



ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

με θέμα

Εφαρμογή Τεχνικών Μηχανικής Μάθησης για τη Διάγνωση της Λευχαιμίας
με Δεδομένα από Εικόνες Μικροσυστοιχιών

ΓΚΟΥΝΤΕΝΟΥΔΗ ΒΑΣΙΛΙΚΗ-ΟΛΥΜΠΙΑ

Π2015090

Επιβλέπων Καθηγητής:

ΕΞΑΡΧΟΣ ΘΕΜΗΣ

Κέρκυρα, Οκτώβριος 2020

Πτυχιακή Εργασία

Εφαρμογή Τεχνικών Μηχανικής Μάθησης για τη Διάγνωση της Λευχαιμίας με Δεδομένα από Εικόνες Μικροσυστοιχιών

Περίληψη

Η εργασία εστιάζεται στην Ανάλυση Δεδομένων από εικόνες μικροσυστοιχιών με Τεχνικές Μηχανικής Μάθησης για την ασθένεια της λευχαιμίας. Γίνεται εκτενής αναφορά στην ασθένεια της λευχαιμίας (συμπτώματα, διάγνωση, θνησιμότητα, ποιότητα ζωής των ασθενών, τρόποι αντιμετώπισης) καθώς και στην συμβολή της τεχνολογίας μικροσυστοιχιών στον τομέα της Βιοπληροφορικής και κατ' επέκταση στην λευχαιμία. Εφαρμόστηκαν τεχνικές εξόρυξης δεδομένων σε τρία σύνολα δεδομένων, από ένα δείγμα 72 ασθενών. Πιο συγκεκριμένα έγινε προ-επεξεργασία των δεδομένων και στη συνέχεια εφαρμογή αλγορίθμων μηχανικής μάθησης, ώστε να αποφασιστεί ποιος από τους επιλεγμένους αλγορίθμους πετυχαίνει καλύτερα αποτελέσματα (προβλέψεις) στην ταξινόμηση των τύπων λευχαιμίας των ασθενών (ALL, AML, B-cell, T-cell, BM, PB). Για την έρευνα χρησιμοποιήθηκαν τρία διαφορετικά σύνολα γονιδιακών δεδομένων και με τη βοήθεια του εργαλείου weka προσεγγίστηκαν τα γονίδια που φαίνεται ότι επιτελούν τον σημαντικότερο ρόλο για την αναγνώριση κατηγορίας/τύπου λευχαιμίας.

Το πείραμα πραγματοποιήθηκε εφαρμόζοντας τους αλγορίθμους μόνο για τα σημαντικά χαρακτηριστικά που είχαν προκύψει αρχικά για τα είκοσι πρώτα, έπειτα για τα πενήντα πρώτα γονίδια. Τα αποτελέσματα αξιολογήθηκαν με τα μέτρα της ακρίβειας, εξειδίκευσης, ευαισθησίας καταλήγοντας στην εύρεση των αποτελεσματικότερων αλγορίθμων και των πιο σημαντικών χαρακτηριστικών που επηρεάζουν τελικά την κατηγοριοποίηση των ασθενών.

Τέλος έγινε μια βιβλιογραφική επισκόπηση σε δέκα παρόμοιες μελέτες που πραγματοποιήθηκαν και δημοσιεύθηκαν επίσης σε επιστημονικά περιοδικά, πηγές ιατρικής επιστημονικής έρευνας. Καταγράφηκαν οι μέθοδοι που χρησιμοποιήθηκαν και τα αποτελέσματα που προέκυψαν από την καθεμία.

Thesis

Application of machine learning techniques for the diagnosis of leukemia from microarray images

Abstract

The research focuses on Data Analysis from microarray images with Machine Learning Techniques for leukemia. Extensive reference is made to leukemia disease (symptoms, diagnosis, mortality, quality of life of patients, ways of treatment) as well as the contribution of microarray technology in the field of Bioinformatics and consequently to leukemia. Data mining techniques were applied to three datasets, from a sample of 72 patients. More specifically, the data were pre-processed and then machine learning algorithms were applied, in order to decide which of the selected algorithms achieves better results (predictions) in the classification of leukemia types. Three different sets of gene data were used for the research and with the help of weka tool we approached the genes that seem to play the most important role in identifying types of leukemia.

The experiment was performed by applying the algorithms only to the important features that had emerged first for the first 20, then for the first 50 genes. The results were evaluated with the measures of accuracy, specialization, sensitivity, leading to the finding of the most effective algorithms and the most important features that ultimately affect the categorization of patients.

Finally, a bibliographic review was made of ten similar studies that were carried out and officially published in scientific journals, sources of medical scientific research. The methods used and the results obtained from each were recorded.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέπων καθηγητή μου Κ. Θέμη Έξαρχο για τη βοήθεια του

Και τους γονείς μου που ήταν πάντα δίπλα μου όλα αυτά τα χρόνια στηρίζοντας την προσπάθεια μου.

Περιεχόμενα

ΠΕΡΙΕΧΟΜΕΝΑ ΠΙΝΑΚΩΝ.....	7
ΠΕΡΙΕΧΟΜΕΝΑ ΕΙΚΟΝΩΝ.....	8
ΕΙΣΑΓΩΓΗ.....	10
ΚΕΦΑΛΑΙΟ 1: ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΑΣΘΕΝΕΙΑΣ	
1.1 Η Ασθένεια.....	11
1.2 Συμπτώματα.....	14
1.3 Διάγνωση.....	15
1.4 Θνησιμότητα.....	16
1.5 Ποιότητα Ζωής.....	17
1.6 Μικροσυστοιχίες DNA.....	18
ΚΕΦΑΛΑΙΟ 2 : ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΕΠΙΣΚΟΠΗΣΗ.....	21
ΚΕΦΑΛΑΙΟ 3 : ΜΕΘΟΔΟΛΟΓΙΑ.....	
3.1 Αλγόριθμοι Μάθησης.....	33
3.2 Μέθοδος.....	34
3.3 Προ-επεξεργασία Δεδομένων.....	35
3.4 Σημαντικά χαρακτηριστικά.....	35
3.5 Αλγόριθμοι Ταξινόμησης.....	37
3.6 Ανάλυση Δεδομένων.....	42
ΚΕΦΑΛΑΙΟ 4 : ΣΥΜΠΕΡΑΣΜΑΤΑ.....	52
ΒΙΒΛΙΟΓΡΑΦΙΑ.....	53
ΠΑΡΑΡΤΗΜΑ.....	68

Περιεχόμενα Πινάκων

Πίνακας 1 : Είκοσι σημαντικότερα γονίδια με CfsSubsetEvaluation.....	35
Πίνακας 2: Είκοσι σημαντικότερα γονίδια με InfoGain.....	36
Πίνακας 3: Είκοσι σημαντικότερα γονίδια με Correlation.....	37
Πίνακας 4: AML-ALL ακρίβεια 20 γονίδια.....	42
Πίνακας 5: AML-ALL ευαισθησία 20 γονίδια.....	42
Πίνακας 6: AML-ALL εξειδίκευση 20 γονίδια	43
Πίνακας 7: AML, B-cell, T-cell ακρίβεια 20 γονίδια	43
Πίνακας 8: AML, B-cell, T-cell ευαισθησία 20 γονίδια	44
Πίνακας 9: AML, B-cell, T-cell εξειδίκευση 20 γονίδια	44
Πίνακας 10: B-cell, T-cell, BM,PB ακρίβεια 20 γονίδια.....	45
Πίνακας 11: B-cell, T-cell, BM,PB ευαισθησία 20 γονίδια	45
Πίνακας 12: B-cell, T-cell, BM,PB εξειδίκευση 20 γονίδια	46
Πίνακας 13: AML-ALL ακρίβεια 50 γονίδια.....	46
Πίνακας 14: AML-ALL ευαισθησία 50 γονίδια.....	47
Πίνακας 15: AML-ALL εξειδίκευση 50 γονίδια.....	47
Πίνακας 16: AML, B-cell, T-cell ακρίβεια 50 γονίδια.....	48
Πίνακας 17: AML, B-cell, T-cell ευαισθησία 50 γονίδια.....	48
Πίνακας 18: AML, B-cell, T-cell εξειδίκευση 50 γονίδια.....	50
Πίνακας 19: B-cell, T-cell, BM, PB ακρίβεια 50 γονίδια.....	50
Πίνακας 20: B-cell, T-cell, BM, PB ευαισθησία 50 γονίδια.....	51
Πίνακας 21: B-cell, T-cell, BM, PB εξειδίκευση 50 γονίδια.....	51

Περιεχόμενα Εικόνων

Figure 1: Συμπτώματα Λευχαιμίας.....	14
Figure 2: Πλακίδιο μικροσυστοιχίας.....	20
Figure 3: Τεχνολογία μικροσυστοιχίας.....	20
Figure 4: Ανακάλυψη κατηγορίας AML-ALL.....	24
Figure 5: Απόδοση των ταξινομητών SVM, KNN και Naïve Bayes σε σύνολο δεδομένων λευχαιμίας.....	26
Figure 6: Πειράματα μόλυνσης πλασμιδίου SMARCA4 σε κυτταρικές σειρές KG1 και U937 για σύγκριση απόκρισης σε ετοποσίδη και μιτοξαντρόνη μεταξύ αυθεντικών και επιμολυσμένων κυττάρων.....	27
Figure 7: Ακρίβεια ταξινόμησης του mrKNN ταξινομητή με ποικίλς μεθόδους επιλογής χαρακτηριστικών χρησιμοποιώντας διαφορετικά σύνολα δεδομένων από μικροσυστοιχίες.....	29
Figure 8: Τελικά μοντέλα για την ομάδα U-CLL (αριστερά) και ομάδα M-CLL (δεξιά). Τα ανοιχτά γκρι βέλη είναι για τις αρχικές σχέσεις. Τα μαύρα βέλη είναι οι σχέσεις του τελικού μοντέλου ενώ τα κόκκινα βέλη είναι για τις κοινές σχέσεις.....	30
Figure 9: Περιπτώσεις AML που ταξινομήθηκαν εσφαλμένα ως T-ALL.....	31
Figure 10: Ανίχνευση της Λευχαιμίας και ταξινόμηση χρησιμοποιώντας SVM.....	32
Figure 11: Ανίχνευση της Λευχαιμίας και ταξινόμηση χρησιμοποιώντας SVM.....	32
Figure 12: Λειτουργία των δέντρων απόφασης.....	40
Figure 13: AML-ALL accuracy 20 genes.....	42
Figure 14: AML-ALL sensitivity 20 genes.....	43
Figure 15: AML-ALL specificity 20 genes.....	43
Figure 16: AML, B-cell, T-cell accuracy 20 genes.....	44

Figure 17: AML, B-cell, T-cell sensitivity 20 genes.....	44
Figure 18: AML, B-cell, T-cell specificity 20 genes.....	45
Figure 19: B-cell, T-cell, BM, PB accuracy 20 genes.....	45
Figure 20: B-cell, T-cell, BM, PB sensitivity 20 genes.....	46
Figure 21: B-cell, T-cell, BM, PB specificity 20 genes.....	46
Figure 22: AML-ALL accuracy 50 genes.....	47
Figure 23: AML-ALL sensitivity 50 genes.....	47
Figure 24: AML-ALL specificity 50 genes.....	48
Figure 25: AML, B-cell, T-cell accuracy 50 genes.....	48
Figure 26: AML, B-cell, T-cell sensitivity 50 genes.....	49
Figure 27: AML, B-cell, T-cell specificity 50 genes.....	49
Figure 28: B-cell, T-cell, BM, PB accuracy 50 genes.....	50
Figure 29: B-cell, T-cell, BM, PB sensitivity 50 genes.....	50
Figure 30: B-cell, T-cell, BM, PB specificity 50 genes.....	50
Figure 31: Confusion matrix.....	51

Εισαγωγή

Η Βιοπληροφορική είναι ένας διεπιστημονικός κλάδος όπου ο συνδυασμός της βιολογίας με την πληροφορική, τα μαθηματικά και τη στατιστική αποσκοπούν στην επίλυση βιολογικών προβλημάτων. Ο όρος βιοπληροφορική αναφέρεται κυρίως στη χρήση αλγορίθμων και υπολογιστικών τεχνικών που επιτρέπουν την απάντηση βιολογικών ερωτημάτων (π.χ. αναζήτηση μιας ακολουθίας σε μια βάση δεδομένων, χειρισμός μεγάλου όγκου ακολουθιών ή δεδομένων γονιδιακής έκφρασης κλπ) (Παντελής Γ.Μπάγκος,2015). Στην παρούσα εργασία μελετήθηκαν δεδομένα από εικόνες μικροσυστοιχιών για τη διάγνωση της λευχαιμίας.

Η λευχαιμία είναι μια αιματολογική κακοήθεια, η οποία χαρακτηρίζεται από τη συνεχή άθροιση νεοπλασματικών αιμοποιητικών κυττάρων στο μυελό των οστών. Στη διαδικασία αυτή αυξάνεται ο αριθμός των ανώριμων λευκών αιμοσφαιρίων πάνω από κάποιο όριο, τα κύτταρα αυτά συσσωρεύονται στο σώμα, χύνονται τελικά στο αίμα και ενώ ο αριθμός των φυσιολογικών ερυθρών, λευκών αιμοσφαιρίων και αιμοπεταλίων μειώνεται, ο αριθμός ανώριμων κυττάρων στο αίμα αυξάνεται (Λουκόπουλος, Δ, & Μαριάννα,2016).

Η αρχή της μεθόδου των μικροσυστοιχιών βασίζεται στη σήμανση μεταγράφων που απομονώνονται από βιολογικά δείγματα και στην υβριδοποίηση τους στους ανιχνευτές της μικροσυστοιχίας με σκοπό να προσδιοριστεί η ποσότητα ή το σχετικό επίπεδο έκφρασης κάθε μεταγράφου. Στα συμβατικά πρωτόκολλα, ολικό RNA απομονώνεται από το βιολογικό δείγμα και κατόπιν υφίσταται αντίστροφη μεταγραφή ώστε να σχηματιστεί συμπληρωματικό DNA. Στη συνέχεια το cDNA σημαίνεται με φθορίζοντα ιχνηλάτη. Το σημασμένο μόριο υβριδοποιείται στην πλάκα της μικροσυστοιχίας. Μετά από τη διαδικασία υβριδοποίησης, η μικροσυστοιχία σαρώνεται ώστε να διαπιστωθούν τα επίπεδα υβριδοποίησης του ανιχνευτή. Αυτά υποδηλώνουν τα επίπεδα γονιδιακής έκφρασης στο εξεταζόμενο δείγμα (Σπηλιώτου Βάια, 2011).

Θα μελετηθούν 3 σύνολα γονιδιακών δεδομένων και θα προσεγγιστούν τα γονίδια που φαίνεται ότι επιτελούν τον σημαντικότερο ρόλο για την αναγνώριση κατηγορίας λευχαιμίας. Επίσης εφαρμόζοντας πειράματα στα δεδομένα με αλγορίθμους μηχανικής μάθησης θα αποφασιστεί ποιος από αυτούς πετυχαίνει καλύτερα αποτελέσματα (προβλέψεις) στην ταξινόμηση των τύπων λευχαιμίας των ασθενών.

Κεφάλαιο 1

1.1 Η Ασθένεια

Μια από τις σημαντικότερες ασθένειες που μαστίζουν τη σύγχρονη κοινωνία είναι ο καρκίνος. Υπάρχουν διάφορες μορφές καρκίνου, οι οποίες αφορούν διαφορετικά όργανα του ανθρώπινου σώματος. Η λευχαιμία είναι ένας τύπος καρκίνου του αίματος ή του μυελού των οστών που χαρακτηρίζεται από ανώμαλο πολλαπλασιασμό κυττάρων που παράγουν τα λευκά αιμοσφαίρια και από την είσοδο των κυττάρων αυτών στην κυκλοφορία του αίματος πριν από την ωρίμανση τους.

Η λευχαιμία υποδιαιρείται σε πολλές διαφορετικές ομάδες. Ο πρώτος διαχωρισμός αφορά την οξεία και τη χρόνια. Κατά την οξεία λευχαιμία αυξάνεται ταχεία και επιθετικά ο αριθμός των ανώριμων λευκών αιμοσφαιρίων, καθιστώντας τον μυελό των οστών μη ικανό να παράγει υγιή ερυθρά αιμοσφαίρια. Σε αυτήν την περίπτωση απαιτείται άμεση θεραπεία καθώς τα κακοήγη κύτταρα εξέρχονται στην κυκλοφορία και η μετάσταση σε άλλες περιοχές και όργανα του σώματος είναι αναπόφευκτη. Ο τύπος αυτός συναντάται συχνά στην παιδική ηλικία.

Κατά την χρόνια λευχαιμία παράγονται υπερβολικά σχετικά ώριμα αλλά μη φυσιολογικά λευκά αιμοσφαίρια σε περίοδο μηνών ή χρόνων με αποτέλεσμα την ύπαρξη πολλών μη φυσιολογικών λευκών αιμοσφαιρίων στο αίμα. Η χρόνια λευχαιμία διαγιγνώσκεται περισσότερο σε ενήλικες, αλλά μπορεί θεωρητικά να αναπτυχθεί σε οποιαδήποτε ηλικιακή ομάδα.

Η δεύτερη ταξινόμηση γίνεται ανάλογα με τον τύπο των προσβεβλημένων λευκών αιμοσφαιρίων. Οι δύο κύριοι τύποι λευκών αιμοσφαιρίων είναι τα λεμφοκύτταρα, τα οποία καταπολεμούν ιογενείς λοιμώξεις και τα μυελοειδή κύτταρα, τα οποία έχουν διαφορετικό ρόλο, όπως καταπολέμηση βακτηριακών λοιμώξεων, υπεράσπιση του σώματος από παράσιτα και πρόληψη της εξάπλωσης βλάβης των ιστών. Έτσι προκύπτει η λεμφοβλαστική ή λεμφοκυτταρική λευχαιμία και η μυελοβλαστική ή μυελογενής λευχαιμία.

Συνδυάζοντας αυτές τις 2 κατηγορίες με τις προηγούμενες 2 προκύπτει ένα σύνολο από 4 κύριες κατηγορίες.

Οξεία λεμφοβλαστική λευχαιμία (Acute lymphoblastic leukemia /ALL) : Η ασθένεια εξελίσσεται γρήγορα και δημιουργεί περισσότερα ανώριμα κύτταρα παρά ώριμα. Αναπτύσσεται από πρώιμες (ανώριμες) μορφές λεμφοκυττάρων σε ποσοστό 20% ή μεγαλύτερο στο μυελό των οστών ή/και στο περιφερικό αίμα. Είναι ο πιο κοινός τύπος λευχαιμίας στα παιδιά (περίπου το 75% των περιπτώσεων αφορά παιδιά κάτω των έξι ετών) και οι θεραπείες δίνουν πολλές πιθανότητες επιβίωσης. Η οξεία λεμφοβλαστική λευχαιμία εμφανίζεται και στους ενήλικες (συνήθως ενήλικες άνω των 65 ετών) αλλά η πιθανότητα ίασης είναι πολύ μειωμένη. Οι κλασικές θεραπείες περιλαμβάνουν χημειοθεραπεία και ραδιοθεραπεία.

Οξεία μυελογενής λευχαιμία (Acute myeloid leukemia /AML): Πρόκειται για πάθηση που οφείλεται στον ανεξέλεγκτο πολλαπλασιασμό των μυελοειδών κυττάρων, τα οποία αρχίζουν να αυτοδιπλασιάζονται ανεξέλεγκτα στο μυελό των οστών, το περιφερικό αίμα ή άλλους ιστούς. Είναι πιο συχνό σε ανθρώπους μεγάλης ηλικίας (65 και άνω). Σε παιδιά μικρότερα των 15 ετών η Οξεία Μυελογενής Λευχαιμία αποτελεί το 15-20% όλων των κρουσμάτων οξείας λευχαιμίας. Όπως και στην Οξεία Λεμφοβλαστική Λευχαιμία έτσι και στην Οξεία Μυελογενή Λευχαιμία το απαιτούμενο ποσοστό μυελοειδών κυττάρων(μυελοβλάστες, μονοβλάστες/ προμυελοκύτταρα και μεγακαρυοκύτταρα) για να τεθεί η διάγνωση είναι 20% ή παραπάνω στο μυελό των οστών ή/και στο περιφερικό αίμα. Η ΟΜΛ θεραπεύεται με χημειοθεραπεία. Υπότυποι της ΟΜΛ περιλαμβάνουν την οξεία προμυελοκυτταρική λευχαιμία, την οξεία μυελοβλαστική λευχαιμία και την οξεία μεγακαρυοβλαστική λευχαιμία.

Χρόνια λεμφοκυτταρική λευχαιμία (Chronic lymphocytic leukemia /CLL): Είναι ένας τύπος καρκίνου στον οποίο ο μυελός των οστών παράγει πάρα πολλά ώριμα Β λεμφοκύτταρα. Συνήθως επιδεινώνεται αργά. Η Χρόνια λεμφοκυτταρική λευχαιμία είναι ένας από τους πιο συνηθισμένους τύπους λευχαιμίας σε ενήλικες. Εμφανίζεται συχνά κατά τη διάρκεια ή μετά τη μέση ηλικία και συμβαίνει σπάνια στα παιδιά. Η θεραπεία για τη χρόνια λεμφοκυτταρική λευχαιμία (CLL) εξαρτάται σε μεγάλο βαθμό από το πόσο προχωρημένη είναι όταν διαγνωστεί.

Ίσως χρειαστεί απλώς παρακολούθηση εάν ανιχνευθεί σε πρώιμο στάδιο. Η χημειοθεραπεία είναι η κύρια θεραπεία σε πιο προχωρημένο στάδιο. Ένας υπότυπος είναι η Β-προ λεμφοβλαστική λευχαιμία που αποτελεί μια πιο επιθετική νόσο.

Χρόνια μυελογενής λευχαιμία (Chronic myelogenous or myeloid leukemia/CML): Στη χρόνια μυελογενή λευχαιμία, ο μυελός των οστών παράγει πάρα πολλά λευκά αιμοσφαίρια. Αρχικά, αυτά τα κύτταρα λειτουργούν σχετικά κανονικά. Ωστόσο, καθώς η κατάσταση εξελίσσεται, τα ανώριμα λευκά αιμοσφαίρια που ονομάζονται μυελοβλάστες (ή εκρήξεις) συσσωρεύονται στο αίμα και στον μυελό των οστών. Η υπερανάπτυξη των μυελοβλαστών εμποδίζει την ανάπτυξη άλλων αιμοσφαιρίων, οδηγώντας σε έλλειψη ερυθρών αιμοσφαιρίων (αναιμία) και αιμοπεταλίων. Η ασθένεια μπορεί να εμφανιστεί σε οποιαδήποτε ηλικία αλλά η μέση ηλικία διάγνωσης είναι μεταξύ 50 και 60 ετών. Περίπου τα μισά άτομα με χρόνια μυελογενή λευχαιμία δεν έχουν αρχικά σημάδια και συμπτώματα και διαγιγνώσκονται όταν πραγματοποιείται εξέταση αίματος για άλλο λόγο. Η κατάσταση αποτελείται από τρεις φάσεις: τη χρόνια φάση, την επιταχυνόμενη φάση και τη φάση έκρηξης. Στη χρόνια φάση, ο αριθμός των ώριμων λευκών αιμοσφαιρίων είναι αυξημένος και οι μυελοβλάστες αντιπροσωπεύουν λιγότερο από το 10% των κυττάρων του αίματος. Τα συμπτώματα της κατάστασης κατά τη διάρκεια αυτής της φάσης είναι συνήθως ήπια ή απύσχα και επιδεινώνονται αργά. Η χρόνια φάση μπορεί να διαρκέσει από μήνες σε χρόνια. Στην επιταχυνόμενη φάση, ο αριθμός των μυελοβλαστών είναι ελαφρώς υψηλότερος, αποτελώντας το 10 έως 29 τοις εκατό των κυττάρων του αίματος. Τα συμπτώματα συνεχίζουν να επιδεινώνονται. Η επιταχυνόμενη φάση διαρκεί συνήθως 4 έως 6 μήνες, αν και παραλείπεται σε ορισμένα άτομα που επηρεάζονται. Στην φάση έκρηξης, 30 τοις εκατό ή περισσότερο του αίματος ή των κυττάρων του μυελού των οστών είναι μυελοβλάστες. Τα συμπτώματα είναι πιο σοβαρά σε αυτήν τη φάση, συμπεριλαμβανομένης μιας μαζικά διογκωμένης σπλήνας, οστικού πόνου και απώλειας βάρους. Οι σοβαρές λοιμώξεις και η ανεξέλεγκτη αιμορραγία μπορεί να είναι απειλητικές για τη ζωή.

Λευχαιμία Τ κυττάρων: Αποτελεί μια επιθετική λευχαιμία που επηρεάζει τους ενήλικες. Περισσότεροι άνδρες παρά γυναίκες διαγιγνώσκονται με αυτή την ασθένεια. Παρότι αρκετά σπάνια, είναι ο πιο κοινός τύπος λευχαιμίας ώριμων Τ λεμφοκυττάρων. Περίπου όλες οι άλλες λευχαιμίες περιλαμβάνουν Β λεμφοκύτταρα. Έχει φτωχή ανταπόκριση στη θεραπεία.

Η **T-λευχαιμία των ενηλίκων**. Για το συγκεκριμένο τύπο λευχαιμίας ενοχοποιείται ένας ογκογόνος RNA ιός ο HTLV-1 (Human T-lymphotropic virus 1), ένας ιός παρόμοιος με τον HIV (ρετροϊός). Όπως ο HIV , ο HTLV μολύνει τα CD4+ T λεμφοκύτταρα και αντιγράφεται μέσα σε αυτά. Παρόλα αυτά όμως , αντίθετα με τον HIV, δεν τα καταστρέφει. Αντίθετα, ο HTLV καθιστά «αθάνατα» τα μολυσμένα T-λεμφοκύτταρα, δίνοντας τους τη δυνατότητα να αναπτύσσονται μη φυσιολογικά. αυξάνεται (Λουκόπουλος, Δ, & Μαριάννα,2016).

1.2 Συμπτώματα

Τα συμπτώματα της λευχαιμίας είναι μη ειδικά. Προκαλούνται από τη διήθηση του μυελού των οστών από τα λευχαιμικά κύτταρα και τη μη φυσιολογική παραγωγή λευκών αιμοσφαιρίων, ερυθρών και αιμοπεταλίων, με αποτέλεσμα :

- ❖ πυρετό, απώλεια βάρους
- ❖ μώλωπες,
- ❖ δύσπνοια,
- ❖ μυϊκή αδυναμία,
- ❖ αναιμία,
- ❖ δυσκολία στην κατάποση
- ❖ απώλεια όρεξης,
- ❖ Αίσθημα κόπωσης,
- ❖ Χλωμό δέρμα,
- ❖ Λοιμώξεις που δεν εξαφανίζονται ή συνεχίζουν να επιστρέφουν
- ❖ Μώλωπες (ή μικρά κόκκινα ή μοβ σημεία) στο δέρμα
- ❖ Αιμορραγία, όπως συχνές ή σοβαρές ρινορραγίες, αιμορραγικά ούλα ή βαριά εμμηνορροϊκή αιμορραγία στις γυναίκες
- ❖ νυχτερινές εφιδρώσεις, εύκολη αιμορραγία
- ❖ διόγκωση σπλήνα και ήπατος

Common symptoms of **Leukemia**

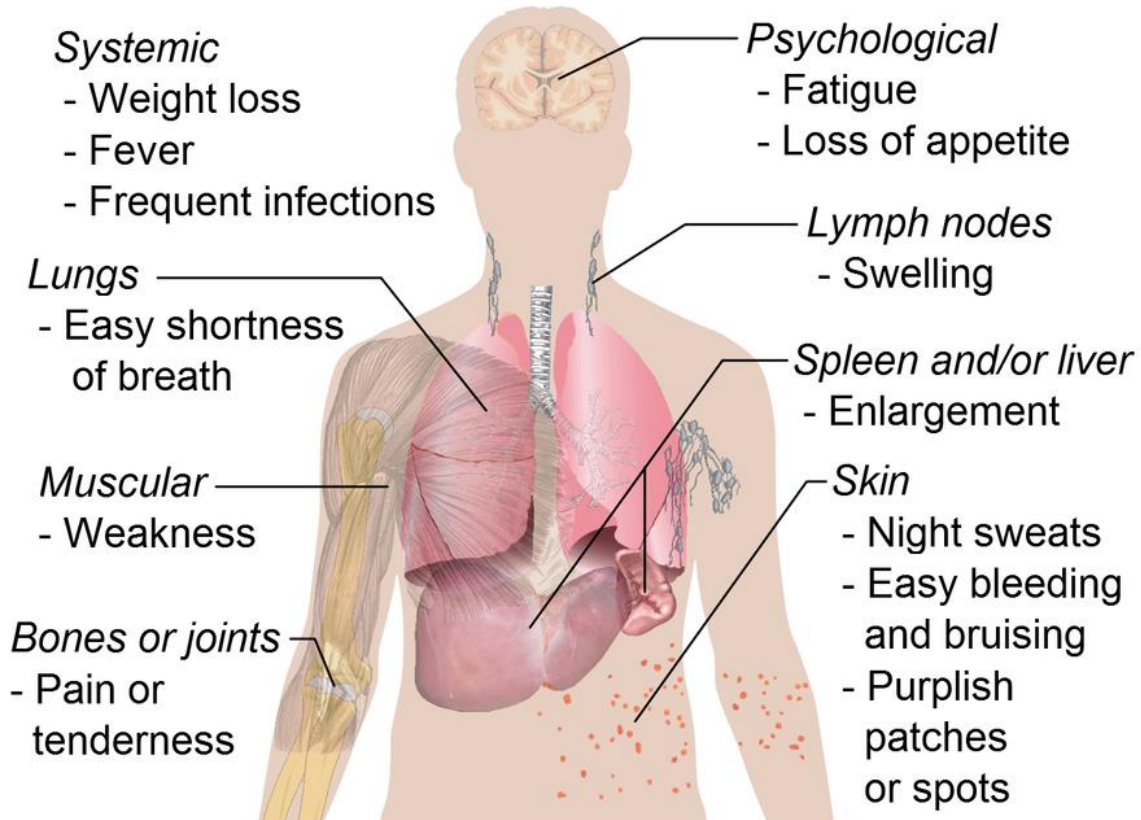


Figure 1: Συμπτώματα Λευχαιμίας

1.3 Διάγνωση

Όλα τα συμπτώματα που σχετίζονται με λευχαιμία μπορεί να αποδοθούν και σε άλλες ασθένειες. Κατά συνέπεια, η διάγνωση της λευχαιμίας γίνεται με γενική εξέταση αίματος, μυελόγραμμα, ανοσοφαινότυπο, απεικονιστικό έλεγχο και βιοχημικές εξετάσεις. Στην γενική εξέταση αίματος μετριοούνται η αιμοσφαιρίνη, τα λευκά αιμοσφαίρια και τα αιμοπετάλια. Η εξέταση μυελογράμματος είναι εξέταση του μυελού των οστών και συνίσταται στη λήψη υλικού με ειδικές βελόνες συνήθως από τα οστά της λεκάνης ή του στέρνου. Η συγκεκριμένη εξέταση είναι επώδυνη και επιβάλλεται τοπική αναισθησία ή βραχεία νάρκωση. Εφόσον η λήψη δείγματος μυελού των οστών αποδεικνύει ότι δεν υπάρχει αντικατάσταση του μυελού από ανώριμα κύτταρα σημαίνει ότι δεν πρόκειται για κακοήγη αιτιολογία.

Τα τελευταία χρόνια, η κατανόηση της μοριακής βάσης των αιματολογικών κακοηθειών έφερε εξελίξεις στη διάγνωση αλλά και στη θεραπεία της ασθένειας. Έτσι αναλύοντας τις διαφορετικές δραστηριότητες των γονιδίων σε μικροσυστοιχίες, μπορεί να γίνει πρόβλεψη του τύπου λευχαιμίας ενός ατόμου. Αυτό θα βοηθήσει και στην εξατομικευμένη θεραπεία. Χαρακτηριστικό παράδειγμα αποτέλεσε το 1999, μια ομάδα από επιστήμονες στην Αμερική η οποία χρησιμοποίησε μικροσυστοιχίες DNA για να διακρίνει ανάμεσα σε φαινομενικά όμοιες μορφές καρκίνου: την οξεία λεμφοβλαστική λευχαιμία (ALL), και την οξεία μυελογενή λευχαιμία (AML), οι οποίες επηρεάζουν κύτταρα του μυελού των οστών. Οι ερευνητές κατάφεραν να ταυτοποιήσουν μια ομάδα 50 γονιδίων που δείχνουν διαφορές στις δραστηριότητες μεταξύ των δυο καρκινικών τύπων.

1.4 Θνησιμότητα-Νοσηρότητα

Σύμφωνα με τις Παγκόσμιες στατιστικές για τον καρκίνο του 2018, η λευχαιμία υπολογίζεται ότι αντιστοιχεί σε 437.033 νέες περιπτώσεις και 309.006 θανάτους σε 185 χώρες, με ποσοστό θνησιμότητας 3.2%. Οι στατιστικές για τον καρκίνο στην Κίνα δείχνουν περίπου 75.300 νέες περιπτώσεις και 53.400 θανάτους ασθενών με λευχαιμία το 2015. Στην Ελλάδα υπολογίζεται πως κάθε χρόνο διαγιγνώσκονται περίπου 1400 νέες περιπτώσεις.

Ερευνητές, οι οποίοι δημοσίευσαν τα ευρήματά τους στην ιατρική επιθεώρηση «Annals of Oncology» σημείωσαν πως η θνησιμότητα από λευχαιμία στην Ευρώπη μειώνεται σε όλες τις ηλικιακές ομάδες, με τα μεγαλύτερα οφέλη να παρατηρούνται στα παιδιά, τους εφήβους, και τους νεαρούς ενήλικες. Η ανάλυση του έδειξε ότι μεταξύ 2009 και 2016 μειώθηκαν κατά 38% στα αγόρια και κατά 20% στα κορίτσια τα ποσοστά θανάτων από λευχαιμία στα παιδιά ηλικίας έως 14 ετών. Αντίστοιχα, στους εφήβους και στους ενήλικες ηλικίας (15-44 ετών), η θνησιμότητα μειώθηκε κατά 26% στους άντρες και κατά 22% στις γυναίκες, ενώ στις ηλικίες 45-69 ετών η αντίστοιχη μείωση και στα δύο φύλλα έφτασε το 19% (Malvezzi, M., Carioli, G., Bertuccio, P., Rosso, T., Boffetta, P., Levi, F., . . . Negri, E., 2020)

1.5 Ποιότητα Ζωής

Η οξεία λευχαιμία (AL) επηρεάζει όχι μόνο την ποιότητα ζωής των ασθενών αλλά επηρεάζει και τους ανθρώπους της οικογένειας οι οποίοι φροντίζουν το άτομο αυτό. Η κόπωση ή η έλλειψη ενέργειας, ο πόνος, η απώλεια όρεξης και η αϋπνία επηρεάζουν το 24-83% των ασθενών με AL. αυτό μπορεί να μειώσει σημαντικά την ποιότητα ζωής που σχετίζεται με την υγεία. Επιπλέον, η κύρια θεραπεία για AL είναι η χημειοθεραπεία. αυτό μπορεί να προκαλέσει μια σειρά ανεπιθύμητων ενεργειών όπως ναυτία και έμετο, ανορεξία, αλωπεκία και μυελοκαταστολή, οι οποίες μπορεί να βλάψουν περαιτέρω την ποιότητα ζωής. Τα διαθέσιμα στοιχεία δείχνουν ότι σχεδόν όλοι οι ασθενείς με καρκίνο, συμπεριλαμβανομένων εκείνων με καρκίνους αίματος, αντιμετωπίζουν μια πολύπλοκη σειρά προκλήσεων που απειλούν τη φυσική, ψυχολογική και πνευματική ευεξία τους.

1.6 Θεραπείες

Τα τελευταία χρόνια η ραγδαία πρόοδος στον τομέα της μοριακής γενετικής και της ανοσολογίας έχει επιφέρει ενθαρρυντικές θεραπείες για την αντιμετώπιση της νόσου. Νέα εξειδικευμένα φάρμακα, θεραπείες με μονοκλωνικά αντισώματα και προφίλ γονιδιακής έκφρασης επιχειρούν να αυξήσουν την συνολική επιβίωση των πασχόντων και να βελτιώσουν την ποιότητα ζωής τους.

1.6.1 Κυτταρική Θεραπεία με Γενετικά Τροποποιημένα με Χιμαιρικούς Αντιγονικούς Υποδοχείς, Τ-κύτταρα (car T-cells)

Σύμφωνα με τον *δρ Γιώργο Βασιλόπουλο*, καθηγητή Παθολογίας-Αιματολογίας στη συγκεκριμένη θεραπεία συλλέγονται από το αίμα του ασθενούς ορισμένα Τ-κύτταρα του ανοσοποιητικού συστήματος. Στη συνέχεια τα κύτταρα αυτά αποστέλλονται σε ειδικό εργαστήριο. Εκεί, τροποποιούνται γενετικά, ώστε να αποκτήσουν στην επιφάνειά τους μια ομάδα πρωτεϊνών που λέγονται χιμαιρικοί αντιγονικοί υποδοχείς (CAR).

Τα τροποποιημένα Τ-κύτταρα καλλιεργούνται και πολλαπλασιάζονται σε εκατοντάδες εκατομμύρια κλώνους στο εργαστήριο. Έπειτα, εγχέονται πίσω στον ασθενή, αφού πρώτα υποβληθεί σε έναν κύκλο ήπιας χημειοθεραπείας. Αν ο ανθρώπινος

οργανισμός τα δεχτεί κανονικά, χωρίς να παρουσιάσει επιπλοκές τα τροποποιημένα T-κύτταρα εξακολουθούν να αναπαράγονται στο σώμα του. Το ερέθισμα για τη συνεχιζόμενη αναπαραγωγή τους είναι η εντόπιση ενός αντιγόνου που «βλέπουν» οι χημαιοκίκοι υποδοχείς στα καρκινικά κύτταρα και του επιτίθενται για να το καταστρέψουν. Τα T-κύτταρα δηλαδή εξαπολύουν επίθεση στα καρκινικά σαν να είναι ένα «ζωντανό» φάρμακο.

Οι CAR-T θεραπείες που έχουν εγκριθεί έως τώρα αναγνωρίζουν ένα αντιγόνο που λέγεται CD19. Σύμφωνα με τις έως τώρα μελέτες, οι θεραπείες έχουν υψηλά ποσοστά ανταπόκρισης και ύφεσης των συγκεκριμένων αιματολογικών κακοηθειών. Δύο θεραπείες έχουν εγκριθεί έως τώρα : Το tisagenlecleucel εγκρίθηκε για παιδιά και νεαρούς ενήλικες (έως 25 ετών) με επιθετική οξεία λεμφοβλαστική λευχαιμία B-κυττάρων (B-ΟΛΛ) και για τη θεραπεία ενηλίκων που πάσχουν από μία συγκεκριμένη μορφή ανθεκτικού ή υποτροπιάζοντος λεμφώματος. Πρόκειται για το επονομαζόμενο διάχυτο λέμφωμα από μεγάλα B-κύτταρα (DLBCL), το οποίο αποτελεί τη συχνότερη μορφή μη-Hodgkin λεμφώματος.

1.6.2 Μικροσυστοιχίες DNA

Η μικροσυστοιχία DNA είναι ένα εργαλείο που χρησιμοποιείται για να προσδιορίσει εάν το DNA από ένα συγκεκριμένο άτομο περιέχει μια μετάλλαξη στα γονίδια, δηλαδή για την ανίχνευση της έκφρασης χιλιάδων γονιδίων ταυτόχρονα. Οι μικροσυστοιχίες DNA είναι διαφάνειες μικροσκοπίου που εκτυπώνονται με χιλιάδες μικροσκοπικές κηλίδες σε καθορισμένες θέσεις, με κάθε κηλίδα να περιέχει μια γνωστή αλληλουχία ή γονίδιο DNA. Συχνά, αυτές οι διαφάνειες αναφέρονται ως τσιπ γονιδίων ή τσιπ DNA. Το τσιπ αποτελείται από ένα μικρό γυάλινο έλασμα που περιβάλλεται από πλαστικό. Ορισμένες εταιρείες κατασκευάζουν microarrays χρησιμοποιώντας μεθόδους παρόμοιες με εκείνες που χρησιμοποιούνται για τη δημιουργία μικροτσίπ υπολογιστών. Τα μόρια DNA που συνδέονται με κάθε διαφάνεια λειτουργούν ως ανιχνευτές για την ανίχνευση γονιδιακής έκφρασης, το οποίο είναι επίσης γνωστό ως το σύνολο μεταγραφών RNA αγγελιοφόρου (mRNA) που εκφράζονται από μια ομάδα γονιδίων.

Για την πραγματοποίηση μιας ανάλυσης μικροσυστοιχιών, μόρια mRNA συλλέγονται τυπικά τόσο από ένα πειραματικό δείγμα όσο και από ένα δείγμα αναφοράς. Για παράδειγμα, το δείγμα αναφοράς μπορεί να συλλεχθεί από ένα υγιές άτομο και το

πειραματικό δείγμα θα μπορούσε να συλλεχθεί από ένα άτομο με μια ασθένεια όπως ο καρκίνος. Τα δύο δείγματα mRNA στη συνέχεια μετατρέπονται σε συμπληρωματικό DNA (cDNA) και κάθε δείγμα επισημαίνεται με έναν φθορίζοντα ιχνηλάτη διαφορετικού χρώματος. Για παράδειγμα, το πειραματικό δείγμα cDNA μπορεί να επισημανθεί με μια κόκκινη φθορίζουσα χρωστική, ενώ το cDNA αναφοράς μπορεί να επισημανθεί με μια πράσινη φθορίζουσα χρωστική ουσία. Τα δύο δείγματα στη συνέχεια αναμειγνύονται μεταξύ τους και αφήνονται να δεσμευτούν με τη διαφάνεια μικροσυστοιχιών. Η διαδικασία στην οποία τα μόρια του cDNA συνδέονται με τους ανιχνευτές DNA στη διαφάνεια ονομάζεται υβριδισμός. Μετά τον υβριδισμό, η μικροσυστοιχία σαρώνεται για να μετρηθεί η έκφραση κάθε γονιδίου που εκτυπώθηκε στη διαφάνεια. Εάν η έκφραση από ένα συγκεκριμένο γονίδιο είναι υψηλότερη στο πειραματικό δείγμα από ότι στο δείγμα αναφοράς, τότε το αντίστοιχο σημείο της μικροσυστοιχίας εμφανίζεται κόκκινο. Αντιθέτως, εάν η έκφραση στο πειραματικό δείγμα είναι μικρότερη από ό, τι στο δείγμα αναφοράς, τότε το σημείο εμφανίζεται πράσινο. Τέλος, αν υπάρχει ίση έκφραση στα δύο δείγματα, τότε το σημείο εμφανίζεται κίτρινο. Τα δεδομένα που συγκεντρώνονται μέσω μικροσυστοιχιών μπορούν να χρησιμοποιηθούν για τη δημιουργία προφίλ γονιδιακής έκφρασης, τα οποία δείχνουν ταυτόχρονες αλλαγές στην έκφραση πολλών γονιδίων σε απόκριση μιας συγκεκριμένης κατάστασης ή θεραπείας. Σήμερα, οι μικροσυστοιχίες DNA χρησιμοποιούνται σε κλινικές διαγνωστικές εξετάσεις για ορισμένες ασθένειες καθώς και για πολύ μεγάλες μελέτες και για κάποιες κλινικές δοκιμές. Η δημιουργία προφίλ γονιδιακής έκφρασης είναι μια ισχυρή τεχνολογία για τη διάγνωση αιματολογικών κακοηθειών με μεγάλη ακρίβεια. Μπορεί να συμπληρώσει τους τρέχοντες διαγνωστικούς αλγόριθμους και να προσφέρει μια αξιόπιστη πλατφόρμα για ασθενείς που δεν έχουν πρόσβαση στον σημερινό τελευταίας τεχνολογίας διαγνωστικό έλεγχο (National Human Genome Research Institute, DNA Microarray Technology Fact Sheet, 2020)



Figure 2: πλακίδιο μικροσυστοιχίας

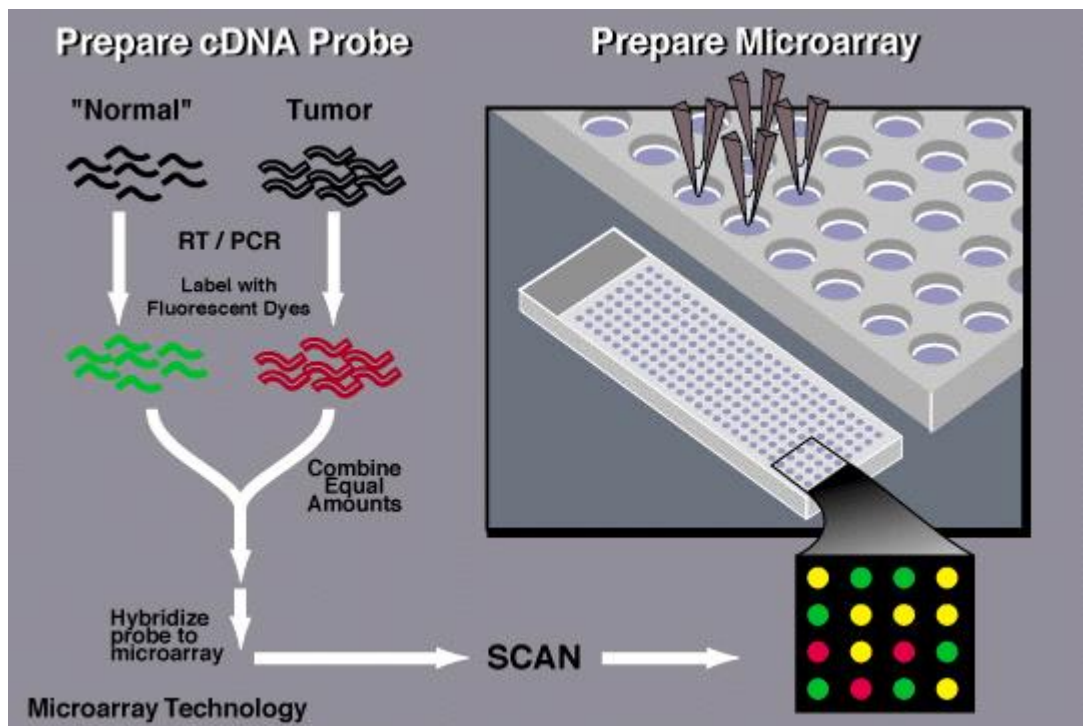


Figure 3: Τεχνολογία Μικροσυστοιχίας

Κεφάλαιο 2

2.1 Βιβλιογραφική Επισκόπηση

Η δημιουργία προφίλ γονιδιακής έκφρασης με βάση μικροσυστοιχίες έχει αναδειχθεί ως αποτελεσματική τεχνική για την ταξινόμηση, την πρόγνωση, τη διάγνωση και τη θεραπεία του καρκίνου. Τα δεδομένα μικροσυστοιχιών ικανοποιούν το ίδιο τις ιδιότητες αξιοπιστίας και ταχύτητας των μεγάλων δεδομένων, καθώς αυτό συνεχίζει να αλλάζει με το χρόνο. Συχνά περιέχουν μεγάλο ποσό έκφρασης, αλλά μόνο ένα κλάσμα από αυτά περιέχει γονίδια που εκφράζονται σημαντικά. Η ακριβής ταυτοποίηση των γονιδίων ενδιαφέροντος που είναι υπεύθυνα για την πρόκληση καρκίνου είναι ζωτικής σημασίας στην ανάλυση δεδομένων μικροσυστοιχιών. Τα περισσότερα υφιστάμενα συστήματα χρησιμοποιούν μια διαδικασία δύο φάσεων όπως η επιλογή / εξαγωγή χαρακτηριστικών ακολουθούμενη από την ταξινόμηση.

Οι Sung-Huai Hsieh, Zhenyu Wang, Po-Hsun Cheng, I-Shun Lee, Sheau-Ling Hsieh και Feipei Lai (2010) πραγματοποίησαν μια μελέτη για την ταξινόμηση της λευχαιμίας. Για την επιλογή των χαρακτηριστικών χρησιμοποιήθηκε η μέθοδος Information Gain. Το μοντέλο χρησιμοποιεί τον αλγόριθμο Support Vector Machine για να αξιολογήσει και να ερμηνεύσει την ταξινόμηση του καρκίνου. Τα πειραματικά αποτελέσματα δείχνουν ότι το μοντέλο SVM απεικονίζει την υψηλότερη ακρίβεια των ταξινομήσεων για την λευχαιμία. Οι γονιδιακές μετρήσεις ελήφθησαν από 63 δείγματα μυελού των οστών και 9 δείγματα περιφερικού αίματος. Το σύνολο δεδομένων περιέχει 72 δείγματα. Όλα τα δείγματα μπορούν να διαιρεθούν σε δύο υποτύπους: 25 δείγματα οξείας μυελοειδούς λευχαιμίας (AML) και 47 δείγματα οξείας λεμφοβλαστικής λευχαιμίας (ALL). Υπάρχουν 7.129 γονίδια στο σύνολο δεδομένων.

Σε άλλη πρόσφατη μελέτη οι S. Rajpurohit, S. Patil, N. Choudhary, S. Gavasane and P. Kosamkar (2018) πρότειναν αυτοματοποιημένα συστήματα για την ανίχνευση της οξείας λεμφοβλαστικής λευχαιμίας (ALL), τα οποία περιλαμβάνουν την εξαγωγή χαρακτηριστικών από εικόνες αίματος χρησιμοποιώντας το MATLAB. Εφάρμοσαν διαφορετικούς ταξινομητές για παραγωγή αποτελεσμάτων, οι οποίοι δίνουν αξιοσημείωτη ακρίβεια. Το σύνολο δεδομένων που χρησιμοποιήθηκε ALL-IDB αποτελείται από δύο σύνολα εικόνων IDB1 και IDB2. Το IDB1 αποτελείται από 108 εικόνες με 510 υποψήφιους λεμφοβλάστες και σκοπός του είναι να δοκιμάσει και την επεξεργασία

εικόνας και την ικανότητα ταξινόμησης του αλγορίθμου. Το IDB2 αποτελείται από 260 εικόνες με τμηματικά κύτταρα από IDB1 εικόνες για να ελέγξει την ικανότητα ταξινόμησης του αλγορίθμου. Για τον CNN, η εκπαίδευση και οι δοκιμές πραγματοποιούνται στα 243 λεμφοκυτταρικά κύτταρα από την ALL IDB2 βάση δεδομένων. Περιέχει 121 κανονικά κύτταρα και 122 κύτταρα έκκρισης. Για τον F-NN, τον K-NN η δοκιμή πραγματοποιήθηκε σε 239 λεμφοκυτταρικά κύτταρα και σε 249 κύτταρα για SVM. Τα κύτταρα εξάγονται από 25 εικόνες της ALL IDB1 βάσης δεδομένων. Το προτεινόμενο σύστημα βελτιώνει περαιτέρω την ακρίβεια ταξινόμησης. Χρησιμοποιεί openCV και skimimage για την επεξεργασία εικόνας ώστε να εξάγει σχετικά χαρακτηριστικά από την εικόνα του αίματος και όχι μόνο τεράστιο αριθμό χαρακτηριστικών και η περαιτέρω ταξινόμηση πραγματοποιείται χρησιμοποιώντας ποικίλους αλγορίθμους: CNN, FNN, SVM και KNN. Ο CNN δίνει την υψηλότερη ακρίβεια 98,33%. Η ακρίβεια που επιτυγχάνεται από άλλους ταξινομητές: FNN, SVM και KNN είναι 95,40%, 91,40% και 93,30% αντίστοιχα.

Οι Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., ... Lander, E. S. (1999) υποστηρίζουν ότι πρόβλεψη κλάσης που προκύπτει αυτόματα ήταν σε θέση να προσδιορίσει την κατηγορία νέων περιπτώσεων λευχαιμίας. Τα αποτελέσματα αποδεικνύουν τη σκοπιμότητα της ταξινόμησης του καρκίνου, βασίζονται μεμονωμένα στην παρακολούθηση της γονιδιακής έκφρασης από DNA μικροσυστοιχίες και προτείνουν μια γενική στρατηγική για την ανίχνευση και την πρόβλεψη του καρκίνου και για άλλους τύπους, ανεξάρτητα από τις προηγούμενες βιολογικές γνώσεις. Το αρχικό σύνολο δεδομένων λευχαιμίας αποτελούνταν από 38 δείγματα μυελού των οστών (27 ALL, 11 AML) που λήφθηκε από ασθενείς με οξεία λευχαιμία στο χρόνο διάγνωσης. RNA που παρασκευάστηκε από μονοπυρηνικά κύτταρα μυελού των οστών υποβλήθηκε σε υβριδοποίηση με μικροακτίνες υψηλής πυκνότητας ολιγονουκλεοτιδίων και περιείχε ανιχνευτές για 6.817 ανθρώπινα γονίδια. Για κάθε γονίδιο, αποκτήθηκε ένα επίπεδο έκφρασης από άποψη ποσότητας. Τα δείγματα υποβλήθηκαν εκ προτέρων σε ποιοτικό έλεγχο προδιαγραφών όσον αφορά την ποσότητα του επισημασμένου RNA και την ποιότητα της εικόνας της μικροσυστοιχίας που έχει σαρωθεί. Το πρώτο ζήτημα ήταν να διερευνηθεί εάν υπήρχαν γονίδια των οποίων το πρότυπο έκφρασης συσχετίζεται έντονα με τη διάκριση κλάσης που μπορεί να προβλεφθεί. Τα 6817 γονίδια ήταν ταξινομημένα κατά βαθμό συσχέτισης. Για να διαπιστωθεί εάν οι παρατηρούμενες συσχετίσεις ήταν ισχυρότερες από ό,τι θα περίμενε κανείς κατά τύχη, αναπτύχθηκε μια μέθοδος που ονομάζεται "Ανάλυση γειτονιάς". Εν συντομία, ορίζεται ένα "εξιδανικευμένο μοτίβο

έκφρασης" που αντιστοιχεί σε ένα γονίδιο που είναι ομοιόμορφα υψηλό σε μία κατηγορία και ομοιόμορφα χαμηλό στην άλλη. Ελέγχεται αν υπάρχει μια ασυνήθιστα υψηλή πυκνότητα γονιδίων "κοντινή" (δηλαδή, παρόμοια) με αυτό το εξιδανικευμένο μοτίβο, σε σύγκριση με ισοδύναμα τυχαία μοτίβα. Για τα 38 δείγματα οξείας λευχαιμίας, η "Ανάλυση γειτονιάς" έδειξε ότι περίπου 1100 γονίδια συσχετίστηκαν περισσότερο με το γονίδιο AML-ALL κλάση από ό, τι θα αναμενόταν τυχαία). Αυτό υπονοούσε ότι η ταξινόμηση θα μπορούσε πράγματι να βασίζεται στα δεδομένα έκφρασης. Το δεύτερο ζήτημα ήταν πώς να χρησιμοποιηθεί μια συλλογή γνωστών δειγμάτων για να δημιουργηθεί μια «πρόβλεψη κλάσης» ικανή να αναθέσει ένα νέο δείγμα σε μία από τις δύο κλάσεις. Αναπτύχθηκε μια διαδικασία που χρησιμοποιεί ένα σταθερό υποσύνολο «πληροφοριακών γονιδίων» (που επιλέγονται με βάση τον συσχετισμό τους με την κλάση διάκρισης) και κάνει μια πρόβλεψη με βάση το επίπεδο έκφρασης αυτών των γονιδίων σε ένα νέο δείγμα. Κάθε πληροφοριακό γονίδιο εκπέμπει μια "σταθμισμένη ψήφο" για μια από τις κλάσεις, με το μέγεθος κάθε ψηφοφορίας να εξαρτάται από το επίπεδο έκφρασης στο νέο δείγμα και το βαθμό της συσχέτισης αυτού του γονιδίου με τη διάκριση κατηγορίας. Οι ψήφοι αθροίστηκαν για να καθοριστεί η επιτυχής κλάση, καθώς και μια "πρόβλεψη δύναμη"(PS), η οποία αποτελεί μέτρο του περιθωρίου νίκης που κυμαίνεται από 0 έως 1. Το δείγμα ανατίθεται στην κατηγορία που κέρδισε εάν το PS υπερβεί ένα προκαθορισμένο όριο και θεωρείται αβέβαιο. Στη βάση της προηγούμενης ανάλυσης, χρησιμοποιήσαμε ένα όριο 0,3. Το τρίτο ζήτημα ήταν ο τρόπος δοκιμής της εγκυρότητας των προγνωστικών κλάσης. Χρησιμοποιήθηκε διαδικασία δύο σταδίων. Η ακρίβεια των προγνωστικών ελέγχθηκε πρώτα με cross-validation στο αρχικό σύνολο δεδομένων.(Συνοπτικά, παρακρατείται ένα δείγμα, δημιουργείται μια πρόβλεψη που βασίζεται μόνο στα υπόλοιπα δείγματα, και προβλέπεται η κλάση του παρακρατηθέντος δείγματος. Η διαδικασία επαναλαμβάνεται για κάθε δείγμα και υπολογίζεται το συσσωρευτικό ποσοστό σφάλματος.) Στη συνέχεια δημιουργείται μια τελική πρόβλεψη βασισμένη στο αρχικό σύνολο δεδομένων και αξιολογείται η ακρίβειά της σε ένα ανεξάρτητο σύνολο δειγμάτων. Αυτή η προσέγγιση εφαρμόστηκε στα 38 δείγματα οξείας λευχαιμίας. Το σύνολο ενημερωτικών γονιδίων που χρησιμοποιήθηκαν για την πρόβλεψη επιλέχθηκε να είναι τα 50 γονίδια που σχετίζονται περισσότερο με AML-ALL διάκριση. Οι παράμετροι της πρόβλεψης προσδιορίστηκαν από τα επίπεδα έκφρασης των 50 αυτών γονιδίων στα γνωστά δείγματα. Η πρόβλεψη χρησιμοποιήθηκε στη συνέχεια για να ταξινομήσει νέα δείγματα, με την εφαρμογή της στα επίπεδα έκφρασης αυτών των γονιδίων στο δείγμα. Οι προβλέψεις των 50 γονιδίων που προέκυψαν από cross-validation test αποδίδουν 36 από τα 38 δείγματα είτε ως AML είτε

ως ALL και τα υπόλοιπα δύο ως αβέβαια (PS, 0,3). Και οι 36 προβλέψεις συμφωνούσαν με την κλινική διάγνωση των ασθενών. Στη συνέχεια δημιουργήθηκε μια πρόβλεψη για 50 γονίδια στη βάση και των 38 δειγμάτων και εφαρμόστηκε σε μια ανεξάρτητη συλλογή 34 δειγμάτων λευχαιμίας. Τα δείγματα αποτελούνταν από 24 μυελού των οστών και 10 περιφερικού αίματος. Συνολικά, ο προγνωστικός παράγοντας έδωσε ισχυρές προβλέψεις για 29 από τα 34 δείγματα και η ακρίβεια ήταν 100%. Η επιτυχία ήταν αξιοσημείωτη επειδή η συλλογή περιελάμβανε ευρύ φάσμα δειγμάτων, συμπεριλαμβανομένων των δειγμάτων από το περιφερικό αίμα παρά από τον μυελό των οστών, από παιδικής ηλικίας ασθενείς με AML και από διαφορετικά εργαστήρια που χρησιμοποιούν διαφορετικά πρωτόκολλα προετοιμασίας δειγμάτων.

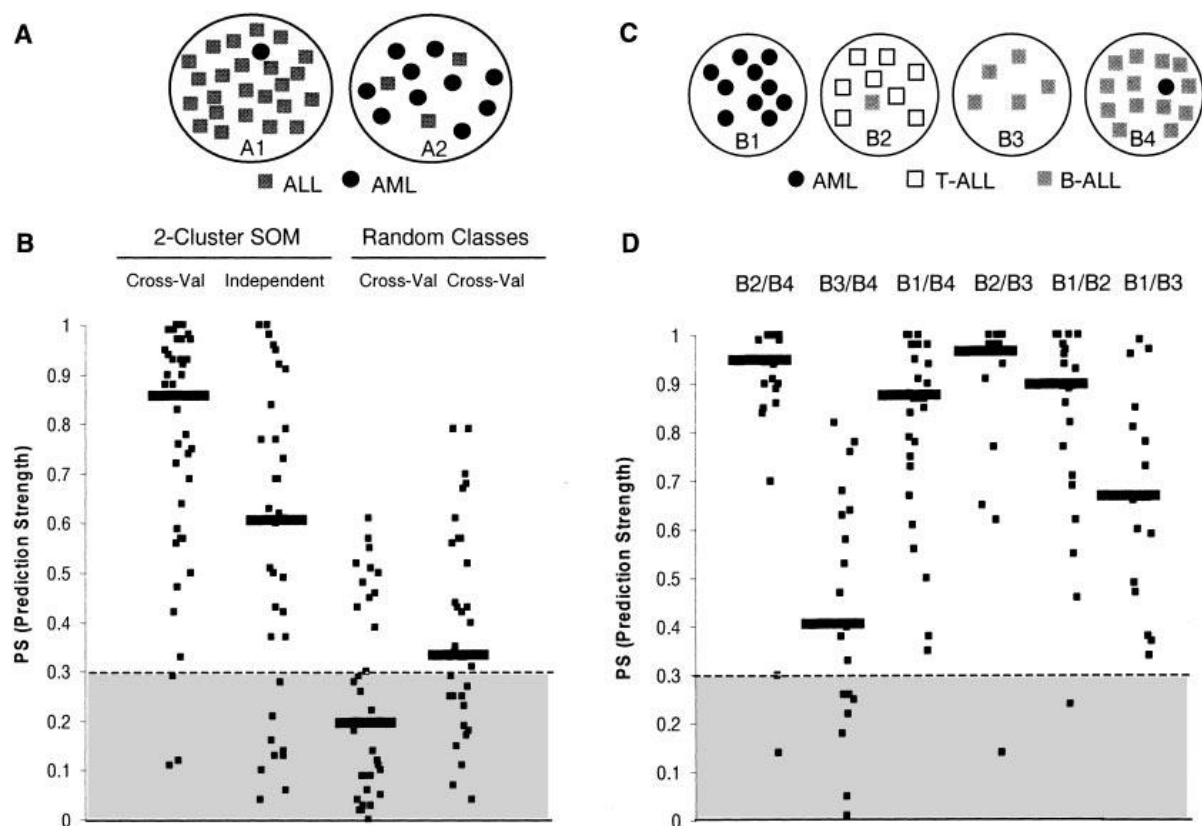


Figure 4: Ανακάλυψη κατηγορίας AML-ALL

(A) Σχηματική αναπαράσταση του SOM δύο συστάδων. Ένας two-cluster SOM δημιουργήθηκε από το αρχικά 38 δείγματα λευχαιμίας. Καθένα από τα 38 δείγματα τοποθετείται έτσι σε μία από τις δύο συστάδες με βάση τα πρότυπα της γονιδιακής έκφρασης για τα 6817 γονίδια που αναλύθηκαν σε κάθε δείγμα. Το σύμπλεγμα A1 περιέχει την πλειονότητα όλων των δειγμάτων (γκρίζα τετράγωνα) και το σύμπλεγμα A2 περιέχει την πλειονότητα των δειγμάτων AML (μαύροι κύκλοι).

B) Κατανομές ισχύος πρόβλεψης (PS). Οι διασπορές δείχνουν την κατανομή των βαθμολογιών ισχύος πρόβλεψης για την πρόγνωση κλάσης. Τα πρώτα δύο γραφήματα δείχνουν την κατανομή για τον προγνωστικό παράγοντα που δημιουργήθηκε για να ταξινομήσουν τα δείγματα ως "τύπου A1" ή "τύπου A2" που έχουν δοκιμαστεί σε cross-validation στο αρχικό σύνολο δεδομένων (διάμεσος PS 5 0,86) και στα ανεξάρτητα σύνολα δεδομένων (μεσαίο PS 5 0,61). Τα υπόλοιπα διαγράμματα δείχνουν την κατανομή για δύο προγνωστικά που αντιστοιχούν σε τυχαίες κλάσεις. Σε αυτές τις περιπτώσεις, οι βαθμολογίες PS είναι πολύ χαμηλότερες (διάμεσος PS 5 0,20 και 0,34, αντίστοιχα), και περίπου τα μισά από τα δείγματα πέφτουν κάτω από το όριο πρόβλεψης (PS 5 0,3). Εξετάστηκαν συνολικά 100 τέτοιοι τυχαίοι προγνωστικοί δείκτες, για τον υπολογισμό της κατανομής των μέσων βαθμολογιών PS για την αξιολόγηση της στατιστικής σημασίας του προγνωστικού για το A1-A2 .

(C) Σχηματική αναπαράσταση του SOM τεσσάρων συστάδων. Τα δείγματα AML εμφανίζονται ως μαύροι κύκλοι, T-lineage ALL ως ανοιχτά τετράγωνα, και B-lineage ALL ως γκριζα τετράγωνα. T- και B-lineages διαφοροποιήθηκαν με βάση τον ανοσοφαινότυπο κυτταρικής επιφάνειας. Η κλάση B1 είναι αποκλειστικά AML, η κλάση B2 περιέχει και τα οκτώ T-ALLs και οι κλάσεις B3 και B4 περιέχουν την πλειονότητα των δειγμάτων B-ALL.

(D) Κατανομή ισχύος πρόβλεψης (PS) για σύγκριση μεταξύ ζευγών μεταξύ κλάσεων. Μελέτες πρόβλεψης με cross-validation δείχνουν ότι οι τέσσερις κλάσεις θα μπορούσαν να διακριθούν με υψηλές βαθμολογίες πρόβλεψης, με εξαίρεση των κατηγοριών B3 και B4. Αυτές οι δύο κλάσεις δεν μπορούσαν να γίνουν εύκολα διακριτές μεταξύ τους, σύμφωνα με το ότι και οι δύο περιέχουν κυρίως δείγματα B-ALL, και υποδηλώνουν ότι τα B3 και B4 θα μπορούσαν να είναι καλύτερα συγχωνευμένα σε μία κλάση.

Μια διαφορετική προσέγγιση για την ανίχνευση καρκινικών βιοδεικτών από δεδομένα λευχαιμίας ακολουθήσαν οι S. Begum, D. Chakraborty and R. Sarkar (2016) χρησιμοποιώντας επιλογή χαρακτηριστικών και επιβλεπόμενη μάθηση. Στο συγκεκριμένο άρθρο διερευνώνται αλγόριθμοι για την ταξινόμηση του συνόλου δεδομένων ασθενών με λευχαιμία. Οι αλγόριθμοι που εξετάζονται είναι οι Naive Bayes (NB), K πλησιέστερου γείτονα (KNN) και Support Vector Machine. Το σύνολο δεδομένων λευχαιμίας έχει συλλεχθεί από δείγματα ασθενών με λευχαιμία από το Golub και θεωρείται σημείο αναφοράς για την ανάλυση μικροσυστοιχιών. Μετράει τα δείγματα οξείας λεμφοβλαστικής λευχαιμίας (ΟΛΛ) και οξείας μυελοειδούς λευχαιμίας (AML) από μυελό των οστών και περιφερικό αίμα αντίστοιχα. Υπάρχουν 72 δείγματα στο σύνολο δεδομένων όπου τα υποσύνολα αποτελούνται από 25 AML δείγματα και 47 ALL. Ταξινομήθηκαν πάνω από 7.169 γονίδια. Από τα 72 δειγματα, τα 40 λαμβάνονται για σύνολο δοκιμών (test set). Έχουν θεωρηθεί τρία διαφορετικά σύνολα εκπαίδευσης μεγέθους 20, 37 και 48 αντίστοιχα. Τα αποτελέσματα δείχνουν την απόδοση των τριών ταξινομητών. Τα πειραματικά αποτελέσματα δείχνουν ότι το ποσοστό ακρίβειας που προέκυψε από τον SVM υπερσχύει των άλλων δύο ταξινομητών. Παρατηρήθηκε ότι ο SVM παρέχει υψηλή προγνωστική ακρίβεια στην ταυτοποίηση συγκεκριμένης υποκατηγορίας καρκίνου. Για το πρώτο σύνολο εκπαίδευσης (20 δείγματα) ο K-NN εμφανίζει ποσοστό επιτυχίας 85%, ο Naive Bayes 80% και ο SVM 92,5%. Για το δεύτερο σύνολο εκπαίδευσης (37 δείγματα) ο

K-NN εμφανίζει ποσοστό επιτυχίας 92,5%, ο Naive Bayes 90% και ο SVM 95%. Για το τρίτο σύνολο εκπαίδευσης (48 δείγματα) ο K-NN εμφανίζει ποσοστό επιτυχίας 95%, ο Naive Bayes 92% και ο SVM 97,5%. Δεδομένου ότι το σύνολο γονιδιακής μικροσυστοιχίας αποτελείται από χιλιάδες γονίδια είναι πολύ σημαντικό να μάθουμε τους δείκτες γονιδίων που επηρεάζουν κυρίως την ταυτοποίηση μιας συγκεκριμένης υποκατηγορίας καρκίνου. Σε αυτή την εργασία χρησιμοποιήθηκε ο αλγόριθμος επιλογής χαρακτηριστικών βασισμένος στη συνοχή (CBFS) για την επιλογή βιοδεικτών. Οι επιλεγμένοι βιοδείκτες είναι: Leukemia M27891 at, Y 07604 at. Οι βιοδείκτες χρησιμοποιήθηκαν για να ελέγξουν την αποτελεσματικότητα του ταξινομητή.

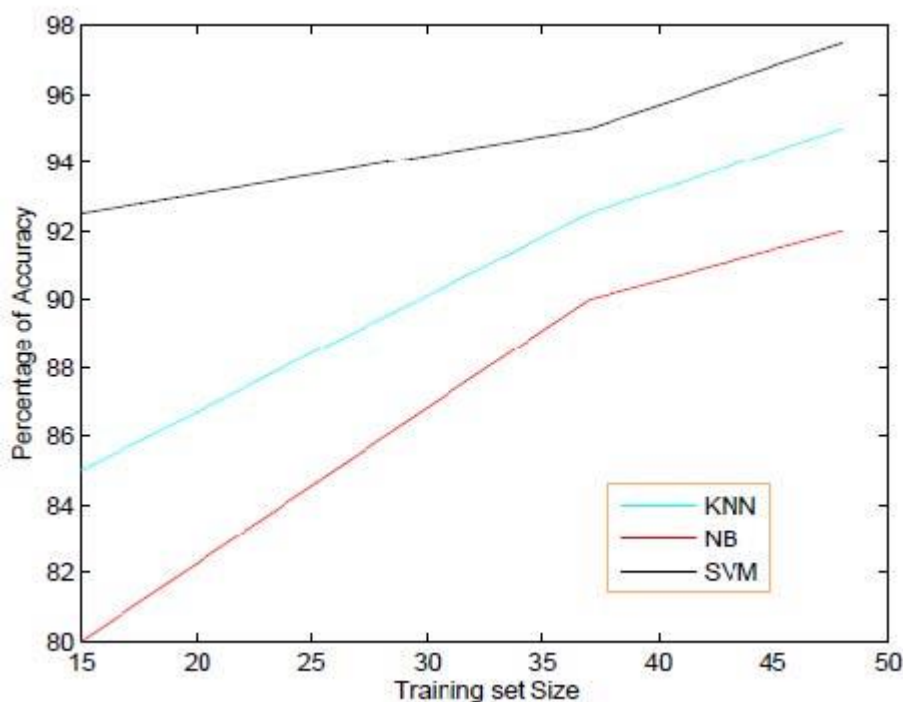


Figure 5: Απόδοση των ταξινομητών SVM, KNN και Naïve Bayes σε σύνολο δεδομένων λευχαιμίας

Οι Lee, S.-I., Celik, S., Logsdon, B. A., Lundberg, S. M., Martins, T. J., Oehler, V. G., ... Becker, P. S. (2018) παρουσίασαν μια υποσχόμενη προσέγγιση μηχανικής μάθησης ώστε να εντοπιστούν ισχυροί μοριακοί δείκτες για στοχευμένη θεραπεία της οξείας μυελοειδούς λευχαιμίας (AML). Εισήχθησαν δεδομένα από 30 ασθενείς με AML, συμπεριλαμβανομένου του προφίλ γονιδιακής έκφρασης και την ευαισθησία *vitro* σε 160 χημειοθεραπευτικά φάρμακα, μια υπολογιστική μέθοδο για τον εντοπισμό αξιόπιστων δεικτών γονιδιακής έκφρασης. Φαίνεται ότι η μέθοδος ξεπερνάει πολλές από τις πλέον

σύγχρονες προσεγγίσεις στην αναγνώριση μοριακών δεικτών. Προσδιόρισαν το SMARCA4 ως δείκτη και οδηγό ευαισθησίας σε αναστολείς τοποϊσομεράσης II, μιτοξαντρόνη, και της ετοποσιδής, στην AML δείχνοντας ότι οι κυτταρικές σειρές που μετασχηματίζονται για να έχουν υψηλή SMARCA4 έκφραση αποκαλύπτουν δραματικά αυξημένη ευαισθησία σε αυτούς τους παράγοντες.

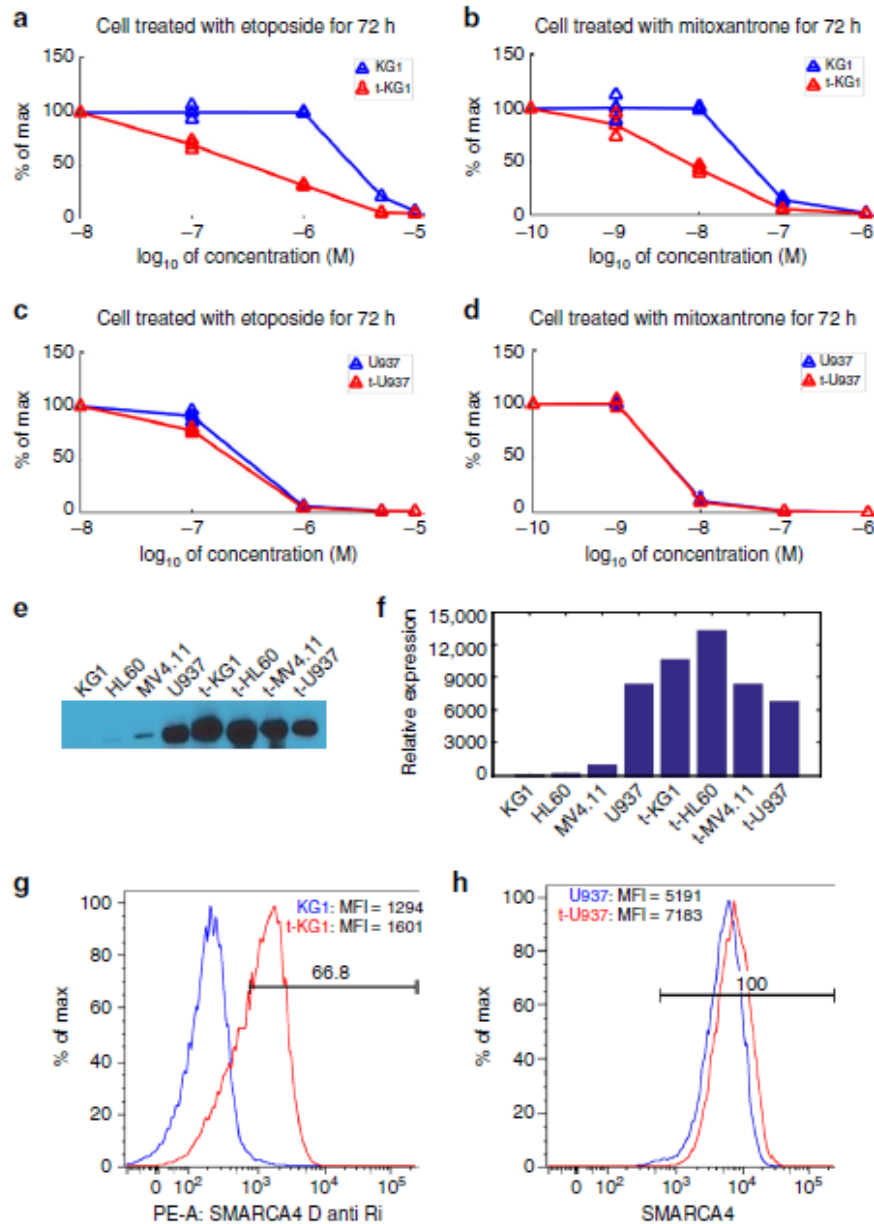


Figure 6: Πειράματα μόλυνσης πλασμιδίου SMARCA4 σε κυτταρικές σειρές KG1 και U937 για σύγκριση απόκρισης σε ετοποσιδή και μιτοξαντρόνη μεταξύ αυθεντικών και επιμολυσμένων κυττάρων.

a, b: Σύγκριση των καμπυλών δόσης-απόκρισης 72 ωρών μεταξύ κυττάρων KG1 (μπλε) και επιμολυσμένων κυττάρων KG1 (ερυθρά) όταν τα κύτταρα υποβάλλονται σε θεραπεία με (α) ετοποσιδή και (β) μιτοξαντρόνη. c, d: Σύγκριση των καμπυλών δόσης-απόκρισης μεταξύ κυττάρων U937 (μπλε) και επιμολυσμένων κυττάρων U937 (ερυθρά) όταν τα κύτταρα υποβάλλονται σε αγωγή με (γ) ετοποσιδή και (δ) μιτοξαντρόνη. Τρία τριγωνικά σημεία σε κάθε σημείο της γραμμής υποδεικνύουν μεμονωμένα σημεία δεδομένων σε διπλότυπα και τον μέσο όρο μεταξύ τους. Η γραμμή συνδέει τους μέσους όρους των διπλότυπων σε κάθε συγκέντρωση που μετράται.

ε Αναπαράσταση δυτικής κηλίδας ελέγχου και επιμολυσμένες γραμμές κυττάρων AML: KG1, U937, HL60 και MV4.11. Η μη περικομμένη έκδοση εμφανίζεται στο Συμπληρωματικό Σχ. 6. f : Ποσοτικοποιήσεις του προτύπου έκφρασης της πρωτεΐνης SMARCA4 καθεμιάς από κυτταρικές γραμμές AML στο (e). g : Κυτταρομετρία ροής δεδομένων έκφρασης επιφάνειας SMARCA4 επιβεβαιώνουν την υπερέκφραση KG1 (μπλε) έναντι επιμολυσμένου KG1 (κόκκινο). h Κυτταρομετρία ροής δεδομένων έκφρασης επιφάνειας SMARCA4 για U937 (μπλε) έναντι διαμολυσμένου U937 (κόκκινο). Σημειώνουμε ότι το U937 ήδη εκφράζει έντονα το SMARCA4, ενώ το KG1 εμφανίζει ελάχιστη έκφραση έως και μετά την επιμόλυνση.

Σύμφωνα με προηγούμενη έρευνα των deAndrés-Galiana, E. J., Fernández-Martínez, J. L., Luaces, O., Coz, J. J. del, Huergo-Zapico, L., Acebes-Huerta, A., ... González-Rodríguez, A. P. (2016) εισάγεται μια απλή μεθοδολογία βασισμένη σε μεθόδους μηχανικής μάθησης ώστε να βοηθήσουν τους γιατρούς στη λήψη αποφάσεων σε διάφορα προβλήματα που σχετίζονται με την χρόνια λεμφοκυτταρική λευχαιμία (CLL). Τα κλινικά δεδομένα που χρησιμοποίησαν ανήκουν σε μια αναδρομική μελέτη μιας ομάδας 265 Καυκάσιων που διαγνώστηκαν με CLL μεταξύ 1997 και 2007 στο νοσοκομείο Cabueñes (Asturias, Ισπανία). Διαφορετικοί μέθοδοι μηχανικής μάθησης εφαρμόστηκαν για να βρεθεί ο συντομότερος κατάλογος των πλέον διακριτικών προγνωστικών μεταβλητών για την πρόβλεψη της ανάγκης για θεραπεία χημειοθεραπείας και για την ανάπτυξη μιας αυτοάνοσης νόσου. Η εμφάνιση αυτοάνοσων νόσων είχε προβλεφθεί με πολύ υψηλή ακρίβεια (> 90%). Η αυτοάνοση ανάπτυξη της νόσου είναι επί του παρόντος μια απρόβλεπτη σοβαρή επιπλοκή της CLL. Η θεραπεία με χημειοθεραπεία έχει προβλεφθεί με χαμηλότερη ακρίβεια (80%). Η ανάλυση κινδύνου έδειξε ότι ο αριθμός των false positive και false negative είναι καλά ισορροπημένα. Η μελέτη υπογραμμίζει τη σημασία των προγνωστικών μεταβλητών που σχετίζονται με τα χαρακτηριστικά των αιμοπεταλίων, των δικτυοερυθροκυττάρων και των φυσικών δολοφόνων, οι οποίοι αποτελούν τους κύριους στόχους της αυτοάνοσης αιμολυτικής αναιμίας και της ανοσοποιητικής θρομβοκυτταροπενίας για την ανάπτυξη αυτοάνοσων νόσων και επίσης τη συνάφεια ορισμένων κλινικών μεταβλητών που σχετίζονται με τα ανοσολογικά χαρακτηριστικά των ασθενών με CLL που δεν λαμβάνονται υπόψη από τους τρέχοντες προγνωστικούς δείκτες για την πρόβλεψη της ανάγκης χημειοθεραπείας.

Οι Kumar, M., Rath, N. K., & Rath, S. K. (2016) χρησιμοποίησαν τρία σύνολα δεδομένων για τη μελέτη τους που αποτελούν μέρος της μελέτης Microarray Innovations in Leukemia (MILE). Το πρώτο σύνολο δεδομένων με τον αριθμό εισαγωγής GSE13159 περιλαμβάνει 2096 δείγματα, στα οποία κάθε δείγμα δεδομένων αντιπροσωπεύεται περαιτέρω από 54,675 διαστάσεις χαρακτηριστικών. Τα δείγματα κατηγοριοποιούνται σε 18 διαφορετικές κλάσεις. Το δεύτερο σύνολο δεδομένων περιλαμβάνει 3248 δείγματα και 1428 χαρακτηριστικά ανά δείγμα. Ομαδοποιείται επίσης σε 18 κλάσεις. Το τρίτο σύνολο

δεδομένων αποτελείται από 870 δείγματα και 54.675 χαρακτηριστικά ανά δείγμα, τα οποία κατατάσσονται σε τρεις διαφορετικές κλάσεις. Στην παρούσα εργασία, διάφορες στατιστικές μέθοδοι (τεστ) βασισμένες στον MapReduce προτείνονται για την επιλογή σχετικών χαρακτηριστικών. Μετά την επιλογή χαρακτηριστικών, ένας ταξινομητής K-πλησιέστερου γείτονα (mrKNN) με βάση τον MapReduce χρησιμοποιείται για την ταξινόμηση δεδομένων μικροσυστοιχιών. Διεξάγεται μια συγκριτική ανάλυση σε αυτά τα μοντέλα βασισμένα στον MapReduce που χρησιμοποιούν σειρές δεδομένων microarray διαφόρων διαστάσεων. Από τα λαμβανόμενα αποτελέσματα, παρατηρείται ότι αυτά τα μοντέλα καταναλώνουν πολύ λιγότερο χρόνο εκτέλεσης από τα συμβατικά μοντέλα επεξεργασία μεγάλων δεδομένων.

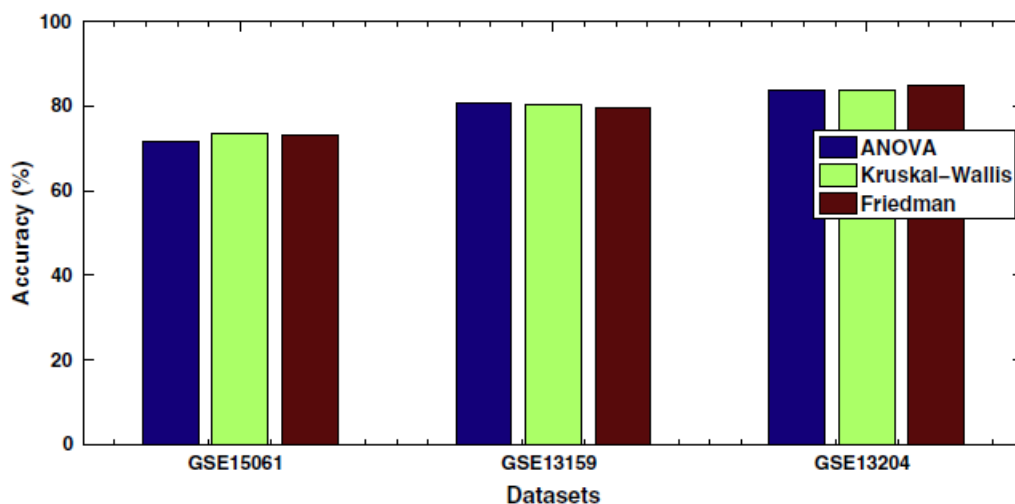


Figure 7: Ακρίβεια ταξινόμησης του mrKNN ταξινομητή με ποικίλες μεθόδους επιλογής χαρακτηριστικών χρησιμοποιώντας διαφορετικά σύνολα δεδομένων από μικροσυστοιχίες

Οι Tsanousa, A., Ntoufa, S., Papakonstantinou, N., Stamatopoulos, K., & Angelis, L. (2019) μελέτησαν τις δομές συσχέτισης των γονιδιακών εκφράσεων σε υποομάδες ασθενών με χρόνια λεμφοκυτταρική λευχαιμία. Στη χρόνια λεμφοκυτταρική λευχαιμία (CLL) η αλληλεπίδραση των λευχαιμικών κυττάρων με το μικροπεριβάλλον επηρεάζει την έκβαση του ασθενούς. Οι περιπτώσεις CLL μπορούν να χωριστούν σε δύο υποομάδες με διαφορετική κλινική πορεία βασισμένη στη μεταλλακτική κατάσταση των γονιδίων βαριάς μεταβλητής ανοσοσφαιρίνης (IGHV): μεταλλαγμένο CLL (M-CLL) και μη μεταλλαγμένο CLL (U-CLL). Δεδομένου ότι στην CLL, η διαφοροποιημένη σχέση γονιδίων μεταξύ των δύο υποομάδων έχει μεγαλύτερη σημασία από τη συμπεριφορά των επιμέρους γονιδίων, η έρευνα διερευνά τις διαφορές μεταξύ των αλληλεπιδράσεων ομάδων γονιδίων

συγκρίνοντας τις δομές συσχετισμού τους. Τα δεδομένα συλλέχθηκαν από το Τμήμα Αιματολογίας του νοσοκομείου Παπανικολάου στην Θεσσαλονίκη. Τα δύο σύνολα δεδομένων περιλάμβαναν προφίλ έκφρασης 74 γονιδίων. Η υπο-ομάδα M-CLL (n = 124) ήταν σχεδόν διπλάσια από την υπο-ομάδα U-CLL (n = 67). Η δοκιμασία Fisher και τα διαστήματα εμπιστοσύνης Zou χρησιμοποιούνται για την ανίχνευση διαφορών συντελεστών συσχέτισης. Στη συνέχεια, τα δίκτυα που δημιουργούνται από τα γονίδια που συμμετέχουν στις περισσότερες διαφορές εκτιμώνται με τη χρήση μοντέλων διαρθρωτικών εξισώσεων (SEM). Η ανάλυση ενισχύεται με γράφημα μοντελοποίησης για την απεικόνιση των διαφορών μεταξύ των ομάδων στις δομές γονιδίων των δύο υποομάδων. Η εφαρμοζόμενη μεθοδολογία αποκάλυψε ισχυρότερους συσχετισμούς μεταξύ γονιδίων σε ασθενείς με U-CLL. Χρησιμοποιώντας μοντέλα διαρθρωτικών εξισώσεων (SEM) για ανάλυση πολλαπλών ομάδων, διαφορετικές δομές γονιδίων μεταξύ των δύο ομάδων ασθενών επιβεβαιώθηκαν. Κατέληξαν στο γεγονός ότι η μελέτη των δομών συσχέτισης μπορεί να διευκολύνει την ανίχνευση των διαφορετικών προφίλ γονιδιακής έκφρασης σε υποομάδες CLL, με πιθανές εφαρμογές σε άλλες ασθένειες.

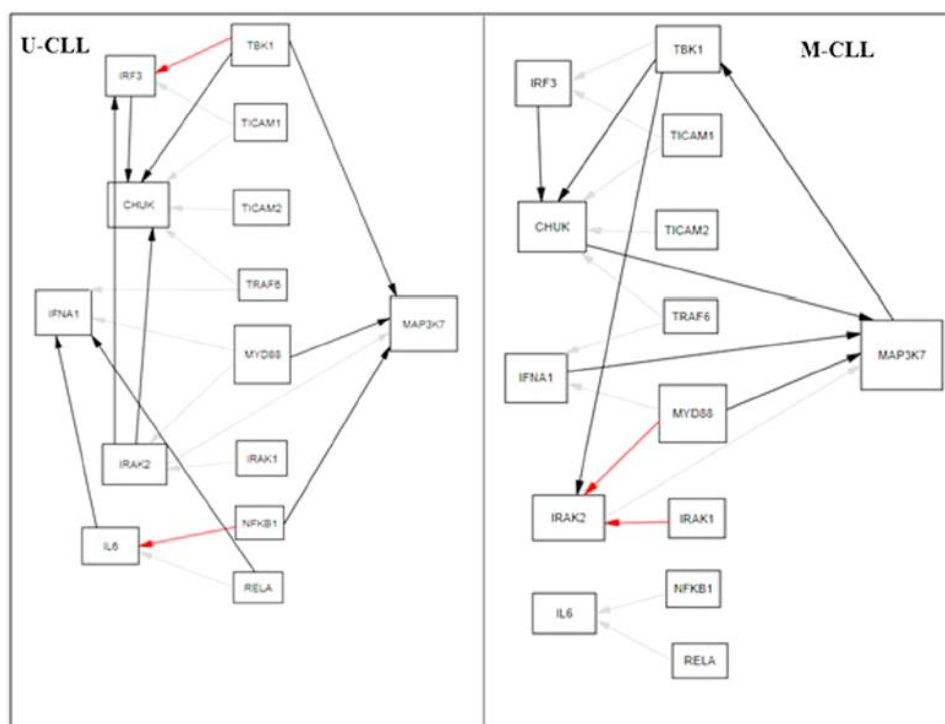


Figure 8 :Τελικά μοντέλα για την ομάδα U-CLL (αριστερά) και ομάδα M-CLL (δεξιά). Τα ανοιχτά γκρι βέλη είναι για τις αρχικές σχέσεις. Τα μαύρα βέλη είναι οι σχέσεις του τελικού μοντέλου ενώ τα κόκκινα βέλη είναι για τις κοινές σχέσεις.

Οι Haferlach, T., Kohlmann, A., Wiczorek, L., Basso, G., Kronnie, G. T., Béné, M.-C., ... Foà, R. (2010) αξιολόγησαν την χρησιμότητα του προφίλ γονιδιακής έκφρασης

ως τη μόνη δοκιμασία να αποτυπωθούν οι λευχαιμίες σε συμβατικές κατηγορίες μυελοειδούς και λεμφοειδών κακοηθειών. Διεξήγαγαν έρευνα σε 11 εργαστήρια σε τρεις ηπείρους, η οποία περιελάμβανε 3.334 ασθενείς. Μια διερευνητική μελέτη αναδρομικής φάσης I σχεδιάστηκε για την ανακάλυψη βιοδεικτών και δημιουργήθηκε προφίλ έκφρασης ολόκληρου του γονιδιώματος από 2.143 ασθενείς με λευχαιμικά και μυελοδυσπλαστικά σύνδρομα. Η διαγνωστική ακρίβεια βασισμένη στο προφίλ της γονιδιακής έκφρασης επικυρώθηκε περαιτέρω σε δεύτερη φάση μελέτης μιας ανεξάρτητης ομάδας 1.191 ασθενών. Με βάση τα 2.096 δείγματα, η μελέτη σταδίου I πέτυχε 92,2% ακρίβεια ταξινόμησης για όλες τις 18 ξεχωριστές κατηγορίες που ερευνήθηκαν (μέση εξειδίκευση 99,7%). Σε μια δεύτερη ομάδα 1.152 ασθενών που συλλέχθηκαν μελλοντικά, ένα σχήμα ταξινόμησης πέτυχε 95,6% ενδιάμεση ευαισθησία και 99,8% ενδιάμεση εξειδίκευση για 14 τυποποιημένους υποτύπους οξείας λευχαιμίας (οκτώ οξείας λεμφοβλαστικής λευχαιμίας και έξι κατηγορίες οξείας μυελογενούς λευχαιμίας, n= 693). Σε 29 (57%) από 51 διαφορετικές περιπτώσεις, τα αποτελέσματα της μικροσυστοιχίας υπερέβησαν τις συνήθεις διαγνωστικές μεθόδους.

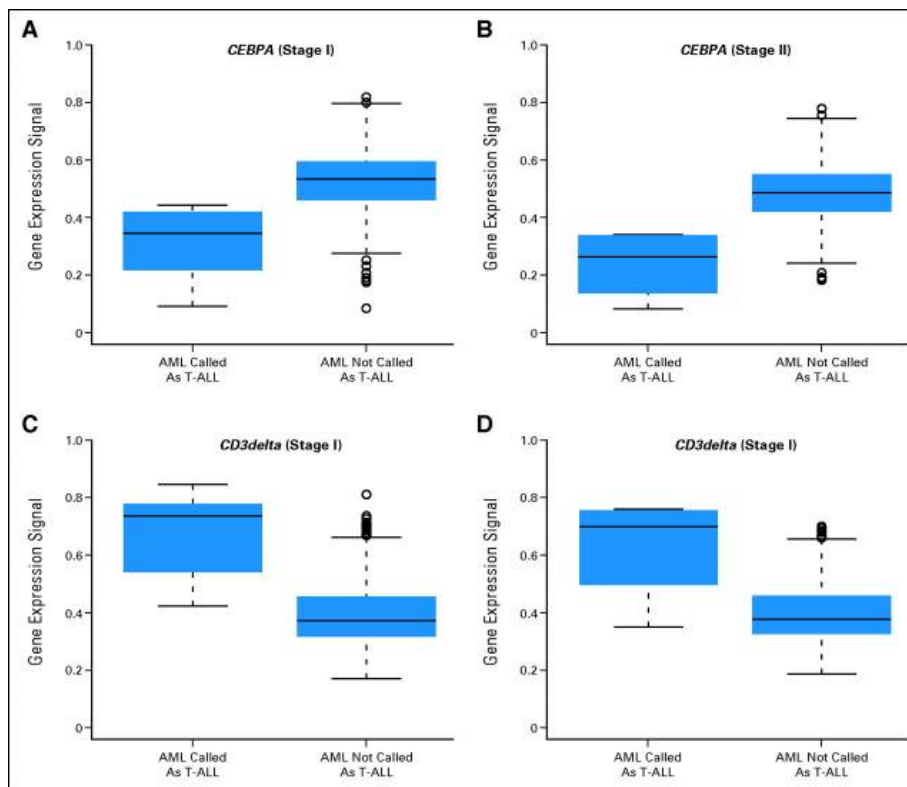


Figure 9: Περιπτώσεις AML που ταξινομήθηκαν εσφαλμένα ως T-ALL

Σε διατριβή που πραγματοποίησαν οι P. Jagadev και H. G. Virani (2017) προσπαθήσαν να επινοήσουν μια μεθοδολογία για την ανίχνευση της λευχαιμίας

χρησιμοποιώντας τεχνικές επεξεργασίας εικόνας, αυτοματοποιώντας έτσι τη διαδικασία ανίχνευσης. Το σύνολο δεδομένων που χρησιμοποιήθηκε περιλαμβάνει 220 εικόνες επιχρισμάτων αίματος λευχαιμικών και μη λευχαιμικών ασθενών. Οι αλγόριθμοι τμηματοποίησης εικόνας που έχουν χρησιμοποιηθεί είναι : k-means clustering, marker controlled watershed και αλγόριθμος τμηματοποίησης βάσει χρώματος HSV. Τα μορφολογικά συστατικά των φυσιολογικών και λευχαιμικών λεμφοκυττάρων διαφέρουν σημαντικά. Ως εκ τούτου, έχουν εξαχθεί διάφορα χαρακτηριστικά από τις τμηματοποιημένες εικόνες λεμφοκυττάρων. Η λευχαιμία ταξινομείται περαιτέρω στους τύπους και τους δευτερεύοντες τύπους της χρησιμοποιώντας τον ταξινομητή Support Vector Machine, ο οποίος είναι ένας ταξινομητής μηχανικής μάθησης. Αυτή η διατριβή στοχεύει στην ανίχνευση της λευχαιμίας και στον προσδιορισμό εάν είναι AML, CML, CLL ή ALL κάνοντας έτσι τη διαδικασία ταξινόμησης ένα βήμα παραπέρα στον τομέα της έρευνας, επειδή οι περισσότερες από τις προηγούμενες μεθόδους περιορίστηκαν στην ανίχνευση της λευχαιμίας ή στην ταξινόμησή της σε έναν ή δύο υπότυπους.

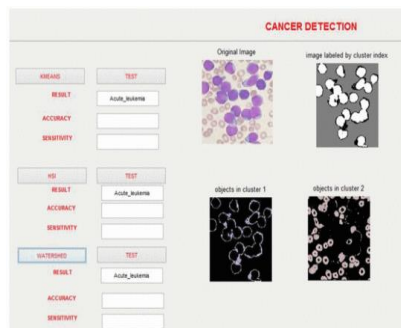


Figure 10: Ανίχνευση της Λευχαιμίας και ταξινόμηση χρησιμοποιώντας SVM

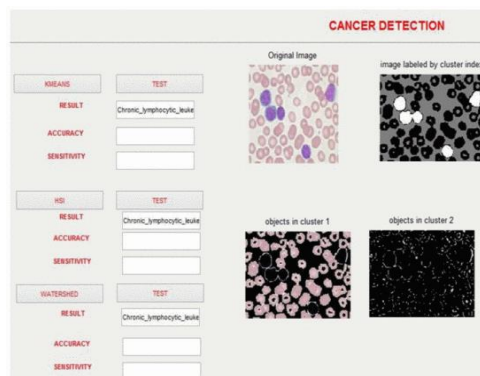


Figure 11: Ανίχνευση της Λευχαιμίας και ταξινόμηση χρησιμοποιώντας SVM

Κεφάλαιο 3

3.1 Αλγόριθμοι Μηχανικής Μάθησης

Η Μηχανική Μάθηση είναι κλάδος της τεχνητής νοημοσύνης και μία διαδικασία για την ανακάλυψη νέας γνώσης όπως σχέσεις, πρότυπα, αλλαγές και σημαντικές δομές από μεγάλα ποσά δεδομένων αποθηκευμένα σε βάσεις δεδομένων κι άλλα αποθετήρια πληροφοριών. Στη διαδικασία της μηχανικής μάθησης πρώιμα δεδομένα ελέγχονται και μελλοντικοί κανόνες μετρώνται για την ανάλυσή τους. Η τεχνική αυτή περιλαμβάνει ένα μεγάλο αριθμό φάσεων. Κατανόηση των δεδομένων, επεξεργασία δεδομένων, μοντελοποίηση, αξιολόγηση, ανάπτυξη. Η μηχανική μάθηση έχει αποδειχθεί ότι είναι πολύ χρήσιμη στο πεδίο της ανάλυσης ιατρικών δεδομένων αφού αυξάνει τη διαγνωστική ακρίβεια έτσι ώστε να μειωθεί το κόστος της θεραπείας του ασθενούς και να αποταμιευτούν ανθρώπινοι πόροι. Υπάρχουν πολλές τεχνικές μηχανικής μάθησης όπως ταξινόμηση, ομαδοποίηση, νευρωνικά δίκτυα, παλινδρόμηση, κανόνες συσχέτισης.

Η Μηχανική Μάθηση αφορά αλγορίθμους οι οποίοι έχουν την ικανότητα να “θυμούνται” ώστε να πραγματοποιήσουν ακριβείς προβλέψεις. Οι αλγόριθμοι αυτοί ‘ανακαλούν στη μνήμη τους’ δεδομένα τα οποία έχουν δοθεί στο σύστημα μάθησης από πριν και συνήθως τα ονομάζουμε σύνολα εκπαίδευσης. Το σύνολο δεδομένων είναι κρίσιμος παράγοντας για την αξιοπιστία των προβλέψεων του συστήματος μάθησης και γι’ αυτό τα δεδομένα πρέπει να είναι γνήσια και η ποσότητα τους ανάλογη με το μέγεθος των δειγμάτων που χρειάζεται ο αλγόριθμος για να μάθει και να πραγματοποιήσει πειράματα.

Βασικές έννοιες των αλγορίθμων μάθησης :

Παρα(δείγματα): Τα δείγματα αντιστοιχούν σε ένα σύνολο ασθενών και υγιών ανθρώπων που θα χρησιμοποιηθούν για εκπαίδευση αλγορίθμων και πειράματα.

Χαρακτηριστικά: Τα χαρακτηριστικά ενός δείγματος είναι το σύνολο των γνωρισμάτων που αφορούν έναν ασθενή ή υγιή άνθρωπο και συνήθως αντιστοιχούν σε αποτελέσματα εργαστηριακών ελέγχων, αλληλουχίες γονιδίων και στοιχεία κλινικών εξετάσεων.

Κλάσεις: Κατηγορίες στις οποίες ταξινομούνται τα δείγματα όπως νοσούντες ή μη νοσούντες, τύποι ασθενειών.

Σύνολο εκπαίδευσης: Καλείται το σύνολο των δειγμάτων που δίνεται στον αλγόριθμο για να εκπαιδευτεί. Ο αλγόριθμος μαθαίνει πως έχουν ταξινομηθεί οι άνθρωποι με συγκεκριμένους δείκτες και χαρακτηριστικά.

Σύνολο ελέγχου: Καλείται το σύνολο των δειγμάτων που χρησιμοποιείται για πειράματα και αξιολόγηση της πρόβλεψης. Η αξιολόγηση γίνεται με συγκεκριμένους δείκτες που ονομάζονται ακρίβεια, εξειδίκευση και ευαισθησία.

3.2 Μέθοδος

Για τη διεξαγωγή συμπερασμάτων χρησιμοποιήθηκε το εργαλείο Weka ως εργαλείο μηχανικής μάθησης βάση παραδειγμάτων, εφαρμόζοντας επτά διαφορετικούς αλγόριθμους. Στη συγκεκριμένη έρευνα εφαρμόστηκαν τεχνικές μηχανικής μάθησης, ώστε να βρεθούν τα σημαντικότερα γονίδια που οδηγούν στην πρόβλεψη της σωστής κατηγορίας λευχαιμίας και αλγόριθμοι ώστε να επιλεγεί αυτός, ο οποίος πετυχαίνει καλύτερα αποτελέσματα (προβλέψεις) στην ταξινόμηση των τύπων λευχαιμίας των ασθενών.

Η έρευνα που παρουσιάζεται χρησιμοποιεί τους εξής ταξινομητές: J48, JRip, K-NN, Naive Bayes, K-Star, Support Vector Machine, Random Forest. Τα σύνολα δεδομένων που μελετώνται για την λευχαιμία περιέχουν 72 παραδείγματα και 7.130 χαρακτηριστικά. Το πρώτο σύνολο δεδομένων περιέχει δύο υποτύπους-κλάσεις (ALL-AML), το δεύτερο τρείς (AML, B-cell, T-cell) και το τρίτο τέσσερις (B-cell, T-cell, BM (bone marrow), PB (peripheral blood)). Η απόδοση των ταξινομητών αξιολογήθηκε με βάση το ποσοστό ακριβείας, ευαισθησίας, εξειδίκευσης.

3.3 Προεπεξεργασία δεδομένων

Το στάδιο της προ-επεξεργασίας των δεδομένων περιλαμβάνει την επιλογή χαρακτηριστικών, το μετασχηματισμό των δεδομένων και την εκκαθάριση των δεδομένων. Στη συγκεκριμένη έρευνα επιλέχθηκαν τρεις διαφορετικές τεχνικές επιλογής χαρακτηριστικών : Information Gain, Correlation Attribute Evaluation και Cfs Subset Evaluation.

Η τεχνική Information Gain αξιολογεί την αξία ενός γονιδίου μετρώντας το κέρδος πληροφοριών σε σχέση με την κλάση. Στο πρώτο σύνολο δεδομένων το γονίδιο U46499 φαίνεται να συνεισφέρει περισσότερη πληροφορία από όλα τα υπόλοιπα, στο δεύτερο το γονίδιο M27891 και στο τρίτο το γονίδιο M23197.

Η τεχνική Correlation attribute evaluation αξιολογεί την αξία ενός γονιδίου μετρώντας τη συσχέτιση μεταξύ αυτού και του υποτύπου. Στο πρώτο σύνολο των δεδομένων με την εφαρμογή αυτής της τεχνικής υποδηλώνεται ότι το γονίδιο AFFX-HSAC07/X00351_5_at έχει την υψηλότερη συσχέτιση με την κλάση εξόδου, στο δεύτερο το γονίδιο U05259_rna1 και στο τρίτο το γονίδιο U05259_rna1.

Εφαρμόζοντας την τεχνική επιλογής χαρακτηριστικών Cfs Subset Evaluation επιλέγεται συγκεκριμένος αριθμός χαρακτηριστικών. Η συγκεκριμένη τεχνική επιλέγει να κρατήσει 82 από τα 7.130 χαρακτηριστικά (γονίδια).

3.4 Σημαντικά Χαρακτηριστικά

Για τον προσδιορισμό των σημαντικότερων γονιδίων αρχικά εκτελούμε τον αλγόριθμο επιλογής χαρακτηριστικών CfsSubsetEvaluation με την προεπιλεγμένη μέθοδο αναζήτησης Best First για κάθε σύνολο δεδομένων. Ο αλγόριθμος κρατάει 82 χαρακτηριστικά και την κλάση. Τα είκοσι σημαντικότερα είναι:

1.D63880_at non-SMC condensin I complex subunit D2	6. D88270_at immunoglobulin lambda locus	11. J04615_at	16. M11722_at Human terminal transferase mRNA
2.D83776_at	7. D89667_at	12. L07633_at	17. M14539_at

terminal uridylyl transferase 4			coagulation factor XIII A chain
3.D86983_at peroxidase	8. HG3088-HT3263_at	13. L35240_at	18. M19507_at myeloperoxidase
4.D87119_at tribbles pseudokinase 2	9. J03798_at small nuclear ribonucleoprotein D1 polypeptide	14. L47738_at	19. M21121_at C-C motif chemokine ligand 5
5. D87434_at sushi domain containing 6	10. J04132_at	15. M11147_at	20. M22960_at cathepsin A

Πίνακας 1 : Είκοσι σημαντικότερα γονίδια με CfsSubsetEvaluation

Στη συνέχεια εκτελούμε τον InfoGain με την μέθοδο αναζήτησης Ranker. Τα είκοσι σημαντικότερα είναι:

1.U46499_at microsomal glutathione S-transferase 1	6.M84526_at complement factor D	11. M83652_s_at complement factor properdin	16.D88270_a
2.M23197_at CD33 molecule	7.M31523_at Human transcription factor (E2A) mRNA	12. M92287_at cyclin D3	17. X59417_at proteasome 20S subunit alpha 6
3.X95735_at zyxin	8.M11722_at Human terminal transferase mRNA	13. X61587_at ras homolog family member G	18. U05259_rna1_at
4.M27891_at	9. M19507_at	14. X62654_rna1_at	19. M55150_at
5.L09209_s_at	10. M63379_at	15. X62320_at	20. M12959_s_at

Πίνακας 2 : Είκοσι σημαντικότερα γονίδια με InfoGain

Τέλος εκτελούμε τον αλγόριθμο επιλογής σημαντικότερων χαρακτηριστικών Correlation με την μέθοδο αναζήτησης Ranker. Τα είκοσι σημαντικότερα είναι:

1. X95735_at zyxin	6. M27891_at	11. M11147_at Human ferritin L chain mRNA	16. M31523_at Human transcription factor (E2A) mRNA
2. M23197_at CD33 molecule	7.M22960_at Human protective protein mRNA	12.M19507_at Human myeloperoxidase mRNA	17. Y07604_at H.sapiens mRNA for nucleoside-diphosphate kinase
3.M84526_at complement factor D	8.M55150_at Human fumarylacetoacetate hydrolase Mrna	13. X59417_at proteasome 20S subunit alpha 6	18. M83652_s_at Homo sapiens complement component properdin mRNA
4.L09209_s_at amyloid beta precursor like protein 2	9.M62762_at Human vacuolar H+ ATPase proton channel subunit mRNA	14. X62654_rna1_at CD63 molecule	19. U05259_rna1_at Human MB-1 gene
5.U46499_at	10.X61587_at H.sapiens rhoG mRNA for GTPase	15. M92287_at cyclin D3	20. S50223_at zinc finger protein 22

Πίνακας 3: : Είκοσι σημαντικότερα γονίδια με Correlation

3.5 Αλγόριθμοι μηχανικής μάθησης

J48: Ο J48 ανήκει στην κατηγορία των δέντρων αποφάσεων. Ένα δέντρο αποφάσεων αποτελείται από κόμβους αποφάσεων και κόμβους φύλλων, όπου ο κόμβος αποφάσεων καθορίζει μια δοκιμή για ένα από τα χαρακτηριστικά και ένας κόμβος φύλλων αντιπροσωπεύει την τιμή κλάσης. Κάθε μονοπάτι από τον ριζικό κόμβο προς τον κόμβο των φύλλων είναι κανόνας. Το σφάλμα ταξινόμησης είναι ο βασικός παράγοντας απόδοσης για το δέντρο αποφάσεων. Στο δέντρο αποφάσεων όταν οι κατηγορίες κλάσεων είναι περισσότερες, η ακρίβεια ταξινόμησης μειώνεται σημαντικά. Πρώτα, ελέγχεται αν όλες οι περιπτώσεις ανήκουν στην ίδια κλάση. Τότε το δέντρο είναι φύλλο και φέρει την

ετικέτα με αυτήν την κατηγορία. Στη συνέχεια για κάθε χαρακτηριστικό, υπολογίζονται οι πληροφορίες και το κέρδος πληροφοριών. Βρίσκεται το καλύτερο χαρακτηριστικό διαχωρισμού. Εξετάζει το κέρδος πληροφοριών (information gain – διαφορά στην εντροπία) που προκύπτει από την επιλογή ενός χαρακτηριστικού για το διαχωρισμό των δεδομένων. Το χαρακτηριστικό με το υψηλότερο ομαλοποιημένο κέρδος είναι αυτό που χρησιμοποιείται για να ληφθεί μια απόφαση. Ο αλγόριθμος επαναλαμβάνεται για μικρότερες υπολίσστες δεδομένων. (Κύρκος, Ε., 2016).

JRip: Ο αλγόριθμος αυτός εφαρμόζει τη μέθοδο Ripper, Repeated Incremental Pruning to Produce Error Reduction. Περνά από 4 στάδια: ανάπτυξη κανόνα, κλάδεμα, βελτιστοποίηση και επιλογή. Τα δεδομένα εκπαίδευσης χωρίζονται σε ένα αναπτυσσόμενο σετ και ένα σετ κλάδεσης. Πρώτον, σχηματίζεται ένα αρχικό σύνολο κανόνων, χρησιμοποιώντας κάποια ευρετική μέθοδο. Αυτό το υπερβολικά μεγάλο σύνολο κανόνων απλουστεύεται στη συνέχεια επανειλημμένα εφαρμόζοντας από ένα σύνολο τελεστών κλαδέματος. Σε κάθε στάδιο απλοποίησης, ο επιλεγμένος τελεστής κλαδέματος είναι αυτός που αποδίδει τη μεγαλύτερη μείωση σφάλματος στο σετ κλάδεσης. Η απλοποίηση τελειώνει όταν εφαρμόζοντας οποιονδήποτε τελεστή κλαδέματος θα αυξήσει το σφάλμα στο κλάδεμα.

K-NN (k-Nearest Neighbors): Η μέθοδος των k-Πλησιέστερων Γειτόνων (k-Nearest Neighbors - kNN) είναι ένας αλγόριθμος της οικογένειας των ταξινομητών βασισμένων σε παραδείγματα. Κάθε δείγμα μπορεί να θεωρηθεί ως ένα σημείο στον χώρο των δύο διαστάσεων. Ένα δείγμα X απέχει από ένα άλλο δείγμα Y, απόσταση $d(X,Y)$ μέσα στον δισδιάστατο χώρο. Η απόσταση $d(X,Y)$ μπορεί να υπολογιστεί ως η Ευκλείδεια απόσταση σύμφωνα με την Εξίσωση 3.1:

$$d(X,Y) = \sqrt{((x_1 - y_1)^2 + (x_2 - y_2)^2)} \blacksquare$$

όπου x_1, y_1 οι τιμές των X και Y για την πρώτη διάσταση και x_2, y_2 οι τιμές των X και Y για την δεύτερη διάσταση. Σύμφωνα με τον αλγόριθμο k-NN, ο χρήστης προκαθορίζει την τιμή της σταθερής παραμέτρου k. Ο αλγόριθμος αναζητά μέσα στον δισδιάστατο χώρο τα k σημεία-δείγματα που βρίσκονται πλησιέστερα στο νέο δείγμα. Ο ταξινομητής εκχωρεί το νέο δείγμα στην κλάση που πλειοψηφεί μεταξύ των k πλησιέστερων γειτόνων. Εάν

οριστεί ότι $k=1$, τότε η νέα παρατήρηση εκχωρείται στην κλάση του πιο όμοιου δείγματος εκπαίδευσης. Ο kNN στο εργαλείο WEKA ονομάζεται IBk (το IB σημαίνει Instance Based, δηλαδή Βάσει Υποθέσεων, και το k επιτρέπει να καθοριστεί ο αριθμός των Γειτόνων που θα εξετάσουμε) (Κύρκος, Ε., 2016).

Naive Bayes: Ο Naïve Bayes για να κατηγοριοποιήσει τα δεδομένα του, χρησιμοποιεί τη θεωρία των πιθανοτήτων. Υπολογίζει την πιθανότητα κάποιο από τα δεδομένα να ανήκει σε μία κατηγορία. Βασίζεται στο θεώρημα του Bayes και κατ' επέκταση, κάνει χρήση δεσμευμένων πιθανοτήτων.

$$P(b|a) = \frac{P(a|b) \times P(b)}{P(a)}$$

Όπου $P(a)$ είναι η πιθανότητα του γεγονότος a και $P(a|b)$ είναι η δεσμευμένη πιθανότητα του a δεδομένου του γεγονότος b . (Κύρκος, Ε., 2016).

K-Star: Ο αλγόριθμος αυτός είναι ένας ταξινομητής που βασίζεται στη μάθηση από παραδείγματα. Διαφέρει από τους ταξινομητές του ίδιου τύπου καθώς χρησιμοποιεί μια συνάρτηση απόστασης βασισμένης στην εντροπία. Στόχος είναι η εύρεση των μικρότερων διαδρομών μεταξύ ενός καθορισμένου ζεύγους κορυφών σε ένα δεδομένο κατευθυνόμενο σταθμισμένο γράφημα (Aljazzar, H., & Leue, S. 2011)

Support Vector Machine: Ο Support Vector Machine (SVM) είναι ένας αλγόριθμος επιβλεπόμενης μηχανικής μάθησης που μπορεί να χρησιμοποιηθεί τόσο για ταξινόμηση όσο και για παλινδρόμηση. Ωστόσο χρησιμοποιείται κυρίως για ταξινόμηση. Βασική ιδέα είναι η κατασκευή ενός υπερεπιπέδου (hyperplane), το οποίο διαχωρίζει τις κλάσεις και λειτουργεί ως συνάρτηση απόφασης. Οι νέες παρατηρήσεις κατηγοριοποιούνται ανάλογα με την πλευρά του υπερεπιπέδου στην οποία βρίσκονται. Στην περίπτωση του SVM αλγόριθμου, ένα σημειακό δεδομένο θεωρείται σαν ένα p -διάστατο διάνυσμα (μία λίστα από p αριθμούς). Το σημαντικό θέμα είναι αν υπάρχει η δυνατότητα διαχωρισμού του συγκεκριμένου σημείου σ' ένα $(p-1)$ -διάστατο υπερπλάνιο. Τα υπερπλάγια είναι όρια απόφασης, που βοηθούν στην ταξινόμηση των νέων σημείων. Τα σημεία που εμπίπτουν σε κάθε πλευρά του υπερπλαισίου μπορούν να αποδοθούν σε διαφορετικές κατηγορίες. Τα διανύσματα στήριξης είναι σημεία, που είναι πλησιέστερα στο υπερπλάνιο και επηρεάζουν τη θέση και τον προσανατολισμό του ίδιου του υπερπλαισίου. Χρησιμοποιώντας αυτούς τους φορείς στήριξης, μεγιστοποιούμε το

περιθώριο (margin) του ταξινομητή. Συνεπώς, η διαγραφή των διανυσμάτων στήριξης θα αλλάξει τη θέση του υπερπλαισίου. Αυτά είναι τα σημεία που μας βοηθούν να κατασκευάσουμε τον SVM. (Rohith Gandhi, 2018)

Random Forest: Για να γίνει κατανοητός ο τρόπος λειτουργίας του αλγορίθμου αυτού πρέπει να εξηγήσουμε πώς λειτουργούν τα δέντρα απόφασης, καθώς είναι τα δομικά στοιχεία του Random Forest.

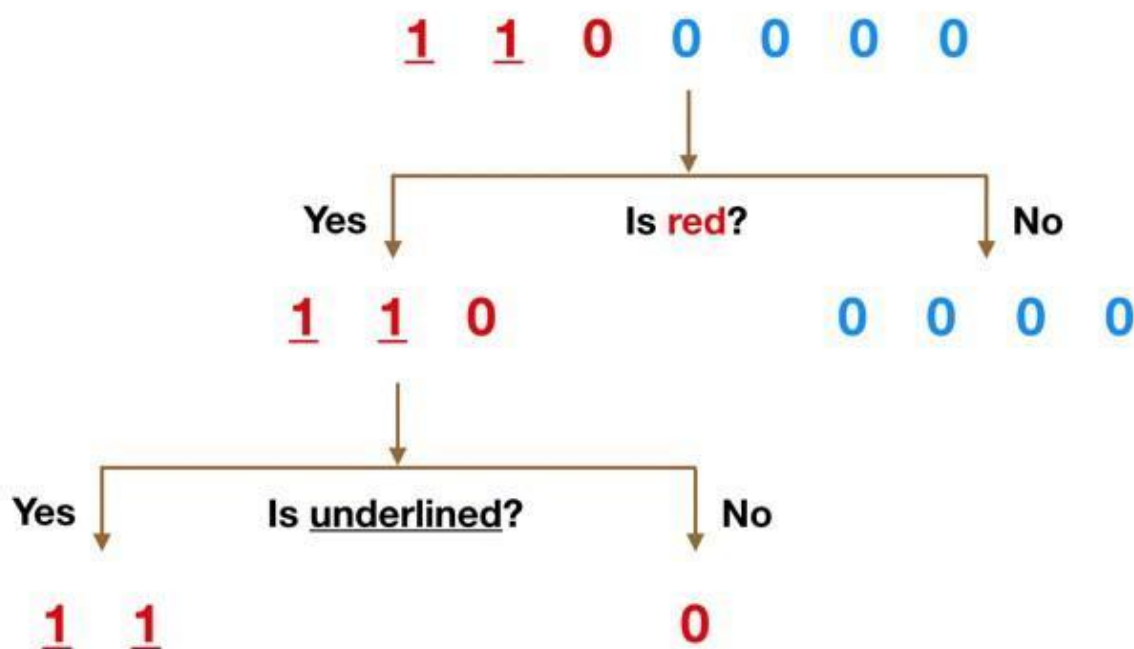


Figure 12: Λειτουργία των δέντρων απόφασης

Ας υποθέσουμε ότι το σύνολο δεδομένων μας αποτελείται από τους αριθμούς στο πάνω μέρος του σχήματος. Έχουμε δύο 1 και πέντε 0 (1 και 0 είναι οι κλάσεις) και επιθυμούμε να διαχωρίσουμε τις κλάσεις χρησιμοποιώντας τα χαρακτηριστικά τους. Τα χαρακτηριστικά είναι το χρώμα (κόκκινο έναντι μπλε) και αν το δείγμα είναι υπογραμμισμένο ή όχι. Το χρώμα μοιάζει ένα αρκετά προφανές χαρακτηριστικό για τον διαχωρισμό αλλά ένα από τα 0 είναι μπλε. Έτσι μπορούμε να χρησιμοποιήσουμε την ερώτηση, "Είναι κόκκινο;" για να χωρίσουμε τον πρώτο κόμβο. Ο κόμβος σε ένα δέντρο μοιάζει ως το σημείο όπου η διαδρομή χωρίζεται σε δύο - δείγματα που πληρούν τα κριτήρια τοποθετούνται κάτω από το κλαδί Ναι και αυτά που δεν πληρούν κάτω από το

Όχι. Το κλαδί Όχι (τα μπλέ) είναι όλα 0 τώρα, έτσι έχουμε τελειώσει, αλλά το κλαδί Ναι μπορεί ακόμα να χωριστεί περαιτέρω. Τώρα μπορεί να χρησιμοποιηθεί το δεύτερο χαρακτηριστικό και να γίνει η ερώτηση "Είναι υπογραμμισμένο;" ώστε να γίνει ένα δεύτερο split. Τα δύο 1 που είναι υπογραμμισμένα πηγαίνουν κάτω από το δευτερεύον τμήμα Yes και το 0 που δεν είναι υπογραμμισμένο πηγαίνει κάτω από το σωστό υποσύστημα. Στην πραγματική ζωή τα δεδομένα δεν είναι τόσο καθαρά, αλλά η λογική που χρησιμοποιεί ένα δέντρο αποφάσεων παραμένει η ίδια.

Σε κάθε κόμβο, θα ρωτήσει -Ποιο χαρακτηριστικό θα μου επιτρέψει να χωρίσω τα δείγματα κατά τρόπο ώστε οι ομάδες που προκύπτουν να είναι όσο το δυνατόν πιο διαφορετικές μεταξύ τους (και τα μέλη κάθε προκύπτουσας υποομάδας να είναι όσο το δυνατόν πιο παρόμοια μεταξύ τους). Ο Random Forest αποτελείται από μεγάλο αριθμό μεμονωμένων δέντρων απόφασης που λειτουργούν ως σύνολο. Κάθε μεμονωμένο δέντρο ξετυλίγει μια πρόβλεψη κλάσης ή αλλιώς δίνει μια κατηγοριοποίηση και η κλάση με τις περισσότερες ψήφους γίνεται η πρόβλεψη του μοντέλου μας. (Κύρκος, E., 2016).

3.6 Ανάλυση δεδομένων

Η έρευνα ασχολείται με την απόδοση των αλγορίθμων ταξινόμησης που προαναφέρθηκαν πάνω στα δεδομένα των τριών συνόλων δεδομένων. Συγκρίνει τα αποτελέσματα από τις μεθόδους ταξινόμησης. Η ανάλυση που πραγματοποιήθηκε βασίστηκε στα αποτελέσματα που προέκυψαν από τον αριθμό των σωστών ταξινομημένων παραδειγμάτων. Η απόδοση και τα αποτελέσματα φαίνονται στους παρακάτω πίνακες. Η πρώτη μελέτη πραγματοποιήθηκε με τα 20 σημαντικότερα γονίδια, η δεύτερη με 50.

	Info Gain	Correlation	CfsSubset evaluation
J48	81,94	84,72	94,44
JRIP	91,66	90,27	91,66
K-NN (IBK)	95,83	94,44	97,22
Naive Bayes	95,83	95,83	98,61
K-Star	95,83	95,83	90,27
SMO	95,83	94,44	97,22
Random Forest	97,22	94,44	98,61

Πίνακας 4: AML-ALL accuracy 20 genes

AML-ALL (20 genes)

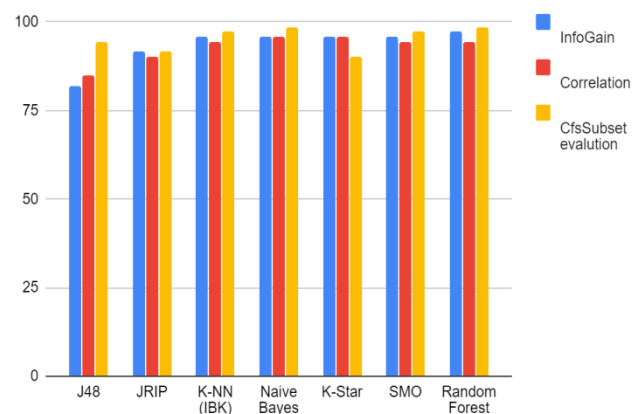


Figure 13: AML-ALL accuracy 20 genes

	Info Gain	Correlation	CfsSubset evaluation
J48	81,9	84,7	81,9
JRIP	91,7	90,3	90,3
K-NN (IBK)	95,8	94,4	95,8
Naive Bayes	95,8	95,8	95,8
K-Star	95,8	95,8	95,8
SMO	95,8	94,4	95,8
Random Forest	97,2	94,4	97,2

Πίνακας 5: AML-ALL sensitivity 20 genes

AML-ALL (20 genes) Sensitivity

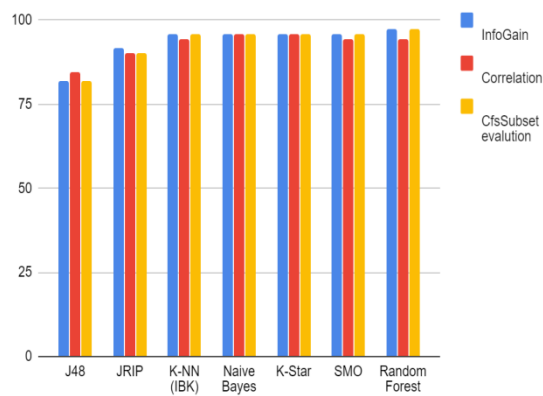


Figure 14: AML-ALL sensitivity 20 genes

	Info Gain	Correlation	CfsSubset evaluation
J48	82,9	82,5	82,9
JRIP	88,1	87,3	85,5
K-NN (IBK)	95,9	91,4	95,9
Naive Bayes	95,9	95,9	95,9
K-Star	94	92,2	94
SMO	94	91,4	94
Random Forest	94,8	93,3	94,8

Πίνακας 6: AML-ALL specificity 20 genes

AML-ALL (20 genes) Specificity

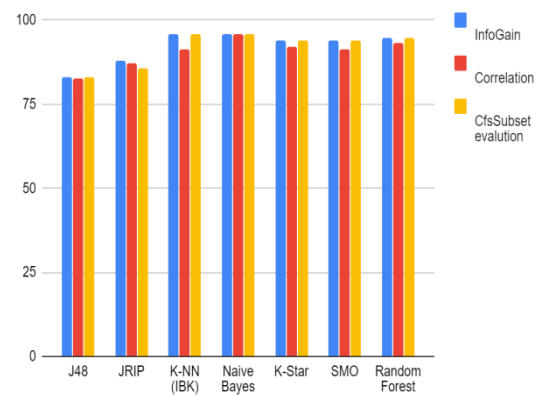


Figure 15: AML-ALL specificity 20 genes

	Info Gain	Correlation	CfsSubset Evaluation
J48	86,11	87,5	87,5
JRIP	88,88	79,16	88,88
K-NN (IBK)	94,44	88,88	91,66
Naive Bayes	94,44	94,44	95,83
K-Star	95,83	91,66	83,33
SMO	94,44	93,05	93,05
Random Forest	97,22	95,83	91,66

Πίνακας 7: AML, B-cell, T-cell accuracy 20genes

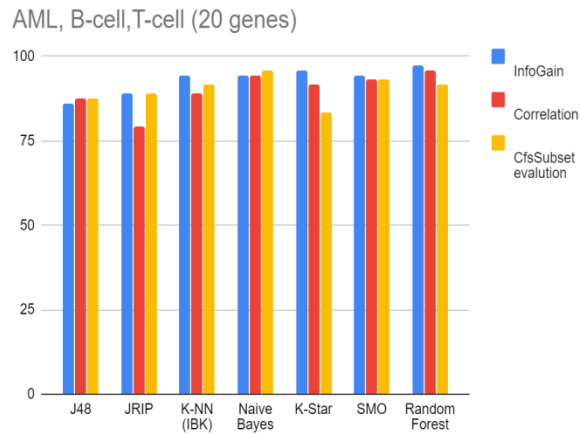


Figure 16: AML, B-cell, T-cell accuracy 20 genes

	Info Gain	Correlation	CfsSubset evaluation
J48	86,1	87,5	87,5
JRIP	88,9	79,2	88,9
K-NN (IBK)	94,4	88,9	91,7
Naive Bayes	94,4	94,4	95,8
K-Star	95,8	91,7	83,3
SMO	94,4	93,1	93,1
Random Forest	97,2	95,8	91,7

Πίνακας 8: AML, B-cell, T-cell sensitivity 20 genes

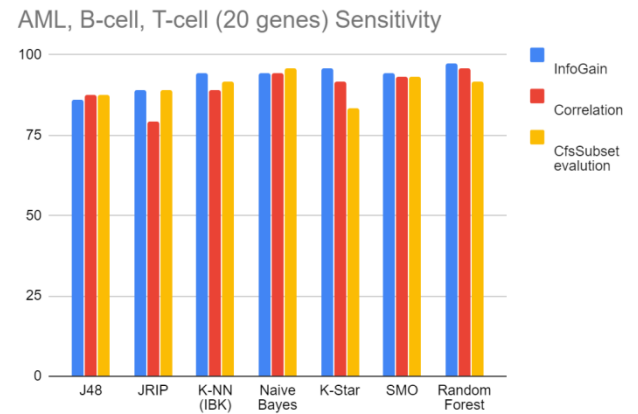


Figure 17: AML, B-cell, T-cell sensitivity 20 genes

	Info Gain	Correlation	CfsSubset evaluation
J48	89,6	93,9	92,5
JRIP	92,2	84,6	94,6
K-NN (IBK)	98,1	95,4	92,3
Naive Bayes	96,8	97,6	97,5
K-Star	98,9	98,3	89
SMO	98,7	97,9	94,7
Random Forest	98,5	98,9	94,5

Πίνακας 9: AML, B-cell, T-cell specificity 20 genes

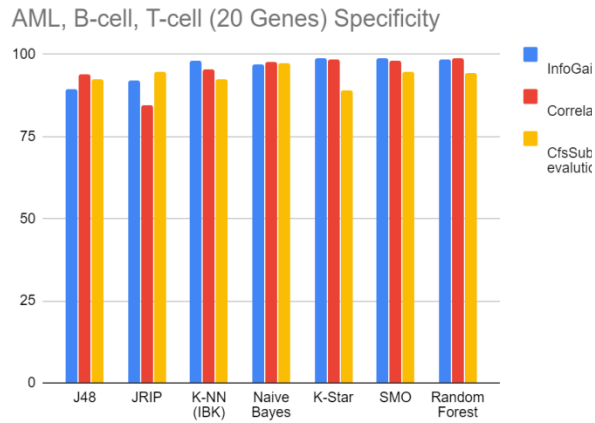


Figure 18: AML, B-cell, T-cell specificity 20 genes

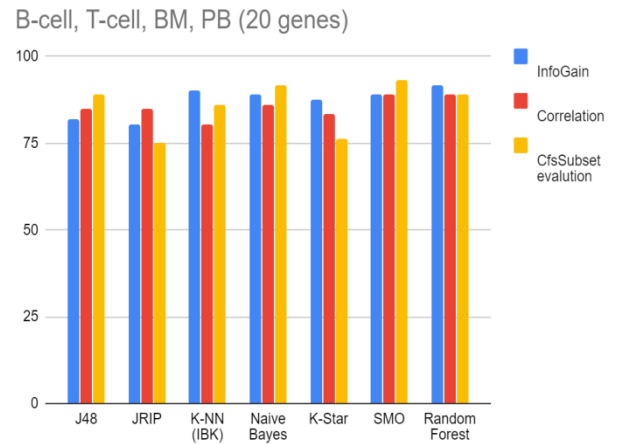


Figure 19: B-cell, T-cell, BM, PB accuracy 20 gene

	Info Gain	Correlation	CfsSubset evaluation
J48	81,94	84,72	88,88
JRIP	80,55	84,72	75
K-NN (IBK)	90,27	80,55	86,11
Naive Bayes	88,88	86,11	91,66
K-Star	87,5	83,33	76,38
SMO	88,88	88,88	93,05
Random Forest	91,66	88,88	88,88

Πίνακας 10: B-cell, T-cell, BM,PB accuracy 20 gene

	Info Gain	Correlation	CfsSubset evaluation
J48	81,9	87,5	87,5
JRIP	83,3	79,2	81,9
K-NN (IBK)	90,3	88,9	94,4
Naive Bayes	88,9	94,4	94,4
K-Star	87,5	91,7	97,2
SMO	88,9	93,1	95,8
Random Forest	94,4	95,8	97,2

Πίνακας 11: : B-cell, T-cell, BM, PB sensitivity 20 genes

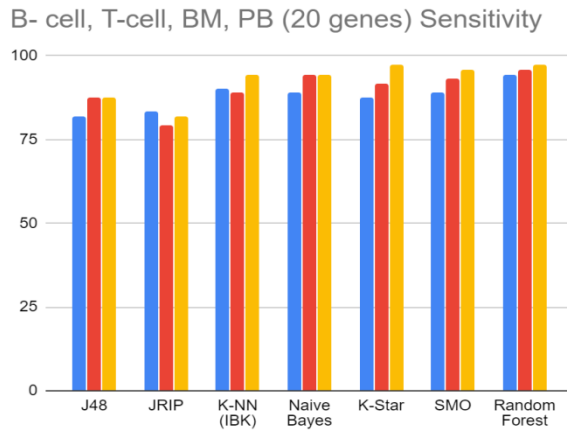


Figure 20: B-cell, T-cell, BM, PB sensitivity 20 genes

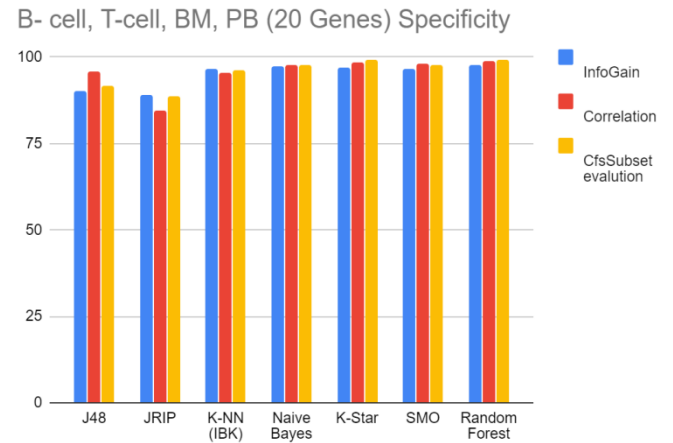


Figure 21: B-cell, T-cell, BM, PB specificity 20 genes

	Info Gain	Correlation	CfsSubset evaluation
J48	90,1	95,9	91,7
JRIP	89	84,6	88,5
K-NN (IBK)	96,7	95,4	96,2
Naive Bayes	97,3	97,6	97,6
K-Star	96,8	98,3	99,1
SMO	96,7	97,9	97,8
Random Forest	97,7	98,9	99,1

Πίνακας 12: B-cell, T-cell, BM, PB specificity 20 genes

	Info Gain	Correlation	CfsSubset evaluation
J48	84,72	81,94	87,5
JRIP	90,27	87,5	87,5
K-NN (IBK)	87,5	95,83	97,22
Naive Bayes	98,61	97,22	100
K-Star	65,27	68,05	98,61
SMO	95,83	97,22	98,61
Random Forest	87,5	95,83	97,22

Πίνακας 13: AML-ALL accuracy 50 genes

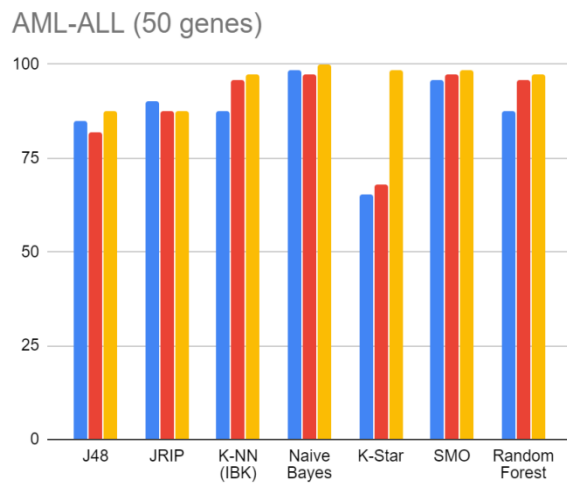


Figure 22: AML-ALL accuracy 50 genes

	Info Gain	Correlation	CfsSubset evaluation
J48	83,3	81,9	87,5
JRIP	91,7	87,5	87,5
K-NN (IBK)	94,4	95,8	97,2
Naive Bayes	97,2	97,2	100
K-Star	77,8	68,1	98,6
SMO	97,2	97,2	98,6
Random Forest	97,2	95,8	97,2

Πίνακας 14: AML-ALL sensitivity 50 genes

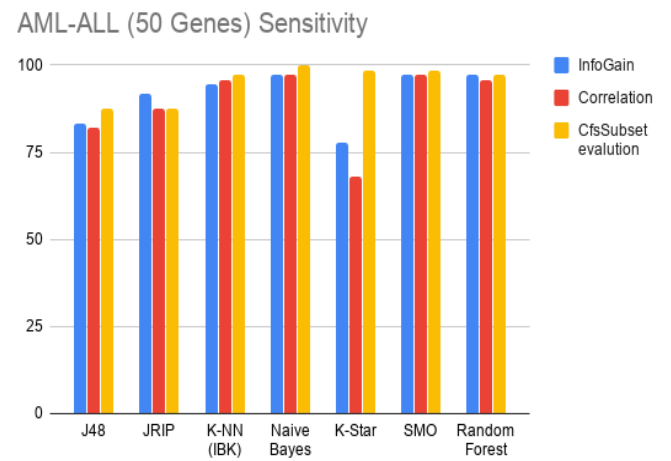


Figure 23: AML-ALL sensitivity 50 genes

	Info Gain	Correlation	CfsSubset evaluation
J48	83,6	82,9	85,9
JRIP	91,8	82,1	82,1
K-NN (IBK)	95,2	94	96,7
Naive Bayes	96,7	96,7	100
K-Star	62	39,9	97,4
SMO	96,7	96,7	97,4
Random Forest	94,8	94	94,8

Πίνακας 15: AML-ALL specificity 50 genes

AML-ALL (50 Genes) Specificity

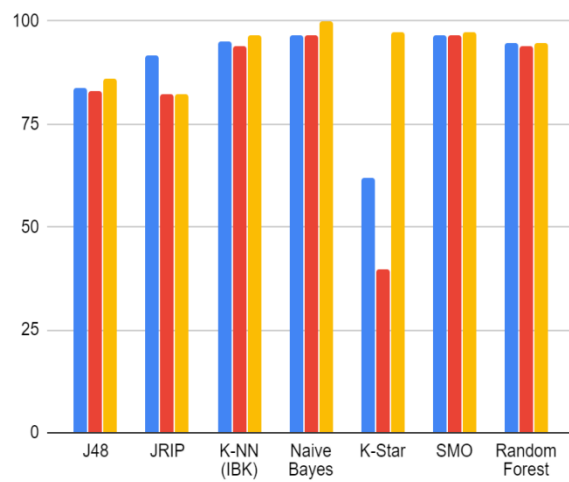


Figure 24: AML-ALL specificity 50 genes

AML,B-cell,T-cell (50 Genes)

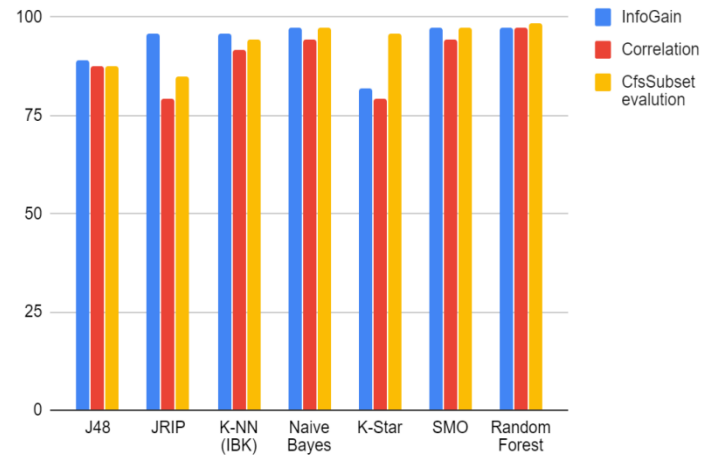


Figure 25: AML, B-cell, T-cell accuracy 50 genes

	Info Gain	Correlation	CfsSubset evaluation
J48	88.88	87.5	87.5
JRIP	95.83	79.16	84.72
K-NN (IBK)	95.83	91.66	94.44
Naive Bayes	97.22	94.44	97.22
K-Star	81.94	79.16	95.83
SMO	97.22	94.44	97.22
Random Forest	97.22	97.22	98.61

Πίνακας 16: AML, B-cell, T-cell accuracy 50 genes

	Info Gain	Correlation	CfsSubset evaluation
J48	88,9	87,5	84,7
JRIP	95,8	79,2	90,3
K-NN (IBK)	95,8	91,7	94,4
Naive Bayes	97,2	94,4	95,8
K-Star	81,9	79,2	95,8
SMO	97,2	94,4	94,4
Random Forest	97,2	97,2	97,2

Πίνακας 17: AML, B-cell, T-cell sensitivity 50 genes

AML, B-cell, T- cell (50 Genes) Sensitivity

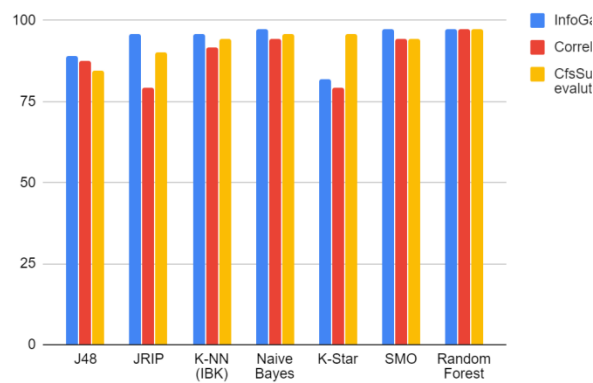


Figure 26: AML, B-cell, T-cell sensitivity 50 genes

AML, B-cell, T-cell (50 Genes) Specificity

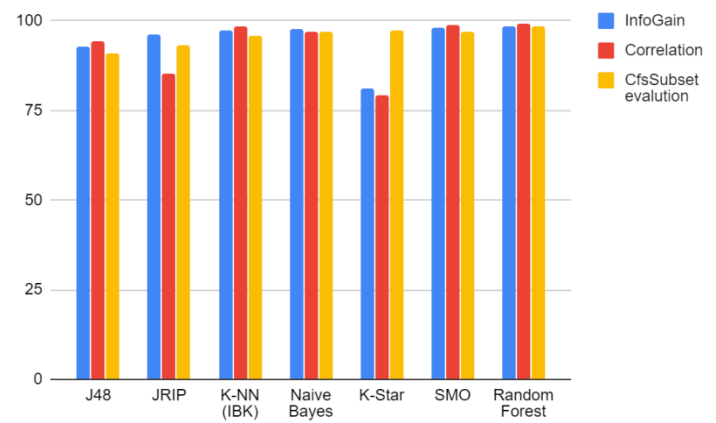


Figure 27: AML, B-cell, T-cell specificity 50 genes

	Info Gain	Correlation	CfsSubset evaluation
J48	93	94,4	91,1
JRIP	96,2	85,4	93,2
K-NN (IBK)	97,5	98,3	96
Naive Bayes	97,7	96,8	97
K-Star	81,2	79,4	97,5
SMO	98,2	98,7	96,8
Random Forest	98,5	99,1	98,5

Πίνακας 18: AML, B-cell, T-cell specificity 50 genes

	Info Gain	Correlation	CfsSubset evaluation
J48	86,11	54,16	83,33
JRIP	93,05	54,16	79,16
K-NN (IBK)	94,44	55,55	97,22
Naive Bayes	94,44	40,27	94,44
K-Star	84,72	50	93,05
SMO	94,44	69,44	97,22
Random Forest	84,72	66,66	98,61

Πίνακας 19: B-cell, T-cell, BM, PB accuracy 50 genes

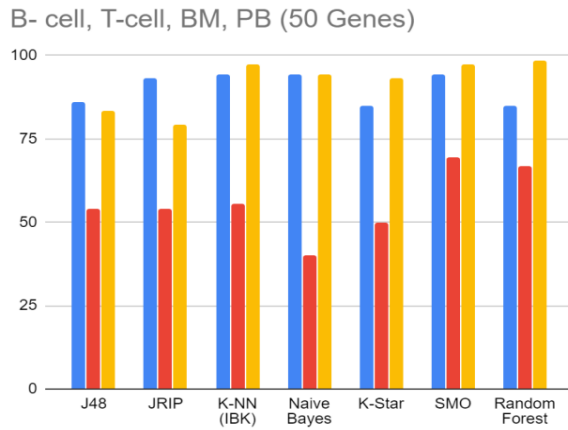


Figure 28: B-cell, T-cell, BM, PB accuracy 50 genes

	Info Gain	Correlation	CfsSubset evaluation
J48	86,1	86,1	83,3
JRIP	93,1	69,4	79,2
K-NN (IBK)	94,4	91,7	97,2
Naive Bayes	94,4	87,5	94,4
K-Star	84,7	69,4	93,1
SMO	94,4	94,4	97,2
Random Forest	91,7	91,7	98,6

Πίνακας 20: B-cell, T-cell, BM, PB sensitivity 50 genes

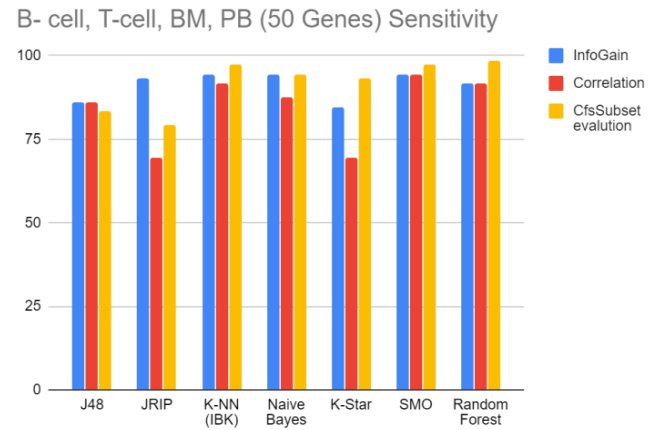


Figure 29: B-cell, T-cell, BM, PB sensitivity 50 genes

	Info Gain	Correlation	CfsSubset evaluation
J48	93,3	93	92,8
JRIP	95,2	80,6	88,4
K-NN (IBK)	97,7	98,1	97,9
Naive Bayes	97,7	94,2	96,7
K-Star	87,2	71,3	94,2
SMO	97,7	98,8	97,9
Random Forest	96,6	96,6	99,4

Πίνακας 21: B-cell, T-cell, BM, PB specificity 50 genes

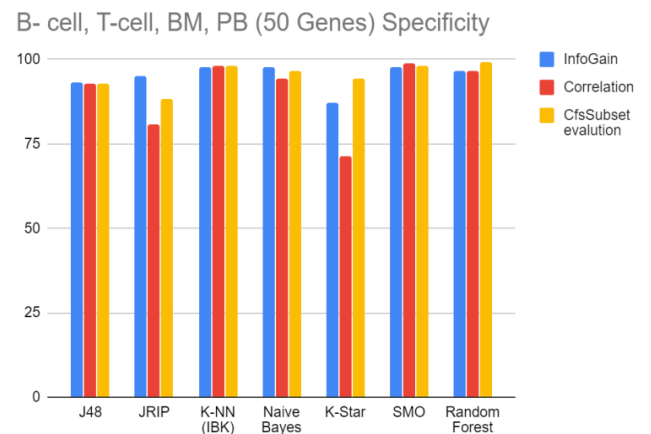


Figure 30: B-cell, T-cell, BM, PB specificity 50 gene

3.7 Αξιολόγηση

Για την αξιολόγηση της απόδοσης των μεθόδων ταξινόμησης χρησιμοποιήθηκε η τεχνική K-fold cross validation (k=10). Με την τεχνική αυτή το σύνολο δεδομένων χωρίζεται σε τυχαία δέκα κομμάτια. Έπειτα 9 από τα κομμάτια χρησιμοποιούνται για εκπαίδευση και το τελευταίο για δοκιμές. Έτσι το κάθε κομμάτι έχει χρησιμοποιηθεί 1 φορά για δοκιμή και 9 για εκπαίδευση. Η απόδοση και τα αποτελέσματα των ταξινομητών φαίνονται στους πίνακες παραπάνω. Υπάρχουν διαφορετικές επιδόσεις και μέτρα που υπολογίζονται βάσει του confusion matrix που παράγεται από τους αλγόριθμους. Η εικόνα απεικονίζει επίσης το μοντέλο του confusion matrix. Σε αυτόν τον πίνακα, κάθε σειρά εμφανίζει την πραγματική κλάση και η στήλη εμφανίζει την προβλεπόμενη κλάση.

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

Figure 31: confusion matrix

TP (True Positive) και TN (True Negative) είναι τα παραδείγματα που είναι σωστά ταξινομημένα ως μια δεδομένη κλάση και FP (False Positive) και FN (False Negative) είναι οι περιπτώσεις που ταξινομούνται ψευδώς ως μια συγκεκριμένη κλάση. Άλλα μέτρα για την αξιολόγηση είναι η ακρίβεια επιλεγμένων στοιχείων που είναι σωστά και υπολογίζεται ως Ακρίβεια (P) = $TP / (TP + FP)$ και Ανάκληση των σωστών στοιχείων που έχουν επιλεγεί και υπολογίζεται ως Recall (R) = $TP / (TP + FN)$.

Κεφάλαιο 4

Συζήτηση-Συμπεράσματα

Οι τεχνικές μηχανικής μάθησης και τα εργαλεία έχουν επιφέρει σημαντικές αλλαγές στο τρόπο που αναλύονται τα δεδομένα αποκαλύπτοντας χρήσιμες πληροφορίες. Στην συγκεκριμένη εργασία έχει αναλυθεί η εφαρμογή και απόδοση επτά αλγορίθμων ταξινόμησης οι οποίοι παράγουν διαφορετικά αποτελέσματα. Εφαρμόστηκαν μεθοδολογίες για επιλογή χαρακτηριστικών και για κατάταξη και βρέθηκαν τα γονίδια που συνεισφέρουν περισσότερη πληροφορία για το συγκεκριμένο πρόβλημα. Επίσης προτάθηκε ένα μοντέλο που κάνει διάγνωση της λευχαιμίας σε δύο, τρεις και τέσσερις κλάσεις. Συγκρίνοντας διαφορετικούς αλγορίθμους ο Naive Bayes αποδίδει καλύτερα όταν διαχωρίζουμε δυο κλάσεις, ο Random Forest όταν διαχωρίζουμε 3 κλάσεις και ο Support Vector Machine στις 4 κλάσεις.

Η συγκεκριμένη ανάλυση υποδεικνύει ότι ο αλγόριθμος Naive Bayes ταξινόμησε σωστά τον μεγαλύτερο αριθμό παραδειγμάτων με μέσο όρο ακρίβειας 97,68%, μέσο όρο ευαισθησίας 96,96 και μέσο όρο εξειδίκευσης 96,8 όταν διαχωρίζουμε δυο κλάσεις. Στην περίπτωση των τριών κλάσεων ο Random Forest ταξινόμησε σωστά τον μεγαλύτερο αριθμό παραδειγμάτων με μέσο όρο ακρίβειας 96,29, μέσο όρο ευαισθησίας 96,05 και μέσο όρο εξειδίκευσης 98%. Τέλος στην περίπτωση των τεσσάρων κλάσεων ο Support Vector Machine ταξινόμησε σωστά τον μεγαλύτερο αριθμό παραδειγμάτων με μέσο όρο ακρίβειας 88,65, μέσο όρο ευαισθησίας 94,21 και μέσο όρο εξειδίκευσης 97,79%.

ΠΑΡΑΡΤΗΜΑ

AML-ALL (20 γονίδια)

Αλγόριθμοι	Without Feature selection	Feature selection InfoGain	Feature selection Correlation	Feature selection CfsSubset Evaluation
1.J48	Acc=83,33% SeA=89,36% SpA=72% SeB=72% SpB=89,36%	Acc=81,94% SeA=80,9% SpA=84% SeB=84% SpB=80,9%	Acc=84,72% SeA=87,2% SpA=80% SeB=80% SpB=87,2%	Acc=94,44% SeA=97,9% SpA=88% SeB=88% SpB=97,9%
JRIP	Acc=87,5% SeA=85,1% SpA=92% SeB=92% SpB=85,1%	Acc=91,66% SeA=95,7% SpA=84% SeB=84% SpB=95,7%	Acc=90,27% SeA=93,6% SpA=84% SeB=84% SpB=93,6%	Acc=91,66% SeA=93,6% SpA=88% SeB=88% SpB=93,6%
K-NN (IBK)	Acc=87,5% SeA=93,6% SpA=76% SeB=76% SpB=93,6%	Acc=95,83% SeA=95,7% SpA=96% SeB=96% SpB=95,7%	Acc=94,44% SeA=97,9% SpA=88% SeB=88% SpB=97,9%	Acc=97,22% SeA=100% SpA=92% SeB=92% SpB=100%
Naive Bayes	Acc=98,61% SeA=100% SpA=96% SeB=96% SpB=100%	Acc=95,83% SeA=95,7% SpA=96% SeB=96% SpB=95,7%	Acc=95,83% SeA=95,7% SpA=96% SeB=96% SpB=95,7%	Acc=98,61% SeA=100% SpA=96% SeB=96% SpB=100%
K-Star	Acc=65,27% SeA= 100% SeB=0% SpA=0%	Acc=95,83% SeA=97,9% SpA=92% SeB=92%	Acc=95,83% SeA=100% SpA=88% SeB=88%	Acc=90,27% SeA=91,5% SpA=88% SeB=88%

	SpB=100%	SpB=97,9%	SpB=100%	SpB=91,5%
SMO	Acc=95,83% SeA=97,9% SpA=92% SeB=92% SpB=97,9%	Acc=95,83% SeA=97,9% SpA=92% SeB=92% SpB=97,9%	Acc=94,44% SeA=97,9% SpA=88% SeB=88% SpB=97,9%	Acc=97,22% SeA=100% SpA=92% SeB=92% SpB=100%
Rando m Forest	Acc=83,33% SeA=89,36% SpA=72% SeB=72% SpB=89,36%	Acc=97,22% SeA=100% SpA=92% SeB=92% SpB=100%	Acc=94,44% SeA=95,7% SpA=92% SeB=92% SpB=95,7%	Acc=98,61% SeA=100% SpA=96% SeB=96% SpB=100%

AML, B-Cell, T-cell (20 γονίδια)

Αλγόριθμοι	Without Feature selection %	Feature selection InfoGain %	Feature selection Correlation %	Feature selection CfsSubset evaluation%
J48	Acc=95,83 SeA=94,7 SpA=97,1 SeB=100 SpB=98,4 SeC=96 SpC=97,9	Acc=86,11 SeA=94,7 SpA=85,3 SeB=66,7 SpB=96,8 SeC=80 SpC=93,6	Acc=87,5 SeA=94,7 SpA=94,1 SeB=66,7 SpB=93,7 SeC=84 SpC=93,6	Acc=87,5 SeA=89,5 SpA=91,2 SeB=77,8 SpB=95,2 SeC=88 SpC=93,6
JRIP	Acc=90,27 SeA=92,1 SpA=91,8 SeB=88,9 SpB=98,4 SeC=88 SpC=93,6	Acc=88,88 SeA=89,5 SpA=91,2 SeB=77,8 SpB=98,4 SeC=92 SpC=91,5	Acc=79,16 SeA=84,2 SpA=82,4 SeB=33,3 SpB=98,4 SeC=88 SpC=83	Acc=88,88 SeA=86,8 SpA=100 SeB=66,7 SpB=98,4 SeC=100 SpC=85,1

K-NN (IBK)	Acc=81,94 SeA=92,1 SpA=73,5 SeB=66,7 SpB=98,4 SeC=72 SpC=93,6	Acc=94,44 SeA=97,4 SpA=100 SeB=88,9 SpB=96,8 SeC=92 SpC=95,7	Acc=88,88 SeA=94,7 SpA=97,1 SeB=88,9 SpB=93,7 SeC=80 SpC=93,6	Acc=91,66 SeA=97,4 SpA=88,2 SeB=77,8 SpB=100 SeC=88 SpC=95,7
Naive Bayes	Acc=94,44 SeA=100 SpA=88,2 SeB=66,7 SpB=100 SeC=96 SeC=100	Acc=94,44 SeA=97,4 SpA=97,1 SeB=77,8 SpB=98,4 SeC=96 SpC=95,7	Acc=94,44 SeA=97,4 SpA=100 SeB=77,8 SpB=98,4 SeC=96 SpC=93,6	Acc=95,83 SeA=97,4 SpA=97,1 SeB=88,9 SpB=98,4 SeC=96 SpC=97,9
K-Star	Acc=52,77 SeA= 100 SeB=0 SpA=0 SpB=100 SeC=0 SeC=100	Acc=95,83 SeA=97,4 SpA=100 SeB=88,9 SpB=96,8 SeC=96 SpC=97,9	Acc=91,66 SeA=97,4 SpA=92,1 SeB=88,9 SpB=92,1 SeC=84 SpC=97,9	Acc=83,33 SeA=84,2 SpA=88,2 SeB=77,8 SpB=96,8 SeC=84 SpC=87,2
SMO	Acc=97,22 SeA=100 SpA=94,1 SeB=88,9 SpB=100 SeC=96 SeC=100	Acc=94,44 SeA=97,4 SpA=100 SeB=100 SpB=95,2 SeC=88 SpC=97,9	Acc=93,05 SeA=94,7 SpA=100 SeB=88,9 SpB=95,2 SeC=92 SpC=95,7	Acc=93,05 SeA=94,7 SpA=94,1 SeB=77,8 SpB=100 SeC=96 SpC=93,6

Rando m Forest	Acc=83,33 SeA=100 SpA=64,7 SeB=33,3 SpB=100 SeC=76 SpC=100	Acc=97,22 SeA=97,4 SpA=100 SeB=88,9 SpB=100 SeC=100 SpC=95,7	Acc=95,83 SeA=97,4 SpA=100 SeB=88,9 SpB=96,8 SeC=96 SpC=97,9	Acc=91,66 SeA=92,1 SpA=94,1 SeB=77,8 SpB=98,4 SeC=96 SpC=93,6
-------------------------------	--	--	--	---

B-cell, T-cell, BM, PB (20 γονίδια)

Αλγόριθμοι	Without Feature selection %	Feature selection InfoGain %	Feature selection Correlation %	Feature selection CfsSubset evaluation%
J48	Acc=87,5 SeA=94,7 SpA=97,1 SeB=100 SpB=98,4 SeC=81 SpC=92,2 SeD=25 SpD=95,6	Acc=81,94 SeA=94,7 SpA=88,2 SeB=44,4 SpB=93,7 SeC=85,7 SpC=90,2 SeD=25 SpD=100	Acc=84,72 SeA=97,4 SpA=91,2 SeB=55,6 SpB=96,8 SeC=90,5 SpC=88,2 SeD=0 SpD=100	Acc=88,88 SeA=92,1 SpA=91,2 SeB=77,8 SpB=98,4 SeC=95,2 SpC=92,2 SeD=50 SpD=100

JRIP	Acc=84,72 SeA=94,7 SpA=88,2 SeB=77,8 SpB=98,4 SeC=85,7 SpC=88,2 SeD=0 SpD=100	Acc=80,55 SeA=86,8 SpA=82,4 SeB=44,4 SpB=98,4 SeC=85,7 SpC=86,3 SeD=75 SpD=100	Acc=84,72 SeA=94,7 SpA=94,1 SeB=66,7 SpB=93,7 SeC=90,5 SpC=90,2 SeD=0 SpD=100	Acc=75 SeA=89,5 SpA=88,2 SeB=22,2 SpB=95,2 SeC=81 SpC=78,4 SeD=25 SpD=100
K-NN (IBK)	Acc=83,33 SeA=92,1 SpA=73,5 SeB=66,7 SpB=100 SeC=76,2 SpC=94,1 SeD=75 SpD=100	Acc=90,27 SeA=97,4 SpA=100 SeB=88,9 SpB=96,8 SeC=90,5 SpC=90,2 SeD=25 SpD=100	Acc=80,55 SeA=92,1 SpA=97,1 SeB=100 SpB=88,9 SeC=66,7 SpC=90,2 SeD=0 SpD=98,5	Acc=86,11 SeA=94,7 SpA=79,4 SeB=88,9 SpB=100 SeC=66,7 SpC=94,1 SeD=100 SpD=100
Naive Bayes	Acc=87,5 SeA=100 SpA=82,4 SeB=66,7 SpB=100 SeC=90,5 SpC=94,1 SeD=0 SpD=100	Acc=88,88 SeA=97,4 SpA=100 SeB=77,8 SpB=98,4 SeC=81 SpC=92,2 SeD=75 SpD=95,6	Acc=86,11 SeA=97,4 SpA=97,1 SeB=66,7 SpB=98,4 SeC=90,5 SpC=86,3 SeD=0 SpD=98,5	Acc=91,66 SeA=94,7 SpA=97,1 SeB=88,9 SpB=98,4 SeC=100 SpC=92,2 SeD=25 SpD=100

K-Star	Acc=52,77 SeA= 100 SpA=0 SeB=0 SpB=100 SeC=0 SpC=100 SeD=0 SpD=100	Acc=87,5 SeA=97,4 SpA=100 SeB=100 SpB=93,7 SeC=81 SpC=92,2 SeD=0 SpD=98,5	Acc=83,33 SeA=94,7 SpA=100 SeB=88,9 SpB=90,5 SeC=76,2 SpC=90,2 SeD=0 SpD=98,5	Acc=76,38 SeA=81,6 SpA=85,3 SeB=44,4 SpB=100 SeC=90,5 SpC=76,5 SeD=25 SpD=100
SMO	Acc=93,05 SeA=97,4 SpA=94,1 SeB=88,9 SpB=100 SeC=95,2 SpC=94,1 SeD=50 SpD=100	Acc=88,88 SeA=97,4 SpA=100 SeB=88,9 SpB=96,8 SeC=85,7 SpC=90,2 SeD=25 SpD=98,5	Acc=88,88 SeA=94,7 SpA=100 SeB=88,9 SpB=96,8 SeC=95,2 SpC=88,2 SeD=0 SpD=100	Acc=93,05 SeA=92,1 SpA=97,1 SeB=88,9 SpB=100 SeC=95,2 SpC=94,1 SeD=100 SpD=98,5
Random Forest	Acc=76,38 SeA=100 SpA=61,8 SeB=11,1 SpB=100 SeC=76,2 SpC=92,2 SeD=0	Acc=91,66 SeA=97,4 SpA=100 SeB=88,9 SpB=98,4 SeC=95,2 SpC=90,2 SeD=25	Acc=88,88 SeA=97,4 SpA=100 SeB=77,8 SpB=98,4 SeC=95,2 SpC=86,3 SeD=0	Acc=88,88 SeA=92,1 SpA=91,2 SeB=88,9 SpB=98,4 SeC=90,5 SpC=92,2 SeD=50

	SpD=100	SpD=100	SpD=100	SpD=100
--	---------	---------	---------	---------

AML-ALL (50 γονίδια)

Αλγόριθμοι	Feature selection InfoGain	Feature selection Correlation	Feature selection CfsSubset evaluation
J48	Acc=84,72% SeA=89,4% SpA=76% SeB=76% SpB=89,4%	Acc=81,94% SeA=80,9% SpA=84% SeB=84% SpB=80,9%	Acc=87,5% SeA=89,4% SpA=84% SeB=84% SpB=89,4%
JRIP	Acc=90,27% SeA=91,5% SpA=88% SeB=88% SpB=91,5%	Acc=87,5% SeA=93,6% SpA=76% SeB=76% SpB=93,6%	Acc=87,5% SeA=93,6% SpA=76% SeB=76% SpB=93,6%
K-NN (IBK)	Acc=87,5% SeA=93,6% SpA=76% SeB=76% SpB=93,4%	Acc=95,83% SeA=97,9% SpA=92% SeB=92% SpB=97,9%	Acc=97,22% SeA=97,9% SpA=96% SeB=96% SpB=97,9%
Naive Bayes	Acc=98,61% SeA=100%	Acc=97,22% SeA=97,9%	Acc=100% SeA=100%

	SpA=96% SeB=96% SpB=100%	SpA=96% SeB=96% SpB=97,9%	SpA=100% SeB=100% SpB=100%
K-Star	Acc=65,27% SeA=100% SpA=0% SeB=0% SpB=100%	Acc=68,05% SeA=100% SpA=8% SeB=8% SpB=100%	Acc=98,61% SeA=100% SpA=96% SeB=96% SpB=100%
SMO	Acc=95,83% SeA=97,9% SpA=92% SeB=92% SpB=97,9%	Acc=97,22% SeA=97,9% SpA=96% SeB=96% SpB=97,9%	Acc=98,61% SeA=100% SpA=96% SeB=96% SpB=100%
Random Forest	Acc=87,5% SeA=100% SpA=64% SeB=64% SpB=100%	Acc=95,83% SeA=97,9% SpA=92% SeB=92% SpB=97,9%	Acc=97,22% SeA=100% SpA=92% SeB=92% SpB=100%

AML, B-cell, T-cell (50 γονίδια)

Αλγόριθμοι	Feature selection InfoGain %	Feature selection Correlation %	Feature selection CfsSubset evaluation%
J48	Acc=88,88 SeA=94,7 SpA=94,1 SeB=66,7 SpB=98,4 SeC=88 SpC=89,4	Acc=87,5 SeA=94,7 SpA=94,1 SeB=66,7 SpB=92,1 SeC=84 SpC=95,7	Acc=87,5 SeA=94,7 SpA=91,2 SeB=66,7 SpB=96,8 SeC=84 SpC=91,5
JRIP	Acc=95,83 SeA=100 SpA=94,1 SeB=88,9 SpB=100 SeC=92 SpC=97,9	Acc=79,16 SeA=92,1 SpA=79,4 SeB=44,4 SpB=93,7 SeC=72 SpC=91,5	Acc=84,72 SeA=94,7 SpA=76,5 SeB=55,6 SpB=98,4 SeC=80 SpC=95,7

K-NN (IBK)	Acc=95,83 SeA=94,7 SpA=97,1 SeB=100 SpB=98,4 SeC=96 SpC=97,9	Acc=91,66 SeA=92,1 SpA=100 SeB=100 SpB=92,1 SeC=88 SpC=97,9	Acc=94,44 SeA=97,4 SpA=91,2 SeB=77,8 SpB=100 SeC=96 SpC=97,9
Naive Bayes	Acc=97,22 SeA=97,4 SpA=97,1 SeB=88,9 SpB=100 SeC=100 SpC=97,9	Acc=94,44 SeA=97,4 SpA=97,1 SeB=77,8 SpB=98,4 SeC=96 SpC=95,7	Acc=97,22 SeA=100 SpA=97,1 SeB=77,8 SpB=100 SeC=100 SpC=97,9
K-Star	Acc=81,94 SeA=97,4 SpA=64,7 SeB=100 SpB=98,4 SeC=52 SpC=100	Acc=79,16 SeA=97,4 SpA=61,8 SeB=100 SpB=96,8 SeC=44 SpC=100	Acc=95,83 SeA=97,4 SpA=94,1 SeB=88,9 SpB=100 SeC=96 SpC=97,9
SMO	Acc=97,22 SeA=97,4 SpA=97,1 SeB=100 SpB=98,4 SeC=96 SpC=100	Acc=94,44 SeA=94,7 SpA=100 SeB=100 SpB=95,2 SeC=92 SpC=97,9	Acc=97,44 SeA=97,4 SpA=100 SeB=88,9 SpB=100 SeC=100 SpC=95,7

Rando m Forest	Acc=97,22	Acc=94,44	Acc=97,22
	SeA=97,4	SeA=97,4	SeA=100
	SpA=100	SpA=100	SpA=100
	SeB=88,9	SeB=88,9	SeB=88,9
	SpB=100	SpB=98,4	SpB=100
	SeC=100	SeC=100	SeC=100
	SpC=95,7	SpC=97,9	SpC=97,9

B-Cell, T-cell, BM, PB (50 γονίδια)

Αλγόριθμοι	Feature selection InfoGain %	Feature selection Correlation %	Feature selection CfsSubset evaluation%
1. J48	Acc=86,11 SeA=94,7 SpA=91,2 SeB=77,8 SpB=93,7 SeC=81 SpC=96,1 SeD=50 SpD=98,5	Acc=54,16 SeA=57,9 SpA=67,6 SeB=55,6 SpB=95,2 SeC=57,1 SpC=70,6 SeD=0 SpD=94,1	Acc=83,33 SeA=94,7 SpA=91,2 SeB=77,8 SpB=95,2 SeC=76,2 SpC=94,1 SeD=25 SpD=95,6
2. JRIP	Acc=93,05 SeA=97,4 SpA=94,1 SeB=88,9 SpB=100 SeC=90,5 SpC=94,1 SeD=75 SpD=100	Acc=54,16 SeA=73,7 SpA=52,9 SeB=11,1 SpB=93,7 SeC=47,6 SpC=78,4 SeD=0 SpD=97,1	Acc=79,16 SeA=84,2 SpA=85,3 SeB=44,4 SpB=92,1 SeC=85,7 SpC=90,2 SeD=75 SpD=100

K-NN (IBK)	Acc=94,44 SeA=94,7 SpA=100 SeB=88,9 SpB=100 SeC=100 SpC=92,2 SeD=75 SpD=100	Acc=55,55 SeA=76,3 SpA=61,8 SeB=22,2 SpB=82,5 SeC=38,1 SpC=86,3 SeD=25 SpD=98,5	Acc=97,22 SeA=97,4 SpA=97,1 SeB=88,9 SpB=100 SeC=100 SpC=98 SeD=100 SpD=100
Naive Bayes	Acc=94,44 SeA=97,4 SpA=100 SeB=88,9 SpB=100 SeC=100 SpC=92,2 SeD=50 SpD=100	Acc=40,27 SeA=44,7 SpA=70,6 SeB=11,1 SpB=77,2 SeC=52,4 SpC=66,7 SeD=0 SpD=97,1	Acc=94,44 SeA=100 SpA=97,1 SeB=77,8 SpB=100 SeC=100 SpC=94,1 SeD=50 SpD=100
K-Star	Acc=84,72 SeA=94,7 SpA=79,4 SeB=88,9 SpB=98,4 SeC=81 SpC=94,1 SeD=0 SpD=100	Acc=50 SeA=71,1 SpA=52,9 SeB=0 SpB=84,1 SeC=42,9 SpC=80,4 SeD=0 SpD=100	Acc=93,05 SeA=97,4 SpA=91,2 SeB=100 SpB=100 SeC=90,5 SpC=96,1 SeD=50 SpD=100

SMO	Acc=94,44 SeA=94,7 SpA=100 SeB=88,9 SpB=100 SeC=100 SpC=92,2 SeD=75 SpD=100	Acc=69,44 SeA=92,1 SpA=58,8 SeB=11,1 SpB=98,4 SeC=66,7 SpC=86,3 SeD=0 SpD=100	Acc=97,22 SeA=97,4 SpA=97,1 SeB=88,9 SpB=100 SeC=100 SpC=98 SeD=100 SpD=100
Rando m Forest	Acc=84,72 SeA=92,1 SpA=88,2 SeB=66,7 SpB=93,7 SeC=85,7 SpC=94,1 SeD=50 SpD=100	Acc=66,66 SeA=92,1 SpA=50 SeB=11,1 SpB=98,4 SeC=57,1 SpC=88,2 SeD=0 SpD=100	Acc=98,61 SeA=100 SpA=100 SeB=100 SpB=100 SeC=100 SpC=98 SeD=75 SpD=100

REFERENCES

- [1] Sung-Huai Hsieh, Zhenyu Wang, Po-Hsun Cheng, I-Shun Lee, Sheau-Ling Hsieh and Feipei Lai, "Leukemia cancer classification based on Support Vector Machine," *2010 8th IEEE International Conference on Industrial Informatics*, Osaka, 2010, pp. 819-824. URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5549638&isnumber=5549388>
- [2] S. Rajpurohit, S. Patil, N. Choudhary, S. Gavasane and P. Kosamkar, "Identification of Acute Lymphoblastic Leukemia in Microscopic Blood Image Using Image Processing and Machine Learning Algorithms," *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Bangalore, 2018, pp. 2359-2363. URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8554576&isnumber=8554361>
- [3] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., ... Lander, E. S. (1999, October 15). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. URL: <https://science.sciencemag.org/content/286/5439/531/tab-article-info>
- [4] S. Begum, D. Chakraborty and R. Sarkar, "Identifying cancer biomarkers from leukemia data using feature selection and supervised learning," *2016 IEEE First International Conference on Control, Measurement and Instrumentation (CMI)*, Kolkata, 2016, pp. 249-253. URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7413749&isnumber=7413696>
- [5] Lee, S.-I., Celik, S., Logsdon, B. A., Lundberg, S. M., Martins, T. J., Oehler, V. G., ... Becker, P. S. (2018, January 3). A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. URL: <https://www.nature.com/articles/s41467-017-02465-5>
- [6] deAndrés-Galiana, E. J., Fernández-Martínez, J. L., Luaces, O., Coz, J. J. del, Huergo-Zapico, L., Acebes-Huerta, A., ... González-Rodríguez, A. P. (2016, March 5). Analysis of clinical prognostic variables for Chronic Lymphocytic Leukemia decision-making problems. URL: <https://www.sciencedirect.com/science/article/pii/S1532046416000393>

[7] Kumar, M., Rath, N. K., & Rath, S. K. (2016, March 11). Analysis of microarray leukemia data using an efficient MapReduce-based K-nearest-neighbor classifier.

URL:<https://www.sciencedirect.com/science/article/pii/S1532046416000423>

[8] Tsanousa, A., Ntoufa, S., Papakonstantinou, N., Stamatopoulos, K., & Angelis, L. (2019, May 18). Study of gene expressions' correlation structures in subgroups of Chronic Lymphocytic Leukemia Patients.

URL:<https://www.sciencedirect.com/science/article/pii/S1532046419301297>

[9] Haferlach, T., Kohlmann, A., Wiczorek, L., Basso, G., Kronnie, G. T., Béné, M.-C., ... Foà, R. (2010). Clinical Utility of Microarray-Based Gene Expression Profiling in the Diagnosis and Subclassification of Leukemia: Report From the International Microarray Innovations in Leukemia Study Group. *Journal of Clinical Oncology*, 28(15), 2529–2537. doi: 10.1200/jco.2009.23.4732

[10] P. Jagadev and H. G. Virani, "Detection of leukemia and its types using image processing and machine learning," 2017 International Conference on Trends in Electronics and Informatics (ICEI), Tirunelveli, 2017, pp. 522-526, doi: 10.1109/ICOEI.2017.8300983.

[11] Yiu, T. (2019, August 14). Understanding Random Forest.

Retrieved from <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>

[12] Νεότερα δεδομένα για τη λευχαιμία: Νοσοκομείο ΥΓΕΙΑ. (2018, October 9).

Retrieved from <https://www.hygeia.gr/neotera-dedomena-gia-ti-leychaimia/>

[12] National Human Genome Research Institute, DNA Microarray Technology Fact Sheet. (n.d.). (2020,15 August). Retrieved from <https://www.genome.gov/about-genomics/fact-sheets/DNA-Microarray-Technology>

[13] (n.d.). Retrieved from <https://www.nature.com/scitable/definition/microarray-202/>

[14] Σπηλιώτου Βάια (2011). Μελέτη της γονιδιακής έκφρασης με τη μέθοδο των μικροσυστοιχιών (microarrays) σε μυ αρουραίου μετά από άσκηση. Διπλωματική Εργασία, Εφαρμογές Μοριακής Βιολογίας – Μοριακή Γενετική – Διαγνωστικοί Δείκτες, Πανεπιστήμιο Θεσσαλίας.

[15] (n.d.). Retrieved from http://www.labtestsonline.gr/condition/Condition_Leukemia.html?idx=2

[16] Improved J48 Classification Algorithm for the Prediction ... (n.d.). Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.678.9273&rep=rep1&type=pdf>

[17] Malvezzi, M., Carioli, G., Bertuccio, P., Rosso, T., Boffetta, P., Levi, F., . . . Negri, E. (2020, January 06). European cancer mortality predictions for the year 2016 with focus on leukaemias. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0923753419357692>

[18] Loukopoulos, D., Politi, M., Λουκόπουλος, Δ, & Μαριάννα, Π. (2016, February 10). HEMATOLOGY LESSONS. Retrieved from <https://repository.kallipos.gr/handle/11419/3081>

[19] Kyrkos, E., & Κύρκος, E. (2016, January 25). Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων. Retrieved from <https://repository.kallipos.gr/handle/11419/1226>

[20]] “Support Vector Machine — Introduction to Machine Learning Algorithms”, Rohith Gandhi, Jun. 2018

[21] Aljazzar, H., & Leue, S. (2011, August 05). K*: A heuristic search algorithm for finding the k shortest paths. Retrieved October 01, 2020, from <https://www.sciencedirect.com/science/article/pii/S0004370211000865>