (Reinforcement) [18].

1) Η Εποπτευόμενη Μάθηση χρησιμοποιείται όταν έχουμε ήδη γνωστά ταξινομημένα δεδομένα σε κλάσεις (με ετικέτες) και θέλουμε να εκπαιδευτούν οι αλγόριθμοι με αυτά τα δεδομένα για να κατασκευαστούν μοντέλα, που θα έχουν την δυνατότητα με ακριβή πρόβλεψη, να διαχωρίσουν καινούργια δεδομένα στις σωστές κλάσεις.

2) Η Μη – Εποπτευόμενη Μάθηση ή ομαδοποίηση (clustering) έχει σκοπό με συγκεκριμένους αλγόριθμους να αποκαλύψει κρυφές πληροφορίες από τα δεδομένα και να τα χωρίσει στον βέλτιστο αριθμό ομάδων. Ουσιαστικά είναι δεδομένα χωρίς κλάσεις (χωρίς ετικέτες).

3) Η Ημί – Εποπτευόμενη Μάθηση διαχειρίζεται δεδομένα που συνδυάζουν τόσο ταξινομημένα δεδομένα όσο και μη ταξινομημένα δεδομένα (με ή χωρίς ετικέτες). Συνήθως χρησιμοποιείται μια μικρή ποσότητα δεδομένων με ετικέτα και μια μεγάλη ποσότητα χωρίς ετικέτα και με αυτό το τρόπο βελτιώνεται η ακρίβεια δραματικά

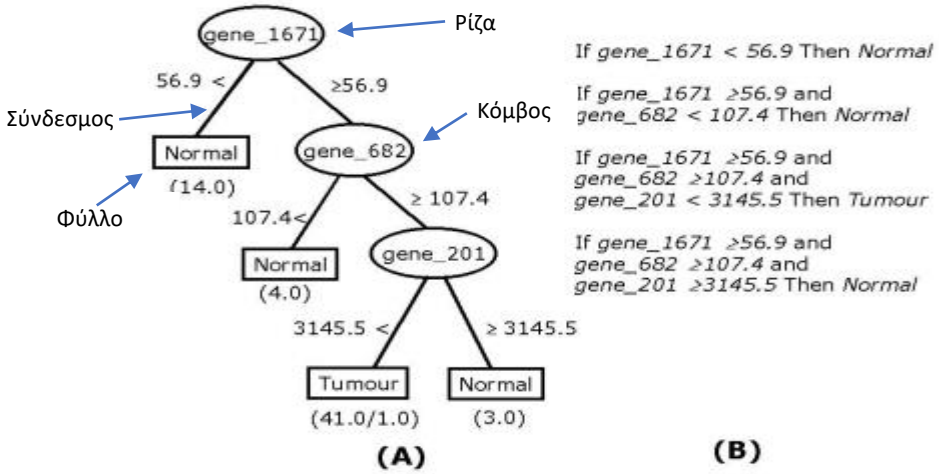
4) Στην Ενισχυτική Μάθηση οι φάσεις εκπαίδευσης και ελέγχου εναλλάσσονται. Οι αλγόριθμοι βασίζονται σε ένα σύστημα δοκιμών και σφαλμάτων όπου ο αλγόριθμος επιβραβεύεται όταν μαθαίνει τα σωστά μοτίβα. Έτσι με αυτό το τρόπο το μοντέλο καθορίζει αυτόματα την μαθησιακή του συμπεριφορά προκειμένου να μεγιστοποιήσει την απόδοση του με ένα σύστημα ανταμοιβής που είναι γνωστό ως σήμα ενίσχυσης.

1.1.4 Τεχνικές λειτουργικής ταξινόμησης

Σε ένα πρόβλημα ταξινόμησης, έχουμε ένα σύνολο δεδομένων που χωρίζονται σε κλάσεις. Πιο αναλυτικά, υπάρχουν δεδομένα καταχωρημένα σε κλάσεις και σύμφωνα με ορισμένες λειτουργίες της ΜΜ και ένα σύνολο κανόνων ταξινόμησης οι αλγόριθμοι προσπαθούν να προσαρμοστούν στα δεδομένα. Σε πολλές καταστάσεις αυτό το σύνολο κανόνων δεν είναι γνωστό και οι μόνες διαθέσιμες πληροφορίες είναι η ετικέτα που έχει η κάθε κλάση και πόσα δείγματα εμπεριέχει. Γενικά οι αλγόριθμοι ταξινόμησης, είναι αυτοί που μαθαίνουν από τα διαθέσιμα δεδομένα με τις μεθόδους ΜΜ και με διάφορα μέτρα αξιολογούνται κατά πόσο είναι ικανοί να ταξινομήσουν άγνωστα δεδομένα στις σωστές κλάσεις. Σε αυτό το υποκεφάλαιο θα παρουσιαστούν πέντε γνωστές τεχνικές ταξινόμησης: 1) τα δέντρα αποφάσεων, 2) οι μηχανές διανυσμάτων υποστήριξης, 3) οι Κ – πλησιέστεροι γείτονες, 4) τα μπεϋζιανά δίκτυα και 5) τα τεχνητά νευρωνικά δίκτυα όπου χρησιμοποιούνται σε πολλούς τομείς αλλά και στην Βιοπληροφορική.

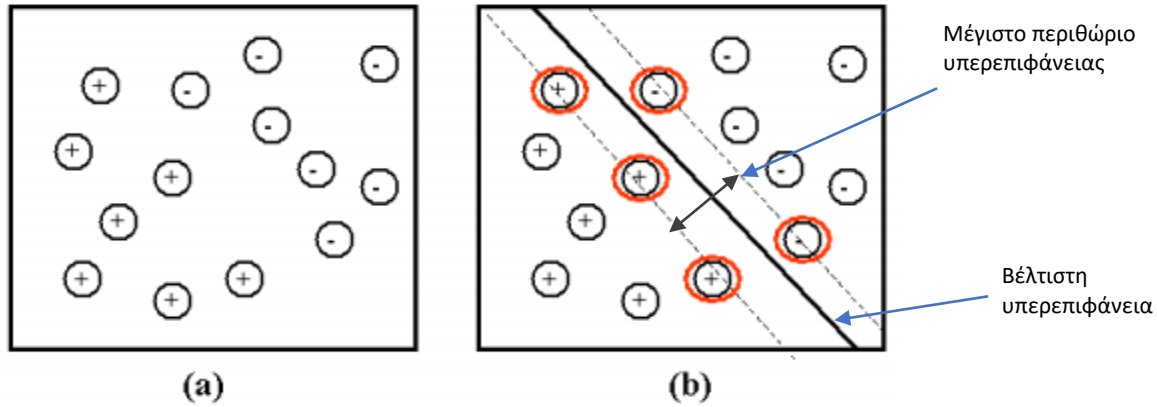
1)Μια από τις πιο διαδιδόμενες τεχνικές ταξινόμησης στην Βιοπληροφορική είναι τα δέντρα αποφάσεων (Decision trees). Είναι επίσης γνωστά ως δέντρα ταξινόμησης (Classification trees) ή δέντρα παλινδρόμησης (Regression trees) και χρησιμοποιούν προσεγγίσεις ιεραρχικών αποφάσεων για την εκτίμηση και την ταξινόμηση των δειγμάτων [3]. Υπάρχει μια ποικιλία δέντρων αποφάσεων, αλλά τα απλά δέντρα αποφάσεων C4.5 είναι τα ευκολότερα απ’ όλα, όπου η ταξινόμηση μπορεί να γίνει κατανοητή σχεδόν με ευκολία. Η έξοδος του αλγορίθμου είναι ένα δέντρο αποφάσεων το οποίο μπορεί εύκολα να αναπαρασταθεί ως ένα σύνολο συμβολικών κανόνων (IF&THEN). Αυτοί οι συμβολικοί κανόνες μπορούν να ερμηνευθούν άμεσα και να παρέχουν χρήσιμες πληροφορίες για τους ερευνητές. Πιο ουσιαστικά είναι ένα δέντρο όπου η κορυφή ονομάζεται ρίζα (root) και λαμβάνει θέση τυχαία ένα χαρακτηριστικό, κάθε κόμβος (node) είναι μια δοκιμή για τις

τιμές ενός χαρακτηριστικού και τα φύλλα (leaves) αντιπροσωπεύουν τις κλάσεις που είναι διαθέσιμες για ταξινόμηση [3].



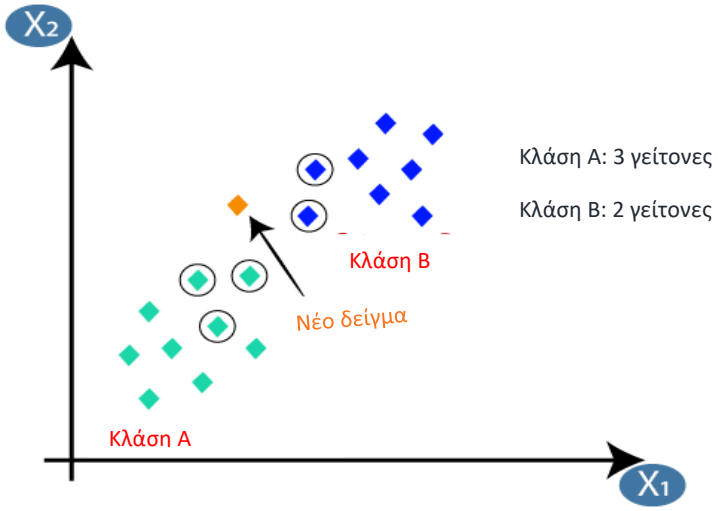
Εικόνα 1.2: **(A)** Δέντρο αποφάσεων που δημιουργήθηκε από σύνολο δεδομένων όγκου του παχέος εντέρου. Οι κόμβοι αντιπροσωπεύουν τα γονίδια και τα κλαδιά είναι οι συνθήκες έκφρασης. Τα φύλλα του δέντρου αντιπροσωπεύουν τα αποτελέσματα της απόφασης (Στην περίπτωση αυτή << καρκινικός ιστός >> ή << φυσιολογικός ιστός>>). Οι αριθμοί από κάτω δηλώνουν τον αριθμό των αποφάσεων που είναι σωστές (TP/FP). **(B)** Οι κανόνες αυτοί ισοδυναμούν με τους κανόνες του δέντρου αποφάσεως [3].

2) Μια άλλη τεχνική ταξινόμησης είναι η μηχανή διανυσμάτων υποστήριξης (Support Vector Machine - SVM). Είναι αρκετά δημοφιλής στο τομέα της Βιοϊατρικής έρευνας και έχει λάβει χρήση σε πολλές εφαρμογές όπως αναγνώριση πτυχών πρωτεΐνης και ταξινόμηση δεδομένων από μικροσυστοιχίες. Η βασική ιδέα τους είναι η κατασκευή μιας υπερεπιφάνειας (hypersurface), το οποίο θα αποτελέσει την επιφάνεια απόφασης, με τέτοιο τρόπο ώστε να μεγιστοποιείται η απόσταση που διαχωρίζει τα πλησιέστερα δείγματα που ανήκουν σε διαφορετικές κλάσεις και έχουν ως στόχο την σχεδίαση ενός υπολογιστικά αποδοτικού τρόπου για την εκμάθηση κατάλληλων διαχωριστικών επιπέδων. Η επιλογή της υπερεπειφάνειας γίνεται με τέτοιο τρόπο ώστε να απέχει όσο το δυνατόν περισσότερο από τα κοντινότερα θετικά και αρνητικά δείγματα (υπερεπιφάνεια μεγίστου περιθώριου). Αν και οι SVM έχουν καλή απόδοση, έχουν μειονέκτημα το ότι είναι μια χρονοβόρα δοκιμή και έχουν έλλειψη εκφραστικής ισχύος. [4]



Εικόνα 1.3: (A) Μη ταξινομημένα δείγματα σε ένα σύνολο δεδομένων (η κλάση «+» και η κλάση «-») (B) Ταξινόμηση με SVM όπου τα κυκλικά κόκκινα σημεία αντιπροσωπεύουν τα δείγματα που είναι μέσα στα όρια ταξινόμησης.[4]

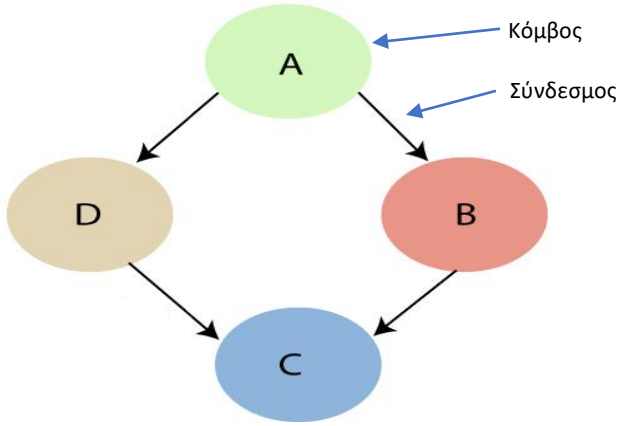
3) Οι αλγόριθμοι Κ – πλησιέστερου γείτονα (K-NN) είναι μια άλλη τεχνική όπου φαίνεται να είναι από τις πιο αξιόπιστες τεχνικές και χρησιμοποιείται ευρέως στην Βιοπληροφορική. Μπορεί να διαχειριστεί οποιοδήποτε σύνολο δεδομένων, όμως η αποτελεσματικότητα ή ο χρόνος εκτέλεσης των αλγορίθμων αυτών είναι άμεσα ανάλογη με το μέγεθος του συνόλου δεδομένων. Επιτυγχάνει ικανοποιητική απόδοση όταν το μέγεθος των δεδομένων εκπαίδευσης είναι μεγάλο και μπορεί να μειωθεί δραματικά σε περιπτώσεις όπου το σύνολο δεδομένων είναι μικρό [14]. Η διαδικασία ταξινόμησης του κάθε δείγματος σε αυτή την περίπτωση γίνεται με βάσει των Κ πλησιέστερων δειγμάτων. Ουσιαστικά μπορεί να χρησιμοποιηθεί π.χ. η ευκλείδεια απόσταση, για την μέτρηση της ομοιότητας των δειγμάτων με βάση όλα τα χαρακτηριστικά και στην συνέχεια για ένα νέο δείγμα που χρειάζεται να ταξινομηθεί βρίσκει τα πλησιέστερα Κ δείγματα και εκχωρείται στην κλάση η οποία εμπεριέχει την πλειοψηφία των δειγμάτων [13].



Εικόνα 1.4: Ταξινόμηση με Κ = 5 πλησιέστερους γείτονες. Οι 3 πλησιέστεροι γείτονες ανήκουν στην Κλάση Α οπότε το νέο δείγμα ταξινομείτε σε αυτήν [19].

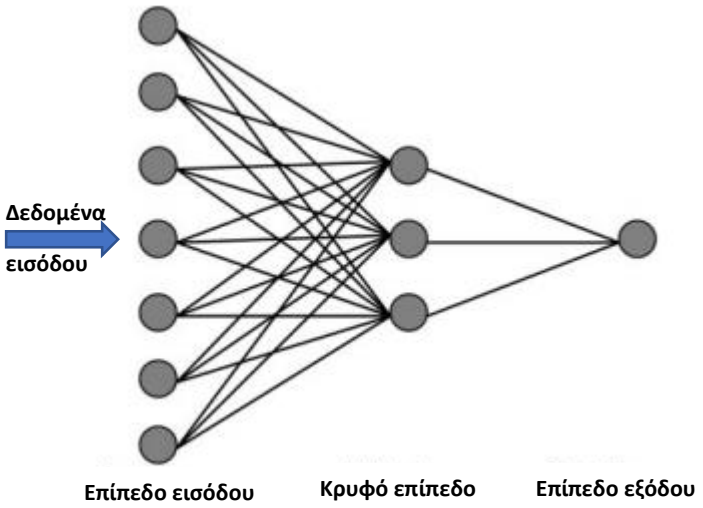
4) Τα μπεϋζιανά δίκτυα (Bayesian Networks - BNs) είναι γραφικά μοντέλα πιθανών σχέσεων ενός συνόλου μεταβλητών ενδιαφέροντος και έχει δοκιμαστεί σε πολλά πεδία όπως σε αλληλουχίες DNAκαι πρόβλεψη πρωτείνης δευτερογενούς δομής. Παρέχουν μια

ευέλικτη προσέγγιση κατά την αντιμετώπιση προβλημάτων ανάλυσης δεδομένων λόγω του ότι έχουν την ικανότητα να χειρίζονται ελλιπή σύνολα δεδομένων. Το δίκτυο αποτελείται από μια δομή που κωδικοποιεί ένα σύνολο ανεξάρτητων ισχυρισμών υπό όρους σχετικά με τις μεταβλητές ενδιαφέροντος και ένα σύνολο τοπικής κατανομής πιθανότητας που σχετίζεται με κάθε μεταβλητή. Η ενοποίηση αυτών την δύο συνόλων παράγει μια κοινή κατανομή για τις μεταβλητές ενδιαφέροντος και με βάσει τις πιθανότητες που συλλέγονται από τις σχέσεις τους το γραφικό μοντέλο μπορεί να πάρει την καλύτερη απόφαση. Αναπαρίσταται με κόμβους που αντιπροσωπεύουν τυχαίες μεταβλητές και συνδέσμους που δείχνουν πως ένας κόμβος επηρεάζει άμεσα τον άλλον. Είναι σημαντικό να επισημανθεί πως η μπεϋζιανή πιθανότητα είναι ο βαθμός εμπιστοσύνης για ένα συμβάν που συμβαίνει, ενώ η κλασική πιθανότητα βασίζεται στην μέτρηση της συχνότητας της εμφάνισης αυτού του συμβάντος σε όλη τη διάρκεια της εμπειρίας [4].



Εικόνα 1.5: Μπεϋζιανό δίκτυο (BN). A,B,C και D αντιπροσωπεύουν τυχαίες μεταβλητές [20].

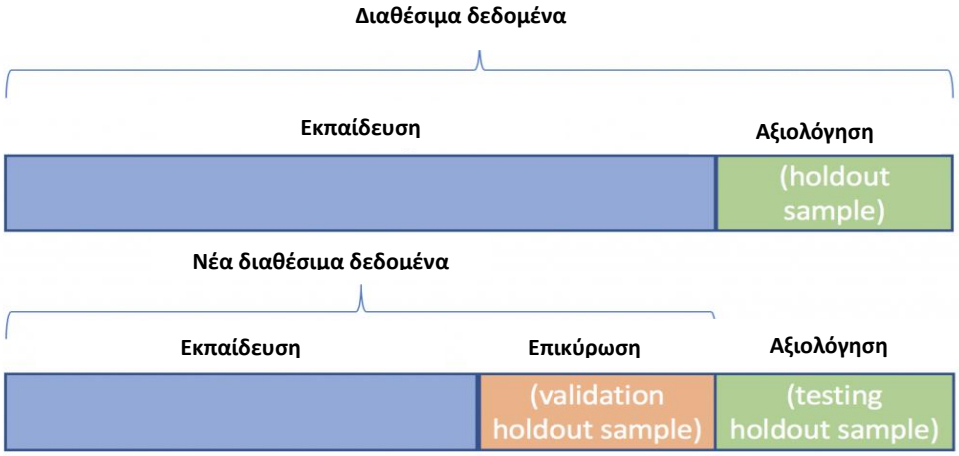
5) Τέλος, Τα τεχνητά νευρωνικά δίκτυα (Artificial Neural Networks - ANNs) χρησιμοποιούνται αρκετά στην Βιοπληροφορική λόγω της ικανότητας τους να αντιμετωπίζουν θορυβώδη, μη γραμμικά και πολύ μεγάλων διαστάσεων δεδομένα. Εμπνέονται από τον τρόπο με τον οποίο ο ανθρώπινος εγκέφαλος μαθαίνουν και επεξεργάζονται πληροφορίες και η ιδέα τους προσομοιώνει την συμπεριφορά ενός βιολογικού νευρωνικού δικτύου. Αποτελούνται από τοεπίπεδο εισόδου (Input layer) όπου λαμβάνει τα αρχικά δεδομένα και στην συνέχεια μέσω συνάψεων και αναλόγως με τις τιμές των σχετικών βαρών (weights) πολλαπλασιάζονται, άρα τροποποιούνται, και μεταβιβάζονται στο πρώτο κρυφό επίπεδο (hidden layer). Αυτές οι τιμές αθροίζονται και τροφοδοτούνται μέσω μιας μη γραμμικής συνάρτησης μεταφοράς (π.χ. σιγμοειδής, υπερβολική εφαπτομένη) η οποία κλιμακώνεται και παράγει έξοδο. Τελικά, αυτή η τροποποιημένη πληροφορία φτάνει στους κόμβους του επίπεδου εξόδου (Output layer), όπου και είναι το αποτέλεσμα του δικτύου. Η επιλογή των κόμβων και των κρυφών επιπέδων είναι ανάλογη με τα δεδομένα και είναι ένα ζήτημα διερεύνησης. Επίσης, συχνά αποκαλούνται «μαύρο – κουτί» λόγω των διαδικασιών που είναι δύσκολο να ερμηνευθούν και να επικυρωθούν [21].



Εικόνα 1.6: Αρχιτεκτονική τεχνητού νευρωνικού δικτύου perceptron [21].

1.1.5 Μέθοδοι επικύρωσης – αξιολόγησης των μοντέλων

Ο απώτερος σκοπός των μοντέλων ΜΜ είναι να εκπαιδευτούν από τα παραδείγματα και να αξιολογηθούν κατά πόσο είναι κατάλληλα να διαχειριστούν το πρόβλημα που υπάρχει. Είναι σημαντικό να γνωρίζει ο χρήστης αν μπορεί να εμπιστευτεί τις προβλέψεις του και αυτό επιτυγχάνεται γενικεύοντας τα μοντέλα σε άγνωστα δεδομένα. Πρέπει δηλαδή τα δεδομένα εκπαίδευσης να διαχωρίζονται σε δεδομένα εκπαίδευσης μοντέλου και δεδομένα αξιολόγησης μοντέλου (Σύνολο επικύρωσης). Επίσης αυτή η διαδικασία μπορεί να γίνει επαναληπτικά για να παρθούν πιο αξιόπιστες προβλέψεις. Οι πιο γνωστές διαδικασίες είναι: Ο διαχωρισμός δεδομένων (datasplit), η επαναδειγματοληψία bootstrap (bootstrap resampling), η Κ φορές αναδιπλωμένη διασταυρωμένη επικύρωση (K-fold cross validation), η Επαναλαμβανόμενη Κ φορές αναδιπλωμένη διασταυρωμένη επικύρωση (Repeated k-fold cross validation) και η επικύρωση με αποκλεισμό ενός δείγματος (leave-one-out validation) [22].



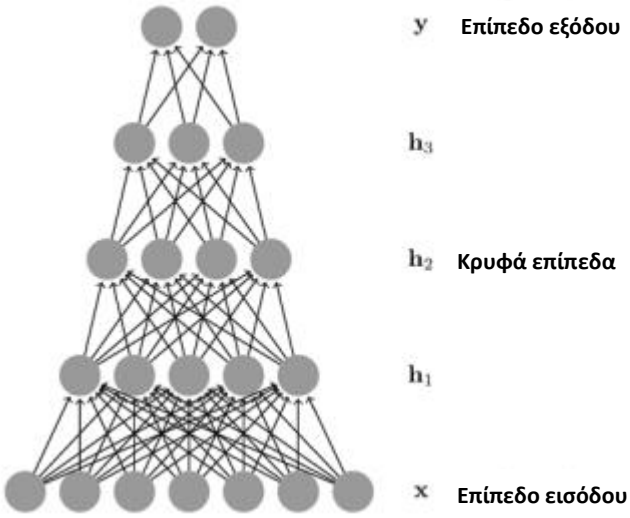
Εικόνα 1.7: Ολική διαδικασία αξιολόγησης μοντέλου [23].

- Ο διαχωρισμός δεδομένων αφορά τον διαχωρισμό για την προετοιμασία του μοντέλου, δηλαδή η εκπαίδευση του μοντέλου από ένα μεγάλο μέρος των δεδομένων και η αξιολόγηση του στα υπόλοιπα δεδομένα π.χ. διαχωρισμός 70%-30% αντίστοιχα. Είναι αρκετά χρήσιμη μέθοδος για μεγάλα σύνολα δεδομένων.
- Στην περίπτωση της επαναδειγματοληψίας bootstrap λαμβάνονται τυχαία δείγματα από το σύνολο δεδομένων (μπορούν να ληφθούν και επαναλαμβανόμενα). Τα αποτελέσματα παρέχουν μια ένδειξη της διακύμανσης της απόδοσης του μοντέλου και οι αρκετές επαναλήψεις είναι χρήσιμες για αξιόπιστα αποτελέσματα. Επίσης, είναι αρκετά αποδοτική μέθοδος σε μικρού μεγέθους συνόλων δεδομένων.
- Στην Κ φορές διασταυρωμένη αναδιπλωμένη επικύρωση χωρίζονται τα δείγματα σε ένα αριθμό Κ- υποσυνόλων και κάθε φορά ένα υποσύνολο χρησιμοποιείται για αξιολόγηση και τα υπόλοιπα για εκπαίδευση. Αυτή η διαδικασία επαναλαμβάνεται μέχρι να χρησιμοποιηθεί το κάθε υποσύνολο για αξιολόγηση και να εκτιμηθεί η συνολική ακρίβεια. Οι πιο δημοφιλές τιμές για το Κ είναι 5 και 10.
- Η διαδικασία Επαναλαμβανόμενης Κ φορές αναδιπλωμένης διασταυρωμένης επικύρωσης είναι ίδια με την Κ φορές αναδιπλωμένης διασταυρωμένης επικύρωσης, με την διαφορά ότι μπορεί να επαναληφθεί όλη η διαδικασία των Κ φορών και η τελική ακρίβεια να καθοριστεί με τον μέσο όρο των επαναλήψεων.
- Τέλος, η μέθοδος επικύρωσης με αποκλεισμό ενός δείγματος χρησιμοποιεί ένα δείγμα για αξιολόγηση και τα υπόλοιπα για εκπαίδευση. Η διαδικασία αυτή τελειώνει όταν όλα τα δείγματα χρησιμοποιηθούν για αξιολόγηση

1.1.6 Βαθιά νευρωνικά δίκτυα

Η βαθιά Μάθηση (Deep Learning - ΒΜ), είναι ένας κλάδος της ΜΜ και έχει μεγάλη ανέλιξη τα τελευταία χρόνια. Έχει ξεπεράσει αρκετούς προηγούμενους περιορισμούς που αντιμετωπίζει γενικότερα η ΜΜ και έχει κινήσει το ενδιαφέρον σε ακαδημαϊκό αλλά και σε βιομηχανικό επίπεδο. Είναι αρκετά υποσχόμενη μέθοδος στην Βιοπληροφορική λόγω της δυνατότητας ανάλυσης δεδομένων υψηλών διαστάσεων και μπορεί επίσης να ανακαλύψει άγνωστα και περίπλοκα μοτίβα και συσχετίσεις για να παρέχουν πληροφορίες κατανόησης της φύσης των δεδομένων [24].

Τώρα, τα Βαθιά νευρωνικά δίκτυα (Deep Neural Networks - DNNs) λειτουργούν όπως και τα απλά τεχνητά νευρωνικά δίκτυα (ANNs) με την διαφορά ότι είναι πιο γενικευμένα. Δηλαδή τα ANNs αποτελούνται από ένα κρυφό επίπεδο ενώ τα DNNs αποτελούνται από λίγα έως και πολλά κρυφά επίπεδα μεταφέροντας δεδομένα μέσω συνάψεων από το ένα κρυφό επίπεδο στο επόμενο. Αυτό έχει ως συνέπεια να μπορούν να επεξεργαστούν καλύτερα τα δεδομένα και να παρθούν καλύτερα αποτελέσματα [24].



Εικόνα 1.8: Αρχιτεκτονική βαθιά νευρωνικών δικτύων με 3 κρυφά επίπεδα [24].

Μια από τις πιο διαδιδόμενες τεχνικές στα DNNs είναι η οπίσθιας διάδοσης (Backpropagation - BP). Αφού επιλέγονται τυχαία τα βάρη του δικτύου αυτή η τεχνική χρησιμοποιείται για τις απαραίτητες διορθώσεις των βαρών. Ουσιαστικά γίνεται τροφοδοσία του δικτύου προς τα εμπρός (δηλ. από το επίπεδο εισόδου προς το επίπεδο εξόδου), μετά γίνεται οπίσθια διάδοση στο επίπεδο εξόδου και στη συνέχεια στα κρυφά επίπεδα [25]. Τέλος, οι τιμές των βαρών ενημερώνονται χρησιμοποιώντας αλγόριθμους βελτιστοποίησης που βασίζονται στην Στοχαστική κατάβαση δυναμικού (Stochastic Gradient Descent - SDG), η οποία παρέχει στοχαστικές προσεγγίσεις εκτελώντας ενημερώσεις των βαρών για κάθε μικρό αριθμό δεδομένων του συνόλου. Παράδειγμα αλγορίθμων με SDG είναι ο RMSProp, ο Adagrad και Adam, όπου τροποποιούν προσαρμοστικά τα βάρη αναλόγως της τιμής που καθορίζεται για τον ρυθμό μάθησης (learning rate) [26].

Περιλαμβάνουν επίσης κάποιες επιπρόσθετες παραμέτρους βελτιστοποίησης του δικτύου όπως dropout και κανονικοποίηση βαρών. Πιο αναλυτικά:

- Το dropout είναι μια τεχνική που χρησιμοποιείται κατά την διάρκεια της εκπαίδευσης του δικτύου, η οποία διαγράφει τυχαία κρυφούς κόμβους και τα βάρη των υπόλοιπων κόμβων εκπαιδεύονται κανονικά με BP. Το κύριο κίνητρο πίσω από αυτή την τεχνική είναι να αποτρέψει προβλήματα όπως την υπερπροσαρμογή (overfitting) και να αναγκάσει το δίκτυο να προσαρμόζεται με την συμπεριφορά των δεδομένων [27].
- Η κανονικοποίηση βαρών είναι ένας τρόπος εξ' αναγκασμού των βαρών να λαμβάνουν πιο μικρές τιμές και έτσι συμβάλλει στην αντιμετώπιση της υπερπροσαρμογής. Υπάρχουν δύο τύποι, η L1 όπου με την τιμή που λαμβάνει επηρεάζει την απόλυτη τιμή των συντελεστών των βαρών και η L2 (ή μείωση βαρών) είναι ανάλογη με το τετράγωνο των συντελεστών των βαρών [28].

Βάσει των τύπων επιπέδων που χρησιμοποιούνται και την μέθοδο εκμάθησης, τα DNN μπορούν να κατηγοριοποιηθούν σε πολλαπλές κατηγορίες perceptron (Multilayer

Perceptron - MLP), σε δίκτυα βαθιάς πεποίθησης (Deep Belief Networks - DBN) και σε εξέλιξη αρχιτεκτονικής συστήματος (System Architecture Evolution - SAE) [24].

- Η αρχιτεκτονική των MLP όπως προαναφέρθηκε είναι παρόμοια με τα ANNs, με την διαφορά ότι μπορεί να έχουν περισσότερους κόμβους ή και κρυφά επίπεδα. Μπορούν να εκπαιδευτούν μόνο με δεδομένα χωρισμένα σε κλάσεις (με ετικέτα) και χρησιμοποιούνται συνήθως όταν διατίθεται μεγάλος αριθμός δεδομένων
- Οι αρχιτεκτονικές των DBN έχουν ως δομικά στοιχεία περιορισμένες μηχανές Boltzmann (Restricted Boltzmann Machine - RBM) και των SAE χρησιμοποιώντας αυτόματους κωδικοποιητές (Auto Encoders - AEs). Η κύρια διαφορά είναι ότι η εκπαίδευση εκτελείται αρχικά με μη ταξινομημένα δεδομένα (χωρίς ετικέτες) και στη συνέχεια για βελτιστοποίηση γίνεται επανεκπαίδευση με τα δεδομένα χωρισμένα στις κλάσεις (με ετικέτες). Επίσης έχουν ισχυρό πλεονέκτημα σε περιπτώσεις που υπάρχουν ανεπαρκή δεδομένα με ετικέτες να λαμβάνουν ικανοποιητικά αποτελέσματα.

1.1.7 Μέτρα αποτίμησης ποιότητας

Για να μπορέσουμε να μετρήσουμε την ποιότητα των μοντέλων έχουν αναπτυχθεί διάφορα μέτρα. Ξεκινώντας από τον πίνακα αληθείας, όπου είναι από τους ποιο σημαντικούς τρόπους απεικόνισης αξιοπιστίας των μοντέλων, μπορούν να παραχθούν όλα τα εξής μέτρα αξιολόγησης:

- **Πίνακας αληθείας:** Ένας πίνακας που απεικονίζει την απόδοση του μοντέλου χρησιμοποιώντας τα δεδομένα σε μήτρα. Χρησιμοποιείται κυρίως για προβλήματα ταξινόμησης και συγκρίνει την προβλεπόμενη ταξινόμηση με την πραγματική ταξινόμηση με Αληθώς Θετικά (ΑΘ), Ψευδώς Θετικά (ΨΘ), Αληθώς Αρνητικά (ΑΑ) και Ψευδώς Αρνητικά (ΨΑ) [29].

Πίνακας 1.1: Πίνακας αληθείας δυαδικής ταξινόμησης

		Πραγματική	
		Θετική κλάση	Αρνητική κλάση
Πρόβλεψη	Θετική κλάση	Αληθώς Θετικά	Ψευδώς Θετικά
	Αρνητική κλάση	Ψευδώς Αρνητικά	Αληθώς Αρνητικά

- **Ακρίβεια:** Το ποσοστό των σωστά ταξινομημένων δειγμάτων [29].

$$Ακρίβεια = \frac{ΑΘ + ΑΑ}{ΑΘ + ΨΘ + ΨΑ + ΑΑ}$$

- **Ευαισθησία:** Το ποσοστό των σωστά ταξινομημένων δειγμάτων από την πραγματική θετική κλάση [29].

$$Ευαισθησία = \frac{\text{ΑΘ}}{\text{ΑΘ} + \text{ΨΑ}}$$

- **Ειδικότητα:** Το ποσοστό των σωστά ταξινομημένων δειγμάτων από την πραγματική αρνητική κλάση [29].

$$Ειδικότητα = \frac{\text{ΑΑ}}{\text{ΑΑ} + \text{ΨΘ}}$$

- **Αληθώς θετικά:** Ο αριθμός των σωστά ταξινομημένων δειγμάτων της θετικής κλάσης [29].

- **Ψευδώς αρνητικά:** Ο αριθμός των λανθασμένα ταξινομημένων δειγμάτων της θετικής κλάσης [29].

- **Ψευδώς θετικά:** Ο αριθμός των λανθασμένα ταξινομημένων δειγμάτων της αρνητικής κλάσης [29].

- **Αληθώς αρνητικά:** Ο αριθμός των σωστά ταξινομημένων δειγμάτων της αρνητικής κλάσης [29].

- **Συντελεστής Kappa:** Χρησιμοποιείται συνήθως για προβλήματα πολλαπλών κλάσεων αλλά και σε προβλήματα ανισορροπίας δεδομένων. Συγκεκριμένα δείχνει πόσο ικανός είναι ο ταξινομητής συγκρίνοντας τον με ένα ταξινομητή που υποθέτει τυχαία για τα δεδομένα [30].

$$K = \frac{Πο - Πε}{1 - Πε}$$

  Όπου *Πο* αφορά την ακρίβεια του ταξινομητή που χρησιμοποιήθηκε και *Πε* την ακρίβεια του τυχαίου ταξινομητή. Η τιμή του είναι πάντα μικρότερη ή ίση με 1 και ορίζεται ως εξής:
  - $K < 0$ → Ανίκανος
  - 0.00 – 0.20 → Κακός
  - 0.21 – 0.40 → Σχεδόν μέτριος
  - 0.41 – 0.60 → Μέτριος
  - 0.61 – 0.80 → Ουσιαστικός
  - 0.81 – 1 → Σχεδόν τέλειος

- **Θετική προγνωστική αξία:** Ποσοστό σωστά ταξινομημένων δειγμάτων στην πρόβλεψη για τη θετική κλάση [29].

$$ΘΠΑ = \frac{\text{ΑΘ}}{\text{ΑΘ} + \text{ΨΘ}}$$

- **Αρνητική προγνωστική αξία:** Ποσοστό σωστά ταξινομημένων δειγμάτων στην πρόβλεψη για τη αρνητική κλάση [29].

$$ΑΠΑ = \frac{\text{ΑΑ}}{\text{ΑΑ} + \Psi\text{Α}}$$

- **Σταθμισμένη Ορθότητα:** Αφορά τον μέσο όρο της ευαισθησίας και της ειδικότητας [31].

$$ΣΟ = \frac{\text{Ευαισθησία} + \text{Ειδικότητα}}{2}$$

- **F1 score:** Είναι ο αρμονικός μέσος όρος μεταξύ θετικής προγνωστικής αξίας και ευαισθησίας. Χρησιμοποιείται για να ενημερώσει πόσες από τις προβλέψεις ταξινομήθηκαν σωστά και συγκλίνει προς το μικρότερο ποσοστό των δύο μέτρων λαμβάνοντας πιο αξιόπιστη βαθμολογία για το μοντέλο [32].

$$F1\ score = 2\ X \frac{ΘΠΑ + Ευαισθησία}{\text{ΘΠΑ} + \text{Ευαισθησία}}$$

- **Ποσοστό ψευδώς αρνητικών:** Είναι το αντίστροφο της ευαισθησίας. Δηλαδή το ποσοστό των δειγμάτων που στην πραγματικότητα είναι θετικά και έχουν προβλεφθεί ως αρνητικά [29].

$$ΠΨΑ = \frac{\Psi\text{Α}}{\Psi\text{Α} + \text{ΑΘ}}$$

- **Ποσοστό αληθώς αρνητικών:** Είναι το αντίστροφο της ειδικότητας. Δηλαδή το ποσοστό των δειγμάτων που στην πραγματικότητα είναι αρνητικά και έχουν προβλεφθεί ως θετικά [29].

$$ΠΑΑ = \frac{\Psi\Theta}{\Psi\Theta + \text{ΑΑ}}$$

- **Ψευδές ποσοστό ανακάλυψης:** Είναι το αντίστροφο της θετικής προγνωστικής αξίας. Δηλαδή το ποσοστό των λάθος ταξινομημένων δειγμάτων στην θετική κλάση [31].

$$ΨΠΑ = \frac{\Psi\Theta}{\Psi\Theta + \text{ΑΘ}}$$

- **Ψευδές ποσοστό παράλειψης:** Είναι το αντίστροφο της αρνητικής προγνωστικής αξίας. Δηλαδή το ποσοστό των λάθος ταξινομημένων δειγμάτων στην αρνητική κλάση [31].

$$\Psi\Pi\Pi = \frac{\Psi A}{\Psi A + AA}$$

- **Συντελεστής συσχέτισης Matthews:** Είναι ένας συντελεστής συσχέτισης μεταξύ πραγματικών και προβλέψιμων τιμών και παράγει υψηλή βαθμολογία εάν το μοντέλο μπόρεσε να προβλέψει σωστά την πλειοψηφία των θετικών και αρνητικών δειγμάτων [32].

$$\Sigma\Sigma M = \frac{A\Theta\ \chi\ AA - \Psi\Theta\ \chi\ \Psi A}{\sqrt{(A\Theta + \Psi\Theta)\ \chi\ (A\Theta + \Psi A)\ \chi\ (AA + \Psi\Theta)\ \chi\ (AA + \Psi A)}}$$

- **Informedness:** Ο δείκτης αυτός δίνει ίσο βάρος στις ΨΘ και ΨΑ τιμές, οπότε έχει την ίδια αναλογία στα συνολικά λανθασμένα αποτελέσματα [33].

$$Informedness = Ευαισθησία + Ειδικότητα - 1$$

- **Markedness:** Ο δείκτης αυτός ποσοτικοποιεί πόσο αξιόπιστη είναι η πρόβλεψη και καθορίζει την πιθανότητα να προβλεφθεί σωστά ένα νέο δείγμα από το μοντέλο [33].

$$Markedness = \Theta\Pi A + A\Pi A - 1$$

1.2 Λευχαιμία και γονιδιακές εκφράσεις

Η ταξινόμηση καρκίνων λευχαιμίας βασίστηκε κυρίως στη μορφολογική εμφάνιση του όγκου, αυτό όμως έχει σοβαρούς περιορισμούς. Οι όγκοι με παρόμοια ιστοπαθολογική εμφάνιση μπορούν να ακολουθούν διαφορετικά κλινικά πρωτόκολλα και να αντιδρούν διαφορετικά σε συγκεκριμένες θεραπείες. Η Οξεία μυελοειδής λευχαιμία (AML) και η οξεία λεμφοβλαστική λευχαιμία (ALL), είναι δύο τύποι λευχαιμίας που έχουν διαφορετικό φαινότυπο. Ωστόσο, υποκατηγοριοποιούνται στον όρο «Λευχαιμία», καθώς έχουν ορισμένα κοινά χαρακτηριστικά. Οι πιο κύριες διαφορές μεταξύ των δύο αναγράφονται στον πιο κάτω πίνακα [34].

Πίνακας 1.2: Διαφορές μεταξύ ALL και AML [34].

	ALL	AML
ΣΗΜΕΙΑ ΑΝΑΡΡΟΦΗΣΗΣ	Βλαστικά κύτταρα περιφερικού αίματος ή με αναρρόφηση μυελού των οστών	Βλαστικά κύτταρα περιφερικού αίματος ή με αναρρόφηση μυελού των οστών
ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ	Πολλά ανώριμα λευκά αιμοσφαίρια στον μυελό των οστών	Μεγάλοι ράβδοι Auer σε μυελοειδή βλαστικά κύτταρα
ΗΛΙΚΙΑΚΑ ΠΟΣΟΣΤΑ	Προσβάλλει κυρίως παιδία και νέους κάτω των 20 ετών	Ενήλικες.
ΣΥΜΠΤΩΜΑΤΑ	Πυρετός, ο λήθαργος, η αιμορραγία, ο μυοσκελετικός πόνος, η ηματοσπληνομεγαλία (διόγκωση του ήπατος και του σπλήνα) και η λεμφαδενοπάθεια	Πυρετός, αιμορραγία, ηματοσπληνομεγαλία, λεμφαδενοπάθεια, απώλεια βάρους, μώλωπες
ΠΟΣΟΣΤΟ ΕΠΙΒΙΩΣΗΣ	<50 ετών: 75% >= 50 ετών: 25%	< 50 ετών: 55% >= 50 ετών: 14%

Η μηχανική μάθηση (ΜΜ) αναδύεται γρήγορα σε διάφορους τομείς της έρευνας για τον καρκίνο. Οι αλγόριθμοι ΜΜ μπορούν να αντιμετωπίσουν τεράστιες ποσότητες ιατρικών δεδομένων και να παρέχουν καλύτερη κατανόηση των κακοηθών. Η ικανότητά επεξεργασίας πληροφοριών από διαφορετικούς διαγνωστικούς τρόπους και λειτουργίες για την πρόβλεψη της πρόγνωσης, προτείνει θεραπευτικές στρατηγικές και δείχνει ότι η ΜΜ στη διαχείριση των τύπων Λευχαιμίας μπορεί να εξασφαλίσει γρήγορη και ακριβή διάγνωση, ποσοστό ρίσκου και βέλτιστη θεραπεία. Ωστόσο, αυτές οι τεχνικές συνοδεύονται από διάφορες παγίδες και χρειάζονται ένα αυστηρό κανονιστικό πλαίσιο για να διασφαλιστεί η ασφαλής χρήση της ΜΜ [35]. Με την αύξηση των αλληλουχιών του DNA στις βάσεις δεδομένων, οι αλγόριθμοι μηχανικής μάθησης έχουν εφαρμοστεί ολοένα και περισσότερο στην ανάλυση γονιδιακής έκφρασης, με στόχο την ταξινόμηση όγκων, την πρόβλεψη επιβίωσης, τον εντοπισμό θεραπευτικών στόχων και την ταξινόμηση γονιδιακών λειτουργιών. Επίσης, αλγόριθμοι ΜΜ έχουν σχεδιαστεί για την ανάλυση μοτίβων έκφρασης γονιδίου που λαμβάνονται μέσω αλληλουχίας RNA (RNA - seq), όπου μπορούν να χρησιμοποιηθούν για την ακριβή πρόβλεψη πιθανότητας πλήρους ύφεσης σε ασθενείς με AML που έχουν λάβει θεραπεία. Ο Ophir G. et. al. χρησιμοποίησαν δεδομένα γονιδιακής έκφρασης ώστε να διερευνήσουν και να αξιολογήσουν διαφορετικούς αλγορίθμους ΜΜ για την πρόβλεψη πλήρους ύφεσης σε ασθενείς με AML πριν από τη θεραπεία. Κατά αυτό τον τρόπο αποκάλυψε μια σημαντική υποκείμενη γενετική διαφορά μεταξύ των ασθενών με αντίθετα αποτελέσματα μετά τη θεραπεία [36].

Η διάγνωση της λευχαιμίας είναι μια χρονοβόρα και απαιτητική διαδικασία, επομένως οι τεχνικές ΜΜ μπορούν να βοηθήσουν στην εξέλιξη των ερευνών αυτών. Τα λάθη λόγω έλλειψης

εμπειρίας στο προσωπικό καθώς και οι περιορισμένες παροχές των νοσοκομείων, έχουν οδηγήσει πολλές φορές σε λανθασμένες αποφάσεις σε πολλές ασθένειες αλλά και στην λευχαιμία. Ωστόσο, η ΜΜ παρέχει αρκετά πλεονεκτήματα, όπως υψηλή ακρίβεια στην διάγνωση, μειωμένο κόστος και γρήγορη ροή στα περιστατικά. Σαφώς και είναι απίθανο η ΜΜ να αντικαταστήσει τους γιατρούς, όμως συμβάλει στην βελτίωση της διάγνωσης; όπως και της περίθαλψης [36].

1.3 Καρκίνος του μαστού και μορφολογικά χαρακτηριστικά

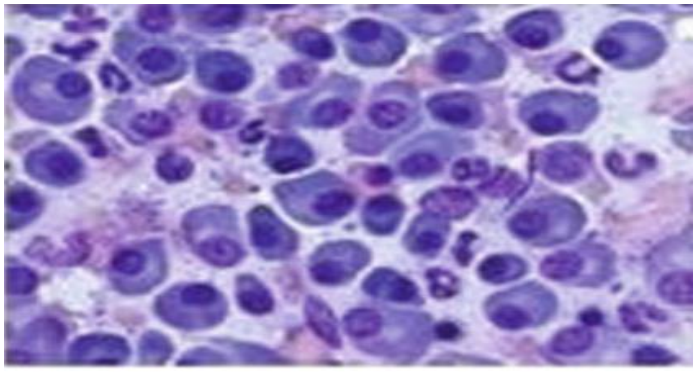
Ο καρκίνος του μαστού είναι η δεύτερη κύρια αιτία θανάτου για γυναίκες σε όλο τον κόσμο και περισσότερο από το 8% περνούν την ασθένεια αυτή κατά την διάρκεια της ζωής τους. Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας (ΠΟΥ) γίνεται διάγνωση σε περίπου ένα εκατομμύριο γυναίκες και πάνω από πεντακόσιες χιλιάδες πεθαίνουν κάθε χρόνο. Επίσης, εκτιμάται ότι συχνότητα εμφάνισης της ασθένειας αυτής θα αυξάνεται με την πάροδο του χρόνου λόγω της καταστροφής του περιβάλλοντος. Πιο αναλυτικά, ο καρκίνος ξεκινά με την ανεξέλεγκτη διαίρεση ενός κυττάρου, με αποτέλεσμα να εμφανίζεται μια ορατή μάζα που ονομάζεται όγκος. Χωρίζεται σε δύο κατηγορίες τον καλοήθη και τον κακοήθη, όπου ο κακοήθης αναπτύσσεται πιο γρήγορα και εισβάλλει στους γύρω ιστούς προκαλώντας την βλάβη τους. Αν και η συχνότητα εμφάνισης του καρκίνου του μαστού αυξάνεται, η θνησιμότητα του έχει μειωθεί σε γυναίκες όλων των ηλικιών. Αυτό μπορεί να σχετίζεται με βελτιώσεις στη θεραπεία και την υιοθέτηση του μαστογραφικού ελέγχου. Ωστόσο, είναι γνωστό ότι οι ειδικοί ακτινολόγοι μπορούν να χάσουν ένα σημαντικό ποσοστό κακοηθειών και να οδηγηθούν σε εσφαλμένη διάγνωση. Πιο συγκεκριμένα, το κλειδί για την μείωση ποσοστού θνησιμότητας (40% ή και περισσότερο), είναι η έγκαιρη διάγνωση του, όσο νωρίτερα ανιχνευθεί η ασθένεια, τόσο καλύτερη θα είναι η θεραπεία. Όμως η έγκαιρη ανίχνευση απαιτεί ακριβή και αξιόπιστη διάγνωση που θα πρέπει να μπορεί να διακρίνει καλοήθεις και κακοήθεις όγκους [37].

Πίνακας 1.3: Κύρια χαρακτηριστικά του καλοήθη και κακοήθη όγκου [38].

Καλοήθης όγκος	Κακοήθης όγκος
Δεν κάνει μετάσταση	Συνήθως κάνει μετάσταση σε άλλα μέρη του σώματος
Αναπτύσσονται αργά	Αναπτύσσονται γρήγορα
Είναι απίθανο να επανεμφανισθεί εάν αφαιρεθεί και απαιτούν χημειοθεραπείες /ή και ακτινοβολίες	συνήθως απαιτεί χημειοθεραπεία /ή και ακτινοβολίες και χειρουργική επέμβαση.
Καθαρά όρια, φυσιολογικοί πυρήνες & ομαλό σχήμα	Ανώμαλο σχήμα, μη φυσιολογικά χρωμοσώματα και DNA, χαρακτηρίζονται από σκοτεινούς μεγάλους πυρήνες

Οι δύο πιο συχνά χρησιμοποιούμενες μέθοδοι για την ανίχνευση καρκίνου του μαστού είναι η φυσική εξέταση και η μαστογραφία. Έτσι, μπορούν να προσφέρουν κατά προσέγγιση πιθανότητα ότι ένα κομμάτι είναι κακοήθη και μπορεί επίσης να εντοπίσει

κάποιες άλλες βλάβες, όπως μια απλή κύστη. Καθώς αυτές οι μέθοδοι είναι ασαφείς, ένας επαγγελματίας υγειονομικής περίθαλψης μπορεί να εκτελέσει μια διαδικασία γνωστή ως αναρρόφηση λεπτής βελόνης ή αναρρόφηση και κυτταρολογία λεπτής βελόνης (Fine Needle Aspiration - FNA) λαμβάνοντας δείγμα υγρού από τον όγκο για ανάλυση με μικροσκόπιο, όπου βοηθά στην αξιοπιστία της διάγνωσης. Η βιοψία αναρρόφησης με λεπτή βελόνα είναι μια αρκετά απλή και γρήγορη διαδικασία για τη διάγνωση της κακοήθειας. Ουσιαστικά, η διαδικασία περιλαμβάνει τη συλλογή ενός δείγματος κυττάρων ή υγρού από μια κύστη που ακολουθείται από εξέταση των κυττάρων με μικροσκόπιο. Εάν το κομμάτι δεν μπορεί να γίνει αισθητό, τότε απαιτείται απεικόνιση για να βρεθεί η ακριβής τοποθεσία. Αυτό μπορεί να γίνει με υπερηχογράφημα ή στερεοτακτική μαστογραφία. Στην τεχνική υπερήχων, ο χειρουργός παρακολουθεί τη βελόνα στην οθόνη υπερήχων και την οδηγεί στην περιοχή και στη στερεοτακτική μαστογραφία (για το στήθος), δύο μαστογραφίες χρησιμοποιούνται σε διαφορετικές γωνίες και με τη βοήθεια ενός υπολογιστή δημιουργούνται ακριβείς συντεταγμένες [39]. Ωστόσο, όλες αυτές οι μέθοδοι είναι πολύ χρονοβόρες οπότε έχουν χρησιμοποιηθεί τεχνικές εξαγωγής χαρακτηριστικών κυτταρικών πυρήνων από εικόνες με καρκίνο του μαστού.



Εικόνα 1.9: Εικόνα FNA για επεξεργασία [39].

Οι συμβατικές μέθοδοι παρακολούθησης και διάγνωσης βασίζονται στην ανίχνευση συγκεκριμένων χαρακτηριστικών από έναν άνθρωπο παρατηρητή. Λόγω του μεγάλου αριθμού ασθενών σε μονάδες εντατικής θεραπείας (ΜΕΘ) όμως αυτό θεωρείται δύσκολο, οπότε με την βοήθεια υπολογιστών έχουν αναπτυχθεί αυτοματοποιημένα διαγνωστικά συστήματα όπου μετατρέπουν τα ως επί το πλείστων ποιοτικά διαγνωστικά κριτήρια σε ένα πιο αντικειμενικό πρόβλημα ποσοτικής ταξινόμησης χαρακτηριστικών [37]. Προκειμένου λοιπόν να βελτιωθεί η ακρίβεια της διάγνωσης έχουν εφαρμοστεί μεθόδοι ΜΜ, όπου δείχνουν να είναι υψηλά ακριβείς σε μορφολογικά χαρακτηριστικά κυττάρων [39].

1.4 Πάρκινσον και σήμα ομιλίας

Η νόσος του Πάρκινον (Parkinson) είναι σπάνια πριν από την ηλικία των 50 ετών, αλλά η επίπτωση αυξάνεται 5-10 φορές από τα 60 εώς τα 90 έτη ζωής. Οι παγκόσμιες εκτιμήσεις νοσηλείας της κυμαίνονται από 5 έως 35 νέες περιπτώσεις ανά 100.000 άτομα ετησίως και με τον παγκόσμιο επιπολασμό να εκτιμάται περίπου στο 0.3% και να

αυξάνεται απότομα σε ηλικίες άνω 80, περίπου 3%. Η θνησιμότητα της νόσου αυτής δεν αυξάνεται την πρώτη δεκαετία αλλά αυξάνεται με την πάροδο του χρόνου. Η βελτίωση της υγειονομικής περίθαλψης οδήγησε σε μεγαλύτερη επιβίωση ασθενών και καλύτερη ποιότητα ζωής αλλά καμία από τις περιθάλψεις δεν έχει οδηγήσει στην πλήρη θεραπεία της νόσου [41]. Παραμένει λοιπόν μια προοδευτική διαταραχή που προκαλεί σοβαρή αναπηρία όπου η περαιτέρω καθυστέρηση της και ο εντοπισμός πρώιμων εκδηλώσεων της νόσου, είναι από τις βασικές ανεκπλήρωτες ανάγκες.

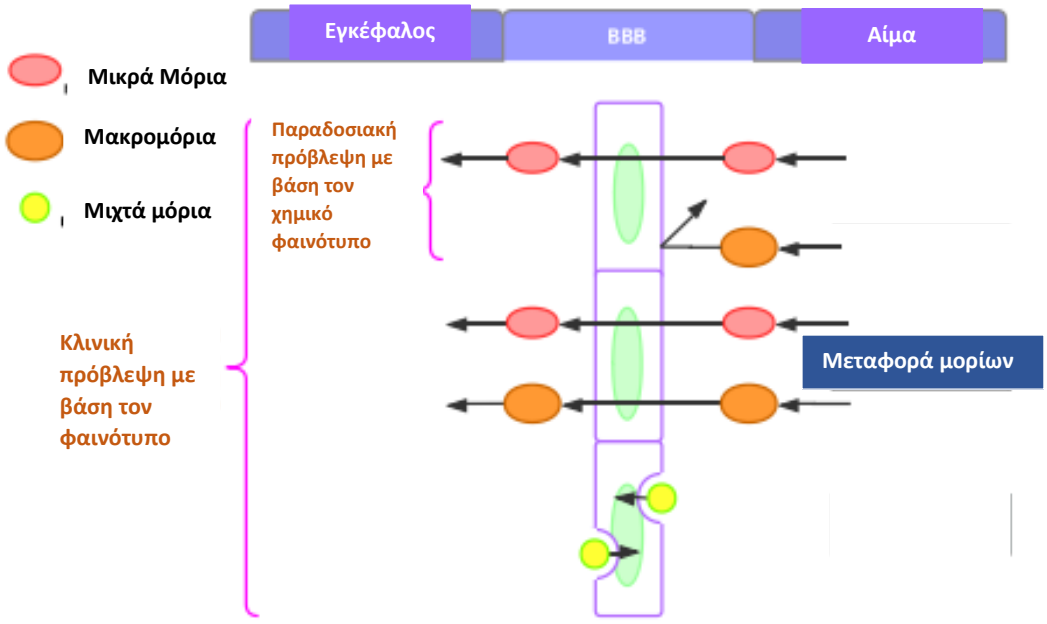
Πίνακας 1.4: Κύρια συμπτώματα της νόσου του Πάρκινσον [41].

Συμπτώματα Πάρκινσον
Τρέμουλο (Συνήθως κατά την ξεκούραση)
Βραδυκινησία
Φωνητικοί τόνοι
Ακαμψία & στάση του σώματος, πόνος
Δυσκολία βάδισης
Γαστρεντερικά προβλήματα
Ψυχικά προβλήματα
Μειωμένη αίσθηση οσμής

Τα συστήματα διάγνωσης στην νόσο του Πάρκινσον βασίζονται στην μέτρηση της σοβαρότητας των συμπτωμάτων χρησιμοποιώντας μη επεμβατικές συσκευές και εργαλεία. Ένα από τα πιο σημαντικά προβλήματα που παρατηρήθηκαν σε ποσοστό περίπου 90% των ασθενών, στα αρχικά στάδια της νόσου, είναι τα φωνητικά προβλήματα. Επομένως έχουν διενεργηθεί μελέτες με διάφορους αλγόριθμους για επεξεργασία σήματος ομιλίας και την εξαγωγή χρήσιμων πληροφοριών [42]. Η ΜΜ έδειξε ότι μπορεί να βοηθήσει αρκετά λόγω της δυνατότητας που έχει να μπορεί να εξετάζει και να συγκρίνει μια ποικιλία αλγορίθμων ταυτόχρονα. Ο χαρακτηρισμός και ο ποσοτικός προσδιορισμός των φωνητικών παραμέτρων είναι χρήσιμοι για την καλύτερη κατανόηση των αλλαγών στις φωνές των ασθενών. Οι εξασθενημένες φωνητικές πτυχές και η έλλειψη ευφράδειας λόγου στους ασθενείς με Πάρκινσον μπορεί να οδηγήσουν σε μεταβαλλόμενες δονήσεις της γλωττίδας, αλλαγές στο εύρος ακοής και να επηρεάσουν την ένταση της φωνής (υποφωνία) [43]. Με αρκετές μεταβλητές φωνητικής συχνότητας, έντασης και παραμέτρων μη γραμμικής δυναμικής που υπολογίζονται από τα ηλεκτρικά σήματα και εντοπίζονται από τυπικές δοκιμές ομιλίας, μπορούν να επιλεχθούν οι πιο διακριτές φωνητικές παράμετροι που ταξινομούν πιο σωστά τους πάσχοντες από τους υγιούς.

1.5 Διαπερατότητα αιματοεγκεφαλικού φραγμού (BBB)

Το φράγμα αίματος εγκεφάλου (BBB) είναι ένα επιλεκτικό ημιδιαπερατό φράγμα μεμβράνης που βρίσκεται στον εγκέφαλο και αποτελείται από επιλεκτικούς σφιχτούς κόμβους για να διαχωρίσει το κεντρικό νευρικό σύστημα (CNS) από το κυκλοφορικό σύστημα. Η διαπερατότητα του BBB από τα φάρμακα σχετίζεται με τις διαφορετικές χημικές ιδιότητες των χημικών ενώσεων των φαρμάκων. Είναι μια σημαντική πρόκληση στην ανακάλυψη νευροθεραπευτικών φαρμάκων γιατί οι συμβατικές προσεγγίσεις για τη μέτρηση της διαπερατότητας BBB είναι δαπανηρές, χρονοβόρες και απαιτούν εργασία. Επί του παρόντος, οι νευρολογικές ασθένειες αντιπροσωπεύουν το 28% των ατόμων με αναπηρίες όλων των ηλικιών. Παρά τον μεγάλο αριθμό ασθενειών που αφορούν το Κεντρικό Νευρικό Σύστημα (Central Nervous System - CNS), τα αποτελεσματικά φάρμακα για την αντιμετώπιση τους είναι ανεπαρκή. Πιο συγκεκριμένα το BBB προστατεύει τον εγκέφαλο από παρεμβατικές ξενοβιοτικές χημικές ενώσεις για να διατηρήσει την ομοιόσταση του εγκεφάλου. Η δομή του κυρίως αποτελείται από εγκεφαλικό ενδοθήλιο, το οποίο εμποδίζει την διαπερατότητα των μεγάλων μορίων 100% και μικρών μορίων 98% στο CNS. Επίσης επιτρέπει τη μεταφορά μόνο των υδατικών λιποδιαλυτών και επιλεκτικά μεταφορικών μορίων. Στις ασθένειες του CNS, το χαμηλό ποσοστό επιτυχίας της ανακάλυψης φαρμάκων οφείλεται στην αποτυχία των σχεδιασμένων ενώσεων να διασχίσουν το BBB, με αποτέλεσμα ανεπαρκή έκθεση στο CNS. Οι ερευνητές έχουν κάνει πολλές προσπάθειες για την ανακάλυψη φαρμάκων. Ωστόσο, πολλές δοκιμασμένες ενώσεις είχαν αποτύχει λόγω της έλλειψης ικανότητας διείσδυσης στο BBB παρά στην έλλειψη δραστικότητας. Το BBB λοιπόν είναι ένας από τους λόγους που έχει επιβραδύνει η ανακάλυψη φαρμάκων για το CNS [44], [45].Έτσι, πρέπει να δοθεί μεγαλύτερη προσοχή για να διασφαλιστεί η εξεύρεση και βελτιστοποίηση κατάλληλων χημικών ενώσεων για την ανάπτυξη νέων φαρμάκων που θα καταφέρνουν να επιδρούν στο CNS.



Εικόνα 1.10: Μηχανισμοί φαρμάκων που περνούν το BBB. Το δεξί μέρος παρουσιάζει το αιμοφόρο αγγείο, το

οποίο δείχνει τους μηχανισμούς για τη διέλευση του φαρμάκου στο BBB, και το αριστερό μέρος είναι ο εγκέφαλος, οποίος δείχνει το εύρος των κλινικών μεθόδων πρόβλεψης διαπερατότητας BBB με βάση το φαινότυπο και των χημικών χαρακτηριστικών [45].

Η ανάπτυξη φαρμάκων για τη θεραπεία εγκεφαλικών παθήσεων απαιτεί τη μελέτη της επίδρασης του φαρμάκου στο BBB καθώς και τη μεταφορά του μέσω του BBB. Συνήθως, αυτές οι δοκιμές πραγματοποιούνται σε μικρά ζώα, δεδομένου ότι επιτρέπουν τον χαρακτηρισμό της φαρμακοδυναμικής και της φαρμακοκινητικής του φαρμάκου καθώς και την πιθανή ανοσοαπόκριση που προκαλείται από το φάρμακο. Ωστόσο αυτές οι δοκιμές είναι δαπανηρές, χρονοβόρες και δημιουργούν ηθικά ζητήματα λόγω της χρήσης ζώων. Επιπλέον, λόγω των διαφορών μεταξύ του BBB σε ζώα και ανθρώπους στην έκφραση της P-γλυκοπρωτείνης (Pgp), των πρωτεϊνών, των μεταφορέων και των διαμεμβρανικών πρωτεϊνών που σχετίζονται με την αντοχή στα φάρμακα, καθώς και της φαρμακοκινητικής του εγκεφάλου των υποστρωμάτων Pgp, πολλά επιτυχημένα φάρμακα σε προ-κλινικές φάσεις απέτυχαν στην εφαρμογή στους ανθρώπους [46].

Τα φάρμακα του CNS εμφανίζουν συγκεκριμένες φυσικοχημικές και φαρμακοκινητικές ιδιότητες, οι οποίες τους επιτρέπουν να διεισδύσουν στο BBB. Επιπλέον, ο προσδιορισμός της ιδιότητας διείσδυσης BBB είναι απαραίτητη προϋπόθεση για την εύρεση υποψήφιων φαρμακευτικών ενώσεων που μπορούν να επηρεάσουν το CNS. Γενικά για να διαπεράσουν το BBB πρέπει να ακολουθούν τον κανόνα των 5 του Λιπίνσκι: Φάρμακα που μοιάζουν με ενώσεις πρέπει να έχουν λιπόφιλη δράση < 5, μοριακό βάρος < 500, αποδέκτες δεσμού υδρογόνου < 10, δότες υδρογόνου < 5 και πολική επιφάνεια < 140 A^2. Ωστόσο, οι ενώσεις του CNS είναι λιγότερο πολικές και λιπόφιλες και έχουν χαμηλότερα μοριακά βάρη (<450), χαμηλότερη περιοχή πολικής επιφάνειας (60-70Å2), δέκτες δεσμών υδρογόνου <7, δότες δεσμού υδρογόνου <3 και λιγότερο εύκαμπτες δομές με λίγους περιστρεφόμενους δεσμούς (<8)[44],[47].

Τα τελευταία χρόνια, με την ανάπτυξη της τεχνητής νοημοσύνης έχουν εμφανιστεί ορισμένες στατιστικές μέθοδοι ή αλγόριθμοι ΜΜ για να γίνεται πιο εύκολα αυτή η πρόβλεψη και όλο και περισσότερο συγκεντρώνουν την προσοχή της επιστημονικής κοινότητας. Σε πολλές τέτοιες εφαρμογές η μηχανική μάθηση έχει δείξει ισχυρές δυνατότητες να ανταγωνιστεί ή ακόμη και να ξεπεράσει συμβατικούς υπολογισμούς. Στις πλείστες έρευνες, οι τιμές των CNS+/ CNS-, LogBBB και surface permeability product (LogPS) χρησιμοποιούνται ως τιμές αναφοράς για να περιγράψουν την ικανότητα διαπερατότητας του BBB σε μια ποσοτική σχέση δομής-δραστικότητας (Quantitative Structure – Activity Relationship - QSAR). Χρησιμοποιούνται ευρέως γνωστοί αλγόριθμοι ΜΜ όπως η μηχανή διανυσμάτων υποστήριξης (SVM), ο συλλογικός ταξινομητής τυχαίου δάσους (Random Forest), ο K-πλησιέστερων γειτόνων (K-NN) και τα τεχνητά νευρωνικά δίκτυα (ANN) [44], [45] για την κατασκευή εποπτευόμενων μοντέλων ταξινόμησης, στα οποία η BBB+ (οι χημικές ενώσεις που μπορούν να διαπεράσουν το BBB) και BBB- (χημικές ενώσεις που δεν μπορούν να διαπεράσουν το BBB) χρησιμοποιούνται ως οι 2 κλάσεις για την εκμάθηση των μοντέλων ταξινόμησης.

2. Εφαρμογές Μηχανικής Μάθησης

Κατά την διάρκεια της βιβλιογραφικής ανασκόπησης παρατηρήθηκε πόσο σημαντικό είναι να επιλεχθούν οι κατάλληλοι παράμετροι και μέθοδοι για την αποτελεσματική εφαρμογή της ΜΜ στα δεδομένα. Υπάρχουν διάφορες τεχνικές ΜΜ, στο κεφάλαιο αυτό συγκεκριμένα θα δοκιμαστούν ευρέως γνωστοί αλγόριθμοι που φαίνεται να χρησιμοποιούνται πιο συχνά σε προβλήματα ταξινόμησης στην Βιοπληροφορική. Έπειτα, θα αναδειχθεί ο τρόπος εφαρμογής ΜΜ σε τέσσερα διαφορετικά πεδία για να μπορέσουμε να διαπιστώσουμε την αποτελεσματικότητα της με διάφορες μετρήσεις αξιολόγησης.

2.1 Υλικά και λογισμικά

Η εφαρμογή ΜΜ υλοποιήθηκε μέσω ενός φορητού υπολογιστή CPU (Intel core i7) με διαθέσιμη μνήμη RAM περίπου στα 3.5 GB και χρήση της γλώσσας προγραμματισμού R. Η γλώσσα R είναι ένα δωρεάν εργαλείο που μπορεί να εγκατασταθεί εύκολα και είναι ευέλικτη για στατιστικούς υπολογισμούς και απεικόνιση γραφημάτων. Παρέχει μια ποικιλία στατιστικών μεθόδων όπως γραμμική και μη γραμμική μοντελοποίηση, κλασικές στατιστικές δοκιμές, ταξινόμηση και συσταδοποίηση. Είναι αρκετά παρόμοια με την γλώσσα προγραμματισμού S, αν και έχουν σημαντικές διαφορές, μπορούν κομμάτια κώδικα της S να τρέξουν αναλλοίωτα σε περιβάλλον R. Επιπρόσθετα, είναι αρκετά χρήσιμη για την επιστημονική κοινότητα λόγω του πλεονεκτήματος της να παράγει με ευκολία γραφικές παραστάσεις, μαθηματικών εξισώσεων και συμβόλων όπου χρειάζεται. Μπορεί να διαχειριστεί δεδομένα να αποθηκεύσει δεδομένα και να προβληθούν στην οθόνη του υπολογιστή είτε σε έντυπη μορφή. Επιπλέον, έχει την δυνατότητα να συνδεθεί με άλλες πλατφόρμες όπως C, C++, και Fortran για υπολογιστικά απαιτητικές εργασίες αλλά και μέσω διάφορων πακέτων που είναι διατεθειμένα στο διαδικτυακό ιστότοπο CRAN για να καλύψει ένα ευρύ φάσμα σύγχρονων στατιστικών [48]. Για την ανάπτυξη του κώδικα της γλώσσας R χρησιμοποιήθηκε το RStudio, το οποίο ένα περιβάλλον ανάπτυξης (Integrated Development Environment - IDE) που μπορεί να συνδυάσει διάφορα κομμάτια (κονσόλα, επεξεργασία πηγής, γραφήματα, ιστορικό, βοήθεια κ.λπ.). Το RStudio εκτελείται σχεδόν σε όλες τις μεγάλες πλατφόρμες Windows, Mac OS X και Linux όπου έχει σχεδιαστεί για την διευκόλυνση εκμάθησης σε νέους χρήστες R αλλά και να παρέχει υψηλή παραγωγικότητα σε προχωρημένους χρήστες [49]. Για την υλοποίηση του ΜΜ σε αυτή την μελέτη χρησιμοποιήθηκαν κυρίως τα πακέτα Caret, Keras, tensorflow, ggplot2, ROCR, pROC, webchem .

- Το πακέτο Caret είναι ένας αναδιπλωτής (wrapper) όπου μπορεί να συνδεθεί με άλλα πακέτα και να δώσει την ευκολία χειρισμού για μεθόδους ΜΜ. Δημιουργήθηκε συγκεκριμένα για να μπορεί ένας χρήστης να αξιολογήσει γρήγορα πολλούς διαφορετικά μοντέλα για να βρεθεί η βέλτιστη λύση στα δεδομένα. Η R παρέχει ποίκιλα πακέτα για μοντελοποίηση, όμως λόγω της δυσκολίας να βρεθούν τα μοντέλα από τα πακέτα και λόγω των συντακτικών αλλαγών που έχουν στον κώδικα ίσως να αποθαρρύνεται ο χρήστης να αξιολογεί ποίκιλα μοντέλα. Έτσι το Caret βοηθάει στην μείωση της πολυπλοκότητας παρέχοντας διεπαφή σε λειτουργίες για μοντέλα (πάνω από 140) και προβλέψεις

ΜΜ. Το πακέτο αυτό παρέχει επίσης πολλές επιλογές για προεπεξεργασία δεδομένων και τεχνικές συντονισμού παραμέτρων [50].

- Το keras δουλεύει μέσω διεπαφής προγραμματισμού εφαρμογών (Application Programming Interface - API) με βαθιά νευρωνικά δίκτυα ψηλού επιπέδου και έχει υλοποιηθεί ούτως ώστε να εκτελείται αναλόγως γρήγορα. Έχει αρκετά φιλικό API και αυτό το κάνει εύκολο προς την χρήση για την δημιουργία των μοντέλων και επιτρέπει την εκτέλεση του ίδιου κώδικα σε CPU ή GPU [51], όπου έγινε χρήση για την δημιουργία των βαθιά νευρωνικών δικτύων.
- Το tensorflow που επίσης χρησιμοποιήθηκε για τα Βαθιά νευρωνικά δίκτυα, αφορά αριθμητικούς υπολογισμούς χρησιμοποιώντας γραφήματα ροής δεδομένων. Έχει και αυτό ευέλικτη αρχιτεκτονική που επιτρέπει υπολογισμούς σε ένα ή και πολλούς υπολογιστές GPU ή CPU, σε διακοσμητές και κινητή συσκευή με ένα μόνο API. Αναπτύχθηκε αρχικά από μηχανικούς και ερευνητές που εργάζονται στην ομάδα Google Brain, κυρίως για σκοπούς μηχανικής μάθησης και βαθιά νευρωνικά δίκτυα αλλά μπορεί να εφαρμοστεί και σε άλλους τομείς αποτελεσματικά [52].
- Ένα άλλο πακέτο που χρησιμοποιήθηκε αρκετά για την παραγωγή γραφικών παραστάσεων είναι το ggplot2. Είναι ισχυρό, επειδή δεν περιορίζεται σε προκαθορισμένα γραφήματα αλλά παρέχει την δυνατότητα να δημιουργηθούν νέα γραφικά μορφοποιημένα για το πρόβλημα. Επιπλέον, είναι βοηθητικό γιατί μπορεί να δημιουργηθεί αρχικά ένα γράφημα και στην συνέχεια να προστεθούν επίπεδα σχολιασμών, υπομνήματα ή και δείκτες [53].
- Τα πακέτα pROC και ROCR αφορούν την ανάλυση (Receiver Operating Characteristic - ROC. Εφαρμόζουν πολλαπλές στατιστικές μεθόδους για να συγκριθούν οι καμπύλες ROC και οι περιοχές κάτω από τις καμπύλες (Area Under the Curve - AUC).
- Τέλος, το πακέτο webchem αλληλοεπιδρά με μια σειρά από διαδικτυακές βάσεις δεδομένων και μπορεί μέσω του να ληφθούν χημικές πληροφορίες [54].

Τα σύνολα δεδομένων συλλέχθηκαν από διάφορες αναφορές και πηγές. Έγινε προσπάθεια συλλογής από διαφορετικά πεδία Βιοπληροφορικής για να αναδειχθεί αν είναι χρήσιμη και ευέλικτη η ΜΜ. Συγκεκριμένα:

I. Το σύνολο δεδομένων λευχαιμίας (ALL – AML) πάρθηκε από τον ιστότοπο Kaggle [55] και αφορά την έρευνα που δημοσιεύτηκε από τους Golub et al. (1999) [56] , όπου παρουσίασαν πως νέες περιπτώσεις καρκίνου του μαστού, μέσω μικροσυστοιχιών DNA, μπορούν να ταξινομηθούν με παρακολούθηση γονιδιακής έκφρασης [54]. Κατ’ επέκταση συλλέξανε τις μετρήσεις από τον μυελό των οστών και επίχρισμα αίματος, στην συνέχεια τοποθέτησαν διαφορετικά ολιγονουκλεοτίδια (probes) DNA σε μια στερεή επιφάνεια σε ομάδες σχηματίζοντας συστοιχίες μικροσκόπησης και μετά εφάρμοσαν φως φθορισμού στα ολιγονουκλεοτίδια για να αποκαλυφθούν τυχόν κοινά γονίδια μεταξύ των δειγμάτων, οι διαφορές καθώς και ελαττωματικά γονίδια. Στο τελικό στάδιο ο φωτοπολλαπλασιαστής μετάτρεψε τα φωτόνια σε μετρήσεις ηλεκτρικών σημάτων. Πιο ουσιαστικά, οι μετρήσεις που περιέχονται σε αυτό το σύνολο δεδομένων αντιστοιχούν σε 47 δείγματα με οξεία μυελογενή λευχαιμία (ALL) και 25 δείγματα με οξεία λεμφοβλαστική λευχαιμία (AML)

II. Το επόμενο σύνολο δεδομένων που αφορά τον καρκίνο του μαστού πάρθηκε και αυτό από το Kaggle [57] και έχει δημιουργηθεί από τους Δρ. Wolberg et al. [58] Σε αυτό το έργο συλλέξανε δείγματα από υγρό μάζας μαστού με αναρρόφηση διά λεπτής βελόνης (FNA), στην συνέχεια το πρόγραμμα Xcyt χρησιμοποιήθηκε για την ανάλυση κυτταρολογικών χαρακτηριστικών βάσει ψηφιακής σάρωσης, όπου γίνεται χρήση ένας αλγόριθμος προσαρμογής καμπύλης για τον υπολογισμό 10 χαρακτηριστικών από κάθε κύτταρο του δείγματος. Και τέλος υπολογίστηκε η μέση τιμή, η ακραία τιμή και το τυπικό σφάλμα για κάθε χαρακτηριστικό στην εικόνα και έτσι η κατάληξη ήταν να έχουν 30 μορφολογικά χαρακτηριστικά [58] που προσδιορίζουν 357 δείγματα για τον καλοήθη και 212 δείγματα κακοήθη καρκίνο του μαστού.

III. Το τρίτο σύνολο δεδομένων είναι για την νόσο του Πάρκινσον, πάρθηκε επίσης από το Kaggle [59] και φτιάχτηκε από τη μελέτη των Sakar et al. (2019) [42] Ουσιαστικά αποτελείται από ηχογραφήσεις ομιλίας από 48 υγιείς και 147 πάσχοντες ανθρώπους. Κατά την διαδικασία συλλογής χρησιμοποίησαν μικρόφωνο το οποίο ρύθμισαν στα 44,1 KHz και ηχογράφησαν τη συνεχές φωνή του φωνήεντος «Α». Στην πορεία εφάρμοσαν διάφορους αλγόριθμους σήματος ομιλίας, συμπεριλαμβανομένων των χαρακτηριστικών συχνότητας χρόνου, των συντελεστών με συχνότητα Mel (MFCC), των χαρακτηριστικών που βασίζονται σε μετασχηματισμό Wavelet, των χαρακτηριστικών Vocal Fold και των συντονισμένων μετασχηματισμών παράγοντα Q (Tunable Q-factor Wavelet Transform – TWQT) [42] για την εξαγωγή των 23 χαρακτηριστικών του συνόλου.

IV. Το τελευταίο σύνολο δεδομένων που σχετίζεται με την διαπερατότητα των φαρμάκων στον αιματοεγκεφαλικό φραγμό (BBB), έγινε συλλογή των δεδομένων που χρησιμοποίησαν οι Shaker et al. (2020) [44] από διάφορες βιβλιογραφίες (Adenot and Lahana (2004) [60], Gao, et al. (2017) [61], Martins, et al. (2012) [62], Plisson and Piggott (2019) [63], Singh, et al. (2020) [64], Wang, et al. (2018) [65], Yuan, et al. (2018) [66]) για τα δεδομένα εκπαίδευσης και (Gao, et al. (2017) [61], Plisson and Piggott, (2019) [63], Singh, et al. (2020) [64]) για τα δεδομένα αξιολόγησης. Αναλυτικά, πάρθηκαν smiles ή ονόματα χημικών ενώσεων συγκεκριμένα για 5729 διαπερατές και 2082 μη διαπερατές στο BBB (BBB+/BBB-). Κατόπιν η βάση δεδομένων PubChem link: https://pubchem.ncbi.nlm.nih.gov/ [67], όπου η χρήση της είναι ελεύθερη και παρέχει δεδομένα για χημικές πληροφορίες όπως μικρά μόρια, μεγαλύτερα μόρια π.χ. νουκλεοτίδια, υδατάνθρακες, λιπίδια, πεπτίδια και χημικά τροποποιημένα μακρομόρια. Τα δεδομένα αυτά προέρχονται από διάφορες πηγές όπως κυβερνητικές υπηρεσίες, πωλητές χημικών και εκδότες περιοδικών κ.τ.λ. [66], Στην μελέτη αυτή χρησιμοποιήθηκε για την συμπλήρωση των ελλιπών πληροφοριών (smiles ή ονόματα). Συγκεκριμένα το smiles είναι ένα σύστημα χημικής σημειογραφίας σχεδιασμένο για σύγχρονη επεξεργασία χημικών πληροφοριών όπου αντιπροσωπεύει τη μοριακή δομή με μια κωδικοποίηση σχεδιασμένη από χημικούς για να περιγράψουν ένα μόριο. Επίσης είναι αρκετά ευέλικτο για χρήση σε εφαρμογές υπολογιστών λόγω της δομής του που αποτελείται από μικρού μήκους σειράς χαρακτήρων για γρήγορη επεξεργασία [69]. Τέλος, το πρόγραμμα padel [70] είναι ένα λογισμικό που χρησιμοποιείται για τον υπολογισμό $1^{ης}$, $2^{ης}$ και $3^{ης}$ τάξης μοριακών χαρακτηριστικών αλλά και αποτυπωμάτων. Πιο αναλυτικά το λογισμικό αυτό υπολογίζει 1875 χαρακτηριστικά (1444 χαρακτηριστικά για $1^{ης}$ και $2^{ης}$ τάξης και 431 για χαρακτηριστικά

3ης τάξης) και 12 τύπους δακτυλικών αποτυπομάτων [64]. Στην παρούσα μελέτη χρησιμοποιήθηκε για την μετατροπή των smiles σε μοριακά χαρακτηριστικά 1ης και 2ης τάξης.

2.2 Μεθοδολογία εργασίας

2.2.1 Σύνολα Εκπαίδευσης

Μετά την συλλογή δεδομένων ακολούθησε η διαδικασία προετοιμασίας των συνόλων δεδομένων. Έγινε προεπεξεργασία για κάθε σύνολο ούτως ώστε να διαγραφούν τα λιγότερο χρήσιμα χαρακτηριστικά και δείγματα και να εντοπιστούν αυτά τα οποία θα έχουν καλύτερη επίδραση στον διαχωρισμό των κλάσεων. Αναλυτικά:

I. Σύνολο εκπαίδευσης ALL – AML: Αρχικά ορίστηκε η ετικέτα για κάθε ασθενή και λόγω των λίγων δειγμάτων που υπήρχαν έγινε συγχώνευση του συνόλου εκπαίδευσης και του συνόλου αξιολόγησης. Στην συνέχεια εφαρμόστηκε στατιστική ανάλυση [71]. Πιο λεπτομερώς, εξετάστηκαν αρχικά οι κατανομές των χαρακτηριστικών για κάθε κλάση (κανονική κατανομή ή μη κανονική κατανομή) με την δοκιμασία Shapiro wilk [72]. Με την δοκιμασία Shapiro wilk ουσιαστικά, ελέγχεται εάν οι τιμές του κάθε χαρακτηριστικού κατανέμονται κανονικά με βάση τον μέσο όρο ή αν έχουν σημαντική διαφορά μεταξύ τους με ένα κατώφλι που ονομάζεται άλφα και συνήθως ορίζεται στο 0.05. Αν η τιμή του p για κάθε χαρακτηριστικό είναι μεγαλύτερη του άλφα τότε ισχύει ότι η κατανομή είναι κανονική, αντίθετα αν η τιμή αυτή είναι μικρότερη του άλφα τότε η κατανομή είναι μη κανονική. Στην συνέχεια έγινε συγχώνευση των δύο κλάσεων και χρησιμοποιήθηκε η παραμετρική δοκιμασία t για όσα χαρακτηριστικά είχαν τιμές κανονικής κατανομής και στις 2 κλάσεις, ενώ στις περιπτώσεις που σε μια από τις δύο κλάσεις οι τιμές των χαρακτηριστικών είχαν μη κανονική κατανομή ή ακόμα σε περιπτώσεις που και οι δύο κλάσεις να είχαν μη κανονική κατανομή χρησιμοποιήθηκε η μη παραμετρική δοκιμασία Wilcoxon. Συγκεκριμένα η δοκιμασία t έγινε για την σύγκριση των κατανομών των δύο κλάσεων και ως μηδενική υπόθεση H_0 θεωρείται ότι έχουν παρόμοια κατανομή ενώ η εναλλακτική υπόθεση H_1 δείχνει ότι έχουν στατιστικά σημαντική διαφορά [73]. Για την δοκιμασία Wilcoxon η μηδενική υπόθεση δείχνει ότι η διαφορά κατά ζεύγους τιμών των κλάσεων έχουν κατανομή πιθανότητας περίπου στο μηδέν άρα ότι δεν έχουν σημαντική διαφορά, αντίθετα η εναλλακτική υπόθεση δείχνει ότι έχουν κατανομή πιθανότητας μεγαλύτερη του μηδενός και έτσι έχουν στατιστικά σημαντική διαφορά [74]. Αφού εντοπίστηκαν λοιπόν όλα τα χαρακτηριστικά που είχαν στατιστικά σημαντική διαφορά, μετέπειτα χρησιμοποιήθηκε ο συντελεστής συσχέτισης Pearson για την μέτρηση της ομοιότητας των χαρακτηριστικών μεταξύ τους. Ο συντελεστής συσχέτισης παίρνει τιμές από -1 έως 1, όσο πιο κοντά στο 1 είναι η τιμή του τόσο πιο ψηλή συσχέτιση υποδηλώνει, ενώ αντίστροφα όσο πιο κοντά στο -1 τόσο πιο λίγη συσχέτιση [75]. Στην δική μας περίπτωση επειδή μας ενδιαφέρουν τα χαρακτηριστικά με καλύτερο διαχωρισμό, λάβαμε τα χαρακτηριστικά τα οποία είχαν χαμηλή συσχέτιση (<0.75). Στην πορεία ακολούθησε χρήση της αναδρομικής εξάλειψης χαρακτηριστικών (Recursive Feature Elimination - RFE), όπου χρησιμοποιεί την ικανότητα του τυχαίου δάσους (Random Forest - RF) με επαναλήψεις για την εύρεση των σημαντικών

χαρακτηριστικών και είναι αρκετά χρήσιμη για προβλήματα λίγων δειγμάτων και πολλών χαρακτηριστικών [76]. Ο RFE κατ' ουσία ψάχνει με μέγιστο όριο το 1/3 της μικρότερης κλάσης και επιδιώκει να βελτιώσει την απόδοση της γενίκευσης, αφαιρώντας τα λιγότερο σημαντικά χαρακτηριστικά των οποίων η διαγραφή θα έχει την μικρότερη επίδραση στον διαχωρισμό των κλάσεων [76].

II. Σύνολο εκπαίδευσης καλοήθη – κακοήθη όγκου μαστού: Για αρχή έγινε αφαίρεση των χαρακτηριστικών που είχαν ελλιπείς τιμές (NAs) και επίσης αφαιρέθηκαν τα ids. Μετέπειτα όπως και στο προηγούμενο σύνολο έγινε χρήση της RFE για την εξαγωγή των πιο σημαντικών χαρακτηριστικών ψάχνοντας με μέγιστο όριο το 1/3 της μικρότερης κλάσης.

III. Σύνολο εκπαίδευσης του Πάρκινσον: Έγιναν οι ίδιες διαδικασίες με το προηγούμενο σύνολο. Δηλαδή, αφαιρέθηκαν χαρακτηριστικά με ελλιπείς τιμές και έγινε χρήση της RFE με μέγιστο όριο το 1/3 της μικρότερης κλάσης.

IV. Σύνολο εκπαίδευσης της διαπερατότητας των φαρμάκων στο BBB: Έγινε αρχικά συγχώνευση των δεδομένων που πάρθηκαν από όλες τις πηγές και παρατηρήθηκε ότι για κάποιες χημικές ενώσεις έλειπαν τα smiles και για άλλες τα ονόματα. Οπότε μέσω του πακέτου webchem πρωτίστως, έγινε λήψη των IDs από την βάση δεδομένων PubChem και ακολούθως βάσει των IDs πάρθηκαν όλα τα smiles που έλειπαν και αρκετά ονόματα φαρμάκων. Αργότερα χρησιμοποιήθηκε το πρόγραμμα padel για την εξαγωγή των χαρακτηριστικών ($1^{ης}$ και $2^{ης}$ τάξης μοριακά χαρακτηριστικά) και στην συνέχεια λόγω του ότι πολλές χημικές ενώσεις είναι ίδιες με διαφορετικό όνομα, εντοπίστηκαν οι κοινές από τα χαρακτηριστικά τους και έγινε η διαγραφή τους βάσει κριτηρίων. Αν οι επαναλαμβανόμενες χημικές ενώσεις είχαν την ίδια ετικέτα τότε έφευγαν όλες και παρέμενε ένα. Σε περιπτώσεις που οι ετικέτες ήταν διαφορετικές έγινε αφαίρεση βάσει της πλειοψηφίας και αν ισούνταν οι ετικέτες τότε διαγράφονταν όλα. Στην πορεία έγινε αφαίρεση των χημικών ενώσεων με ελλιπείς πληροφορίες αλλά και χαρακτηριστικών που εμπεριείχαν τα IDs και τα ονόματα των χημικών ενώσεων. Λόγω όμως της χαμηλής μνήμης RAM χρησιμοποιήθηκε το 1/3 των αρχικών δεδομένων, συγκεκριμένα τα σύνολα δεδομένα των (Gao, et al. (2017) [61], Plisson and Piggott, (2019) [63], Singh, et al. (2020) [64]). Εν συνεχεία αφαιρέθηκαν τα χαρακτηριστικά με χαμηλή διακύμανση [37] και εφαρμόστηκε η RFE με μέγιστο όριο το 1/3 της μικρότερης κλάσης για την αποβολή των λιγότερο σημαντικών χαρακτηριστικών.

2.2.2 Σύνολα Επικύρωσης

Αφού λοιπόν έγινε η προετοιμασία των δεδομένων εκπαίδευσης, ακολούθησαν τα σύνολα επικύρωσης για την αξιολόγηση διάφορων αλγορίθμων στα υπάρχοντα δεδομένα:

I. Σύνολο επικύρωσης ALL – AML: Επιλέχθηκε να χρησιμοποιηθεί η μέθοδος επαναδειγματοληψίας bootstrap λόγω των λίγων δειγμάτων που υπήρχαν και εφαρμόστηκε για κάθε συνδυασμό από τα χαρακτηριστικά που επέλεξε η RFE, για την αξιολόγηση πέντε αλγορίθμων:
    1) Η Γραμμική Διαχωριστική Ανάλυση (Linear Discriminant Analysis - LDA), όπου η βασική του λειτουργία είναι να φτιάξει ένα μικρότερο χώρο διαστάσεων από τα δεδομένα που του δίνονται, με τα πιο διαχωρίσιμα δεδομένα [77].

2) Η μηχανή διανυσμάτων υποστήριξης (SVM)
3) Το τυχαίο δάσος (RF), ο οποίος δουλεύει με τα δέντρα αποφάσεων και έχει πλεονεκτήματα αντιμετώπισης μικρού μεγέθους δειγμάτων και μεγάλων διαστάσεων χαρακτηριστικών [78].
4) Τα νευρωνικά δίκτυα (ANN) και αποτελούνται συγκεκριμένα από ένα κρυφό επίπεδο [79].
5) Ο απλοϊκός μπεϋζιανός (NB) που δουλεύει με την ανεξαρτησία υπό όρους, αποφεύγοντας την άμεση εκτίμηση κάθε σχετικής πιθανότητας που είτε λείπουν από τα δεδομένα, είτε δεν έχουν επαρκείς τιμές [80].

Αφού λοιπόν ορίστηκαν οι αλγόριθμοι που θα αξιολογηθούν, κατόπιν αυξήθηκε η χαμηλή AML ούτως ώστε να γίνει ισόποση με την ALL με επαναλήψεις τυχαίων δειγμάτων της. Έτσι αφού υπήρχε ίσος αριθμός δειγμάτων σε κάθε κλάση εφαρμόστηκε η επαναδειγματοληψία bootstrap όλων των δεδομένων μαζί για να δημιουργηθεί το νέο σύνολο δεδομένων και μετέπειτα πραγματοποιήθηκε κανονικοποίηση δεδομένων z-score [81] . Αργότερα περίπου το 70% των συνολικών δειγμάτων χρησιμοποιήθηκε για εκπαίδευση και το υπόλοιπο 30% για αξιολόγηση. Για να ληφθούν πιο αξιόπιστα αποτελέσματα η διαδικασία αυτή επαναλήφθηκε 5 φορές και οι μέσοι όροι με την τυπική απόκλιση των επαναλήψεων αυτών, έγιναν τα κριτήρια για την επιλογή του καλύτερου μοντέλου.

II. Σύνολο επικύρωσης Καλοήθη – Κακοήθη όγκου μαστού: Στην περίπτωση αυτή επειδή είχαμε περισσότερα δείγματα από πριν έγινε χρήση της επαναλαμβανόμενης Κ φορές αναδιπλωμένης διασταυρωμένης επικύρωσης για τους συνδυασμούς όλων των χαρακτηριστικών της RFE, με τους 5 εξής αλγορίθμους:
   1) Η μηχανή ενίσχυσης κλίσης (GBM), όπου και αυτός βασίζεται στα δέντρα αποφάσεων. Η βασική του ιδέα η κατασκευή αδύναμων διαδοχικών μοντέλων, με κάθε μοντέλο να μαθαίνει και να βελτιώνεται από το προηγούμενο βάσει του σφάλματος ολόκληρου του συνόλου που έχει μάθει μέχρι στιγμής [82].
   2) Η μηχανή υποστήριξης διανυσμάτων (SVM)
   3) Το τυχαίο δάσους (RF)
   4) Η ανάλυση τετραγωνικών διακρίσεων (Quadratic Discriminant Analysis - QDA), είναι παρόμοιος με τον LDA αν και προσαρμόζεται καλύτερα στα δεδομένα συνήθως, λόγω της περισσότερης ευελιξίας που επιτρέπει στον πίνακα συνδιακύμανσης [83].
   5) Κ – πλησιέστερων γειτόνων (K-NN)

   Στην συνέχεια, για να υπάρχει ισορροπία ανάμεσα στις 2 κλάσεις αφαιρέθηκαν τυχαία από την μεγαλύτερη κλάση δείγματα ώστε να γίνουν ισόποσες και πραγματοποιήθηκε κανονικοποίηση ελαχίστου – μεγίστου [75]. Έτσι ο αριθμός Κ πήρε τον αριθμό 5, οπότε χωρίστηκαν τα δείγματα σε 5 ισόποσες ομάδες με μια ομάδα κάθε φορά να χρησιμοποιείται για αξιολόγηση και οι υπόλοιπες για εκπαίδευση. Τέλος, η διαδικασία που περιεγράφηκε έγινε άλλες 5 φορές και όπως πριν βρέθηκε ο μέσος όρος και η τυπική απόκλιση για κάθε μοντέλο σε κάθε συνδυασμό χαρακτηριστικών για την επιλογή του καλύτερου.

III. Σύνολο επικύρωσης διάγνωσης Πάρκινον: Εφαρμόστηκε η μέθοδος της επαναδειγματοληψίας bootstrap σε κάθε συνδυασμό από τα χαρακτηριστικά της

RFE για την σύγκριση 5 αλγόριθμων που χρησιμοποιήθηκαν και στα προηγούμενα σύνολα δεδομένων:

1) Η μηχανή υποστήριξης διανυσμάτων (SVM)
2) Η μηχανή ενίσχυσης κλίσης (GBM)
3) Το τυχαίο δάσους (RF)
4) Τα νευρωνικά δίκτυα (ANN)
5) Κ – πλησιέστερων γειτόνων (K-NN)

Συγκεκριμένα, έγιναν αρχικά τυχαίες επαναλήψεις δειγμάτων για την δημιουργία του καινούργιου συνόλου, στην συνέχεια πραγματοποιήθηκε η κανονικοποίηση z-score [81] και μετέπειτα έγινε διαχωρισμός των δειγμάτων σε εκπαίδευση και αξιολόγηση 67% – 33% αντίστοιχα. Αυτή η διαδικασία επαναλήφθηκε 10 φορές και βρέθηκε ο μέσος όρος και η τυπική απόκλιση όπως και προηγουμένως.

IV. Σύνολο επικύρωσης διαπερατότητας BBB: Έγινε και πάλι χρήση της μεθόδου επαναδειγματοληψίας bootstrap, για τον συνδυασμό των σημαντικών χαρακτηριστικών που εντόπισε η RFE, για τους ίδιους αλγόριθμους και την ίδια διαδικασία που χρησιμοποιήθηκε στο σύνολο επικύρωσης του Πάρκινσον. Με τις διαφορές ότι ο διαχωρισμός των δειγμάτων ήταν 70% για εκπαίδευση - 30% αξιολόγηση και οι επαναλήψεις της διαδικασίας ήταν 5.

2.2.3 Σύνολα Αξιολόγησης

Σύνολο αξιολόγησης χρησιμοποιήθηκε μόνο για το σύνολο BBB λόγω των ικανοποιητικών δειγμάτων που υπήρχαν στην περίπτωση αυτή. Για την προετοιμασία του συνόλου αξιολόγησης διαγράφηκαν τα δείγματα που περιείχαν ελλιπείς τιμές και στην συνέχεια δοκιμάστηκε το καλύτερο μοντέλο από το σύνολο επικύρωσης για την αξιολόγηση του.

2.2.4 Μοντέλο Βαθιάς Μάθησης

Προσπάθεια κατασκευής μοντέλου Βαθιά Νευρωνικών Δικτύων (DNN) έγινε και πάλι για το σύνολο δεδομένων BBB και αυτό το επέτρεψε ο ικανοποιητικός αριθμός δειγμάτων που υπήρχαν. Για το σύνολο εκπαίδευσης ακολούθησαν οι ίδιες διαδικασίες με την προηγούμενη ενότητα με την διαφορά ότι η κανονικοποίηση z-score [81] πραγματοποιήθηκε πριν την διαδικασία του συνόλου επικύρωσης.

Αρχικά η αρχιτεκτονική του δικτύου χτίστηκε με ένα κρυφό επίπεδο για την επικύρωση του έγινε διαχωρισμός με το 70% των δεδομένων να εκπαιδεύεται και το υπόλοιπο 30% να αξιολογείται. Αφού λοιπόν τοποθετήθηκαν το επίπεδο εισόδου που ήταν ο αριθμός διαθέσιμων χαρακτηριστικών του συνόλου εκπαίδευσης, το επίπεδο εξόδου που το καθορίζει ο αριθμός των διαθέσιμων κλάσεων και 1 κρυφό επίπεδο με χαμηλό αριθμό κόμβων λήφθηκαν κάποιες παρατηρήσεις σχετικά με την συμπεριφορά του δικτύου στα δεδομένα. Μετέπειτα έγινε σταδιακή αύξηση των κόμβων και των κρυφών επιπέδων και κάθε φορά που άλλαζε κάποια από τους παραμέτρους γινόταν επικύρωση του μοντέλου για την απεικόνιση της απόδοσης του. Έτσι, παρατηρήθηκε ο αριθμός κρυφών επιπέδων και κόμβων, που κυμαινότανε η απόδοση τους καλύτερα να ήταν μέχρι 4 και μέχρι 2048 αντίστοιχα. Επίσης δοκιμάστηκαν οι συμπεριφορές των αλγορίθμων στοχαστικής κατάβασης δυναμικού (SDG) με τιμές ρυθμού μάθησης από 0,000001 έως και 0,1. Στην

πορεία δοκιμάστηκαν συνδυασμοί από 1 μέχρι 4 κρυφά επίπεδα και από 2 μέχρι 4096 κόμβους για κάθε κρυφό επίπεδο. Όσον αφορά την διάρκεια εκπαίδευσης καθορίστηκε γύρω στις 150 εποχές αρχικά και αφού φάνηκε περίπου στις 30 εποχές να μην συγκλίνουν περαιτέρω καθορίστηκαν στις 30. Στην συνέχεια τοποθετήθηκαν οι παράμετροι βελτίωσης dropout και κανονικοποίηση βαρών L2 στο επίπεδο εισόδου και στο πρώτο κρυφό επίπεδο. Στην τεχνική dropout συγκεκριμένα δοκιμάστηκαν τιμές από 0.1 έως 0.5 και στην κανονικοποίηση βαρών L2 από 0.000001 έως 0.1. Έτσι λοιπόν με κριτήριο την απόδοση του κάθε συνδυασμού στα δεδομένα, εντοπίστηκε το μοντέλο με τον καλύτερο συνδυασμό κόμβων και παραμέτρων.

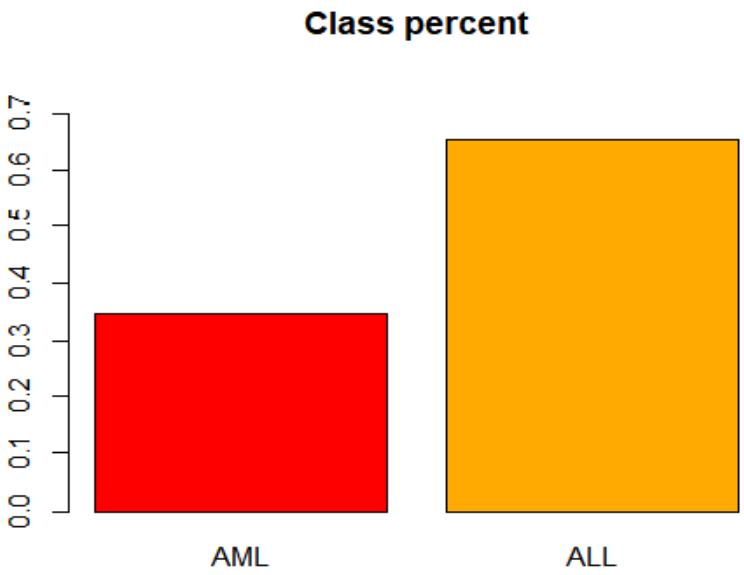
Τέλος, χρησιμοποιώντας τα ίδια δεδομένα έγινε επανεκπαίδευση του μοντέλου (fine tuning) με τα ίδια επίπεδα και κόμβους, όπου δοκιμάστηκαν διάφοροι συνδυασμοί παραμέτρων όπως και πριν. Επιπροσθέτως χρησιμοποιήθηκε η τεχνική "παγώματος – ξεπαγώματος" (freeze – unfreeze), χωρίς να αφήνει όλα τα βάρη του μοντέλου να ανανεωθούν και ξεκινώντας την επανεκπαίδευση τους από τα τελευταία επίπεδα σταδιακά μέχρι το επίπεδο εισόδου. Με αυτό τον τρόπο λοιπόν έγινε μια μικρή αύξηση στις επιδόσεις του μοντέλου αυτού.

3. Αποτελέσματα μελέτης

Στην ενότητα αυτή θα παρουσιαστούν τα αποτελέσματα των συνόλων δεδομένων που χρησιμοποιήθηκαν στην μεθοδολογία εργασίας. Συγκεκριμένα θα αξιολογηθούν οι μέθοδοι για κάθε πεδίο με μετρήσεις που χρησιμοποιούνται στην ΜΜ.

3.1 Σύνολο δεδομένων ALL - AML

Το πρώτο σύνολο δεδομένων σύνολο δεδομένων αποτελείτο από 7129 χαρακτηριστικά και 73 δείγματα. Συγκεκριμένα 47 ασθενείς με ALL και 25 ασθενείς με AML. Η αναλογία ποσοστού φαίνεται στην **εικόνα 3.1**



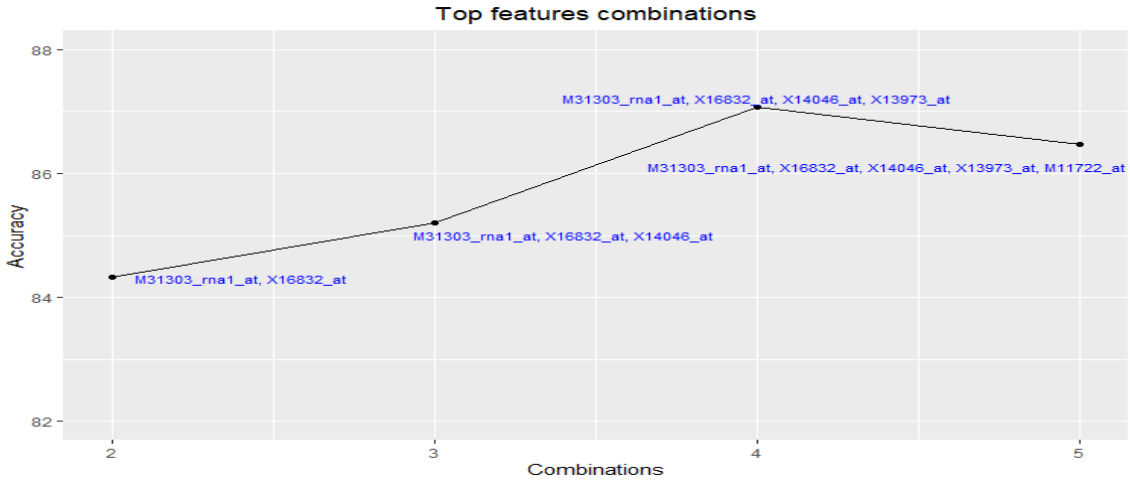
Εικόνα 3.1: Ποσοστιαία αναλογία κάθε κλάσης (AML – ALL)

Τα χαρακτηριστικά που βρέθηκαν να έχουν στατιστικά σημαντική διαφορά ήταν 2091 και από αυτά χαμηλή συσχέτιση είχαν τα 1542. Στην συνέχεια με την εφαρμογή της RFE στα χαμηλά συσχετισμένα επίλεξε τα 5 καλύτερα χαρακτηριστικά: **«M31303_rna1_at», «X16832_at», «X14046_at», «X13973_at», «M11722_at»**. Βάσει λοιπόν αυτών των χαρακτηριστικών χρησιμοποιήθηκε η μέθοδος της επαναδειγματοληψίας bootstrap για να λάβουμε την καλύτερη απόδοση του μοντέλου μαζί με τους συνδυασμούς όπως φαίνεται στον Πίνακα 7, όπου κάθε τιμή αντιπροσωπεύει την μέση τιμή των επαναλήψεων και την τυπική απόκλιση.

Πίνακας 3.1: Καλύτερο αποτέλεσμα για κάθε μοντέλο (AML – ALL)

Μοντέλο	Συνδυασμός χαρακτηριστικών	Ακρίβεια	Ευαισθησία	Ειδικότητα
LDA*	«M31303_rna1_at», «X16832_at», «M11722_at»	84.00 ±7	91.67 ±15	78.89 ±10
SVM*2	«M31303_rna1_at», «X16832_at», «X14046_at», «X13973_at»	89.67 ±4	94.17 ± 8	86.67 ± 6
RF*3	«M31303_rna1_at», «X16832_at», «X14046_at», «X13973_at»	91.33 ±4	96.67 ± 4	87.78 ± 7
NNET*4	«M31303_rna1_at», «X16832_at», «X14046_at»	87.67±9	95.00±6	80.33±15
NB*5	«M31303_rna1_at», «X16832_at», «X14046_at»	85.67±6	91.67±18	81.67±7

Πιο κάτω στο **γράφημα 3.2** παρουσιάζεται ο ψηλότερος μέσος όρος μοντέλων για τους καλύτερους συνδυασμούς σε κάθε αριθμό συνδυασμών και στον Πίνακα 8 οι αποδόσεις του κάθε μοντέλου ξεχωριστά στον καλύτερο συνδυασμό.

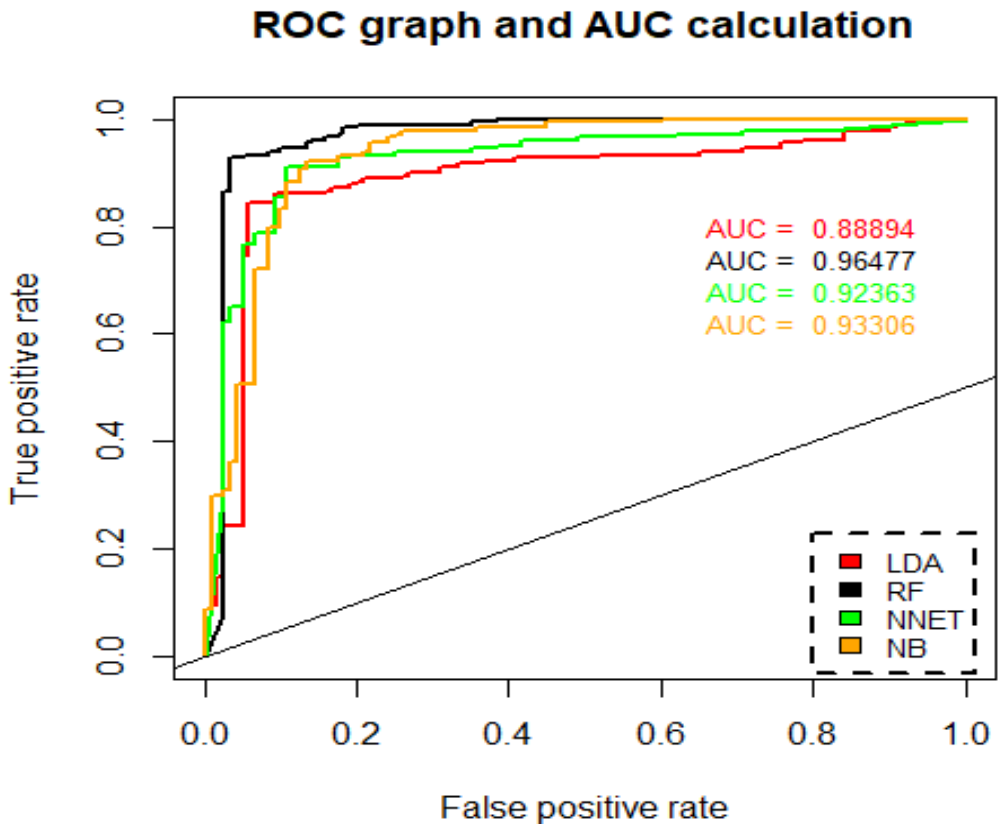


Εικόνα 3.2: Μέσος όρος ακρίβειας μοντέλων για τους καλύτερους συνδυασμούς (ALL – AML)

Πίνακας 3.2: Αποτελέσματα κάθε μοντέλου για τον καλύτερο συνδυασμό: «M31303_rna1_at», «X14046_at», «X16832_at», «X13973_at» (ALL – AML)

	SVM	LDA	RF	NNET	NB
Ακρίβεια	89.67 ± 4	82.33 ± 6	91.33 ± 4	87.67 ± 4	84.33 ± 5
Ευαισθησία	94.17 ± 8	94.17 ± 9	96.67 ± 4	89.17 ± 10	91.67 ± 9
Ειδικότητα	86.67 ± 6	74.44 ± 7	87.78 ± 7	86.67 ± 5	79.44 ± 6
Αληθώς θετικά	11 ± 1	11 ± 1	12 ± 1	11 ± 1	11 ± 1
Ψευδώς αρνητικά	2 ± 1	5 ± 1	2 ± 1	2 ± 1	4 ± 1
Ψευδώς θετικά	1 ± 1	1 ± 1	0 ± 1	1 ± 1	1 ± 1
Αληθώς αρνητικά	16 ± 1	13 ± 1	16 ± 1	16 ± 1	14 ± 1
Συντελεστής kappa	78.98 ± 8	65.09 ± 11	82.48 ± 8	74.61 ± 9	68.55 ± 10
Θετική προγνωστική αξία	82.99 ± 6	71.24 ± 6	84.68 ± 8	82.01 ± 6	75.19 ± 6
Αρνητική προγνωστική αξία	96.03 ± 5	95.41 ± 7	97.64 ± 3	92.84 ± 7	93.91 ± 6
Σταθμισμένη ορθότητα	90.42 ± 4	84.31 ± 6	92.22 ± 4	87.92 ± 5	85.56 ± 5
F1 score	87.91 ± 5	80.98 ± 6	90.07 ± 5	85.07 ± 5	82.36 ± 5
Ποσοστό ψευδώς αρνητικών	5.83 ± 8	5.83 ± 9	3.33 ± 4	10.83 ± 10	8.33 ± 9
Ποσοστό αληθώς αρνητικών	13.33 ± 6	25.56 ± 7	12.22 ± 7	13.33 ± 5	20.56 ± 6
Ψευδές ποσοστό ανακάλυψης	17.01 ± 6	28.76 ± 6	15.32 ± 8	17.99 ± 6	24.81 ± 6
Ψευδές ποσοστό παράλειψης	3.97 ± 5	4.59 ± 7	2.36 ± 3	7.16 ± 7	6.09 ± 6
Συντελεστής συσχέτισης Matthews	79.91 ± 8	67.62 ± 11	83.37 ± 8	75.33 ± 9	70.1 ± 10
Informedness	80.83 ± 8	68.61 ± 12	84.44 ± 8	75.83 ± 10	71.11 ± 10
Markedness	79.02 ± 7	66.65 ± 11	82.32 ± 8	74.85 ± 9	69.1 ± 10

Οι καμπύλες ROC είναι ένας άλλος τρόπος που χρησιμοποιείται για την αξιολόγηση των μοντέλων. Παράγεται υπολογίζοντας και σχεδιάζοντας το ποσοστό αληθώς θετικών έναντι του ποσοστού του ποσοστού ψευδώς θετικών. Επιπροσθέτως, το AUC αφορά την περιοχή κάτω από την καμπύλη και όσο μεγάλη είναι τόσο καλύτερη είναι η αξιολόγηση του μοντέλου.



Εικόνα 3.3: Ανάλυση ROC και AUC (ALL – AML)

Όπως προκύπτει από τον **πίνακα 3.1 και 3.2** και από τα **γραφήματα 3.2 και 3.3** ο RF είναι το καλύτερο μοντέλο με 91.33% ακρίβεια, 96.67% ευσαισθησία και 87.78% ειδικότητα με τα εξής χαρακτηριστικά: **«M31303_rna1_at», «X16832_at», «X14046_at», «X13973_at».** Τα αποτελέσματα του καλύτερου μοντέλου (RF) φαίνονται αναλυτικά στον πίνακα 9 και στον **πίνακα 10** όπου είναι ένας πίνακας αληθείας.

Πίνακας 3.3: Απόδοση τυχαίου δάσους (RF) (ALL – AML)

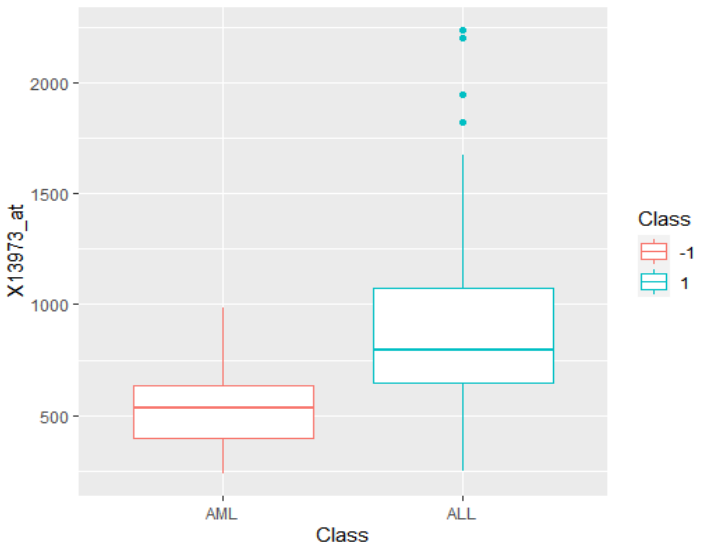
	RF
Ακρίβεια	91.33 ± 4
Ευαισθησία	96.67 ± 4
Ειδικότητα	87.78 ± 7
Αληθώς θετικά	12 ± 1
Ψευδώς αρνητικά	2 ± 1
Ψευδώς θετικά	0 ± 1
Αληθώς αρνητικά	16 ± 1
Συντελεστής kappa	82.48 ± 8

Θετική προγνωστική αξία	84.68 ± 8
Αρνητική προγνωστική αξία	97.64 ± 3
Σταθμισμένη ορθότητα	92.22 ± 4
F1 score	90.07 ± 5
Ποσοστό ψευδώς αρνητικών	3.33 ± 4
Ποσοστό αληθώς αρνητικών	12.22 ± 7
Ψευδές ποσοστό ανακάλυψης	15.32 ± 8
Ψευδές ποσοστό παράλειψης	2.36 ± 3
Συντελεστής συσχέτισης Matthews	83.37 ± 8
Informedness	84.44 ± 8
Markedness	82.32 ± 8

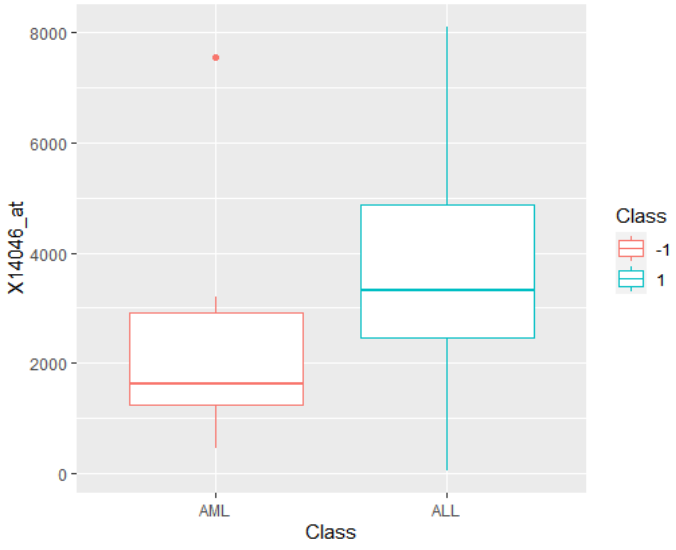
Πίνακας 3.4: Πίνακας αληθείας καλύτερου μοντέλου (RF) (ALL – AML)

		Πραγματική		
	N=30	AML	ALL	Σύνολο
Πρόβλεψη	AML	**12**	**2**	**14**
	ALL	**0**	**16**	**16**

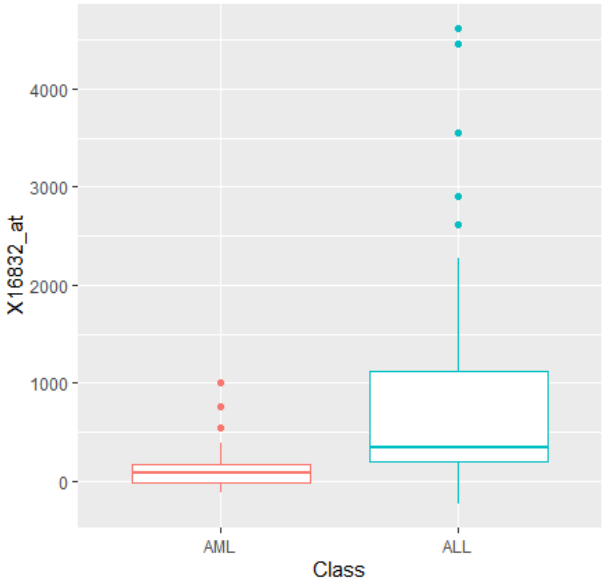
Τέλος, το θηκόγραμμα για κάθε χαρακτηριστικό (ολιγονουκλεοτίδιο) από τα 4 που εντοπίστηκαν.



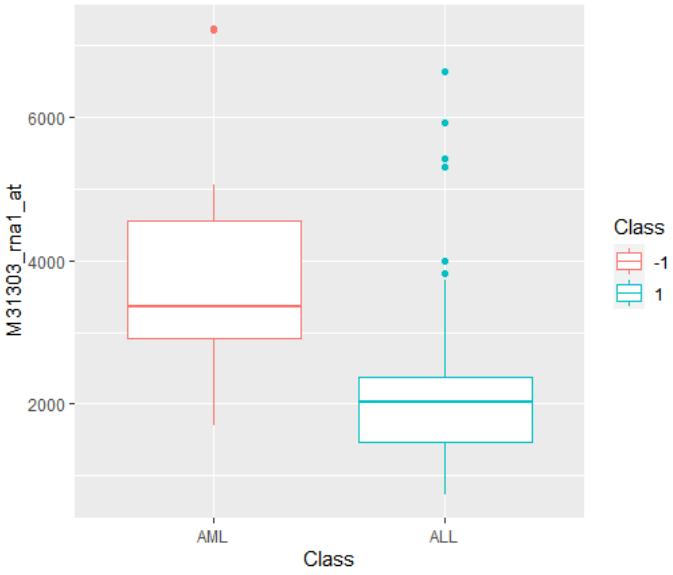
Εικόνα 3.4: Θηκόγραμμα για «X13973_at»



Εικόνα 3.5: Θηκόγραμμα για «X14046_at»



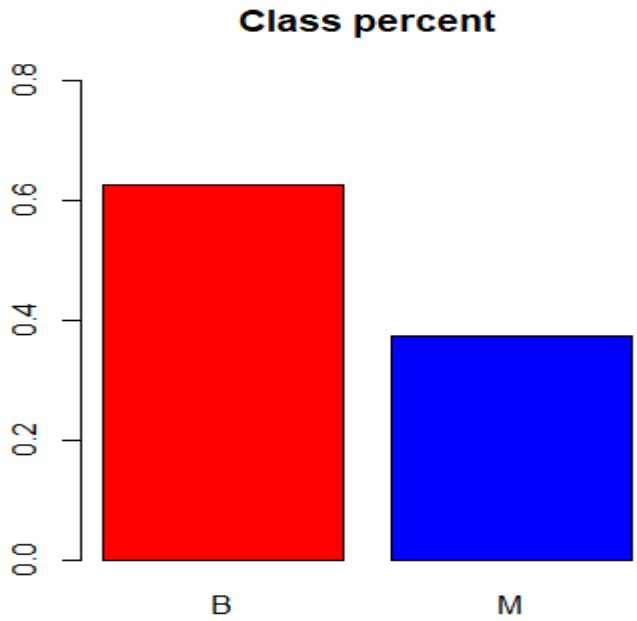
Εικόνα 3.6: Θηκόγραμμα για «X16832_at»



Εικόνα 3.7: Θηκόγραμμα για «M31303_ma1_at»

3.2 Σύνολο δεδομένων καλοήθη – κακοήθη όγκου του μαστού

Αρχικά το σύνολο δεδομένων αυτό αποτελείτο από 30 χαρακτηριστικά και 569 δείγματα. Αναλυτικά, υπήρχαν 357 δείγματα για καλοήθη και 212 για κακοήθη.



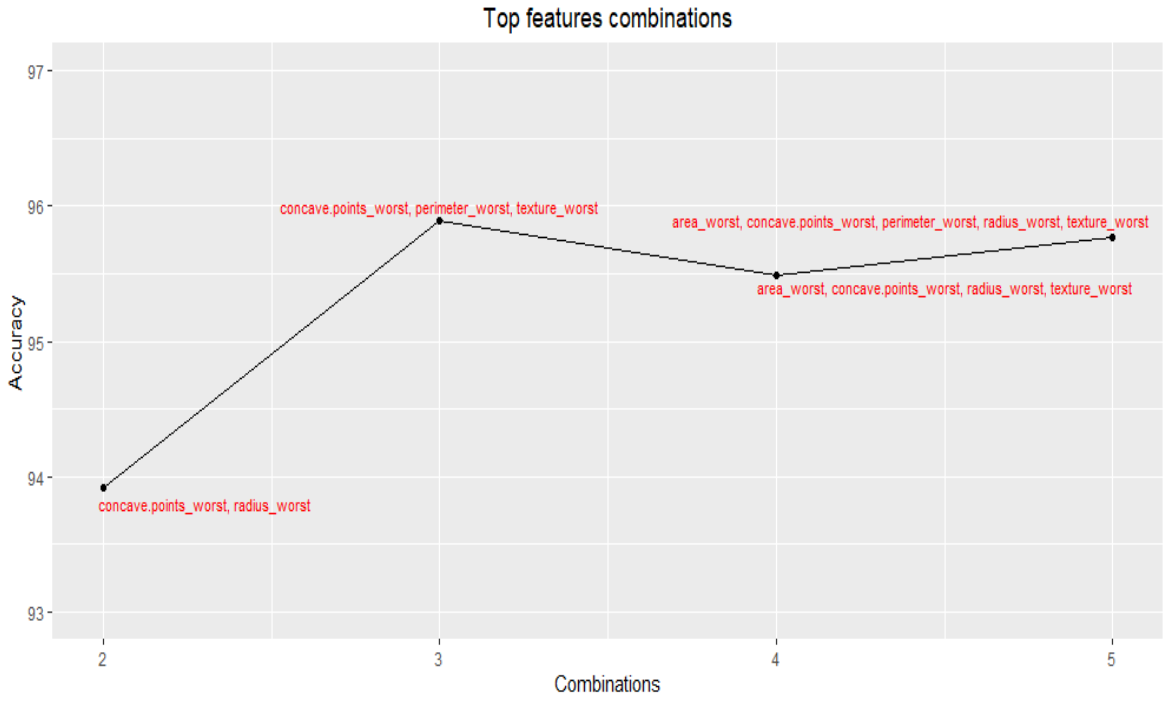
Εικόνα 3.8: Ποσοστιαία αναλογία κάθε κλάσης (Β: καλοήθη – Μ: κακοήθη)

Έγινε λοιπόν εφαρμογή της RFE επιλέγοντας τα 5 καλύτερα χαρακτηριστικά → **«area_worst», «concave.points_worst», «perimeter_worst», «radius_worst», «texture_worst»** και έπειτα με τη χρήση της επαναλαμβανόμενης K – φορές διασταυρωμένης επικύρωσης λήφθηκαν τα πιο ψηλά αποτελέσματα κάθε μοντέλου για την επιλογή του καλυτέρου.

Πίνακας 3.5: Καλύτερο αποτέλεσμα για κάθε μοντέλο (B – M)

Μοντέλο	Συνδυασμός χαρακτηριστικών	Ακρίβεια	Ευαισθησία	Ειδικότητα
GBM	**«concave.points_worst» , «perimeter_worst», «texture_worst»**	**96.70 ±1**	**96.23 ±2**	**95.29±2**
SVM	**«concave.points_worst» , «perimeter_worst», «texture_worst»**	**96.45±1**	**97.16 ± 2**	**95.75 ± 3**
RF	**«area_worst», «concave.points_worst», «perimeter_worst», «radius_worst», «texture_worst»**	**95.85 ±2**	**95.94 ± 3**	**95.75± 3**
QDA	**«area_worst», «concave.points_worst», «texture_worst»**	**96.13±2**	**96.70±3**	**95.55±3**
K-NN	**«concave.points_worst», «radius_worst», «texture_worst»**	**96.27±1**	**97.17±2**	**95.38±3**

Στο **γράφημα 3.9** παρουσιάζεται ο ψηλότερος μέσος όρος μοντέλων για τους καλύτερους συνδυασμούς σε κάθε αριθμό συνδυασμών και στον Πίνακα 12 οι αποδόσεις του κάθε μοντέλου ξεχωριστά στον καλύτερο συνδυασμό.

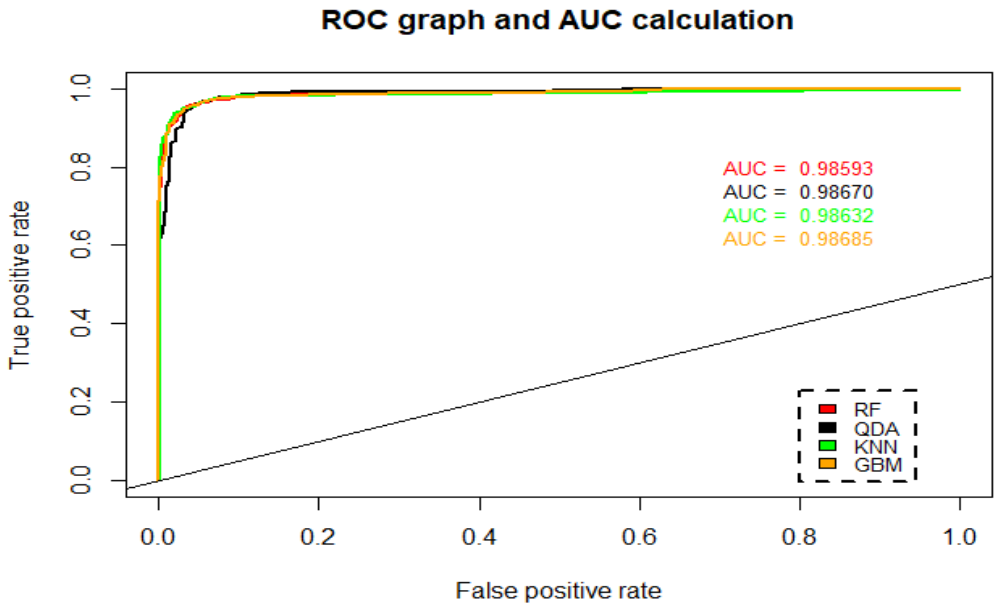


Εικόνα 3.9: Μέσος όρος ακρίβειας μοντέλων για τους καλύτερους συνδυασμούς (B – M)

Πίνακας 3.6: Αποτελέσματα κάθε μοντέλου για τον καλύτερο συνδυασμό: «concave.points_worst» ,«perimeter_worst», «texture_worst» (B – M)

	SVM	GBM	RF	QDA	K-NN
Ακρίβεια	96.45 ± 1	96.7 ± 1	95.76 ± 1	94.8 ± 2	95.75 ± 1
Ευαισθησία	97.17 ± 2	96.71 ± 2	96.23 ± 2	94.81 ± 1	96.22 ± 2
Ειδικότητα	95.75 ± 3	96.69 ± 2	95.29 ± 2	94.8 ± 4	95.28 ± 3
Αληθώς θετικά	41 ± 1	41 ± 1	41 ± 1	40 ± 1	41 ± 1
Ψευδώς αρνητικά	2 ± 1	1 ± 1	2 ± 1	2 ± 1	2 ± 1
Ψευδώς θετικά	1 ± 1	1 ± 1	2 ± 1	2 ± 0	2 ± 1
Αληθώς αρνητικά	41 ± 1	41 ± 1	40 ± 1	40 ± 2	40 ± 1
Συντελεστής kappa	92.91 ± 3	93.4 ± 3	91.52 ± 2	89.61 ± 3	91.5 ± 2
Θετική προγνωστική αξία	95.89 ± 3	96.77 ± 2	95.39 ± 2	94.93 ± 3	95.44 ± 3
Αρνητική προγνωστική αξία	97.16 ± 2	96.77 ± 2	96.24 ± 2	94.82 ± 1	96.24 ± 2
Σταθμισμένη ορθότητα	96.46 ± 1	96.7 ± 1	95.76 ± 1	94.8 ± 2	95.75 ± 1
F1 score	96.49 ± 1	96.7 ± 1	95.78 ± 1	94.83 ± 2	95.78 ± 1
Ποσοστό ψευδώς αρνητικών	2.83 ± 2	3.29 ± 2	3.77 ± 2	5.19 ± 1	3.78 ± 2
Ποσοστό αληθώς αρνητικών	4.25 ± 3	3.31 ± 2	4.71 ± 2	5.2 ± 4	4.72 ± 3
Ψευδές ποσοστό ανακάλυψης	4.11 ± 3	3.23 ± 2	4.61 ± 2	5.07 ± 3	4.56 ± 3
Ψευδές ποσοστό παράλειψης	2.84 ± 2	3.23 ± 2	3.76 ± 2	5.18 ± 1	3.76 ± 2
Συντελεστής συσχέτισης Matthews	92.98 ± 3	93.47 ± 3	91.58 ± 2	89.67 ± 3	91.59 ± 2
Informedness	92.91 ± 3	93.4 ± 3	91.53 ± 2	89.6 ± 3	91.51 ± 2
Markedness	93.05 ± 3	93.55 ± 3	91.63 ± 2	89.75 ± 3	91.68 ± 2

Στη συνέχεια απεικονίζεται οι καμπύλες ROC και ο υπολογισμός AUC



Εικόνα 3.10: Ανάλυση ROC και AUC (B – M)

Όπως προκύπτει από τον **πίνακα 3.5 και 3.6** και από τα **γραφήματα 3.9 και 3.10** ο GBM είναι το καλύτερο μοντέλο με 96.7% ακρίβεια, 96.71% ευσαισθησία και 96.69% ειδικότητα με τα εξής χαρακτηριστικά: **«concave.points_worst» , «perimeter_worst»,**

«texture_worst». Τα αποτελέσματα του καλύτερου μοντέλου παρουσιάζονται στον **πίνακα 3.7** και στον **πίνακα 3.8** αναλυτικά **.**

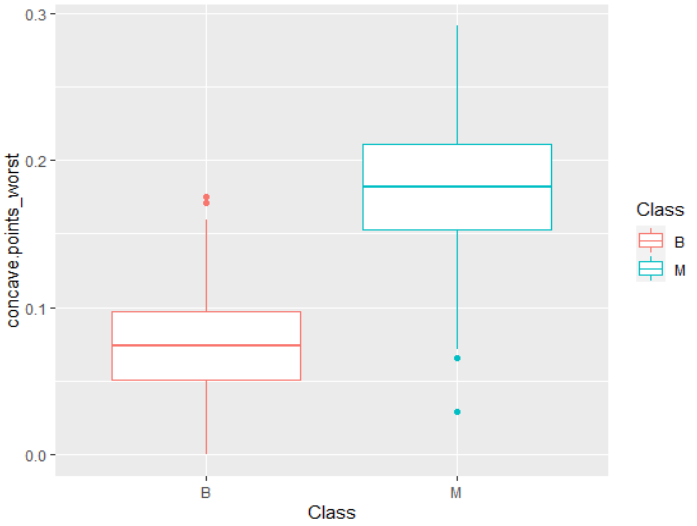
Πίνακας 3.7: Απόδοση μηχανής ενίσχυσης κλίσης (GBM) (B – M)

	GBM
Ακρίβεια	96.7 ± 1
Ευαισθησία	96.71 ± 2
Ειδικότητα	96.69 ± 2
Αληθώς θετικά	41 ± 1
Ψευδώς αρνητικά	1 ± 1
Ψευδώς θετικά	1 ± 1
Αληθώς αρνητικά	41 ± 1
Συντελεστής kappa	93.4 ± 3
Θετική προγνωστική αξία	96.77 ± 2
Αρνητική προγνωστική αξία	96.77 ± 2
Σταθμισμένη ορθότητα	96.7 ± 1
F1 score	96.7 ± 1
Ποσοστό ψευδώς αρνητικών	3.29 ± 2
Ποσοστό αληθώς αρνητικών	3.31 ± 2
Ψευδές ποσοστό ανακάλυψης	3.23 ± 2
Ψευδές ποσοστό παράλειψης	3.23 ± 2
Συντελεστής συσχέτισης Matthews	93.47 ± 3
Informedness	93.4 ± 3
Markedness	93.55 ± 3

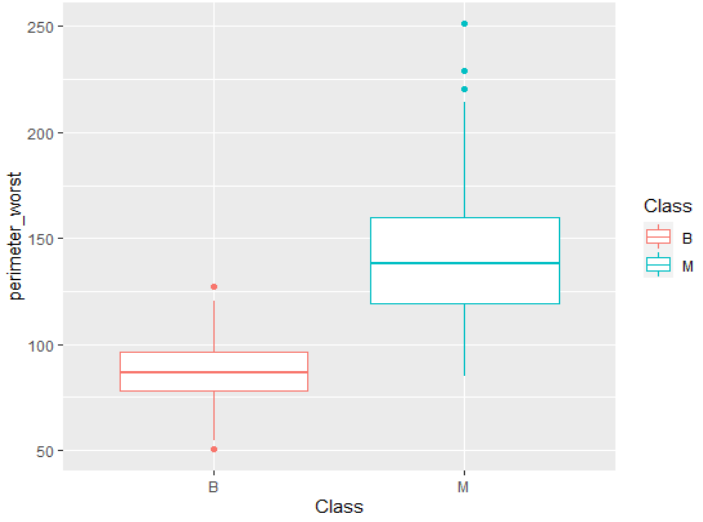
Πίνακας 3.8: Πίνακας αληθείας καλύτερου μοντέλου (GBM) (B – M)

		Πραγματική		
	N=84	M	B	Σύνολο
Πρόβλεψη	**M**	**41**	**1**	**42**
	B	**1**	**41**	**42**

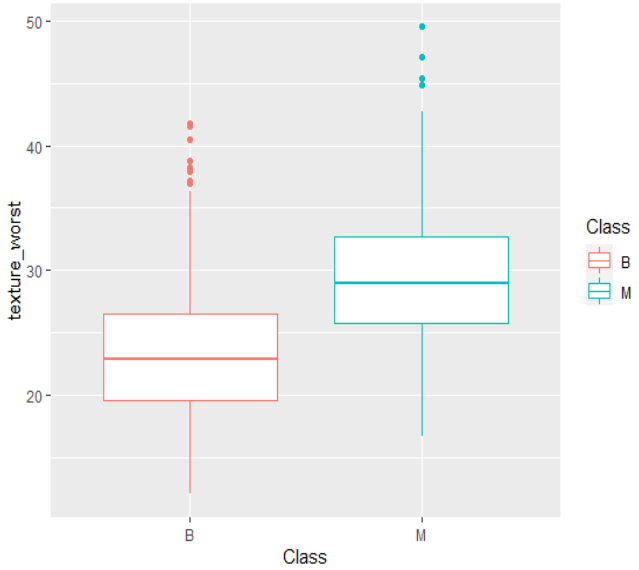
Το θηκόγραμμα κάθε χαρακτηριστικού αλλά και ο διαχωρισμός των δειγμάτων στον τρισδιάστατο χώρο δίνονται στα γραφήματα παρακάτω



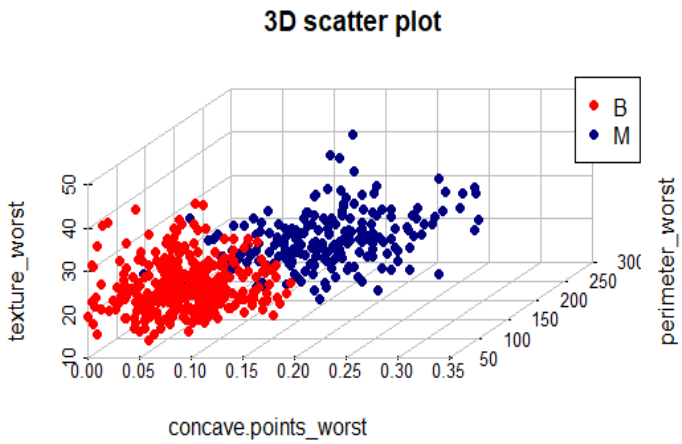
Εικόνα 3.11: θηκογράμμα για «concave.points_worst»



Εικόνα 3.12: Θηκόγραμμα για «perimeter_worst»



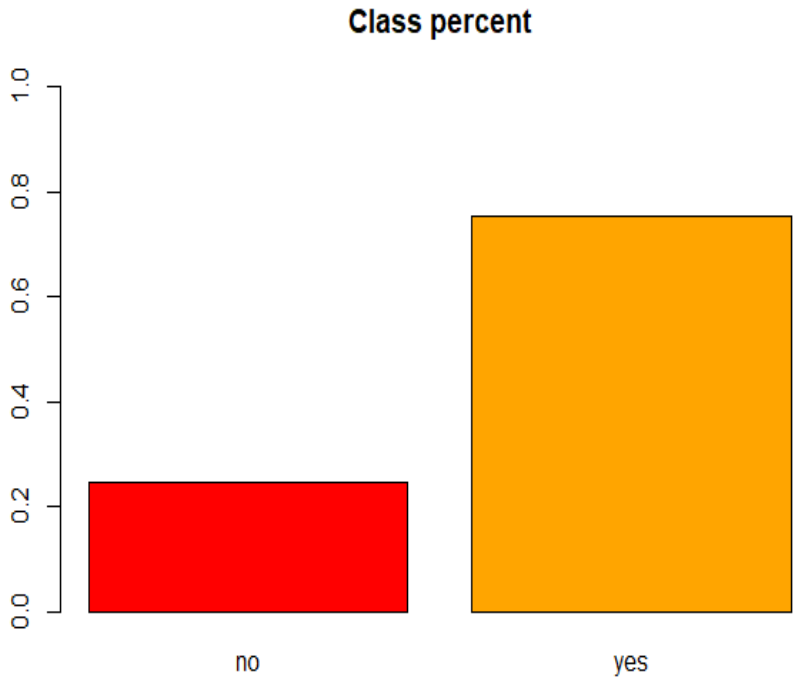
Εικόνα 3.13: Θηκόγραμμα για «texture_worst»



Εικόνα 3.14: 3D γράφημα καλύτερων χαρακτηριστικών

3.3 Σύνολο δεδομένων διάγνωσης Πάρκινσον

Το σύνολο δεδομένων αυτό, εμπεριείχε 21 χαρακτηριστικά και 195 δείγματα (οι 147 να νοσούν από Πάρκινσον και οι 48 να είναι υγιείς).



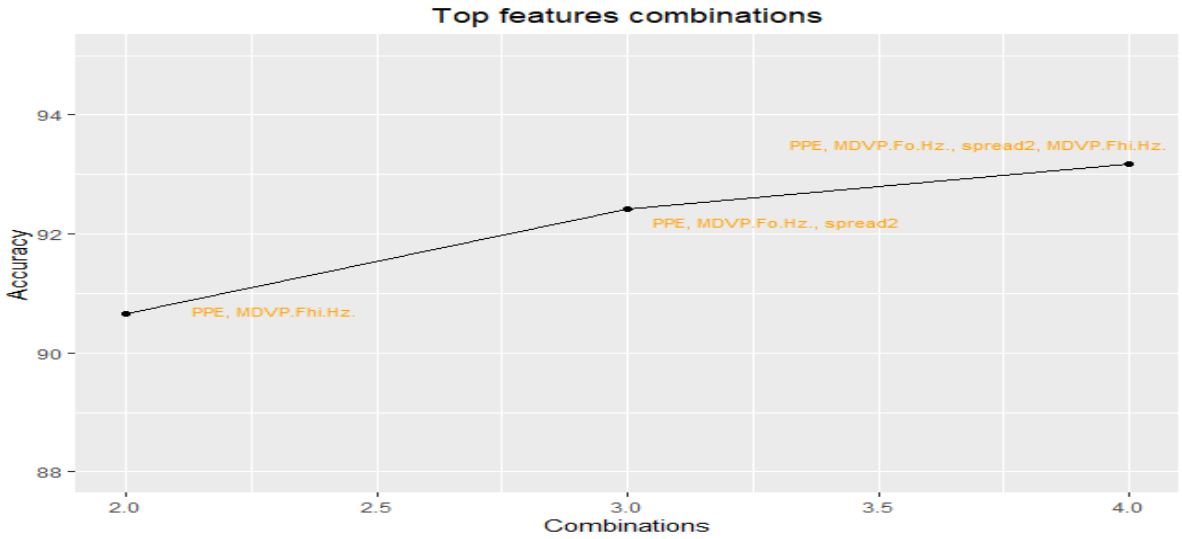
Εικόνα 3.15: Ποσοστιαία αναλογία κάθε κλάσης (no: υγιείς – yes: νοσούν με πάρκινσον)

Κατόπιν η RFE επέλεξε 4 χαρακτηριστικά ως τα πιο σημαντικά → **«PPE», « MDVP.Fo.Hz.», «MDVP.Fhi.Hz.», «spread2»** και με την μέθοδο της επαναδειγματοληψίας bootstrap λήφθηκαν τις αξιολογήσεις των μοντέλων

Πίνακας 2.9: Καλύτερο αποτέλεσμα για κάθε μοντέλο (Πάρκινσον)

Μοντέλο	Συνδυασμός χαρακτηριστικών	Ακρίβεια	Ευαισθησία	Ειδικότητα
GBM	**« PPE », « MDVP.Fo.Hz. », « spread2 », « MDVP.Fhi.Hz. »**	**95.47 ±1**	**97.55 ±2**	**93.60±3**
SVM	**« PPE », « MDVP.Fo.Hz. », « spread2 », « MDVP.Fhi.Hz. »**	**91.68±3**	**92.22± 4**	**91.2 ± 5**
RF	**« PPE », « MDVP.Fo.Hz. », « spread2 », « MDVP.Fhi.Hz. »**	**96.3 ±2**	**97.77 ± 4**	**95.00± 3**
NNET	**« PPE », « MDVP.Fo.Hz. », « spread2 », « MDVP.Fhi.Hz. »**	**91.68±3**	**95.77±3**	**88.00±5**
K-NN	**« PPE », « MDVP.Fo.Hz. », « spread2 »**	**91.07±3**	**95.82±4**	**86.62±5**

Στο **γράφημα 3.16** παρουσιάζεται ο ψηλότερος μέσος όρος μοντέλων για τους καλύτερους συνδυασμούς σε κάθε αριθμό συνδυασμών και στον Πίνακα 16 οι αποδόσεις του κάθε μοντέλου ξεχωριστά στον καλύτερο συνδυασμό.

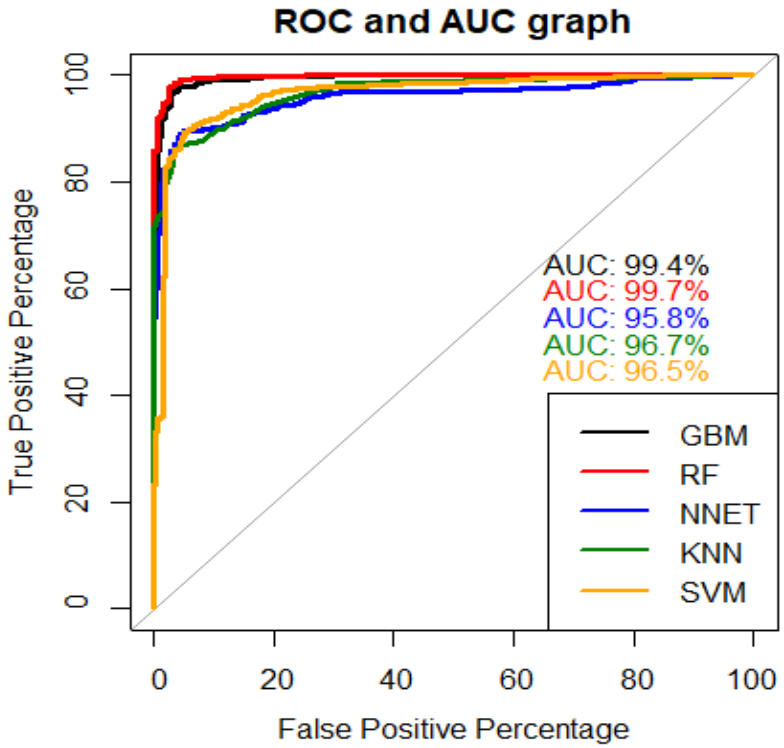


Εικόνα 3.16: Μέσος όρος ακρίβειας μοντέλων για τους καλύτερους συνδυασμούς (Πάρκινσον)

Πίνακας 3.10: Αποτελέσματα κάθε μοντέλου για τον καλύτερο συνδυασμό: « PPE », « MDVP.Fo.Hz. », « spread2 »,

	SVM	GBM	RF	NNET	KNN
Ακρίβεια	92.6 ± 2	96.32 ± 2	96.96 ± 2	88.68 ± 5	90.38 ± 3
Ευαισθησία	92.43 ± 3	94.61 ± 3	96.21 ± 3	87.07 ± 9	87.26 ± 4
Ειδικότητα	92.8 ± 2	98.22 ± 3	97.78 ± 3	90.44 ± 11	93.84 ± 5
Αληθώς θετικά	42 ± 1	45 ± 2	44 ± 2	41 ± 6	43 ± 3
Ψευδώς αρνητικά	4 ± 2	3 ± 1	2 ± 1	6 ± 5	6 ± 2
Ψευδώς θετικά	3 ± 1	1 ± 1	1 ± 1	4 ± 5	3 ± 2
Αληθώς αρνητικά	46 ± 2	48 ± 2	48 ± 2	44 ± 5	44 ± 2
Συντελεστής kappa	85.18 ± 3	92.64 ± 4	93.9 ± 3	77.34 ± 10	80.78 ± 6
Θετική προγνωστική αξία	93.45 ± 2	98.39 ± 3	98.06 ± 3	92.03 ± 8	94.21 ± 4
Αρνητική προγνωστική αξία	91.85 ± 3	94.34 ± 3	96 ± 3	87.27 ± 7	87.05 ± 3
Σταθμισμένη ορθότητα	92.62 ± 2	96.42 ± 2	96.99 ± 2	88.76 ± 5	90.55 ± 3
F1 score	92.9 ± 2	96.43 ± 2	97.07 ± 1	88.93 ± 5	90.5 ± 3
Ποσοστό ψευδώς αρνητικών	7.57 ± 3	5.39 ± 3	3.79 ± 3	12.93 ± 9	12.74 ± 4
Ποσοστό αληθώς αρνητικών	7.2 ± 2	1.78 ± 3	2.22 ± 3	9.56 ± 11	6.16 ± 5
Ψευδές ποσοστό ανακάλυψης	6.55 ± 2	1.61 ± 3	1.94 ± 3	7.97 ± 8	5.79 ± 4
Ψευδές ποσοστό παράλειψης	8.15 ± 3	5.66 ± 3	4 ± 3	12.73 ± 7	12.95 ± 3
Συντελεστής συσχέτισης Matthews	85.27 ± 3	92.78 ± 4	94.02 ± 3	78.39 ± 9	81.18 ± 6
Informedness	85.23 ± 3	92.83 ± 4	93.99 ± 3	77.51 ± 10	81.1 ± 6
Markedness	85.3 ± 3	92.73 ± 4	94.05 ± 3	79.29 ± 8	81.26 ± 6

Στη συνέχεια απεικονίζονται οι καμπύλες ROC και ο υπολογισμός AUC



Εικόνα 3.17: Ανάλυση ROC και AUC (Πάρκινσον)

Όπως προκύπτει από τον **πίνακα 3.9 και 3.10** και από τα **γραφήματα 3.16 και 3.17** ο GBM είναι το καλύτερο μοντέλο με 96.32% ακρίβεια, 97.78% ευσαισθησία και 95% ειδικότητα με τα εξής χαρακτηριστικά: **« PPE », « MDVP.Fo.Hz. », « spread2 », « MDVP.Fhi.Hz. ».** Τα αποτελέσματα του καλύτερου μοντέλου παρουσιάζονται στον **πίνακα 3.11** και στον **πίνακα 3.12** αναλυτικά.

Πίνακας 3.11: Απόδοση τυχαίου δάσους (RF) (Πάρκινσον)

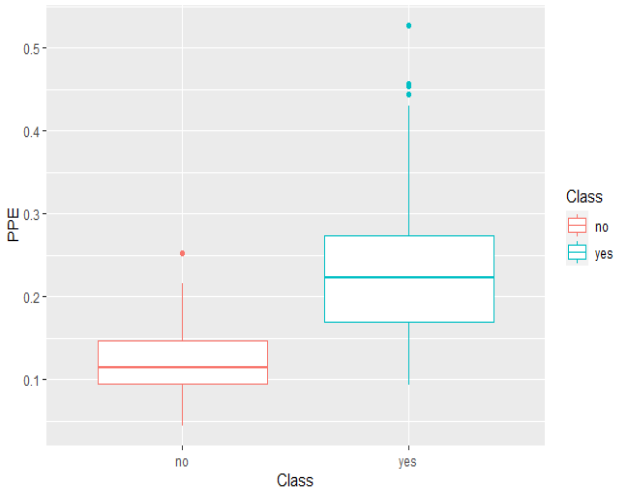
	RF
Ακρίβεια	96.96 ± 2
Ευαισθησία	96.21 ± 3
Ειδικότητα	97.78 ± 3
Αληθώς θετικά	44 ± 2
Ψευδώς αρνητικά	2 ± 1
Ψευδώς θετικά	1 ± 1
Αληθώς αρνητικά	48 ± 2
Συντελεστής kappa	93.9 ± 3
Θετική προγνωστική αξία	98.06 ± 3
Αρνητική προγνωστική αξία	96 ± 3
Σταθμισμένη ορθότητα	96.99 ± 2
F1 score	97.07 ± 1
Ποσοστό ψευδώς αρνητικών	3.79 ± 3
Ποσοστό αληθώς αρνητικών	2.22 ± 3
Ψευδές ποσοστό ανακάλυψης	1.94 ± 3

Ψευδές ποσοστό παράλειψης	4 ± 3
Συντελεστής συσχέτισης Matthews	94.02 ± 3
Fowlkes - mallows index	93.99 ± 3
Informedness	94.05 ± 3
Markedness	96.96 ± 2

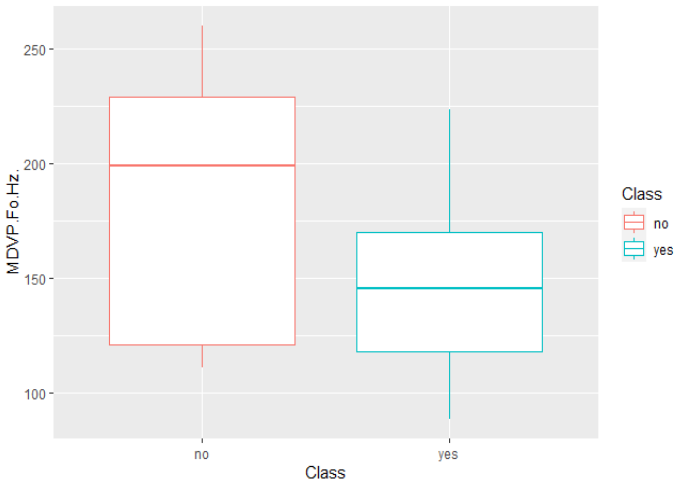
Πίνακας3.12: Πίνακας αληθείας καλύτερου μοντέλου (RF) (Πάρκινσον)

		Πραγματική		
	N=95	YES	NO	Σύνολο
Πρόβλεψη	YES	**44**	**1**	**45**
	NO	**2**	**48**	**50**

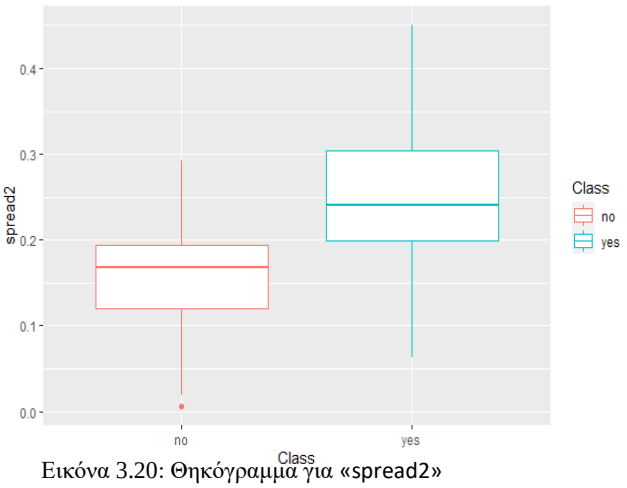
Το θηκόγραμμα κάθε χαρακτηριστικού του καλύτερου συνδυασμού δίνεται πάρακατω



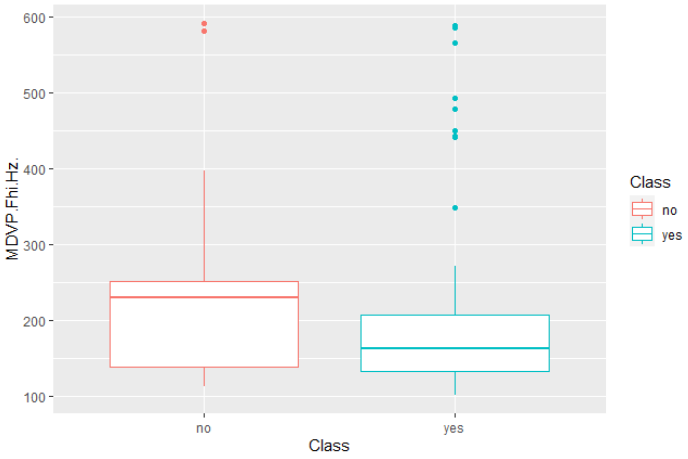
Εικόνα 3.18: Θηκογράμμα για «PPE»



Εικόνα 3.19: Θηκόγραμμα για «MDVP.Fo.Hz»



Εικόνα 3.20: Θηκόγραμμα για «spread2»

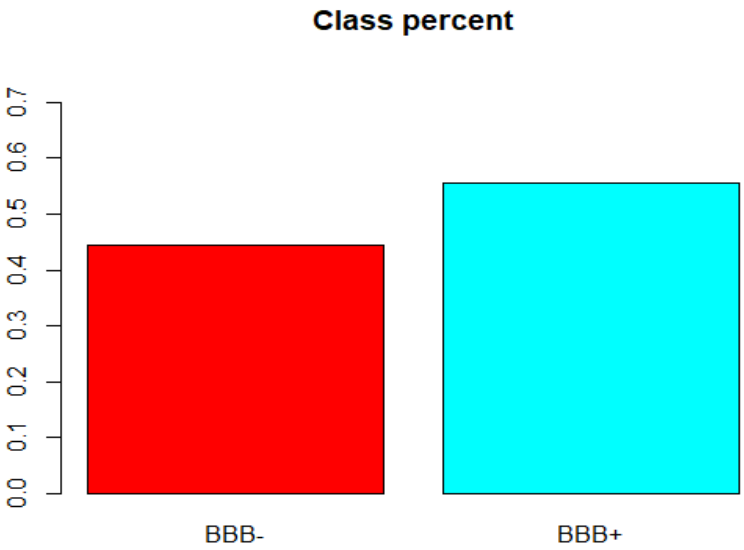


Εικόνα 3.21: Θηκόγραμμα για «MDVP.Fhi.Hz»

3.4 Σύνολο δεδομένων διαπερατότητας BBB

3.4.1 Μηχανική Μάθηση

Υπήρχαν αρχικά 1311 δείγματα και με βάση των κριτηρίων που προαναφέρθηκαν αφαιρέθηκαν 406 επαναλαμβανόμενα δείγματα και άλλα 89 με ελλιπείς πληροφορίες. Για τα χαρακτηριστικά υπήρχαν για αρχή 1444 και αφαιρέθηκαν αυτά που είχαν χαμηλή διακύμανση (τυπική απόκλιση < 0.01). Οπότε συνοπτικά, το σύνολο εκπαίδευσης αποτελείτο από 816 δείγματα (363 να μην διαπερνούν και 453 να διαπερνούν το BBB) και 1192 χαρακτηριστικά.



Εικόνα 3.22: Ποσοστιαία αναλογία κάθε κλάσης (BBB-: μη διαπερατά – BBB+: διαπερατά)

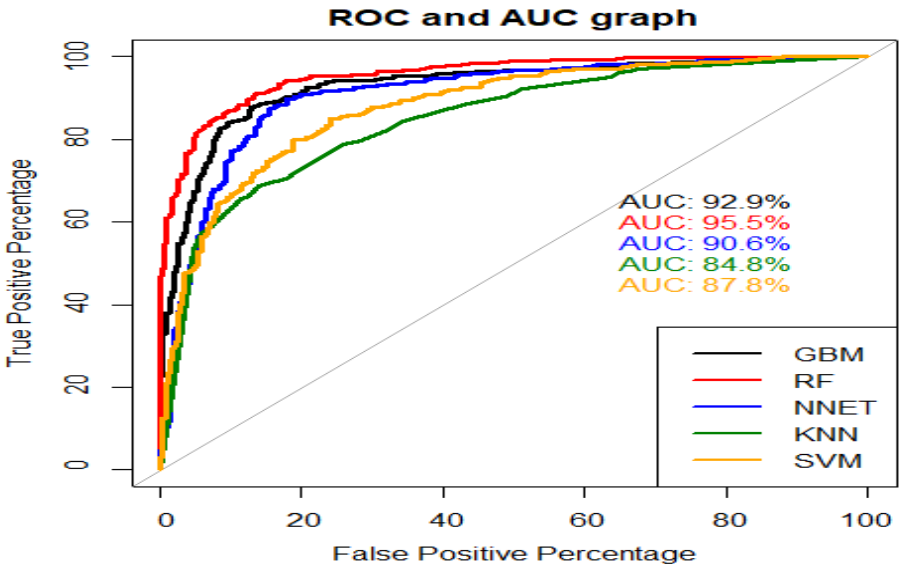
Έτσι, η RFE επέλεξε 120 χαρακτηριστικά ως τα πιο σημαντικά: **"TopoPSA", "SHBd", "MDEO.11", "ATSC1c", "MLFER_A", "SdO", "WTPT.4", "MATS1c", "nHBAcc_Lipinski",**

"maxHBd", "maxdO", "nO", "ATSC0c", "mindO", "ATSC0s", "SpMax3_Bhs", "nHBAcc", "ETA_dEpsilon_D", "ATS0s", "ndO", "SHsOH", "SpMax4_Bhs", "SHBint2", "MAXDN2", "AATS8s", "AATS0s", "ATSC2s", "WTPT.3", "SHBint5", "maxHsOH", "maxHBint2", "meanI", "mindssC", "DELS2", "AATSC0s", "GATS2s", "MLFER_S", "MATS2e", "maxHBint5", "DELS", "AATS7s", "BCUTc.1h", "minHBint2", "AATSC0i", "SpMax1_Bhs", "GATS3i", "MATS2s", "AATSC2s", "GATS2e", "MATS2c", "SM1_Dze", "SpMAD_Dzs", "gmin", "SM1_Dzv", "ETA_Eta_B", "MAXDN", "minHBint4", "minHsOH", "nHBd", "SHBa", "SpMax5_Bhs", "ATSC2e", "WTPT.5", "minHBd", "nHBDon", "CrippenLogP", "AATS1i", "ATSC0e", "SM1_Dzp", "GATS2c", "minwHBa", "nBondsD2", "AATS6e", "BCUTc.1l", "AATSC2e", "nBondsD", "AATS2s", "SpMax2_Bhs", "ATSC2c", "GATS5e", "GATS3s", "AATS6s", "ETA_Epsilon_4", "SHssNH", "minHBint5", "AATS0e", "ETA_Epsilon_2", "SsOH", "SpMin8_Bhi", "Mse", "SdssC", "ETA_dEpsilon_C", "AATSC1p", "gmax", "MAXDP", "minssCH2", "AATS3s", "nHBDon_Lipinski", "AATSC0e", "GATS1e", "GATS2i", "ATS5s", "ATS4m", "AATS4s", "AATS5s", "ATSC1s", "AATSC3v", "maxsOH", "AMR", "nHBint2", "ETA_Eta_B_RC", "MATS3v", "MAXDP2", "EE_Dzs", "maxHssNH", "nAcid", "GATS1s", "sumI", "AATSC1i", "ASP.0" και με την μέθοδο της επαναδειγματοληψίας bootstrap λήφθηκαν τις αξιολογήσεις των μοντέλων

Πίνακας 3.13: Αποτελέσματα κάθε μοντέλου για τα σημαντικά χαρακτηριστικά. (BBB)

	SVM	GBM	RF	NNET	KNN
Ακρίβεια	79.97 ± 4	86.54 ± 2	88.83 ± 2	85.64 ± 3	76.36 ± 3
Ευαισθησία	76.44 ± 5	83.21 ± 5	86.87 ± 5	82.14 ± 6	69.95 ± 4
Ειδικότητα	82.78 ± 4	89.09 ± 3	90.43 ± 2	88.33 ± 3	81.43 ± 4
Αληθώς θετικά	82 ± 7	90 ± 8	94 ± 5	89 ± 9	75 ± 5
Ψευδώς αρνητικά	23 ± 6	15 ± 4	13 ± 3	16 ± 4	25 ± 5
Ψευδώς θετικά	25 ± 6	18 ± 5	14 ± 5	19 ± 7	32 ± 4
Αληθώς αρνητικά	112 ± 7	121 ± 6	123 ± 5	120 ± 7	111 ± 7
Συντελεστής kappa	59.3 ± 7	72.56 ± 5	77.33 ± 4	70.73 ± 7	51.72 ± 7
Θετική προγνωστική αξία	77.99 ± 5	85.9 ± 3	87.86 ± 2	84.94 ± 4	75 ± 4
Αρνητική προγνωστική αξία	81.64 ± 4	87.18 ± 3	89.72 ± 3	86.35 ± 4	77.32 ± 3
Σταθμισμένη ορθότητα	79.61 ± 4	86.15 ± 3	88.65 ± 2	85.23 ± 4	75.69 ± 3
F1 score	77.11 ± 4	84.46 ± 3	87.29 ± 2	83.42 ± 4	72.36 ± 4
Ποσοστό ψευδώς αρνητικών	23.56 ± 5	16.79 ± 5	13.13 ± 5	17.86 ± 6	30.05 ± 4
Ποσοστό αληθώς αρνητικών	17.22 ± 4	10.91 ± 3	9.57 ± 2	11.67 ± 3	18.57 ± 4
Ψευδές ποσοστό ανακάλυψης	22.01 ± 5	14.1 ± 3	12.14 ± 2	15.06 ± 4	25 ± 4
Ψευδές ποσοστό παράλειψης	18.36 ± 4	12.82 ± 3	10.28 ± 3	13.65 ± 4	22.68 ± 3
Συντελεστής συσχέτισης Matthews	59.42 ± 7	72.69 ± 5	77.43 ± 4	70.87 ± 7	51.85 ± 7
Informedness	59.22 ± 7	72.29 ± 5	77.29 ± 4	70.47 ± 7	51.38 ± 7
Markedness	59.63 ± 7	73.09 ± 5	77.58 ± 3	71.29 ± 7	52.32 ± 7

Στη συνέχεια απεικονίζονται οι καμπύλες ROC και ο υπολογισμός AUC του συνόλου επικύρωσης



Εικόνα 3.23: Ανάλυση ROC και AUC συνόλου επικύρωσης (BBB)

Όπως προκύπτει από τον **πίνακα 3.13** και από το **γράφημα 3.23** ο GBM είναι το καλύτερο μοντέλο με 88.51% ακρίβεια, 86.05% ευσαισθησία και 90.91% ειδικότητα με τα χαρακτηριστικά που επέλεξε η RFE**.**

Στους παρακάτω πίνακες φαίνονται τα αποτελέσματα του καλύτερου μοντέλου (RF) στο σύνολο αξιολόγησης

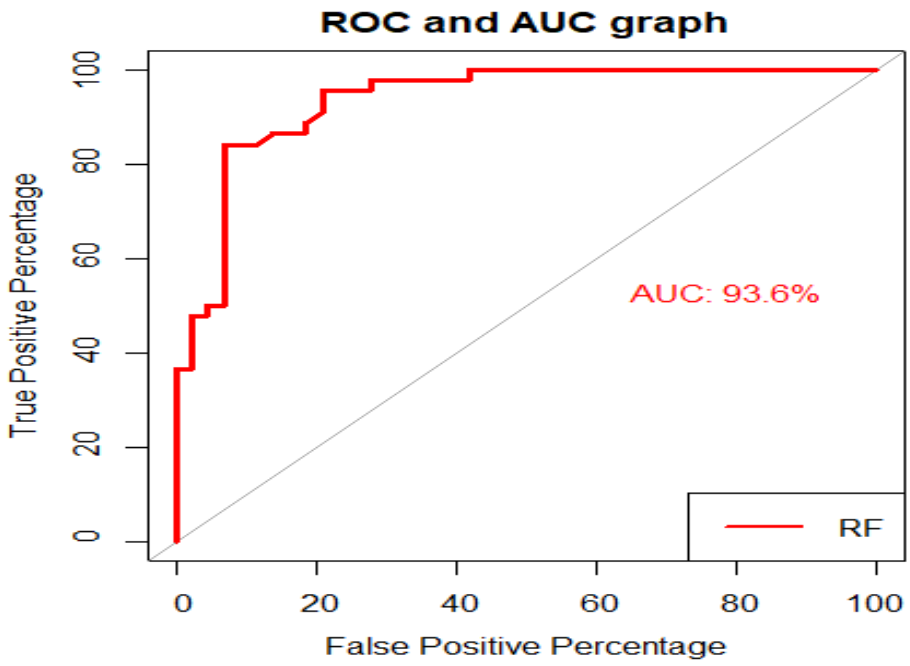
Πίνακας 3.14: Απόδοση τυχαίου δάσους στο σύνολο αξιολόγησης (RF) (BBB)

	RF
Ακρίβεια	88.51
Ευαισθησία	86.05
Ειδικότητα	90.91
Αληθώς θετικά	37
Ψευδώς αρνητικά	4
Ψευδώς θετικά	6
Αληθώς αρνητικά	40
Συντελεστής kappa	77
Θετική προγνωστική αξία	90.24
Αρνητική προγνωστική αξία	86.96
Σταθμισμένη ορθότητα	88.48
F1 score	88.1
Ποσοστό ψευδώς αρνητικών	13.95
Ποσοστό αληθώς αρνητικών	9.09
Ψευδές ποσοστό ανακάλυψης	9.76
Ψευδές ποσοστό παράλειψης	13.04
Συντελεστής συσχέτισης Matthews	77.08
Informedness	76.96
Markedness	77.2

Πίνακας 3.15: Πίνακας αληθείας καλύτερου μοντέλου (RF) (BBB)

		Πραγματική		
	N=87	BBB-	BBB+	Σύνολο
Πρόβλεψη	BBB-	**37**	**4**	**41**
	BBB+	**6**	**40**	**46**

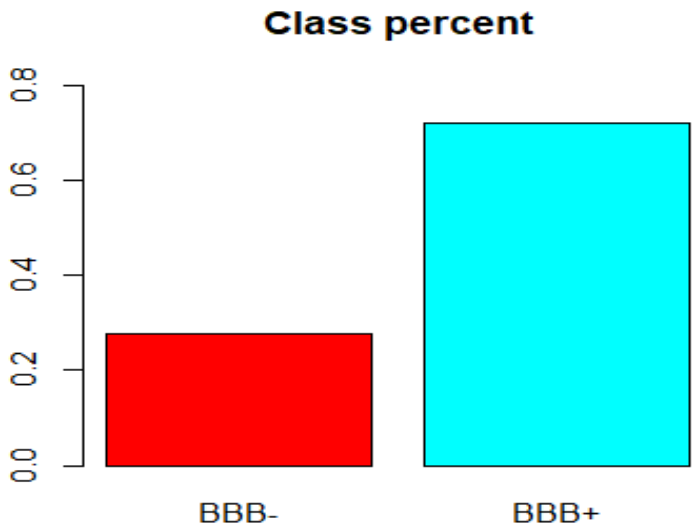
Εδώ, απεικονίζεται η καμπύλη ROC και ο υπολογισμός AUC του καλύτερου μοντέλου για το σύνολο αξιολόγησης



Εικόνα 3.24: Ανάλυση ROC και AUC συνόλου αξιολόγησης (BBB)

3.4.2 Βαθιά Μάθηση

Για το μοντέλο BM, υπήρχαν αρχικά 7811 δείγματα και με βάση των κριτηρίων που προαναφέρθηκαν αφαιρέθηκαν 4247 επαναλαμβανόμενα δείγματα και άλλα 415 με ελλιπείς πληροφορίες. Για τα χαρακτηριστικά υπήρχαν για αρχή 1444 και αφαιρέθηκαν αυτά που είχαν χαμηλή διακύμανση (τυπική απόκλιση < 0.01). Οπότε συνοπτικά, το σύνολο εκπαίδευσης αποτελείτο από 2048 δείγματα (669 να μην διαπερνούν και 1739 να διαπερνούν το BBB) και 1222 χαρακτηριστικά.

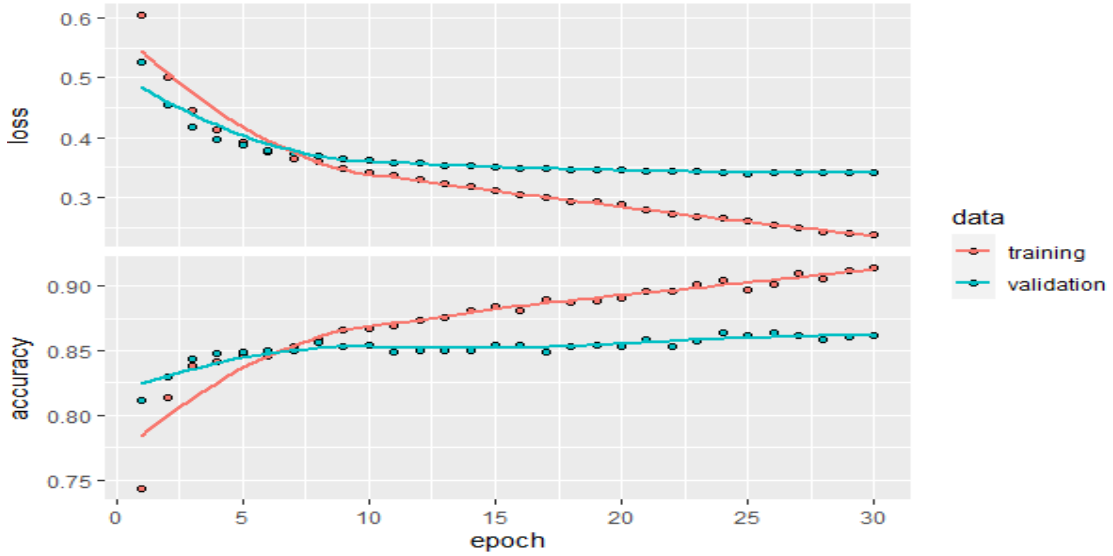


Εικόνα 3.25: Ποσοστιαία αναλογία κάθε κλάσης (BBB-: μη διαπερατά – BBB+: διαπερατά) (DNN)

Μετά, η RFE επέλεξε 223 χαρακτηριστικά ως τα πιο σημαντικά: **"SHBd", "minssssNp", "maxssssNp", "SssssNp", "TopoPSA", "MLFER_BH", "MLFER_BO", "SHsOH", "mindssC", "MDEO.11", "MLFER_S", "WTPT.3", "maxHsOH", "minHsOH", "nHBAcc_Lipinski", "nAcid", "ATSC4v", "maxHBint8", "maxHBd", "ATSC1c", "nHBd", "SHBint8", "nHBDon", "SCH.5", "nHBAcc", "nHBint3", "SHBint6", "MDEN.22", "ALogP", "ETA_Shape_Y", "nHBDon_Lipinski", "minHBint2", "nssssNp", "ATSC0c", "MDEC.33", "minHBa", "ETA_dEpsilon_D", "SRW5", "maxHBint7", "WTPT.5", "SHBint7", "SHBint3", "AATSC1c", "VE3_Dt", "ATSC4m", "GATS4v", "VCH.5", "SHBint5", "AATS1i", "n4HeteroRing", "nT4HeteroRing", "SdssC", "SsOH", "minHBint4", "maxHBint5", "minHBint3", "minHBd", "SHBint10", "GATS2i", "MATS1p", "minHBint8", "WTPT.4", "MLFER_E", "nHBa", "MATS1c", "maxHBint6", "AATSC4v", "GATS1e", "SpMax4_Bhm", "MATS2c", "nHBAcc2", "maxsOH", "maxHBint2", "VE3_Dzi", "minsOH", "VE3_Dze", "MLFER_A", "MLFER_L", "ATSC3v", "n4Ring", "minwHBa", "maxHBint10", "MATS4v", "nT4Ring", "SHBint2", "MATS1s", "SssNH", "maxHBint3", "ATSC1e", "GATS2c", "nHBint6", "SHssNH", "MATS2s", "ATSC3p", "AATSC1p", "MATS4m", "AATS2i", "nHBint8", "IC1", "VE3_Dzv", "SHBint4", "MDEC.24", "GATS4i", "SpMax5_Bhs", "VCH.6", "SM1_DzZ", "SpMax2_Bhe", "SM1_Dzm", "Kier3", "SCH.6", "AATSC0c", "DELS", "AATS7v", "AATSC4m", "nN", "MATS1e", "ATSC2s", "SpMax2_Bhi", "VC.5", "GATS5m", "JGI2", "nO", "BCUTc.1l", "AATS5s", "MATS8c", "SpMax6_Bhs", "VCH.7", "SHaaNH", "MAXDN", "GGI3", "nHsOH", "minHBint7", "SM1_Dze", "SM1_Dzv", "AATS8s", "nsOH", "ATSC0s", "ALogp2", "SpMax5_Bhp", "VC.6", "ASP.7", "VE1_Dt", "MATS7c", "VE3_DzZ", "ATSC1s", "SpMax3_Bhi", "VE3_Dzp", "minHaaNH", "SpMax5_Bhv", "VE3_Dzm", "GATS1m", "SpMax5_Bhm", "SpMax5_Bhi", "AATS7s", "VPC.4", "ASP.6", "GATS6e", "MATS1i", "GATS8p", "ATSC8c",**

"GATS1s", "GATS5e", "DELS2", "AATS5e", "VPC.5", "SpMin1_Bhm", "SpMax7_Bhs", "maxHBa", "SpMax3_Bhs", "GATS2p", "nHBint4", "SpMax2_Bhp", "SpMax4_Bhs", "maxHaaNH", "MAXDN2", "nBase", "nHBint10", "MATS3v", "MAXDP2", "ATS0s", "SdO", "ndssC", "minHBint6", "MAXDP", "ATSC8s", "maxHBint4", "SC.6", "ATS5s", "ETA_Eta_F", "AMR", "maxssNH", "AATS6v", "Kier2", "ETA_EtaP_F", "SpMin5_Bhp", "SpMin2_Bhv", "ATSC7c", "minddssS", "GATS6i", "MDEC.44", "MATS5s", "VPC.6", "SpMin2_Bhp", "SRW10", "minsssN", "maxHCsats", "ETA_Eta_B", "MATS8s", "SpMax5_Bhe", "LipoaffinityIndex", "AATS6e", "GATS1c", "AATSC3v", "AATS5v", "ATSC1i", "GATS1i", "VC.3", "ATSC5c", "JGI3", "ASP.4", "ETA_Eta_F_L", "MATS1v", "AATS0i"

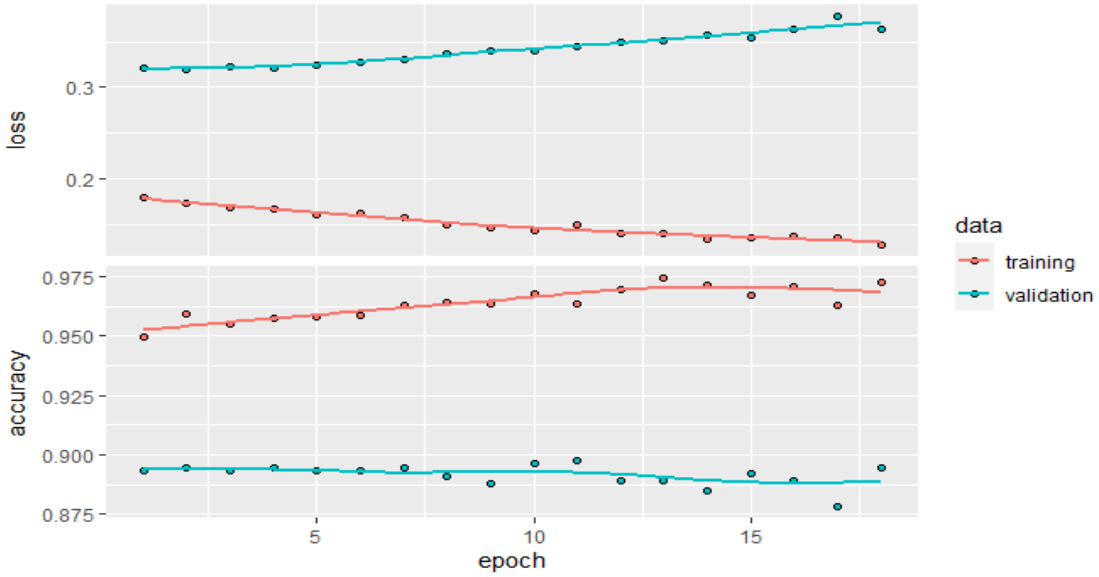
Στην συνέχεια χρησιμοποιήθηκαν τα χαρακτηριστικά αυτά για την κατασκευή της αρχιτεκτονικής του μοντέλου, όπου στο **γράφημα 3.26** απεικονίζεται η ακρίβεια και η απώλεια για κάθε εποχή κατά την διάρκεια της εκπαίδευσης του.



Εικόνα 3.26: Επίδοση μοντέλου κατά την διάρκεια εκπαίδευσης του για κάθε εποχή σε σύνολο εκπαίδευσης και επικύρωσης. (BBB) (DNN)

Η αρχιτεκτονική του μοντέλου αυτού φτιάχτηκε με 4 κρυφά επίπεδα, με το πρώτο να περιέχει 2048 κόμβους, το δεύτερο 4096 κόμβους, το τρίτο με 2048 κόμβους και το τέταρτο με 2048 κόμβους, χρησιμοποιώντας τον Adam με ρυθμό μάθησης 0.00001. Συγκεκριμένα το μοντέλο αυτό είχε 91.52% ακρίβεια στα δεδομένα εκπαίδευσης, 87.95% ακρίβεια στα δεδομένα επικύρωσης και 87.36% στα δεδομένα αξιολόγησης

Στο **γράφημα 3.27** φαίνεται η ακρίβεια και η απώλεια για κάθε εποχή κατά την διάρκεια της επανεκπαίδευσης του.



Εικόνα 3.27: Επίδοση μοντέλου κατά την διάρκεια επανεκπαίδευσης του για κάθε εποχή σε σύνολο εκπαίδευσης και επικύρωσης. (BBB) (DNN)

Έτσι λοιπόν μετά την επανεκπαίδευση του μοντέλου, στα δεδομένα εκπαίδευσης είχαμε 98.75% ακρίβεια, 96.69% ευαισθησία και 99.58% ειδικότητα. Στα δεδομένα επικύρωσης είχαμε 91.69% ακρίβεια, 85.48% ευαισθησία και 88.51%. Τέλος, στα δεδομένα αξιολόγησης είχαμε 88.51% ακρίβεια, 79.07 ευαισθησία και 97.73 ειδικότητα.

Στον **πίνακα 3.16** φαίνονται αναλυτικά οι επιδόσεις του μοντέλου και στον **πίνακα 3.17** φαίνεται ο πίνακας αληθείας του συνόλου αξιολόγησης.

Πίνακας 3.16: Αποτελέσματα μοντέλου σε κάθε σύνολο (εκπαίδευσης, επικύρωσης και αξιολόγησης).

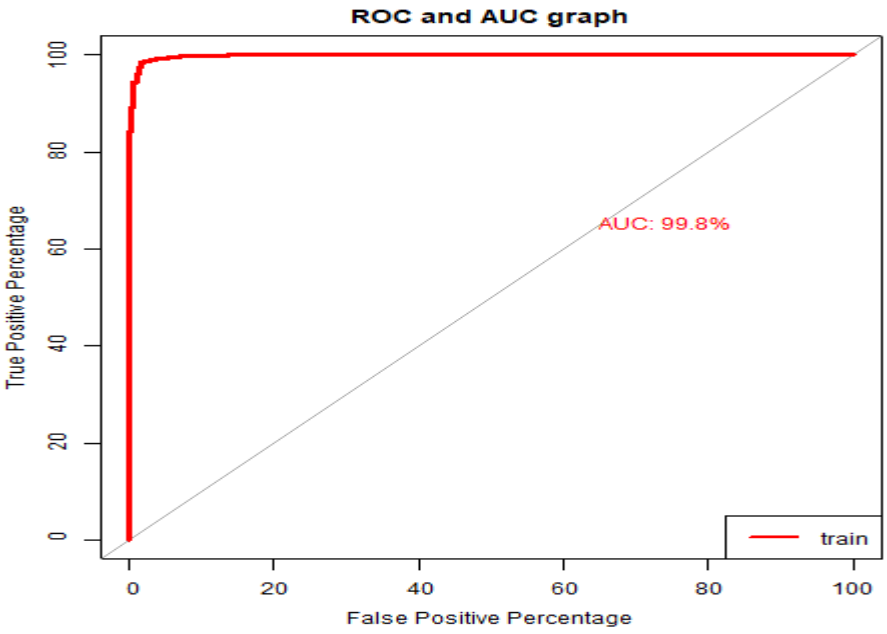
	Train	Validation	Test
Ακρίβεια	98.75	91.69	88.51
Ευαισθησία	96.69	85.48	79.07
Ειδικότητα	99.58	93.84	97.73
Αληθώς θετικά	467	159	34
Ψευδώς αρνητικά	5	33	1
Ψευδώς θετικά	16	27	9
Αληθώς αρνητικά	1198	503	43
Συντελεστής kappa	96.93	78.5	76.96
Θετική προγνωστική αξία	98.94	82.81	97.14
Αρνητική προγνωστική αξία	98.68	94.91	82.69
Σταθμισμένη ορθότητα	98.14	89.66	88.4
F1 score	97.8	84.13	87.18
Ποσοστό ψευδώς αρνητικών	3.31	14.52	20.93
Ποσοστό αληθώς αρνητικών	0.42	6.16	2.27

Ψευδές ποσοστό ανακάλυψης	1.06	17.19	2.86
Ψευδές ποσοστό παράλειψης	1.32	5.09	17.31
Συντελεστής συσχέτισης Matthews	96.94	78.52	78.3
Informedness	96.27	79.33	76.8
Markedness	97.62	77.72	79.84

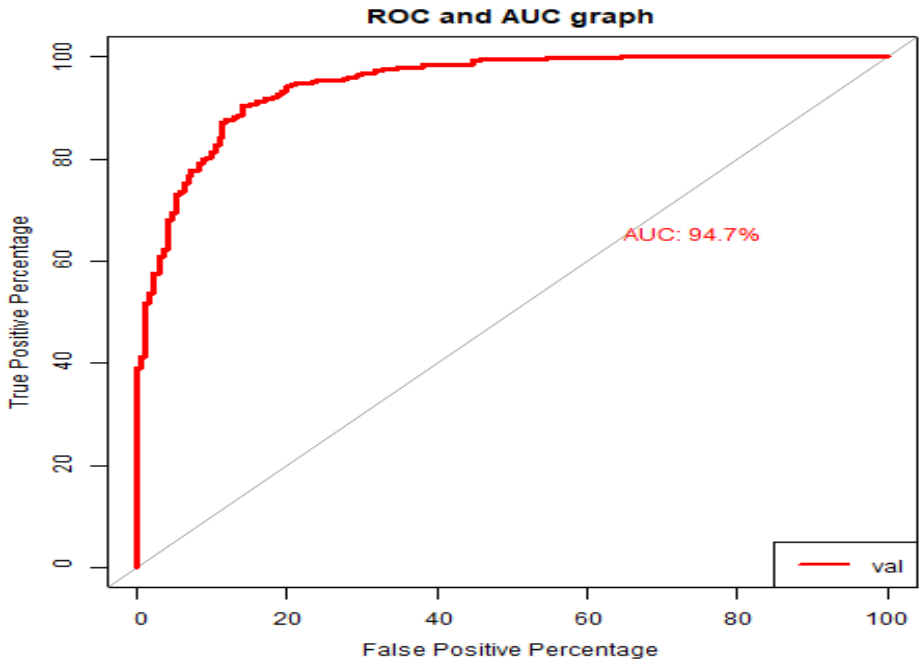
Πίνακας 3.17: Πίνακας αληθείας μοντέλου BM

		Πραγματική		
	N=87	BBB-	BBB+	Σύνολο
Πρόβλεψη	BBB-	**34**	**1**	**41**
	BBB+	**9**	**43**	**46**

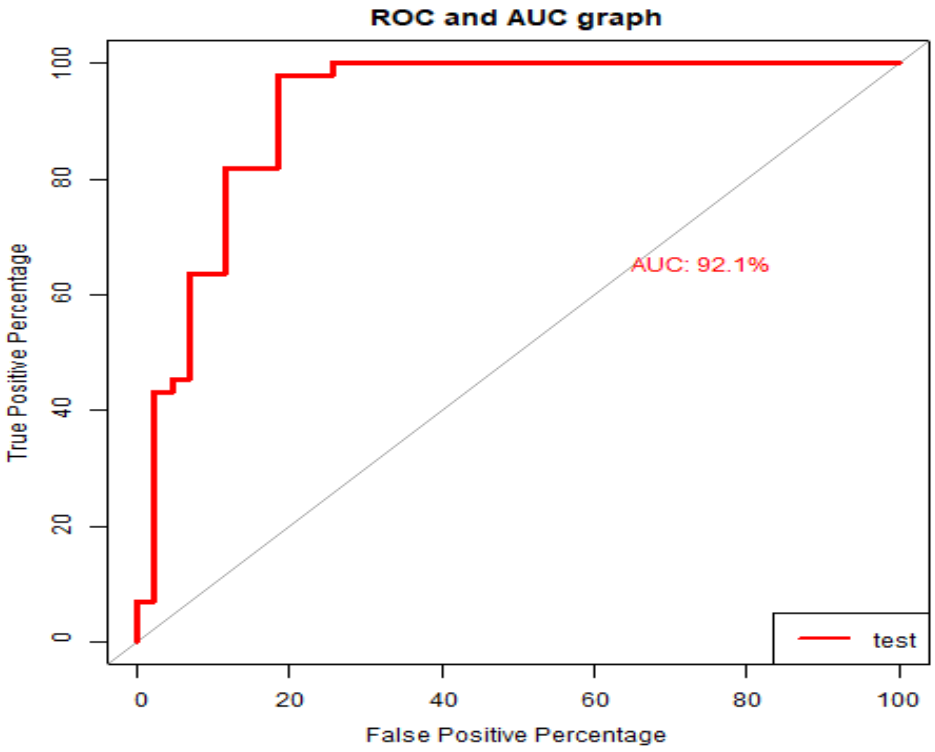
Πιο κάτω, απεικονίζονται σε σειρά οι καμπύλες ROC και ο υπολογισμός AUC του συνόλου εκπαίδευσης, επικύρωσης και αξιολόγησης



Εικόνα 3.28: Ανάλυση ROC και AUC συνόλου εκπαίδευσης (BBB) (DNN)



Εικόνα 3.29: Ανάλυση ROC και AUC συνόλου επικύρωσης (BBB) (DNN)



Εικόνα 3.30: Ανάλυση ROC και AUC συνόλου αξιολόγησης (BBB) (DNN)

4. Σχολιασμός αποτελεσμάτων και συμπεράσματα

Έχοντας παρουσιάσει τα αποτελέσματα και τις μεθόδους ΜΜ που ακολουθήθηκαν για την διάκριση των συνόλων δεδομένων, θα γίνει σχολιασμός αποτελεσμάτων μέσω παρατηρήσεων και βιβλιογραφικής ανασκόπησης. Επίσης, θα συζητηθεί κατά πόσο είναι ικανές και αξιόπιστες αυτοί οι μέθοδοι και θα αναφερθούν διάφοροι περιορισμοί και προβλήματα που προέκυψαν κατά την διαδικασία. Ουσιαστικά, σε αυτό το κεφάλαιο θα γίνει αναφορά σε κάθε σύνολο δεδομένων ξεχωριστά και θα αξιολογηθεί η ΜΜ για κάθε πεδίο που εφαρμόστηκε.

4.1 Σύνολο δεδομένων ALL - AML

Αποφασίστηκε αρχικά να γίνει συγχώνευση του συνόλου εκπαίδευσης και αξιολόγησης λόγω των λιγοστών δειγμάτων για να μπορούν οι αλγόριθμοι να γενικεύσουν καλύτερα με τα δεδομένα. Έτσι, η μέθοδος της επαναδειγματοληψίας boostrap επιλέχθηκε λόγω του ελάχιστου αριθμού δεδομένων που υπήρχαν και φαίνεται να λειτουργεί αποδοτικά. Κάτι που έπαιξε καθοριστικό ρόλο στα αποτελέσματα αυτά ήταν η μικρή ποσότητα δειγμάτων που υπήρχε και ένας άλλος παράγοντας που μεγαλώνει το σφάλμα ακρίβειας είναι η ανισορροπία που υπήρχε ανάμεσα στις δύο κλάσεις. Οπότε αν υπήρχε περισσότερος αριθμός δειγμάτων και καλύτερη ισορροπία θα είχαμε πιθανότατα πιο ακριβές αποτελέσματα. Το θέμα της ανισορροπίας αντιμετωπίστηκε χρησιμοποιώντας αύξηση δειγμάτων της μικρής κλάσης (AML) με bootstrap ώστε να γίνουν και οι 2 ισόποσες. Σημαντικό ρόλο έχει επίσης και ο θόρυβος ο οποίος θεωρείται να υπάρχει αρκετός σε δεδομένα γονιδιακών εκφράσεων. Επιπρόσθετα οι επαναλήψεις που έγιναν έδωσαν μια τυπική απόκλιση ακρίβειας περίπου στο 4, αυτό ίσως να οφείλεται στο τυχαίο διαχωρισμό (εκπαίδευσης – αξιολόγησης) κάθε επανάληψης. Δηλαδή το μοντέλο λαμβάνει κάποια δείγματα που μπορεί να τα μαθαίνει καλύτερα ή το αντίθετο. Παρόλο των εμποδίων που αναφέρθηκαν, φαίνεται η μέθοδος της ΜΜ να ταξινομεί με ικανοποιητική ακρίβεια και ιδιαίτερα τα δείγματα της θετικής κλάσης (AML). Ως θετική κλάση ορίστηκαν οι ασθενείς που πάσχουν από AML και ως αρνητική αυτοί που πάσχουν από ALL. Άρα η ευαισθησία αφορά το ποσοστό των σωστά ταξινομημένων ασθενών με AML και η ειδικότητα το ποσοστό το σωστά ταξινομημένων για πάσχοντες από ALL. Έτσι λοιπόν, ο αλγόριθμος του τυχαίου δάσους (RF) σύμφωνα με τις επιδόσεις του μαθαίνει καλύτερα τα δεδομένα αυτά λόγω της ιδιότητας που έχει να εκπαιδεύεται με bagging (επαναδειγματοληψία bootstrap στους αλγόριθμους).

Σύμφωνα με τα αποτελέσματα του συνόλου αυτού παρατηρείται 4 ολιγονουκλεοτίδια από τα 7129 να διακρίνουν καλύτερα την ασθένεια AML από την ALL, όπου αναγράφονται στο παρακάτω πίνακα μαζί με τις γονιδιακές τους εκφράσεις.

Πίνακας 4.1: Γονιδιακές εκφράσεις κάθε ολιγονουκλεοτιδίου. [84], [85], [86], [87], [88]

Ολιγονουκλεοτίδιο (probe)	**Γονιδιακές εκφράσεις**
M31303_rna1_at	ONCOPROTEIN 18/ STMN1
X16832_at	CATHEPSIN H / CTSH
X14046_at	CD37/ CD37 MOLECULE/ CD37 ANTIGEN

X13973_at	RNH1 / RIBONUCLEASE / ANGIOGENIN INHIBITOR 1

Αναλυτικά, το Oncoprotein 18/STMN1 σχετίζεται με ασθένειες όπως την οξεία Λευχαιμία, με παγκρεατικά αδενώματα, αγοραφοβία και καρκίνο των ωοθηκών [89]. Το Cathepsin H βρίσκεται σε ασθένειες όπως ναρκοληψία, μελαγχρωστική αμφιβληστροειδοπάθεια, προοδευτική μυοκλονία – επιληψία, οστική δυσπλασία και κρίση του θυροειδούς [90]. Ακολούθως, το CD37 εμπλέκεται σε λέμφωμα από μεγάλα Β – κύτταρα, ατελής οστεογένεση και χρόνια λεμφοκυτταρική λευχαιμία [91]. Το τελευταίο ολιγονουκλεοτίδιο RNH1 φαίνεται να παρουσιάζεται σε δερματίτιδες, απλή μερική επιληψία, λαρυγγοτραχειίτιδα, πνευμονική δυσλειτουργία καθώς και καρκίνο του εγκεφάλου [92], [93].

Η Melhem et al. σε μία έρευνα το 1991, σχετικά με τα χαρακτηριστικά, την δομή και την έκφραση του γονιδίου Oncoprotein 18 (Op18) στα λευχαιμικά κύτταρα, ανακάλυψαν ότι η γονιδιακή έκφραση Op18 εμφανίζεται σε αρκετά υψηλά ποσοστά στους πλείστους ασθενής με οξεία λευχαιμία [94]. Πιο συγκεκριμένα εντόπισαν ότι εμπλέκεται στην λευχαιμία στο σημείο της μεταγραφής RNA. Έπειτα, ο Zhang et al. το 2020 ανίχνευσαν ότι η γονιδιακή έκφραση CD37 συναντάτε σε πιο ψηλά ποσοστά στους ασθενείς με AML παρά στους υγιείς [95]. Παρόλο που στην βιβλιογραφία δεν βρέθηκαν στοιχεία συσχετισμού με την Λευχαιμία των υπόλοιπων δύο ολιγονουκλεοτιδίων, η ΜΜ έδειξε ότι συμβάλουν στον συγκεκριμένο δειγματικό χώρο που μελετήθηκε.

4.2 Σύνολο δεδομένων καλοήθη – κακοήθη όγκου του μαστού

Παρατηρείται η μέθοδος επαναλαμβανόμενης Κ – φορές διασταυρωμένης επικύρωσης να ταξινομεί με αρκετά υψηλή ακρίβεια και ο κυριότερος λόγος που επιλέχθηκε είναι τα ικανοποιητικά δείγματα που υπήρχαν. Επίσης είναι ξεκάθαρο από τα ποσοστά της ευαισθησίας και της ειδικότητας τα οποία είναι πανομοιότυπα, πως η ΜΜ μαθαίνει αρκετά καλά και τις 2 κλάσεις. Αναλυτικά το ποσοστό την ευαισθησίας αφορά τα σωστά ταξινομημένα δείγματα για καλοήθη και το ποσοστό της ειδικότητας τα σωστά ταξινομημένα δείγματα για κακοήθη. Αν και όλα τα μοντέλα φαίνεται να έχουν πολύ καλές αποδόσεις, το μοντέλο που έλαβε την πιο ψηλή απόδοση ήταν ο GBM. Ο κυριότερος λόγος όμως που πετυχαίνουν όλα τόσο καλή ακρίβεια, είναι η φυσική τάση των δεδομένων να διαφοροποιούνται όπως φαίνεται στο **γράφημα 3.14**. Εάν τα δεδομένα δεν είχαν αυτή τη φυσική διαφοροποίηση ίσως να μην είχαμε τόσο καλό διαχωρισμό από όλα τα μοντέλα.

Στην περίπτωση αυτή βρέθηκαν 3 μορφολογικά χαρακτηριστικά να διαχωρίζουν καλύτερα τον καλοήθη από τον κακοήθη όγκο. Πιο συγκεκριμένα, αποτελούνται και τα 3 στην κατηγορία των ακραίων τιμών. Το **«concave.points_worst»** αναφέρεται στις ανωμαλίες των κυττάρων, όπου ο καλοήθης έχει λίγες ανωμαλίες ή καθόλου στο σχήμα του ενώ αντίθετα ο κακοήθης έχει αρκετές. Μετά το **«perimeter_worst»** είναι η περίμετρος γύρω από το Πυρήνα και ο κακοήθης έχει μεγαλύτερη περίμετρο από τον καλοήθη λόγω της πιο γρήγορης του αύξησης. Τέλος, το **«texture_worst»** είναι για την τυπική απόκλιση των τιμών του γκρι και συνήθως στον καλοήθη είναι πιο συμμετρικές ενώ στον κακοήθη πιο άτυπες [96]. Έπειτα, όπως απεικονίζεται και στον τρισδιάστατο

χώρο του **γραφήματος 28,** φαίνεται αυτά τα 3 χαρακτηριστικά να διαχωρίζουν αρκετά καλά τα δεδομένα**.** Όπως αναφέρεται στην μελέτη του W. Nick Street et al. οι ακραίες τιμές είναι οι πιο χρήσιμες για την διάκριση των όγκων, καθώς μόνο μερικά κακοήθη κύτταρα μπορούν να εμφανιστούν σε ένα δείγμα [97]. Επίσης στην έρευνα της A. Derangula et al. όπου έχουν εφαρμόσει τεχνικές ΜΜ και έχουν χρησιμοποιήσει 3 αλγόριθμους: τον LGBM, τον CATBOOST και τον XGB. Στην συνέχεια έλαβαν τα 10 καλύτερα διαχωρίσιμα χαρακτηριστικά που επέλεξε ο κάθε αλγόριθμος [98]. Παρατηρείται λοιπόν ότι και στις 3 περιπτώσεις υπάρχουν τα και τα 3 μορφολογικά χαρακτηριστικά που εντοπίστηκαν στην δική μας έρευνα.

4.3 Σύνολο δεδομένων διάγνωσης Πάρκινσον

Η εφαρμογή της επαναδειγματοληψίας bootstrap έγινε λόγω των χαμηλών δειγμάτων που υπήρχαν στα δεδομένα. Ως θετική καθορίστηκε η παθολογική κλάση (πάσχοντες από Πάρκινσον) και ως αρνητική η φυσιολογική κλάση (υγιείς ασθενείς). Άρα η ευαισθησία δείχνει το ποσοστό τον σωστά ταξινομημένων δειγμάτων των ασθενών με Πάρκινσον ενώ το ποσοστό της ειδικότητας δείχνει το ποσοστό των σωστά ταξινομημένων δειγμάτων για τους υγιείς ασθενείς. Για την αντιμετώπιση της μεγάλης ανισορροπίας που υπήρχε στα δεδομένα, όπως και στο σύνολο δεδομένων ALL - AML, πραγματοποιήθηκε αύξηση της χαμηλής κλάσης (υγιείς ασθενείς) ώστε να γίνει ισόποση με την κλάση των πασχόντων από Πάρκινσον. Έτσι λοιπόν, από τον **Πίνακα 15** διαπιστώνεται ότι στην περίπτωση αυτή οι αλγόριθμοι που χρησιμοποιούν τα δέντρα αποφάσεων (GBM και RF), μαθαίνουν καλύτερα τα δεδομένα αυτά. Συγκεκριμένα, σύμφωνα με τις εκτιμήσεις της μεθοδολογίας καλύτερος φαίνεται να είναι ο RF με αρκετά ψηλά αποτελέσματα.

Από την ανάλυση των χαρακτηριστικών ηχογραφήσεων ομιλίας καταλήγετε ότι 4 χαρακτηριστικά είναι πιο αποτελεσματικά για την αναγνώριση της ασθένειας του Πάρκινσον. Προσδιοριστικά, το **«PPE»** αφορά τα "σκαμπανεβάσματα" στην φωνή και συνήθως όταν ο ασθενής νοσεί με Πάρκινσον είναι αυξημένα ενώ αντίθετα όταν είναι υγιείς μειωμένα. Το **«MDVP.Fo.Hz.»** και το **«MDVP.Fhi.Hz.»** είναι η μέση και η μέγιστη συχνότητα φωνής αντίστοιχα και συνήθως έχουν πιο χαμηλές τιμές όταν υπάρχει Πάρκινσον ενώ όταν δεν υπάρχει είναι πιο υψηλές. Το τέταρτο χαρακτηριστικό **«spread2»** σχετίζεται με ένα μη γραμμικό μέτρο της θεμελιώδους διακύμανσης της συχνότητας όπου στην νόσο αυτή συνήθως είναι πιο αυξημένο και αντίθετα όταν είναι υγιείς μειωμένο. Επιπρόσθετα, στην έρευνα του Max A. Little et al. χρησιμοποιώντας την επαναδειγματοληψία bootstrap με 50 επαναλήψεις και τον αλγόριθμο ταξινόμησης SVM αναφέρουν πως το **«PPE»** είναι από τα πιο σημαντικά χαρακτηριστικά για την διάγνωση του Πάρκινσον, λόγω του ότι αφαιρεί φυσικές παραλλαγές στην ανθρώπινη φωνή [97]. Σε μια άλλη έρευνα του Hamid Karimi Rouzbahani και Mohammad Reza Daliri χρησιμοποίησαν την τεχνική SBS όπου εντόπισαν τον καλύτερο συνδυασμό των επτά χαρακτηριστικών με βασικό κριτήριο την απόκλιση και φαίνεται τα τέσσερα από τα επτά να είναι τα ίδια με την δική μας μελέτη [100].

4.4 Σύνολο δεδομένων διαπερατότητας BBB

Στην περίπτωση αυτή αρχικά έγινε προσπάθεια χρησιμοποίησης ολόκληρου του συνόλου με την μέθοδο της K – φορές διασταυρωμένης επικύρωσης. Λόγω όμως της μη επαρκούς μνήμης RAM που υπήρχε στον υπολογιστή μας, αποφάνθηκε να χρησιμοποιηθεί το 1/6 των δειγμάτων με χρήση της επαναδειγματοληψίας bootstrap. Η ευαισθησία στο σύνολο αυτό ορίζει το ποσοστό των σωστά ταξινομημένων δειγμάτων που δεν διαπερνάνε το BBB (BBB-) ενώ η ειδικότητα το ποσοστό των σωστά ταξινομημένων δειγμάτων που διαπερνάνε το BBB (BBB+). Αναλογικά οι δύο κλάσεις στο συγκεκριμένο σύνολο είχαν καλή ισορροπία δεν χρησιμοποιήθηκε αύξηση η μείωση δειγμάτων όπως στις προηγούμενες περιπτώσεις. Από τον **Πίνακα 19** αλλά και από το **γράφημα 37**, παρατηρείται στην διαδικασία επικύρωσης, ο RF να ξεπερνά τα υπόλοιπα μοντέλα με αρκετά καλή απόκλιση στις επαναλήψεις. Έτσι αφού η διαδικασία επικύρωσης επέλεξε τον RF ως το καλύτερο μοντέλο και είχαμε και επαρκείς δεδομένα, χρησιμοποιήθηκε για την αξιολόγηση του σε ανεξάρτητα δεδομένα (Σύνολο αξιολόγησης) όπου φαίνεται να τα ταξινομεί με αρκετά καλή ακρίβεια και με προτέρημα ότι τα αποτελέσματα του είναι πανομοιότυπα με την διαδικασία επικύρωσης.

Κάτι που έπαιξε σημαντικό ρόλο είναι η προεπεξεργασία που έγινε στα δεδομένα. Επιπροσθέτως αν υπήρχε περισσότερη RAM ίσως μπορούσε να αναλυθούν όλα τα φάρμακα που είχαμε στην διάθεση μας και να είχαμε πιο αξιόπιστα αποτελέσματα αλλά και να δοκιμαστούν όλοι οι συνδυασμοί των χαρακτηριστικών που επέλεξε η RFE. Όμως αν και χρησιμοποιήθηκαν πιο λίγα δείγματα, τα αποτελέσματα δείχνουν να είναι παρόμοια με αυτά του Shaker et al. [44]. Αναλυτικά χρησιμοποίησαν την εφαρμογή dragon για την μετατροπή των smiles σε μοριακά χαρακτηριστικά όπου παράγει περισσότερα από το Padel και εκτενέστερα την K φορές αναδιπλωμένη διασταυρωμένη επικύρωση για 10 φορές με τον αλγόριθμο lightGBM. Έτσι έλαβαν αποτελέσματα ακρίβειας 89% στο σύνολο επικύρωσης και 90% στο σύνολο αξιολόγησης [44].

Βάσει των αποτελεσμάτων μας, εντοπίστηκαν 120 μοριακά χαρακτηριστικά 1ης και 2ης να είναι πιο αποτελεσματικά από τα υπόλοιπα για το διαχωρισμό των φαρμάκων, δηλαδή αν μπορούν να διαπεράσουν το BBB ή όχι. Λόγω όμως του μεγάλου τους όγκου τα διαχωρίσαμε σε 3 κατηγορίες όπως παρουσιάζονται στον **Πίνακα 20**.

Πίνακας 4.2: Χαρακτηριστικά που εντοπίστηκαν να διαχωρίζουν τη διαπερατότητα των φαρμάκων στο BBB (ΜΜ) [101]

Μοριακά Χαρακτηριστικά 1ης τάξης	
Απλά	"nO", "nBondsD2", "nBondsD", "CrippenLogP", "Mse", "AMR", "nAcid"
Μοριακά Χαρακτηριστικά 2ης τάξης	
	"ETA_Epsilon_2", "MDEO.11", "WTPT.4", "ETA_dEpsilon_D", "WTPT.3", "SM1_Dze", "SpMAD_Dzs", "WTPT.5", "SM1_Dzp", "ETA_Epsilon_4", "ETA_dEpsilon_C", "ETA_Eta_B_RC", "EE_Dzs“, "TopoPSA", "ATSC1c", "MATS1c", "ATSC0c", "ATS0s", "AATS8s", "AATS0s", "AATS2s", "AATSC0s", "GATS2s", "MATS2e",

Τοπολογικά	"AATS7s", "AATSC0i", "GATS3i", "MATS2s", "AATSC2s", "GATS2e", "MATS2c", "ATSC2e", "AATS1i", "ATSC0e", "GATS2c", "AATS6e", "AATSC2e", "AATS2s", "ATSC2c", "GATS5e", "GATS3s", "AATS6s", "AATS0e", "AATSC1p", "AATS3s", "AATSC0e", "GATS1e", "GATS2i", "ATS5s", "ATS4m", "AATS4s", "AATS5s", "ATSC1s", "AATSC3v", "MATS3v", "GATS1s", "AATSC1i“, "BCUTc.1h", "BCUTc.1l“, "MLFER_A", "MLFER_S“, "SpMax3_Bhs", "SpMax4_Bhs", "SpMax1_Bhs", "SpMax5_Bhs", "SpMax2_Bhs", "SpMin8_Bhi", "ASP.0", "nHBAcc_Lipinski", "nHBAcc", "nHBDon", "nHBDon_Lipinski“, "SHBd",
Ηλεκτροτοπολογικά	"SsOH", "minwHBa", "SdO", "maxHBd", "maxdO", "ndO", "SHsOH", "SHBint2", "MAXDN2", "SHBint5", "maxHsOH", "maxHBint2", "meanI", "mindssC", "DELS2", "maxHBint5", "DELS", "minHBint2", "gmin", "SHBa", "minHBd", "SHssNH", "minHBint5", "SdssC", "gmax", "MAXDP", "minssCH2", "maxsOH", "nHBint2", "MAXDP2", "maxHssNH", "sumI", "minHsOH", "nHBd", "minHBint4"

Για το συγκεκριμένο σύνολο δεδομένων λόγω των αρκετών δεδομένων έγινε εφαρμογή BM για όλα τα δεδομένα που υπήρχαν στην διάθεση μας. Αφού λοιπόν χρησιμοποιήσαμε περισσότερα δεδομένα από την εφαρμογή της ΜΜ, η RFE επέλεξε περισσότερα μοριακά χαρακτηριστικά $1^{ης}$ και $2^{ης}$ τάξης. Επίσης παρατηρείται η ακρίβεια του συνόλου εκπαίδευσης και του συνόλου επικύρωσης να είναι πιο ψηλά από την ΜΜ αλλά στο σύνολο αξιολόγησης να είναι περίπου η ίδια. Αυτό ενδεχομένως να οφείλεται στο 1/6 των δεδομένων που χρησιμοποιήθηκαν για ΜΜ και ίσως να προσαρμόζονται καλύτερα με τα ανεξάρτητα δεδομένα (σύνολο αξιολόγησης) από τα υπόλοιπα 5/6 που αφαιρέθηκαν. Παρόλο λοιπόν που η ακρίβεια του συνόλου αξιολόγησης είναι ίση, λόγω του ότι έγινε χρήση περισσότερων δεδομένων το μοντέλο αυτό θεωρείται πιο έμπιστο.

Ένα πρόβλημα που προέκυψε αρκετά σε αυτή τη περίπτωση είναι η υπερπροσαρμογή (overfitting), όπου το μοντέλο κατά την εκπαίδευση του συγκλίνει σχεδόν τέλεια με τα δεδομένα εκπαίδευσης ενώ στο σύνολο επικύρωσης ή και στο σύνολο αξιολόγησης η επίδοση του να είναι χαμηλή με μεγάλη απόκλιση. Αυτό αντιμετωπίστηκε τοποθετώντας dropout και κανονικοποίηση βαρών L2. Επιπρόσθετα για την εύρεση του μοντέλου αυτού ήταν μια χρονοβόρα διαδικασία λόγω της χαμηλής μνήμης RAM αλλά και των πολλών παραμέτρων που υπήρχαν για να δοκιμαστούν. Μελλοντικά αν υπάρχει στην διάθεση μας περισσότερη RAM θα μπορούσε να δοκιμαστούν εξαντλητικά πιο πολλές παράμετροι για την εύρεση του βέλτιστου μοντέλου.

Βάσει των αποτελεσμάτων μας με το μοντέλο BM, εντοπίστηκαν 223 μοριακά χαρακτηριστικά 1ης και 2ης να είναι πιο αποτελεσματικά από τα υπόλοιπα για το διαχωρισμό των φαρμάκων, δηλαδή αν μπορούν να διαπεράσουν το BBB ή όχι. Λόγω όμως του μεγάλου τους όγκου τα χωρίστηκαν σε 3 κατηγορίες όπως και πριν, όπως φαίνονται στον **Πίνακα 21**.

Πίνακας 4.3: Χαρακτηριστικά που εντοπίστηκαν να διαχωρίζουν τη διαπερατότητα των φαρμάκων στο BBB (BM) [101].

Μοριακά Χαρακτηριστικά 1ης τάξης	
Απλά	"nO", "AMR", "nAcid", "n4HeteroRing", "nT4HeteroRing", "n4Ring" "nT4Ring", “nN", "ALogp2", "nBase"
Μοριακά Χαρακτηριστικά 2ης τάξης	
Τοπολογικά	"MDEO.11", "WTPT.4", "ETA_dEpsilon_D", "WTPT.3", "SM1_Dze", "WTPT.5", "TopoPSA", "ATSC1c", "ATSC0c", "ATS0s", "AATS8s", "AATS7s", "MATS2s", "MATS2c", "AATS1i", "GATS2c", "AATS6e", "GATS5e", "AATSC1p", "GATS1e", "GATS2i", "ATS5s", "AATS5s", "ATSC1s", "AATSC3v", "GATS1s", "BCUTc.1l“, "MLFER_A", "MLFER_S", "SpMax3_Bhs", "SpMax4_Bhs", "SpMax5_Bhs", "nHBAcc_Lipinski", "nHBAcc", "nHBDon", "SHBd", "MLFER_BH", "MLFER_BO", "ATSC4v", "SCH.5", "MDEN.22", "SRW5", "AATSC1c", "VE3_Dt", "ATSC4m", "GATS4v", "VCH.5", "MATS1p", "MLFER_E", "MATS1c", "AATSC4v", "SpMax4_Bhm", "nHBAcc2", "VE3_Dzi", "VE3_Dze", "MLFER_L", “ATSC3v”, “ATSC3p”, “MATS4m”, “AATS2i", “IC1”, “VE3_Dzv", “MDEC.24", “GATS4i", “VCH.6", “SM1_DzZ”, “SpMax2_Bhe”, “SM1_Dzm", “Kier3”, “SCH.6”, “AATSC0c", “AATS7v", “AATSC4m", “MATS1e", “ATSC2s", “SpMax2_Bhi”, “VC.5”, “GATS5m”, "JGI2", “MATS8c", “SpMax6_Bhs", “VCH.7”, “GGI3", “SM1_Dzv”, “ATSC0s", “SpMax5_Bhp”, “VC.6", “ASP.7”, “VE1_Dt", “MATS7c", “VE3_DzZ", “SpMax3_Bhi", “VE3_Dzp”, “SpMax5_Bhv”, “VE3_Dzm", “GATS1m”, “SpMax5_Bhm", “SpMax5_Bhi”, “VPC.4”, “ASP.6", “GATS6e", “MATS1i", “GATS8p”, “ATSC8c", “AATS5e", “VPC.5", “SpMin1_Bhm", “SpMax7_Bhs", “GATS2p", “SpMax2_Bhp", “MATS3v", “ATSC8s", “SC.6”, “ETA_Eta_F", “AATS6v", “ETA_EtaP_F”, “SpMin5_Bhp”, "SpMin2_Bhv", “ATSC7c", “GATS6i", “MDEC.44”, “MATS5s", “VPC.6", “SpMin2_Bhp", “SRW10”, “ETA_Eta_B",

	“MATS8s", “SpMax5_Bhe”, “GATS1c”, “AATS5v", “ATSC1i", “GATS1i", “VC.3", “ATSC5c", “JGI3", “ASP.4", “ETA_Eta_F_L”, “MATS1v", “AATS0i".
Ηλεκτροτοπολογικά	"SsOH", "minwHBa", "SdO", "maxHBd", "SHsOH", "SHBint2", "MAXDN2", "SHBint5", "maxHsOH", "maxHBint2", "meanI", "mindssC", "DELS2", "maxHBint5", "DELS", "minHBd", "SHssNH", "SdssC", "MAXDP", "maxsOH", "MAXDP2", minssssNp", "maxssssNp","SssssNp", "minHsOH", "maxHBint8", "nHBd", "SHBint8", "nHBint3", "SHBint6", "minHBa", "maxHBint7", "SHBint7", "SHBint3", "minHBint4", "maxHBint3", "SHBint10", "minHBint8", "nHBa", "minHBint6","minsOH", “maxHBint10”, “SssNH”, “nHBint6", “nHBint8", “SHBint4”, “SHaaNH", “MAXDN”, “nHsOH", “minHBint7", “nsOH”, “minHaaNH", “maxHBa", “nHBint4", “maxHaaNH", “nHBint10”, “ndssC", “maxHBint4", “maxssNH", “minddssS", “minsssN”, “maxHCsats", “LipoaffinityIndex"

Όσον αφορά τα μοριακά χαρακτηριστικά 1ης τάξης (Απλά) παρουσιάζουν πληροφορίες υπολογίζονται από την μοριακή φόρμουλα των μορίων, οι οποίες εμπεριέχουν το ποσό και τον τύπο των ατόμων σε ένα μόριο και το μοριακό βάρος [102] π.χ. το **«nO»** είναι ο αριθμός ατόμων οξυγόνου και το **«nBondsD2»** ο αριθμός των διπλών δεσμών. Τα μοριακά χαρακτηριστικά 2ης τάξης είναι πιο πολύπλοκα και μας δείχνουν πληροφορίες σχετικά με το μέγεθος, το σχήμα και την ηλεκτρονική κατανομή του μορίου [102]. Χωρίζονται σε δύο κατηγορίες, η πρώτη είναι τα τοπολογικά που αφορούν την περιγραφή και την πρόβλεψη της μοριακής δομής, επεξηγούν πως τα μόρια φτάνουν στο τελικό τους σχήμα και πως επιτυγχάνουν τις δραστηριότητες τους [103] π.χ. το **«ETA_Epsilon_2»** είναι ένα μέτρο του αριθμού των ηλεκτροαρνητικών ατόμων. Η δεύτερη κατηγορία που ονομάζεται ηλεκτροτοπολογικά κωδικοποιούν το ηλεκτρονικό και το τοπολογικό περιβάλλον των ατόμων. Επίσης κωδικοποιούν των αριθμό ηλεκτρονίων και έχουν ως ιδιότητα να προσδιορίζουν τα βασικά δομικά χαρακτηριστικά που συμβάλλουν στις μοριακές δραστηριότητες [103] π.χ. το **«SsOH»** είναι το άθροισμα της ενδογενής ηλεκτρονικής κατάστασης ενός ατόμου με δεσμό υδροξειδίου

4.5 Γενικό συμπέρασμα

Σαν ένα γενικό συμπέρασμα για την ΜΜ βάσει των αποτελεσμάτων που λήφθηκαν, μπορεί να λεχθεί ότι είναι αρκετά αξιόπιστη και πολλές φορές μπορεί να μας αποκαλύπτει κρυφές πληροφορίες οι οποίες δεν έχουν εντοπιστεί από την επιστημονική κοινότητα. Δείχνει να είναι αρκετά ευέλικτη και ένας από τους κυριότερους λόγους είναι ότι μας δίνει την δυνατότητα αξιολόγησης πολλαπλών αλγορίθμων ταυτόχρονα.

Καθοριστικό ρόλο έχουν επίσης οι ενέργειες που γίνονται για προεπεξεργασία των δεδομένων όπου συμβάλουν αρκετά στην αύξηση των επιδόσεων της. Σε αυτή τη διπλωματική λοιπόν είδαμε πόσο χρήσιμες είναι αλλά και πως επιδρούν αυτές οι μέθοδοι σε διαφορετικά πεδία της Βιοϊατρικής έρευνας όπου υπάρχουν ανάγκες για προβλέψεις και επίλυση προβλημάτων ταξινόμησης. Από τις επιδόσεις της είναι εμφανές ότι είναι μια τεχνική χαμηλού κόστους που έχει προοπτικές να γίνει αρκετά χρήσιμη και να αντικαταστήσει χρονοβόρες και ελαττωματικές μεθόδους αλλά και επεμβατικές τεχνικές που είναι επιβλαβές σε ασθενείς. Συνεπώς διαφαίνεται ότι οι μέθοδοι ΜΜ μπορούν να φανούν χρήσιμες και μελλοντικά να παρέχουν πολύτιμη βοήθεια στην βελτίωση των υπηρεσιών υγείας αλλά και στην ανακάλυψη νέων γνώσεων για την Βιοϊατρική έρευνα. Για να διασφαλιστεί όμως η ποιότητα της ΜΜ στους τομείς αυτούς καλύτερα θα ήταν σε μελλοντικές έρευνες να γίνουν αναλύσεις με περισσότερα δείγματα ούτως ώστε να υπάρχουν πιο αξιόπιστα αποτέλεσμα. Στο μέλλον λοιπόν θα μπορούσαμε να επεκτείνουμε τα σύνολα δεδομένων που έχουμε στην διάθεση μας με περισσότερα δεδομένα και με χρήση περισσότερης RAM να καταφέρουμε να λάβουμε πιο αξιόπιστα αποτελέσματα. Επίσης, έχουμε ως στόχο να δοκιμαστούν οι τεχνικές αυτές και σε άλλα βιολογικά πεδία για να σχηματίσουμε μια ολική εικόνα για τις εφαρμογές ΜΜ στην Βιοϊατρική έρευνα.

Αναφορές – Πηγές

[1] P. Hogeweg, "The Roots of Bioinformatics in Theoretical Biology", PLoS Computational Biology, vol. 7, no. 3, p. e1002021, 2011. Available: 10.1371/journal.pcbi.1002021.

[2] S. V, N. S and P. M, "AN EMPIRICAL SCIENCE RESEARCH ON BIOINFORMATICS IN MACHINE LEARNING", JOURNAL OF MECHANICS OF CONTINUA AND MATHEMATICAL SCIENCES, vol. 7, no. 1, 2020. Available: 10.26782/jmcms.spl.7/2020.02.00006.

[3] Tan and D. Gilbert, "Ensemble machine learning on gene expression data for cancer classification", *Bura.brunel.ac.uk*, 2021. [Online]. Available: https://bura.brunel.ac.uk/handle/2438/3013.

[4] A. Choon Tan and D. Gilbert, "Machine Learning and its Application to Bioinformatics: An Overview", *Bioinformatics*, 2003.

[5] P. Larrañaga et al., "Machine learning in bioinformatics", Briefings in Bioinformatics, vol. 7, no. 1, pp. 86-112, 2006. Available: 10.1093/bib/bbk007 [Accessed 20 May 2021].

[6] M. Libbrecht and W. Noble, "Machine learning applications in genetics and genomics", Nature Reviews Genetics, vol. 16, no. 6, pp. 321-332, 2015. Available: 10.1038/nrg3920.

[7] S. Kogenaru, O. Yan, Y. Guo and N. Wang, "RNA-seq and microarray complement each other in transcriptome profiling", bmcgenomics.biomedcentral.com, 2012. [Online]. Available: https://doi.org/10.1186/1471-2164-13-629. [Accessed: 22- Aug- 2021].

[8] Α. Μαλατράς, Core.ac.uk, 2010. [Online]. Available: https://core.ac.uk/download/pdf/132817731.pdf. [Accessed: 28- Apr- 2021].

[9] S. Zhao, W. Fung-Leung, A. Bittner, K. Ngo and X. Liu, "Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells", journals.plos.org, 2014. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0078644. [Accessed: 23- Aug- 2021].

[10] A. Jabeen, N. Ahmad and K. Raza, "Machine Learning-Based State-of-the-Art Methods for the Classification of RNA-Seq Data", link.springer.com, 2017. [Online]. Available: https://link.springer.com/chapter/10.1007%2F978-3-319-65981-7_6. [Accessed: 22- Aug- 2021].

[11] A. Swan, A. Mobasheri, D. Allaway, S. Liddell and J. Bacardit, "Application of Machine Learning to Proteomics Data: Classification and Biomarker Identification in Postgenomics Biology", OMICS: A Journal of Integrative Biology, vol. 17, no. 12, pp. 595-610, 2013. Available: 10.1089/omi.2013.0017.

[12] M. Rafea, P. Elkafrawy, M. Nasef, R. Elnemr and A. Jamal, "Applying Machine Learning of Erythrocytes Dynamic Antigens Store in Medicine", Frontiers in Molecular Biosciences, vol. 6, 2019. Available: 10.3389/fmolb.2019.00019.

[13] Yao, Z. and Ruzzo, W., 2006. *A Regression-based K nearest neighbor algorithm for gene function prediction from heterogeneous data*. [online]

BMCbioinformatics. Available at: <https://doi.org/10.1186/1471-2105-7-S1-S11>

[14] Srinivasa, K., Siddesh, G. and Manisekhar, S., n.d. *Statistical modelling and machine learning principles for bioinformatics techniques, tools, and applications*. pp.63-73.

[15] S. Louca and M. Doebeli, "Efficient comparative phylogenetics on large trees", *Bioinformatics*, 2018. [Online]. Available: http://doi: 10.1093/bioinformatics/btx701. [Accessed: 14- May- 2021].

[16] A. Zheng and A. Casari, *Feature engineering for machine learning*. pp. 1-2.

[17] R. Wallis et al., "Tuberculosis biomarkers discovery: developments, needs, and challenges", *Elsevier*, 2013. [Online]. Available: http://doi:10.3390/metabo10060243. [Accessed: 14- May- 2021].

[18] I. Inza, B. Calvo, R. Armananzas, E. Bengoetxea, P. Larranaga and J. Lozano, "Machine Learning: An Indispensable Tool in Bioinformatics", *Springer*, 2021. [Online]. Available: https://doi.org/10.1007/978-1-60327-194-3_2. [Accessed: 13- May- 2021].

[19] www.javatpoint.com. n.d. *K-Nearest Neighbor(KNN) Algorithm for Machine Learning - Javatpoint*. [online] Available at: https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

[20] www.javatpoint.com. n.d. *Bayesian Belief Network in Artificial Intelligence - Javatpoint*. [online] Available at: https://www.javatpoint.com/bayesian-belief-network-in-artificial-intelligence

[21] Lancashire, L., Lemetre, C. and Ball, G., 2009. *An introduction to artificial neural networks in bioinformaticsçapplication to complex microarray and mass spectrometry datasets in cancer studies*. [online] Academic. Available at: <http://doi:10.1093/bib/bbp012>. [18]

[22] J. Brownlee, *Machine Learning Mastery with R. pp 70-75*

[23] G. Bland, "Train/Test Split and Cross Validation - A Python Tutorial - AlgoTrading101 Blog", *Quantitative Trading Ideas and Guides - AlgoTrading101 Blog*. [Online]. Available: https://algotrading101.com/learn/train-test-split/. [Accessed: 10- May- 2021].

[24] S. Min, B. Lee and S. Yoon, "Deep Learning in Bioinformatics", *academic.oup.com*, 2017. [Online]. Available: https://doi.org/10.1093/bib/bbw068. [Accessed: 31- May- 2021].

[25] M. Cilimkovic, "Neural Networks and Back Propagation Algorithm", *academia.edu*, 2015. [Online]. [Accessed: 02- Jun- 2021].

[26] D. Shalabi and D. Shaaban, "Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix", *Ieeexplore.ieee.org*, 2006. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/4024051. [Accessed: 02- Jun- 2021].

[27] P. Baldi and P. Sadowski, "Understanding Dropout", *Papers.nips.cc*, 2021. [Online]. Available: https://papers.nips.cc/paper/2013/file/71f6278d140af599e06ad9bf1ba03cb0-Paper.pdf. [Accessed: 05- Jun- 2021].

[28] F. Chollet and J. Allaire, *Deep Learning with R*, 1st ed. https://livebook.manning.com/book/deep-learning-with-r/about-this-book/, 2017, pp. 103-104.

[29] M. Awad and R. Khanna, Efficient Lerning Machines. https://library.oapen.org/bitstream/handle/20.500.12657/28170/1001824.pdf?sequence=1, pp. 52-53.

[30] J. Landis and G. Koch, "The Measurement of Observer Agreement for Categorical Data on JSTOR", Jstor.org, 2021. [Online]. Available: https://www.jstor.org/stable/2529310. [Accessed: 16- Aug- 2021].

[31] D. Chicco, N. Totsch and G. Jurman, "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation", ncbi.nlm.nih.gov, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7863449/. [Accessed: 16- Aug- 2021].

[32] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation", link.springer.com, 2020. [Online]. Available: https://link.springer.com/article/10.1186/s12864-019-6413-7. [Accessed: 16- Aug- 2021].

[33] D. Powers, "(PDF) Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation", ResearchGate, 2014. [Online]. Available: https://www.researchgate.net/publication/228529307_Evaluation_From_Precision_Recall_and_F-Factor_to_ROC_Informedness_Markedness_Correlation. [Accessed: 17- Aug- 2021].

[34] H. Inaba and C. Mullighan, "Pediatric acute lymphoblastic leukemia", *Haematologica*, vol. 105, no. 11, pp. 2524-2539, 2020. Available: 10.3324/haematol.2020.247031 [Accessed 11 May 2021].

[35] H. Salah, I. Muhsen, M. Salama, T. Owaidah and S. Hashmi, "Machine learning applications in the diagnosis of leukemia: Current trends and future directions", *International Journal of Laboratory Hematology*, vol. 41, no. 6, pp. 717-725, 2019. Available: 10.1111/ijlh.13089 [Accessed 13 May 2021].

[36] O. Gal, N. Auslander, Y. Fan and D. Meerzaman, "Predicting Complete Remission of Acute Myeloid Leukemia: Machine Learning Applied to Gene Expression", *Cancer Informatics*, vol. 18, p. 117693511983554, 2019. Available: 10.1177/1176935119835544

[37] Osareh, A. and Shadgar, B., 2010. Machine learning techniques to diagnose breast cancer. [online] Ieeexplore.ieee.org. Available at: <https://ieeexplore.ieee.org/abstract/document/5478895> [Accessed 6 May 2021].

[38] L. Fayed, "Differences Between a Malignant and Benign Tumor", *Verywell Health*, 2021. [Online]. Available: https://www.verywellhealth.com/what-does-malignant-and-benign-mean-514240. [Accessed: 06- May- 2021].

[39] S. Sadhukhan, N. Upadhyay and P. Chakraborty, "Breast Cancer Diagnosis Using Image Processing and Machine Learning", *Springer*, 2020. [Online]. Available: https://doi.org/10.1007/978-981-13-7403-6_12. [Accessed: 06- May- 2021].

[40] Lhotská, L., Sukupova, L., Lacković, I. and Ibbott, G., 2018. World Congress on Medical Physics and Biomedical Engineering 2018. Springer, pp.197-200.

[41] W. Poewe et al., "Parkinson disease", *nature*, 2017. [Online]. Available: https://www.nature.com/articles/nrdp201713. [Accessed: 06- May- 2021].

[42] C. Sakar et al., "A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform", *Elsevier*, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1568494618305799?via%3Dihub. [Accessed: 06- May- 2021].

[43] Y. Wu et al., "Dysphonic Voice Pattern Analysis of Patients in Parkinson's Disease Using Minimum Interclass Probability Risk Feature Selection and Bagging Ensemble Learning Methods", *Hindawi*, 2017. [Online]. Available: https://www.hindawi.com/journals/cmmm/2017/4201984/. [Accessed: 07- May- 2021].

[44] B. Shaker et al., "LightBBB: Computational prediction model of blood-brainbarrier penetration based on LightGBM", *Oxford Academic*, 2020. [Online]. Available: https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/btaa918/5942084?redirectedFrom=fulltext. [Accessed: 08- May- 2021].

[45] R. Miao, L. Xia, H. Chen, H. Huang and Y. Liang, "Improved Classification of Blood-Brain-Barrier Drugs Using Deep Learning", *Scientific Reports*, 2019. [Online]. Available: https://www.nature.com/articles/s41598-019-44773-4. [Accessed: 08- May- 2021].

[46] L. Ferreira, "What human blood-brain barrier models can tell us about BBB function and drug discovery?", *Taylor & Francis*, 2019. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/17460441.2019.1646722. [Accessed: 08- May- 2021].

[47] C. Lipinski, "Lead- and drug-like compounds: the rule-of-five revolution", *Elsevier*, 2004. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1740674904000551. [Accessed: 07- May- 2021].

[48] R-project.org. n.d. *R: What is R?*. [online] Available at: <https://www.r-project.org/about.html> [Accessed 9 May 2021]

[49] Rstudio.com. n.d. *RStudio | Open source & professional software for data science teams*. [online] Available at: <https://www.rstudio.com/> [Accessed 9 May 2021].nb

[50] Kuhn, M., Johnson, K. and Analytics, A., n.d. *Applied Predictive Modeling*. Springer, pp.563-565.

[51] "R Interface to Keras", *Keras.rstudio.com*. [Online]. Available: https://keras.rstudio.com/. [Accessed: 31- May- 2021].

[52] J. Allaire, "TensorFlow for R", *Tensorflow.rstudio.com*. [Online]. Available: https://tensorflow.rstudio.com/. [Accessed: 09- Jun- 2021].

[53] Wickham, H. and Sievert, C., 2016. *Ggplot2*. 2nd ed. Dordrecht [etc.]: Springer, pp.3-4.

[54] E. Szöcs, "Chemical Information from the Web [R package webchem version 1.1.1]", *Cran.r-project.org*, 2021. [Online]. Available: https://cran.r-project.org/web/packages/webchem/index.html. [Accessed: 25- May- 2021].

[55] C. Crawford, "Gene expression dataset (Golub et al.)", *Kaggle.com*, 2017. [Online]. Available: https://www.kaggle.com/crawford/gene-expression. [Accessed: 20- May- 2021].

[56] T. Golub et al., "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring", *Science*, 1999. [Online]. Available: http://DOI: 10.1126/science.286.5439.531. [Accessed: 20- May- 2021].

[57] "Breast Cancer Wisconsin (Diagnostic) Data Set", *Kaggle.com*, 2016. [Online]. Available: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data. [Accessed: 20- May- 2021].

[58] "UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set", *Archive.ics.uci.edu*, 1995. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29. [Accessed: 20- May- 2021].

[59] D. Biswas, "Parkinson's Disease (PD) classification", *Kaggle.com*, 2019. [Online]. Available: https://www.kaggle.com/dipayanbiswas/parkinsons-disease-speech-signal-features. [Accessed: 20- May- 2021].

[60] M. Adenot and R. Lahana, "Blood-Brain Barrier Permeation Models: Discriminating between Potential CNS and Non-CNS Drugs Including P-Glycoprotein Substrates", *Pubs.acs.org*, 2004. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/ci034205d. [Accessed: 25- May- 2021].

[61] Z. Gao, Y. Chen, X. Cai and R. Xu, "Predict drug permeability to blood–brain-barrier from clinical phenotypes: drug side effects and drug indications", *Bioinformatics*, 2016. [Online]. Available: https://doi.org/10.1093/bioinformatics/btw713. [Accessed: 25- May- 2021].

[62] I. Martins, A. Teixeira, L. Pinheiro and A. Falcao, "A Bayesian Approach to in Silico Blood-Brain Barrier Penetration Modeling", *Pubs.acs.org*, 2012. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/ci300124c. [Accessed: 25- May- 2021].

[63] F. Plisson and A. Piggott, "Predicting Blood–Brain Barrier Permeability of Marine-Derived Kinase Inhibitors Using Ensemble Classifiers Reveals Potential Hits for Neurodegenerative Disorders", *mdpi.com*, 2019. [Online]. Available: https://www.mdpi.com/1660-3397/17/2/81. [Accessed: 25- May- 2021].

[64] M. Singh, R. Divakaran, L. Konda and R. Kristam, "A classification model for blood brain barrier penetration", *Elsevier*, 2020. [Online]. Available: https://doi.org/10.1016/j.jmgm.2019.107516. [Accessed: 25- May- 2021].

[65] Z. Wang et al., "In Silico Prediction of Blood–Brain Barrier Permeability of Compounds by Machine Learning and Resampling Methods", *Chemistry Europe*, 2018. [Online]. Available: https://doi.org/10.1002/cmdc.201800533. [Accessed: 25- May- 2021].

[66] Y. Yuan, F. Zheng and C. Zhan, "Improved Prediction of Blood–Brain Barrier Permeability Through Machine Learning with Combined Use of Molecular

Property-Based Descriptors and Fingerprints", *Springer*, 2018. [Online]. Available: https://doi.org/10.1208/s12248-018-0215-8. [Accessed: 25- May- 2021].

[67] "PubChem", *Pubchem.ncbi.nlm.nih.gov*. [Online]. Available: https://pubchem.ncbi.nlm.nih.gov/. [Accessed: 26- May- 2021].

[68] "About PubChem", *Pubchemdocs.ncbi.nlm.nih.gov*. [Online]. Available: https://pubchemdocs.ncbi.nlm.nih.gov/about. [Accessed: 26- May- 2021].

[69] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules", pubs.acs.org, 1988 [Online]. Available: https://pubs.acs.org/doi/pdf/10.1021/ci00057a005. [Accessed: 20- May- 2021].

[70] Y. Wei, "PaDEL-Descriptor", *Yapcwsoft.com*, 2014. [Online]. Available: http://www.yapcwsoft.com/dd/padeldescriptor/. [Accessed: 25- May- 2021].

[71] J. Lu, S. Hardi, W. Tao, S. Muse, B. Weir and S. Spruill, "CLASSICAL STATISTICAL APPROACHES TO MOLECULAR CLASSIFICATION OF CANCER FROM GENE EXPRESSION PROFILING", 2002. [Online]. Available: https://doi.org/10.1007/978-1-4615-0873-1_8. [Accessed: 16- May- 2021].

[72] S. Shapiro and M. Wilk, "An Analysis of Variance Test for Normality (Complete Samples)", *jstor.org*, 1965. [Online]. Available: https://www.jstor.org/stable/2333709?seq=1. [Accessed: 25- May- 2021].

[73] T. Kim, "T test as a parametric statistic", *NCBI*, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4667138/. [Accessed: 17- May- 2021].

[74] R. Woolson, "Wilcoxon Signed-Rank Test", *Wiley Online Library*, 2008. [Online]. Available: https://doi.org/10.1002/9780471462422.eoct979. [Accessed: 17- May- 2021].

[75] "Pearson's correlation coefficient", *thebmj*, 2012. [Online]. Available: https://www.bmj.com/content/345/bmj.e4483. [Accessed: 17- May- 2021].

[76] Q. Chen, Z. Meng, X. Liu, Q. Jin and R. Su, "Decision Variants for the Automatic Determination of Optimal Feature Subset in RF-RFE", *mdpi.com*, 2018. [Online]. Available: https://doi.org/10.3390/genes9060301. [Accessed: 25- May- 2021].

[77] P. Xanthopoulos, P. Pardalos and T. Trafalis, "Linear Discriminant Analysis", 2013. [Online]. Available: https://doi.org/10.1007/978-1-4419-9878-1_4. [Accessed: 18- May- 2021].

[78] Y. Qi, "Random Forest for Bioinformatics", *Springer*, 2012. [Online]. Available: https://doi.org/10.1007/978-1-4419-9326-7_11. [Accessed: 18- May- 2021].

[79] B. Ripley, "Feed-Forward Neural Networks and Multinomial Log-Linear Models [R package nnet version 7.3-16]", *Cran.r-project.org*, 2021. [Online]. Available: https://cran.r-project.org/web/packages/nnet/index.html. [Accessed: 19- May- 2021].

[80] S. Chen, G. Webb, L. Liu and X. Ma, "A novel selective naïve Bayes algorithm", *Elsevier*, 2020. [Online]. Available: https://doi.org/10.1016/j.knosys.2019.105361. [Accessed: 19- May- 2021].

[81] D. Shalabi and D. Shaaban, "Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix", *Ieeexplore.ieee.org*, 2006. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/4024051. [Accessed: 02- Jun- 2021].

[82] A. Naketin and A. Knoll, "Gradient boosting machines, a tutorial", *Frontiers*, 2013. [Online]. Available: https://doi.org/10.3389/fnbot.2013.00021. [Accessed: 19- May- 2021].

[83] A. Tharwat, "Inderscience Publishers - linking academia, business and industry through research", *Inderscience.com*, 2016. [Online]. Available: http://www.inderscience.com/offer.php?id=79050. [Accessed: 19- May- 2021].

[84] D. Ling Tong and G. R Ball, "Exploration of leukemia gene regulatory networks using a systems biology approach", *Ieeexplore.ieee.org*, 2014. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6999250. [Accessed: 10- Jul- 2021].

[85] T. Hellem Bø and I. Jonassen, "New feature subset selection procedures for classification of expression profiles", *genomebiology.biomedcentral.com*, 2002. [Online]. Available: https://doi.org/10.1186/gb-2002-3-4-research0017. [Accessed: 10- Jul- 2021].

[86] S. Leydold et al., "Peroxireduxin-4 is Over-Expressed in Colon Cancer and its Down-Regulation Leads to Apoptosis - Sandra M. Leydold, Michael Seewald, Christian Stratowa, Klaus Kaserer, Wolfgang Sommergruber, Norbert Kraut, Norbert Schweifer, 2011", *SAGE Journals*, 2011. [Online]. Available: https://journals.sagepub.com/doi/10.4137/CGM.S6584. [Accessed: 10- Jul- 2021].

[87] B. Charles E. MD, Z. Lizhi MD and G. Joel K. MD, "The Molecular Basis of Pancreatic Fibrosis: Common Stromal... : Pancreas", *LWW*, 2004. [Online]. Available: https://journals.lww.com/pancreasjournal/Abstract/2004/11000/The_Molecular_Basis_of_Pancreatic_Fibrosis__Common.3.aspx. [Accessed: 10- Jul- 2021].

[88] W. Yu and T. Park, "AucPR: An AUC-based approach using penalized regression for disease prediction with high-dimensional omics data", *bmcgenomics.biomedcentral.com*, 2014. [Online]. Available: https://doi.org/10.1186/1471-2164-15-S10-S1. [Accessed: 10- Jul- 2021].

[89] *Genecards.org*. [Online]. Available: https://www.genecards.org/cgi-bin/carddisp.pl?gene=STMN1#diseases. [Accessed: 10- Jul- 2021].

[90] *Genecards.org*. [Online]. Available: https://www.genecards.org/cgi-bin/carddisp.pl?gene=CTSH&keywords=CTSH#diseases. [Accessed: 10- Jul- 2021].

[91] *Genecards.org*. [Online]. Available: https://www.genecards.org/cgi-bin/carddisp.pl?gene=CD37&keywords=CD37#diseases. [Accessed: 10- Jul- 2021].

[92] *Genecards.org*. [Online]. Available: https://www.genecards.org/cgi-bin/carddisp.pl?gene=RNH1&keywords=RIBONUCLEASE#diseases. [Accessed: 10- Jul- 2021].

[93] "RNH1 Antibodies, cDNA Clones Research Reagents | Sino Biological", *Kr.sinobiological.com*. [Online]. Available:

https://kr.sinobiological.com/category/rnh1?fbclid=IwAR3H5uS83c4VMcOMmt9rwBBxF8j8BYd1VD1SAe_aXQliw6hzZnLw849hqgE. [Accessed: 10- Jul- 2021].

[94] R. Melhem, X. Zhu, N. Hailat, J. Strahler and S. Hanash, "Characterization of the gene for a proliferation-related phosphoprotein (oncoprotein 18) expressed in high amounts in acute leukemia.", *Journal of Biological Chemistry*, vol. 266, no. 27, pp. 17747-17753, 1991. Available: 10.1016/s0021-9258(18)55189-9.

[95] Q. Zhang, Q. Han, J. Zi, C. Song and Z. Ge, "CD37 high expression as a potential biomarker and association with poor outcome in acute myeloid leukemia", *Bioscience Reports*, vol. 40, no. 5, 2020. Available: 10.1042/bsr20200008.

[96] A. Narasimha, B. Vasani and H. Kumar, "Significance of nuclear morphometry in benign and malignant breast aspirates", *ncbi.nlm.nih.gov*, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3678677/. [Accessed: 14- Jun- 2021].

[97] W. Street, W. Wolberg and O. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis", *spiedigitallibrary.org*, 2015. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/1905/1/Nuclear-feature-extraction-for-breast-tumor-diagnosis/10.1117/12.148698.short?SSO=1. [Accessed: 25- Jun- 2021].

[98] A. Derangula, P. Edara and P. Karri, "Feature Selection of Breast Cancer Data Using Gradient Boosting Techniques of Machine Learning", 2020. [Online]. Available: https://ejmcm.com/article_2569_a7c3766658c69272f70820a964a9cd36.pdf. [Accessed: 26- Jun- 2021].

[99] M. Little, P. McSharry, E. Hunter, J. Spielman and L. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease", *Nature.com*, 2008. [Online]. Available: https://www.nature.com/articles/npre.2008.2298.1. [Accessed: 29- Jun- 2021].

[100] H. Rouzbahani and M. Daliri, "Diagnosis of Parkinson's Disease in Human Using Voice Signals", *research-management.mq.edu.au*, 2011. [Online]. Available: https://research-management.mq.edu.au/ws/portalfiles/portal/92726924/92697802.pdf. [Accessed: 05- Jul- 2021].

[101] *chemdes*. [Online]. Available: http://www.scbdd.com/padel_desc/descriptors/. [Accessed: 29- Mar- 2021].

[102] B. Chandrasekaran, S. Nidal Abed, O. Al-Attraqchi, K. Kuche and R. K. Tekade, "Computer-Aided Prediction of Pharmacokinetic (ADMET) Properties", *sciencedirect.com*, 2018. [Online]. Available: https://doi.org/10.1016/B978-0-12-814421-3.00021-X. [Accessed: 30- Jun- 2021].

[103] S. Basak and G. Restrepo, "Editorial: Topological and electrotopological descriptors of molecules: fundamental principles and applications to computer aided molecular design - Part I", *pubmed.ncbi.nlm.nih.gov*, 2012. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/22497464/. [Accessed: 29- Mar- 2021].