



Ιόνιο Πανεπιστήμιο

Σχολή Επιστήμης της Πληροφορίας & Πληροφορικής

Τμήμα Πληροφορικής

ΠΜΣ «Ψηφιακές Εφαρμογές και Καινοτομία»

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Μοντέλα πρόβλεψης καρκίνου
με χρήση γενετικών δεδομένων
και τεχνικές μηχανικής μάθησης

Μάντσιος Τριαντάφυλλος (ΠΚΜ201907)

Επιβλέπων Καθηγητής : **Θεμιστοκλής Έξαρχος**

Κέρκυρα, Φεβρουάριος 2021

Περίληψη

Ο καρκίνος των ωοθηκών παρουσιάζεται ως μία από τις πιο ύπουλες μορφές καρκίνου διότι δεν εμφανίζει έγκαιρα συμπτώματα στον ασθενή. Αυτό καθιστά δύσκολο έργο την διάγνωσή του καθώς και την έγκαιρη θεραπεία. Στην παρούσα εργασία γίνεται προσπάθεια εύρεσης συγκεκριμένων μοτίβων, που θα βοηθήσουν στην έγκαιρη διάγνωση του καρκίνου των ωοθηκών. Τα δεδομένα που μελετώνται προέρχονται από μετρήσεις φασματομετρίας μάζας ορού (M/Z), τα οποία επεξεργάζονται και εφαρμόζονται πάνω τους αλγόριθμοι ταξινόμησης. Τα επεξεργασμένα δεδομένα μα δίνουν ένα υποσύνολο 32 χαρακτηριστικών και πάνω σε αυτό το υποσύνολο μετριούνται οι αποδόσεις των αλγορίθμων. Σκοπός της παρούσας μελέτης είναι η παρουσίαση των επιμέρους αποτελεσμάτων με γραφήματα και πίνακες. Τέλος, με την βοήθεια συγκεντρωτικών πινάκων αλλά και γραφημάτων γίνεται η σύγκριση όλων των αποτελεσμάτων.

Abstract

Ovarian cancer presents as one of the most insidious forms of cancer because it does not show early symptoms to the patient. This makes it difficult to diagnose and treat early. In the present work, an attempt is made to find specific patterns that will help in the early diagnosis of ovarian cancer. The data studied are derived from serum mass spectrometry (M / Z) measurements which are processed and applied to the classification algorithms. The processed data give a subset of 32 characteristics and on this subset the performance of the algorithms is measured. The purpose of this study is to present the individual results with graphs and tables. Finally, with the help of summary tables and graphs, all the results are compared.

Πίνακας περιεχομένων

Περίληψη.....	2
Εισαγωγή.....	4
1. Περιγραφή της ασθένειας.....	4
1.1 Στάδια του καρκίνου των ωοθηκών.....	5
1.2 Αντιμετώπιση και θεραπεία του καρκίνου των ωοθηκών.....	6
1.3 Παράγοντες κινδύνου ανάπτυξης των ωοθηκών.....	6
1.4 Κοινωνική διάσταση.....	7
2. Βιβλιογραφική Ανασκόπηση.....	9
2.1 Εκπαίδευση μοντέλων για εξόρυξη βιολογικών δεδομένων. Μια εφαρμογή για την ανίχνευση του καρκίνου των ωοθηκών.....	9
2.2 Συγκριτική μελέτη της μηχανικής μάθησης και των αλγορίθμων βαθιάς μάθησης για ταξινόμηση τύπων καρκίνου με βάση το γονίδιο μικροσυστοιχιών.....	12
2.3 Ανάλυση μεγάλων δεδομένων για την δημιουργία γνώσης. Σύστημα έγκαιρης ανίχνευσης του καρκίνου των ωοθηκών.....	15
2.4 Μηχανική μάθηση και μοντέλα πληροφορικής για την αναγνώριση της γονιδιακής έκφρασης σε πρότυπα του καρκίνου των ωοθηκών που σχετίζονται με την εξέλιξη της νόσου και την θνησιμότητα.....	17
2.5 Ανασκόπηση δεδομένων και επανεξέταση της πρωτεϊνικής μορφολογίας του καρκίνου των ωοθηκών.....	21
3. Περιγραφή Δεδομένων, Αλγορίθμων και Μεθοδολογίας.....	22
3.1 Μηχανική Μάθηση.....	22
3.2 Εργαλείο Επεξεργασίας.....	23
3.3 Περιγραφή Δεδομένων.....	24
3.4 Αλγόριθμοι μείωσης χαρακτηριστικών.....	25
3.5 Αλγόριθμοι Ταξινόμησης.....	27
4. Πειραματικά Αποτελέσματα.....	32
4.1 Κριτήρια αξιολόγησης (Ακρίβεια, Ευαισθησία, Εξειδίκευση, Πίνακας Σύγχυσης).....	32
4.2 Αποτελέσματα σχήματος Naïve Bayes.....	33
4.3 Αποτελέσματα Random Forest.....	36
4.4 Αποτελέσματα Neural Net.....	38
4.5 Αποτελέσματα Support Vector Machine.....	39
5. Συμπεράσματα.....	41
6. Βιβλιογραφία.....	43

Εισαγωγή

Αντικείμενο της παρούσας εργασίας είναι η μελέτη του καρκίνου των ωοθηκών (Ovarian cancer) κατασκευάζοντας μοντέλα πρόβλεψης, με την βοήθεια της μηχανικής μάθησης.

Ο καρκίνος των ωοθηκών είναι ένας κακοήθης όγκος που προσβάλλει τη μία ή και τους δύο ωοθήκες. Υπάρχουν δύο τύποι καρκίνου των ωοθηκών και αυτοί είναι ο επιθηλιακός και ο μη επιθηλιακός.

Όσον αφορά την μηχανική μάθηση, είναι υποπεδίο της επιστήμης των υπολογιστών που αναπτύχθηκε από την μελέτη αναγνώρισης προτύπων και υπολογιστικής θεωρίας μάθησης στην τεχνητή νοημοσύνη. Η μηχανική μάθηση διερευνά τη μελέτη και την κατασκευή αλγορίθμων που μπορούν να μαθαίνουν από τα δεδομένα κάνοντας προβλέψεις σχετικά με αυτά. Αλγόριθμοι κατασκευάζουν μοντέλα από πειραματικά δεδομένα, προκειμένου να δημιουργήσουν προβλέψεις βασιζόμενες στα δεδομένα ή να εξάγουν αποφάσεις που αποτελούν το αποτέλεσμα. Τώρα στο πεδίο της ανάλυσης των δεδομένων, η μηχανική μάθηση είναι μια μέθοδος που χρησιμοποιείται για την επινόηση πολύπλοκων μοντέλων που οδηγούν στην πρόβλεψη.

Επιπλέον, γίνεται μια βιβλιογραφική ανασκόπηση σχετικά με την ανάπτυξη μοντέλων πρόβλεψης του καρκίνου των ωοθηκών. Τέλος, περιγράφεται το σύνολο των δεδομένων που αναλύονται και γίνεται η εφαρμογή των αλγορίθμων με την χρήση του RapidMiner.

Σκοπός της παρούσας εργασίας είναι η διερεύνηση των παραγόντων που επηρεάζουν την εμφάνιση καρκίνου των ωοθηκών και η δημιουργία του μοντέλου πρόβλεψης.

1.Περιγραφή της ασθένειας

Οι ωοθήκες είναι οι περιοχές παραγωγής του ωαρίου στις γυναίκες. Είναι η κύρια πηγή των ορμονών οιστρογόνου και προγεστερόνης σε προ εμμηνόπαυσιακές γυναίκες. Υπάρχουν τρεις τύποι ιστών που μπορούν να προκαλέσουν καρκίνο των ωοθηκών : επιθηλιακά κύτταρα, τα οποία καλύπτουν την ωοθήκη, στρωματικά κύτταρα τα οποία παράγουν ορμόνες και τα γεννητικά κύτταρα τα οποία γίνονται ωάρια. Περίπου το 85 έως 90 τοις εκατό των καρκίνων των ωοθηκών προκαλούνται από επιθηλιακά καρκινώματα. Σύμφωνα με τον παγκόσμιο οργανισμό έρευνας καρκίνων (WCRF) ο

καρκίνος των ωοθηκών είναι ο έβδομος πιο κοινός καρκίνος τους γυναίκες και ο 18^{ος} πιο κοινός καρκίνος συνολικά παγκοσμίως. Περίπου 239,000 περιπτώσεις καταγράφηκαν το 2012 αντιπροσωπεύοντας, σχεδόν το 4% όλων των νέων κρουσμάτων καρκίνου στις γυναίκες. Ο καρκίνος είναι συνήθως θανατηφόρος και είναι η όγδοη πιο συχνή αιτία θανάτου από καρκίνο σε γυναίκες. Ο καρκίνος των ωοθηκών συχνά δεν έχει συμπτώματα στα αρχικά στάδια, επομένως η ασθένεια προχωρά πριν διαγνωστεί. Το ποσοστό επιβίωσης κυμαίνεται από περίπου 30 έως 50 τοις εκατό.

Προδιαθεσικοί παράγοντες : Οι μηχανισμοί που οδηγούν στην ανάπτυξη καρκίνου των ωοθηκών είναι άγνωστοι. Οι πιο γνωστοί τους προδιαθεσικοί παράγοντες είναι το οικογενειακό ιστορικό και η κληρονομικότητα, η μεγάλη ηλικία και προϋπάρχουν καρκίνος του μαστού. Αυξημένη πιθανότητα εμφάνισης καρκίνου των ωοθηκών εμφανίζεται σε γυναίκες όπου η μητέρα τους είχε καρκίνο ωοθηκών. Μεταλλάξεις στα γονίδια BRCA1 ή BRCA2 αυξάνουν τον κίνδυνο εμφάνισης των ωοθηκών κατά 23-54%.

Κλινική εικόνα : Τα αρχικά συμπτώματα τους νόσου δεν είναι παθογνωμικά, γεγονός που καθυστερεί την διάγνωση. Γι' αυτό το 70% των περιστατικών διαγιγνώσκονται σε προχωρημένο στάδιο τους νόσου. Τα συνηθέστερα συμπτώματα είναι το αίσθημα βάρους και πληρότητας στην κοιλιά, καθώς και οι διαταραχές τους εντερικής λειτουργίας ή ενοχλήσεις κατά την ούρηση. Λιγότερο συχνά συμπτώματα είναι η απώλεια όρεξης, ο κοιλιακός πόνος, οι κοιλιακές αιμορραγίες και οι διαταραχές κύκλου.

Διάγνωση : Έπειτα από κλινική εξέταση, έλεγχος του καρκινικού δείκτη CA-125, υπέρηχος ωοθηκών καθώς και μαγνητική τομογραφία.

(American Cancer Society, 2020)

1.1 Στάδια του καρκίνου των ωοθηκών

Ο κάθε ασθενής που διαγιγνώσκεται με καρκίνο των ωοθηκών αποδίδεται με ένα στάδιο. Το κάθε στάδιο περιγράφει το μέγεθος του καρκίνου αλλά και το πόσο έχει εξαπλωθεί. Τα στάδια του καρκίνου των ωοθηκών είναι τέσσερα και αποδίδονται παρακάτω :

- Στάδιο 1 : Ο όγκος βρίσκεται στη μία η και τους δύο ωοθήκες.
- Στάδιο 2 : Ο όγκος βρίσκεται στη μία η και τους δύο ωοθήκες και έχει εξαπλωθεί σε άλλα κοντινά όργανα.

- Στάδιο 3 : Ο όγκος έχει εξαπλωθεί στα εσωτερικά τοιχώματα της κοιλιάς, των εντέρων ή των λεμφαδένων της κοιλιάς και τους λεκάνες.
- Στάδιο 4 : Ο όγκος έχει εξαπλωθεί έξω από την κοιλιά π.χ. στο συκώτι, τους πνεύμονες ή σε μακρινούς λεμφαδένες, δηλαδή έχει κάνει μετάσταση.

Στο καρκίνο των ωοθηκών αποδίδεται τους και τους βαθμούς. Ο βαθμός τους αποτυπώνει το πόσο επιθετικός είναι ο καρκίνος και τον ρυθμό εξάπλωσής του. Ο βαθμός κυμαίνεται από το 1 έως το 3. Όσο μεγαλύτερος ο βαθμός τόσο γρηγορότερη είναι και η ανάπτυξή του.

(National Health Service UK, 2019)

1.2 Αντιμετώπιση και θεραπεία του καρκίνου των ωοθηκών.

Η αρχική θεραπεία είναι συνήθως η χειρουργική, κατά την οποία βγαίνει και η οριστική διάγνωση του καρκίνου με την ιστολογική εξέταση. Σκοπός τώρα της χειρουργικής θεραπείας είναι η εξαίρεση του ή των όγκων που υπάρχουν της ωοθήκης. Απαραίτητη είναι η σταδιοποίηση και η κυτταρομείωση (αφαίρεση των μεγάλων και των μικρών εστιών καρκίνου). Η σωστή χειρουργική επέμβαση έχει μεγάλη σημασία για την περαιτέρω εξέλιξη της νόσου. Με την βοήθεια της χειρουργικής θεραπείας και της ιστολογικής εξέτασης καθορίζεται και το στάδιο το οποίο βρίσκεται ο καρκίνος.

Η χημειοθεραπεία, συνήθως ενδοφλέβια, ακολουθεί σχεδόν πάντα τη χειρουργική θεραπεία και εξαρτάται από τον ιστολογικό τύπο. Τις περισσότερες φορές γίνονται τουλάχιστον 6 κύκλοι θεραπείας, ανά τρεις με τέσσερις εβδομάδες. Στις χημειοθεραπείες χρησιμοποιούνται συνήθως δύο φάρμακα παράγωγα της πλατίνας και της ταξόλης.

Τέλος, η ακτινοθεραπεία στον καρκίνο των ωοθηκών χρησιμοποιείται λιγότερο συχνά, συνήθως σε υποτροπή της νόσου.

(National Health Service UK, 2019)

1.3 Παράγοντες κινδύνου ανάπτυξης των ωοθηκών.

Οι παράγοντες που επιδρούν στον κίνδυνο ανάπτυξης του καρκίνου των ωοθηκών χωρίζονται σε παράγοντες αύξησης κινδύνου και ελάττωσης :

Αύξηση κινδύνου :

- Κληρονομικότητα
- Καθυστερημένη εμμηνόπαυση
- Οιστρογονική ορμονοθεραπεία
- Στεριότητα
- Πληθυσμιακός έλεγχος

Ελάττωση κινδύνου :

- Αντισύλληψη
- Τεκνοποίηση
- Υστερεκτομή
- Θηλασμός

Πρέπει να γίνει ιδιαίτερη αναφορά σε περιπτώσεις όπου ο καρκίνος των ωοθηκών καθορίζεται κληρονομικά και αποδίδονται σε μεταλλάξεις των γονιδίων BRCA 1 και BRCA 2 όπου ο κίνδυνος ανάπτυξης κακοήθειας ανέρχεται σε 40-60 % και 16-27% αντίστοιχα. Οπότε η παρακολούθηση και η αντιμετώπιση των γυναικών αυτών διαφοροποιείται από το γενικό πληθυσμό.

(American Cancer Society, 2020)

1.4 Κοινωνική διάσταση

Ο καρκίνος των ωοθηκών είναι ο 7^{ος} πιο κοινός καρκίνος και ο 8^{ος} πιο θανατηφόρος καρκίνος των γυναικών παγκοσμίως. Το 2012 η μελέτη του Globocan υπολόγισε ότι υπήρχαν 239,000 περιπτώσεις καρκίνου των ωοθηκών και 152,000 θάνατοι (αντιπροσωπεύουν το 3,6% των περιπτώσεων καρκίνου και το 4,3% των θανάτων). Παγκοσμίως υπάρχουν σχεδόν 600,000 γυναίκες που ζούν έχοντας τον καρκίνο των ωοθηκών πέντε χρόνια. Αυτή η μορφή καρκίνου αναπτύσσεται κυρίως σε μεγαλύτερης ηλικίας γυναίκες και είναι συχνότερος σε γυναίκες τους λευκής φυλής. Η μελέτη της Globocan προβλέπει ότι έως το 2035 θα υπάρξει παγκόσμια αύξηση 55% σε περιστατικά και 67% σε θανάτους. Τα στοιχεία υπολογίστηκαν χρησιμοποιώντας τους προοπτικές του παγκόσμιου πληθυσμού του ΟΗΕ. Τώρα όσον αφορά τους αριθμούς που επηρεάζονται από την ήπειρο έχουν ως εξής για το έτος 2012.

OVARIAN CANCER	INCIDENCE 2012	MORTALITY 2012	5-YEAR PREVALENCE
ASIA	111,887	66,215	276,073
EUROPE	65,584	42,749	157,198
NORTH AMERICA	23,529	16,995	58,702
LATIN AMERICA AND CARIBBEAN	17,921	11,471	48,439
AFRICA	17,755	13,085	41,052
OCEANIA	2,043	1,402	5,160

Πίνακας 1 Παγκόσμια θνησιμότητα καρκίνου

(THE WORLD OVARIAN CANCER COALITION ATLAS ©The World Ovarian Cancer Coalition 2018)

Οικονομικά στατιστικά

Ο καρκίνος παραμένει η κύρια αιτία θανάτου παγκοσμίως. Η φροντίδα του καρκίνου είναι από τους ταχύτερα αναπτυσσόμενους τομείς στην υγειονομική περίθαλψη. Το κόστος της φροντίδας του καρκίνου το 2010 ήταν 124,6 δισεκατομμύρια δολάρια και αναμένεται να αυξηθεί στα 173 δισεκατομμύρια έως το 2022. Η θεραπεία του καρκίνου των ωοθηκών είναι πολυτροπική και εντατική. Μετά τη διάγνωση, οι γυναίκες συνήθως υποβάλλονται σε χειρουργική επέμβαση και λαμβάνουν χημειοθεραπεία. Οι εκτιμώμενες ετήσιες δαπάνες για την φροντίδα των γυναικών μετά την χειρουργική επέμβαση τους Ηνωμένες Πολιτείες Αμερικής ανέρχονται τους 100,000 \$.

Στις μεγαλύτερες αναπτυσσόμενες χώρες (Βραζιλία, Ρωσία, Κίνα) το συνολικό κόστος τους απώλειες λόγω πρόωρης θνησιμότητας από καρκίνο εκτιμάται ότι έφτασε τα 46,3 δισεκατομμύρια δολάρια το 2012.

Τέλος, το παγκόσμιο κόστος του καρκίνου αναμένεται να αυξηθεί λόγω της αύξησης του αριθμού των νέων περιπτώσεων, καθώς και του συνεχώς αυξανόμενου κόστους θεραπείας.

(Bercow AS, Chen L, Chatterjee S, et al. Cost of Care for the Initial Management of Ovarian Cancer. *Obstet Gynecol.* 2017)

2. Βιβλιογραφική Ανασκόπηση

Στο κεφάλαιο αυτό παρουσιάζονται διάφορες μελέτες που έχουν πραγματοποιηθεί για την ανίχνευση και την πρόβλεψη του καρκίνου των ωοθηκών. Η έγκαιρη ανίχνευση του καρκίνου έχει την δυνατότητα να μειώσει δραστικά την θνησιμότητα. Η χρήση αλγορίθμων εξόρυξης δεδομένων με δεδομένα φασματομετρίας μάζας μπορεί να αναδειχθεί ως μια μέθοδος με λαμπρό μέλλον. Στο πεδίο αυτό γίνονται τους μελέτες ανά τον κόσμο και αναλύουν διαφορετικές τεχνικές ανίχνευσης του καρκίνου.

2.1 Εκπαίδευση μοντέλων για εξόρυξη βιολογικών δεδομένων. Μια εφαρμογή για την ανίχνευση του καρκίνου των ωοθηκών.

(*Learning Causal Models for Noisy Biological Data Mining: An Application to Ovarian Cancer Detection*)

Το συγκεκριμένο άρθρο επικεντρώνεται στην πιθανότητα εμφάνισης ανεπιθύμητων σφαλμάτων τους μετρήσεις μικροσυστοιχιών DNA και στην φασματοσκοπία πρωτεϊνών, τα οποία θα μπορούσαν να επηρεάσουν την διαγνωστική αξιοπιστία. Τα διαγνωστικά μοντέλα, εκτός από την υψηλή ανθεκτικότητα έναντι τέτοιων σφαλμάτων, πρέπει να είναι πιο κατανοητά και να κατανοούν βαθύτερα τους αλληλεπιδράσεις μεταξύ των βιολογικών οντοτήτων τους τα γονίδια και οι πρωτεΐνες. Στο άρθρο αυτό εισάγεται μια ισχυρή προσέγγιση εξόρυξης γνώσης, που αντιμετωπίζει αυτές τους προκλήσεις. Οι αιτιώδεις αλληλεπιδράσεις μεταξύ γονιδίων και πρωτεϊνών στα θορυβώδη δεδομένα ανακαλύπτονται αυτόματα μέσω τους δικτύου Bayes. Στην συνέχεια, πραγματοποιείται διάγνωση τους ασθενείας που βασίζεται στο δίκτυο χρησιμοποιώντας μία νέα διαδικασία χειρισμού σφαλμάτων, η οποία προσδιορίζει αυτόματα τους θορυβώδεις μετρήσεις και εξηγεί την αβεβαιότητα τους. Συγκεκριμένη εφαρμογή δείχνει ότι στο πρόβλημα τους ανίχνευσης του καρκίνου των ωοθηκών, η προσέγγιση είναι αποτελεσματική και ανακαλύπτει αιτιώδεις αλληλεπιδράσεις μεταξύ ειδικών πρωτεϊνών για τον καρκίνο. Με την προτεινόμενη διαδικασία αντιμετώπισης σφαλμάτων, το δίκτυο διακρίνει τέλεια τα ασθενή με τα υγιή άτομα.

Το σημαντικότερο πρόβλημα που αποτυπώνεται στην εργασία αυτή είναι ότι, ενώ υπάρχουν διάφορα μοντέλα πρόβλεψης, η δυνατότητα διάγνωσης με ακρίβεια σε συγκεκριμένα σύνολα δεν μπορεί να εγγυηθεί ότι τα μοντέλα αποδίδουν το ίδιο και στον γενικό πληθυσμό. Διαπιστώνεται ότι οι ταξινομητές που μαθαίνουν από

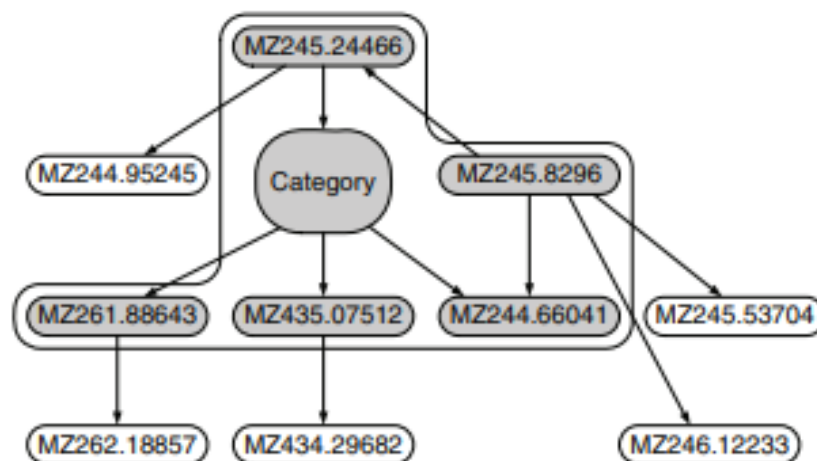
διαφορετικά σύνολα δεδομένων για τον καρκίνο των ωοθηκών διαφέρουν πολύ. Επιπλέον, για ισχυρές διαγνώσεις υπάρχει η ανάγκη για διαχείριση των σφαλμάτων που προκύπτουν και η ανίχνευση αγνώστων ποσοστών σφαλμάτων που υποκρύπτουν οι μετρήσεις κατά την πρόβλεψη.

Στην παρούσα εργασία εξετάζονται 253 περιπτώσεις με 162 ασθενείς και 91 υγιείς ανθρώπους με 15,154 M/Z μετρήσεις για την κάθε περίπτωση. Γίνεται τυχαία επιλογή 81 περιπτώσεων ασθενών και 45 υγιών για την δημιουργία του συνόλου εκμάθησης του αλγορίθμου. Τους γίνεται χρήση τους μεθόδου επιλογής χαρακτηριστικών με βάση την εντροπία και ξεχωρίζουν 4796 τιμές M/Z από τους 15154. Έτσι αποδεικνύεται ότι μόλις το $4796/15154=31,6\%$ των χαρακτηριστικών είναι ουσιαστικά ενώ τα υπόλοιπα είναι αμελητέα. Στην συνέχεια κατατάσσονται αυτά τα χαρακτηριστικά με βάση το όφελος τους πληροφορίας που προσφέρουν και επιλέγονται οι πρώτες 10 τιμές τους φαίνεται στο παρακάτω πίνακα.

Protein (M/Z)	Information Gain	Cutting Points
MZ261.88643	0.8190	0.3771,0.4011,0.4388
MZ262.18857	0.7627	0.4317, 0.5107
MZ245.24466	0.7381	0.4508
MZ244.95245	0.7381	0.4350
MZ245.8296	0.7219	0.3908
MZ245.53704	0.7219	0.5052
MZ244.66041	0.7202	0.2098, 0.3608
MZ246.12233	0.6903	0.3478
MZ435.07512	0.6693	0.3647
MZ434.29682	0.6650	0.4461, 0.5249

Πίνακας 2 Κατάταξη χαρακτηριστικών με βάση την πληροφορία

Στην συνέχεια, με βάση αυτές τους 10 τιμές εκπαιδεύεται ο αλγόριθμος Bayes(Bayesian Network), ο οποίος παράγει το μοντέλο κατηγοριοποίησης τους φαίνεται στο παρακάτω σχήμα.



Σχήμα 1 Μοντέλο Bayesian Network

Αφού εφαρμόστηκαν οι κανόνες αυτοί στα δεδομένα διαπιστώθηκε ότι το μοντέλο είναι αρκετά ακριβές, διότι ταξινομεί εσφαλμένα μόνο έναν από τους 81 καρκίνους και μόνο ένα από τα 46 φυσιολογικά δείγματα. Αυτό δίνει ευαισθησία (Sensitivity) 98,77%, εξειδίκευση (Specifcity) 97.83% και τιμή πρόβλεψης (Predictive value) 98.77%.

Στη συνέχεια, με βάση τα ποσοστά σφάλματος εφαρμόστηκε αλγόριθμος ανίχνευσης σφάλματος ώστε να εντοπιστούν η πιθανότητες σφάλματος για κάθε πρωτεΐνη ξεχωριστά. Τα αποτελέσματα που προκύπτουν είναι καλύτερα σε σχέση με τα προηγούμενα διότι εμφανίζεται βελτίωση στην ακρίβεια πρόβλεψης. Οι πιθανότητες εμφάνισης σφάλματος παρουσιάζονται στον παρακάτω πίνακα.

Name of the Protein (M/Z)	The Estimated Error Rate
MZ244.66041	0.2063
MZ434.29682	0.1429
MZ261.88643	0.0952
MZ262.18857	0.0952
MZ435.07512	0.0476
MZ245.24466	0.0079
MZ244.95245	0.0079
MZ246.12233	0.0079
MZ245.53704	0.0
MZ245.8296	0.0

Πίνακας 3 Πιθανότητες εμφάνισης σφάλματος

Συμπερασματικά έχει παρουσιαστεί ο συστηματικός μηχανισμός για μάθηση, εξαγωγή και εκμετάλλευση της γνώσης των αιτιωδών αλληλεπιδράσεων που περιέχονται σε θορυβώδη βιολογικά δεδομένα για την ανίχνευση του καρκίνου των ωοθηκών. Παρά τον μικρό αριθμό των πρωτεϊνών (μόλις 10), τα αποτελέσματα δείχνουν ότι ο καρκίνος θα μπορούσε να διακριθεί με πολύ μεγάλη ακρίβεια. Εκτός από την προγνωστική ακρίβεια, η ικανότητα δημιουργίας ερμηνεύσιμης γνώσης είναι το βασικό πλεονέκτημα της συγκεκριμένης προσέγγισης. Το δίκτυο Bayesian αποδεικνύεται το πιο σημαντικό, αφού μπορεί εύκολα να μάθει από δεδομένα και να επαληθευτεί με τους προηγούμενες γνώσεις από ιατρούς και βιολόγους.

2.2 Συγκριτική μελέτη της μηχανικής μάθησης και των αλγορίθμων βαθιάς μάθησης για ταξινόμηση τύπων καρκίνου με βάση το γονίδιο μικροσυστοιχιών.

(A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data)

Η μελέτη αυτή πραγματεύεται την ταξινόμηση του καρκίνου, αποτελεί θέμα μείζονος ενδιαφέροντος στην ιατρική, εφόσον το επιτρέπει η ακριβής και αποτελεσματική διάγνωση. Οι μελέτες έχουν ταξινομήσει ανθρώπινους όγκους σε μεγάλη κλίμακα RNA με την χρήση εποπτευόμενων αλγορίθμων μηχανικής μάθησης, οι οποίοι καταφέρνουν να κατασκευάσουν μοριακές ταξινομήσεις καρκινικών κυττάρων. Αυτά τα σύνολα δεδομένων είναι γνωστά με την ονομασία 11_Tumor βάσεις δεδομένων και έχουν χρησιμοποιηθεί σε πολλά έργα.

Οι τεχνικές που εφαρμόστηκαν στην συγκεκριμένη μελέτη είναι οι τεχνικές ML και DL οι οποίες μπορούν να μάθουν τα χαρακτηριστικά του δεδομένου προβλήματος από μία συγκεκριμένη ποσότητα δεδομένων (Eraslan et al., 2019). Αυτά τα δεδομένα συνήθως υποδιαιρούνται τυχαία σε δύο ομάδες: εκπαίδευσης και επικύρωσης. Το σύνολο της εκπαίδευσης χρησιμοποιείται για την βαθμονόμηση των παραμέτρων και το σύνολο της επικύρωσης για την αξιολόγηση της απόδοσης του μοντέλου. Στο άρθρο αυτό αξιολογήθηκαν δύο μη εποπτευόμενες μέθοδοι. Η πρώτη είναι ο αλγόριθμος K-Means, στον οποίο δόθηκε δεδομένος αριθμός πρωτοτύπων και η δεύτερη μέθοδος είναι η ιεραρχική ομαδοποίηση και ταιριάζει καλύτερα σε ακανόνιστα σχήματα από την

πρώτη μέθοδο. Στην συνέχεια, δοκιμάστηκαν οκτώ διαφορετικές ταξινομήσεις αλγορίθμων και παρουσιάστηκαν οι αποδόσεις τους.

Τα δεδομένα που χρησιμοποιήθηκαν αντιπροσωπεύουν μετρήσεις έκφρασης γονιδίου χρησιμοποιώντας μικροσυστοιχίες καρκίνου αλλά και φυσιολογικές βιοψίες (Statnikov et al., 2005; Bolón-Canedo et al., 2014) και είναι ενοποιημένα στην βάση δεδομένων 11 Tumors. Η βάση αυτή αποτελείται από 174 δείγματα με 12,533 μικροσυστοιχίες γονιδιακής έκφρασης για 11 διαφορετικούς τύπους καρκίνου. Οι 12,533 μικροσυστοιχίες γενετικής έκφρασης είναι ακέραιοι με θετικές και αρνητικές τιμές. Οι τιμές αυτές αντιπροσωπεύουν τα χαρακτηριστικά που επιτρέπουν να μάθουν στους αλγορίθμους ML και DL πως να ταξινομούνται ανάλογα με τον τύπο του καρκίνου.

Τα είδη του καρκίνου και ο αριθμός των ασθενών παρουσιάζονται στον παρακάτω πίνακα.

Class	Cancer type	Number of patients
0	Ovary	27
1	Bladder/Ureter	8
2	Breast	26
3	Colorectal	23
4	Gastroesophagus	12
5	Kidney	11
6	Liver	7
7	Prostate	26
8	Pancreas	6
9	Adenocarcinoma	14
10	Lung squamous cell carcinoma	14

Πίνακας 4 Είδη καρκίνου και αριθμός ασθενών

Για την πραγματοποίηση των πειραμάτων το σύνολο των δεδομένων χωρίζεται τυχαία σε δύο υποσύνολα (80% εκπαίδευση και 20% επικύρωση), συμπεριλαμβανομένων 139 δειγμάτων για εκπαίδευση και 35 δειγμάτων για επικύρωση. Έπειτα, εφαρμόστηκαν δύο τεχνικές προ επεξεργασίας των δεδομένων. Η πρώτη τεχνική είναι η κλιμάκωση (Scaling Geron 2017) και η δεύτερη τεχνική είναι η αρχή ανάλυσης των συστατικών (PCA Wold, Esbensen & Geladi, 1987). Η πρώτη τεχνική εγγυάται ότι τα δεδομένα βρίσκονται σε μια σειρά κατάλληλων τιμών για την βαθμονόμηση και η δεύτερη βελτιώνει την στατιστική σημασία των δεδομένων και επίσης εξαλείφεται ο θόρυβός τους. Τέλος, δημιουργήθηκαν τέσσερα διαφορετικά σύνολα δεδομένων για την εκπαίδευση και την επικύρωση κάθε ML ή DL αλγορίθμου.

Στο πλαίσιο των πειραμάτων τώρα εφαρμόστηκαν εποπτευόμενες και μη μαθησιακές προσεγγίσεις των ML και DL για την ταξινόμηση 11 τύπων καρκίνου με βάσι το σύνολο δεδομένων Microarray. Στον παρακάτω πίνακα παρουσιάζονται τα αποτελέσματα των πειραμάτων που διενεργήθηκαν μέσω των διαφορετικών αλγορίθμων που δοκιμάστηκαν.

Algorithm	Conditions on the dataset	Tuning parameters	% Accuracy
Results on validation data (the best result)			
K-Nearest Neighbors	Any	Neighbors=1	88.57
	Scaling	Neighbors: 1	71.43
	PCA	Neighbors: 1	82.86
	Scaling + PCA	Neighbors: 4	48.57
Support Vector Classifier	Any	C=10	8.57
	Scaling	C=70	94.29
	PCA	C=10	8.57
	Scaling + PCA	C=40	91.43
Logistic regression	Any	C=0,1	100.00
	Scaling	C=0,1	97.14
	PCA	C=0,1	94.29
	Scaling + PCA	C=0,1	94.29
Linear discriminant analysis	Any	Default	91.43
	Scaling	Default	91.43
	PCA	Default	97.14
	Scaling + PCA	Default	82.86
Gaussian NB	Any	Default	85.71
	Scaling	Default	85.71
	PCA	Default	80.00
	Scaling + PCA	Default	71.43
Random forest	Any	n_estimators=81, max_depth=91, min_samples_split=10, max_features=50	97.14
	Scaling	n_estimators=91, max_depth=81, min_samples_split=10, max_features=60	97.14
	PCA	n_estimators=91, max_depth=21, min_samples_split=10, max_features=30	94.28
	Scaling + PCA	n_estimators=61, max_depth=11, min_samples_split=10, max_features=20	85.71
Decision tree	Any	max_depth=71, min_samples_split=10, max_features=40	68.57
	Scaling	max_depth=51, min_samples_split=10, max_features=60	68.57
	PCA	max_depth=81, min_samples_split=10, max_features=30	82.85
	Scaling + PCA	max_depth=51, min_samples_split=20, max_features=60	74.28
Multi-layer perceptron	Any	Neurons=800	85.71
	Scaling	Neurons=50	91.43
	PCA	Neurons=300	97.14
	Scaling + PCA	Neurons=50	91.43
K-means	Any	Clusters=16	76.97
	Scaling	Clusters=14	68.34
	PCA	Clusters=16	73.38
	Scaling + PCA	Clusters=11	58.99

Πίνακας 5 Αποτελέσματα πειραμάτων

Τα αποτελέσματα της έρευνας προβλέπουν ότι ο καρκίνος προβλέπεται να γίνει η πιο θανατηφόρα ασθένεια για του ανθρώπους στο μέλλον (Dagenais et al., 2019). Οι

τεχνικές ML και DL είναι πολλά υποσχόμενα εργαλεία για την ταξινόμηση των τύπων του καρκίνου. Σε αυτή την μελέτη παρατηρήθηκαν προβλέψεις με ακρίβεια 93,52% και 94,46%, κάτι που θα επιτρέψει σε ασθενείς την ακριβή και έγκαιρη διάγνωσή τους. Τέλος θα συμβάλει στην ανακάλυψη νέων επιλεκτικών φαρμάκων για την θεραπεία των συγκεκριμένων καρκινικών όγκων.

2.3 Ανάλυση μεγάλων δεδομένων για την δημιουργία γνώσης. Σύστημα έγκαιρης ανίχνευσης του καρκίνου των ωοθηκών.

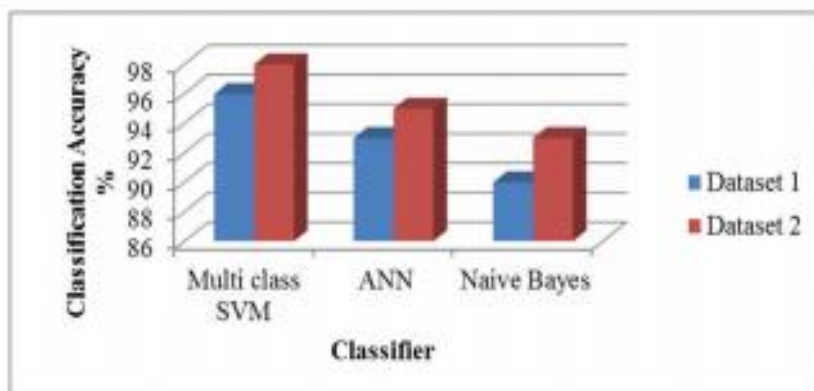
(Analysing Big Data to Build Knowledge Based System for Early Detection of Ovarian Cancer P. Yasodha^{1*} and N. R. Anathanarayanan)

Η ανάλυση των δεδομένων παίζει καθοριστικό ρόλο στην υγειονομική περίθαλψη για την έγκαιρη διάγνωση των θανατηφόρων ασθενειών. Αυτό το άρθρο χρησιμοποιεί μεθόδους εξόρυξης δεδομένων, όπως η επιλογή χαρακτηριστικών και η ταξινόμηση για την κατασκευή προγνωστικού μοντέλου του καρκίνου των ωοθηκών. Συγκεντρώνεται ένα τεράστιο σύνολο δεδομένων για την κατασκευή συστήματος βασισμένου στην γνώση. Χρησιμοποιείται η βελτιστοποίηση του γενετικού σμήνους υβριδικών σωματιδίων (PGSO) για την βελτιστοποίηση των επιλεγμένων χαρακτηριστικών για την αποτελεσματική ταξινόμηση του καρκίνου των ωοθηκών, είτε φυσιολογικός, είτε πρώιμος, είτε διαφορετικό στάδιο του καρκίνου των ωοθηκών. Υιοθετείται ως ταξινομητής ο Multi Class SVM για την ταξινόμηση των φυσιολογικών ή των διαφορετικών σταδίων του καρκίνου χρησιμοποιώντας το βέλτιστο σύνολο επιλεγμένων χαρακτηριστικών.

Τα γονίδια και η εξέταση αίματος χρησιμοποιούνται συνήθως ως δείκτες για τον καρκίνο των ωοθηκών, οπότε το συγκεκριμένο άρθρο περιέχει δύο σύνολα δεδομένων που λαμβάνονται υπόψη. Το πρώτο σύνολο δεδομένων αποτελείται από 493 παραδείγματα και 12,042 τιμές αυτών. Το σύνολο των δεδομένων ταξινομήθηκε σε 39 διαφορετικές περιπτώσεις ως πρώιμο στάδιο (Στάδιο I και II) και σε 454 περιπτώσεις ως όψιμο στάδιο (Στάδιο III και IV). Το δεύτερο σύνολο δεδομένων βασίζεται σε αποτελέσματα εξετάσεων αίματος και περιέχει 172 παραδείγματα για 28 διαφορετικές περιπτώσεις. Το σύνολο αυτό περιέχει 23 φυσιολογικά, 10 οριακά, 19 από πρώιμο στάδιο και 42 παραδείγματα του τελευταίου σταδίου.

Η απόδοση του συστήματος υπολογίζεται με βάση τους όρους απόλυτου σφάλματος (absolute error), την ακρίβεια ταξινόμησης (classification accuracy), την ευαισθησία

(sensitivity), την ειδικότητα (specificity) και την αντιστοιχία (respectively). Πραγματοποιήθηκαν πειράματα από τρεις διαφορετικούς αλγορίθμους οι οποίοι είναι ο Multiclass SVM, ο ANN και ο Naïve Bayes και παρουσιάζονται το τετραγωνικό σφάλμα για τα δύο σύνολα καθώς και η ακρίβεια αυτών.



Σχήμα 2 Απόδοση αλγορίθμων

Classifier	Dataset I			Dataset II		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
Multiclass SVM	97	95.4	96	99	96.7	98
ANN	95	91.6	93	97	92.9	95
Naive Bayes	93	87	90	96	91.4	93

Πίνακας 6 Μετρικές αλγορίθμων

Η ακρίβεια ταξινόμησης του προτεινόμενου συστήματος συγκρίνεται με το ANN και τον Naïve Bayes και όπως φαίνεται πέτυχε ακρίβεια 96% για το πρώτο σετ δεδομένων και 98% για το δεύτερο σετ δεδομένων. Ο παραπάνω πίνακας δείχνει την ευαισθησία (Sensitivity), την ειδικότητα (Specificity) και την ακρίβεια (Accuracy) για την ταξινόμηση του πρώιμου σταδίου και του τελευταίου σταδίου του καρκίνου. Γίνεται αντιληπτό ότι το προτεινόμενο σύστημα πετυχαίνει μεγαλύτερη ακρίβεια από το ANN και τον Naïve Bayes.

Η διάγνωση του καρκίνου των ωοθηκών είναι μια σημαντική έρευνα λαμβάνοντας υπόψη τη σημαντική αύξηση του ποσοστού των θανάτων. Η πρόωρη ανίχνευση και η

ακριβής σταδιοποίηση του, θα βοηθήσουν για την έγκαιρη και σωστή διαδικασία διάγνωσης του. Η απόδοση του συστήματος που έχει δοκιμαστεί χρησιμοποιώντας τους δύο διαφορετικούς καρκίνους των ωοθηκών και το προτεινόμενο σύστημα βασισμένο στην γνώση πέτυχε την μεγαλύτερη ακρίβεια σε σχέση με τους υπόλοιπους ταξινομητές. Συμπερασματικά, το σύστημα έχει την τάση να παρέχει καλύτερη πρόταση στον γιατρό για την πραγματοποίηση της διαγνωστικής διαδικασίας.

2.4 Μηχανική μάθηση και μοντέλα πληροφορικής για την αναγνώριση της γονιδιακής έκφρασης σε πρότυπα του καρκίνου των ωοθηκών που σχετίζονται με την εξέλιξη της νόσου και την θνησιμότητα.

(Machine learning and bioinformatics models to identify gene expression patterns of ovarian cancer associated with disease progression and mortality)

Ο καρκίνος των ωοθηκών είναι μια κοινή αιτία θανάτου από καρκίνο μεταξύ των γυναικών παγκοσμίως. Επομένως υπάρχει επιτακτική ανάγκη για τον εντοπισμό των παραγόντων που επηρεάζουν την θνησιμότητα από τον καρκίνο των ωοθηκών. Πολλά κλινικά δεδομένα ασθενών με καρκίνο των ωοθηκών είναι προσβάσιμα στο κοινό μέσω του Broad Institute Genome Atlas που περιλαμβάνουν την ηλικία του ασθενούς, τη θέση του καρκίνου, το στάδιο καθώς και το προφίλ μεταγραφής γονιδίου. Αυτά επιτρέπουν μελέτες που συσχετίζουν την επιβίωση ασθενών με καρκίνο των ωοθηκών με γονιδιακή έκφραση για τον εντοπισμό νέων βιοδεικτών για την πρόβλεψη της θνησιμότητας των ασθενών.

Σε αυτή την έρευνα ενσωματώνονται κλινικά και ιστογραφικά δεδομένα από ασθενείς από την πύλη TCGA. Προσδιορίζονται τα επίπεδα έκφρασης mRNA σε σύγκριση με τον κανονικό ιστό των ωοθηκών 41 γονιδίων που έχουν ήδη εμπλακεί σε καρκίνο των ωοθηκών. Επίσης, με την βοήθεια της COX για την χρήση μοντέλων παλινδρόμησης κινδύνου για την ανάλυση των παραγόντων και των μεταγραφικών πληροφοριών σχετικά με τις επιπτώσεις στην επιβίωση του κάθε ασθενή. Η μεθοδολογία ανάλυσης πολλαπλών σταδίων που χρησιμοποιήθηκε σε αυτή την μελέτη περιλαμβάνει δεδομένα TCGA (Μεταγραφικά και κλινικά). Επίσης, περιλαμβάνει δεδομένα για τον καρκίνο των ωοθηκών τα οποία λήφθηκαν από το cBioPortal και στην συνέχεια γίνεται συνδυασμός επιλεγμένων γονιδιακών βιοδεικτών τα οποία λαμβάνονται από την βάση

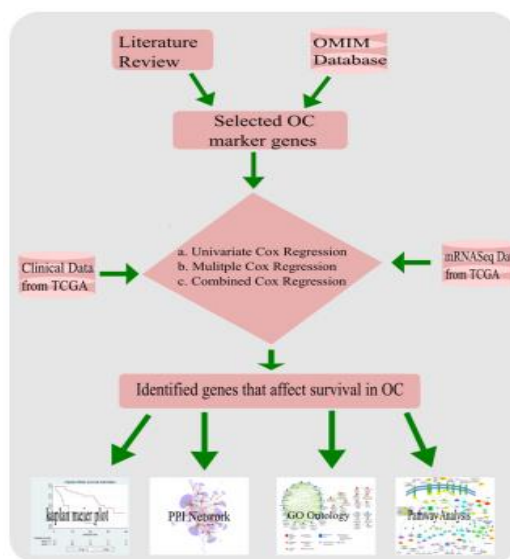
δεδομένων OMIM. Τέλος, έγινε αίτηση για συνδυασμένη παλινδρόμηση Cox για την προσδιορισμό των βασικών γονιδιακών βιοδεικτών. Επίσης τα 41 γονίδια που έχουν εμπλακεί σε καρκίνο των ωοθηκών, καθώς και οι πηγές από της οποίες αντλήθηκαν φαίνονται στον παρακάτω πίνακα :

Ovarian cancer associated significant 41 genes and their sources.

Gene Name	Resource	Ref.
BMP15, FMR1, FOXL2, CDH1, NR5A1, OPCML, DIAPH2, RAD51C, RAD51D, MSH2, FSHR, PARK2, POF1B, RRAS2, SF1, BRCA1, BRCA2, CTNNB1, ERBB2, PIK3CA and MSH2	Significant gene of OC from OMIM database https://www.omim.org/	[29]
BRCA1, BRCA2, BSL2, CTNNB1, ERBB2, MSH2, MUC16, PIK3CA, KRAS, TP53	http://www.cancer-genetics.org/X1003.htm	[33]
KRAS, PIK3CA, and TP53	Mutation profile and clinical outcome of mixed endometrioid-serous endometrial carcinomas are different from that of pure endometrioid or serous carcinomas	[35]
KLK6, CP, SLP1, EZH2, HPN, SCGB2A1	Predicting biomarkers for ovarian cancer using gene-expression microarrays	[31]
TLR4	The inflammatory microenvironment in epithelial ovarian cancer: a role for TLR4 and MyD88 and related proteins	[36]
WFDC2	HE4 (WFDC2) gene overexpression promotes ovarian tumor growth	[37]
RXRG, DVLI, MEF2B, MEF2C, BSL2, CD36, RXRG, E2F1 and TLR4	The impact on high-grade serous ovarian cancer of obesity and lipid metabolism-related gene expression patterns: the underestimated driving force affecting prognosis	[38]

Πίνακας 7 Εμπλεκόμενα γονίδια καρκίνου των ωοθηκών

Για αυτήν την μελέτη η διαδικασία ανάλυσης πολλαπλών βημάτων που αναπτύχθηκε και εφαρμόστηκε φαίνεται στο παρακάτω σχήμα και το πώς ενσωματώθηκαν μεταγραφικά και κλινικά δεδομένα ασθενών με καρκίνο των ωοθηκών.

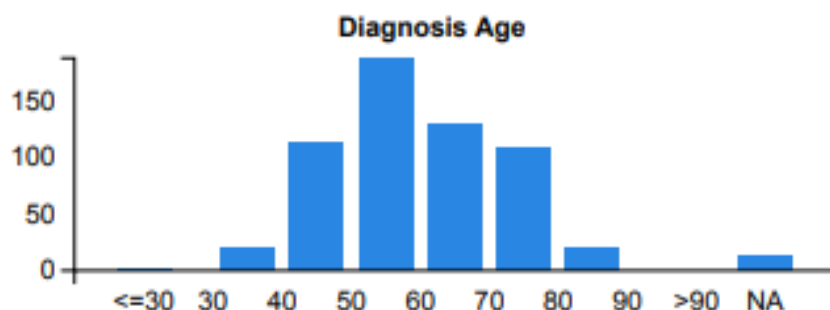


Σχήμα 3 Ενσωμάτωση στοιχείων ασθενών

Οι μέθοδοι που χρησιμοποιήθηκαν σχετικά με την επιβίωση των ασθενών με καρκίνο των ωοθηκών είναι οι εξής. Αρχικά χρησιμοποιήθηκε εκτιμητής ορίου για την εκτίμηση της συνάρτησης επιβίωσης και στην συνέχεια πραγματοποιήθηκε δοκιμή κατάταξης για τον προσδιορισμό των λειτουργιών επιβίωσης των δύο διαφορετικών ομάδων. Στην συνέχεια χρησιμοποιήθηκε παλινδρόμηση Cox για τον προσδιορισμό της σημασίας των γονιδίων καθώς και τον κλινικών παραγόντων. Τέλος πραγματοποιήθηκαν λειτουργικές αναλύσεις των πιο σημαντικών γονιδίων που βρέθηκαν από την ανάλυση. Επιλέχθηκαν σημαντικές κλινικές και δημογραφικές μεταβλητές που επηρεάζουν τον καρκίνο των ωοθηκών. Επιλέχθηκαν επίσης 41 γονίδια που είχαν πειραματικά καθορισμένη σημασία στο πλαίσιο του καρκίνου των ωοθηκών και πραγματοποιήθηκε ξεχωριστά για κάθε ένα γονίδιο ταυτόχρονη και πολυπαραγοντική παλινδρόμηση.

Τα αποτελέσματα της έρευνας συμπεριλαμβάνουν 6 κλινικούς παράγοντες και δεδομένα έκφρασης mRNA για 41 γονίδια. Όλοι οι κλινικοί παράγοντες ήταν κατηγορηματικοί εκτός από την ηλικία.

Η ηλικιακή κατανομή των ασθενών φαίνεται στο παρακάτω σχήμα :



Σχήμα 4 Ηλικιακή κατανομή

Η ηλικία των ασθενών κατά τη στιγμή της διάγνωσής τους είναι τα 59,7 έτη και κυμαίνεται μεταξύ 41 και 89 ετών. Μετά την ανάλυση παλινδρόμησης πολλαπλών μεταβλητών Cox PH με κλινικά δεδομένα και δεδομένα mRNA, ταυτόχρονα παρουσιάζονται οι εκτιμώμενοι συντελεστές παλινδρόμησης μαζί με τις αναλογίες κινδύνου (HR) καθώς και τις τιμές P του συνδυασμού COX.

Variables	β	HR	P-Value
AMERICAN INDIAN OR ALASKA NATIVE	2.78E+00	1.61E+01	0.00074
ASIAN	-2.73E-02	9.73E-01	0.95317
BLACK OR AFRICAN AMERICAN	7.37E-01	2.09E+00	0.01637
NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER	1.54E+00	4.68E+00	0.16029
age	2.60E-02	1.03E+00	3.30E-05
anatomic_site Peritoneum ovary	3.83E-02	1.04E+00	0.97248
histologic_grade G2	6.95E-01	2.00E+00	0.40454
G3	8.01E-01	2.23E+00	0.33189
G4	2.65E+00	1.42E+01	0.08085
GB	3.36E-01	1.40E+00	0.83559
GX	-1.91E-01	8.26E-01	0.84474
tumour_site Left	-3.71E-01	6.90E-01	0.05113
tumour_site Right	2.69E-01	1.31E+00	0.19474
Stage IB	1.22E+00	3.40E+00	0.25561
Stage IC	-7.38E-01	4.78E-01	0.49448
Stage IIA	-1.58E+01	1.43E-07	0.99185
Stage IIB	6.35E-02	1.07E+00	0.93521
Stage IIC	-1.18E+00	3.09E-01	0.00483
Stage IIIA	-9.63E-02	9.08E-01	0.90753
Stage IIIB	-2.79E-01	7.57E-01	0.42995
Stage IIIC	-3.10E-01	7.33E-01	0.08836
Stage IV	NA	NA	NA

Πίνακας 8 Παρουσίαση αναλογιών κινδύνου καθώς και των τιμών P

Από τον παραπάνω πίνακα παρατηρείται ότι εάν η ηλικία του ασθενούς αυξάνεται κατά ένα έτος τότε η πιθανότητα κινδύνου αυξάνεται κατά 1,03 φορές. Επίσης παρατηρήθηκε ότι ο αμερικανικός και ο μαύρος πληθυσμός είχε μεγαλύτερο κίνδυνο θανάτου. Συγκεκριμένα δεν βρέθηκαν στοιχεία που να αποδεικνύουν ότι η εθνικότητα και ο ιστολογικός βαθμός έχουν κάποια σημαντική σχέση με τον κίνδυνο θανάτου. Παραδόξως, η ανάλυση δείχνει ότι οι ασθενείς με όγκο ωοθήκης δεξιά έχουν περισσότερες πιθανότητες να επιβιώσουν σε σύγκριση με ασθενής με διμερή όγκο, αλλά οι ασθενείς που έχουν αριστερή και διμερή όγκο εμφανίζουν την ίδια πιθανότητα κινδύνου. Το στάδιο του όγκου είναι ένας σημαντικός παράγοντας που επηρεάζει τον κίνδυνο θανάτου σε πολλούς καρκίνους και σύμφωνα με αυτό, βρέθηκε ότι οι ασθενείς με στάδιο IIC OC είναι 0,31 φορές λιγότερο πιθανό να επιβιώσουν σε σύγκριση με το στάδιο IA, αν και δεν υπήρχαν ενδείξεις.

Σε αυτή τη μελέτη η ανάλυση επιβίωσης 577 ασθενών με καρκίνο των ωοθηκών αποκάλυψε ότι μόνο 10 από τα 41 γονίδια που επιλέχθηκαν προβλέπουν μειωμένη επιβίωση από τον καρκίνο. Στις περιπτώσεις των γονιδίων TLR4, BSCL2, CDH1, ERBB2 και SCGB2A1, οι ασθενείς με αυτές τις τροποποιημένες εκφράσεις αυτών των πέντε γονιδίων είχαν σημαντικά χαμηλότερο χρόνο επιβίωσης από τους ασθενείς με μη μεταβολή αυτών των πέντε γονιδίων. Λαμβάνοντας υπόψη το αμοιβαίο αποτέλεσμα [20]

όλων των επιλεγμένων 41 γονιδίων σε πολυπαραγοντική ανάλυση, TLR4, BSCL2, CDH1, ERBB2 και PTPRE εμφάνισαν σημαντική επιρροή στην επιβίωση του καρκίνου των ωοθηκών, υποδηλώνοντας ότι αυτό μπορεί να εξηγήσει μερικές από τις αλλαγές επιβίωσης που προβλέπονται από το SCGB2A1.

2.5 Ανασκόπηση δεδομένων και επανεξέταση της πρωτεϊνικής μορφολογίας του καρκίνου των ωοθηκών.

(A data review and re-assessment of ovarian cancer serum proteomic profiling)

Στο άρθρο αυτό οι ερευνητές μελετούν τις μετρήσεις φασματογραφίας μάζας του καρκίνου των ωοθηκών και προσπαθούν να μετρήσουν την αξιοπιστία τους για την έγκαιρη διάγνωση του. Αναλύεται ένα σύνολο δεδομένων παρόμοιο με αυτό που επεξεργάζεται σε αυτή την εργασία και εφαρμόζονται διάφορες στατιστικές μέθοδοι, με σκοπό τον εντοπισμό δεικτών που θα βοηθήσουν στον εντοπισμό των ασθενών από τους υγιείς ανθρώπους. Επιλέγεται μια απλή στρατηγική χρήσης της Wilcoxon στα πρωτογενή δεδομένα με στόχο την καλύτερη κατανόηση των ιδιοτήτων του συνόλου των δεδομένων. Χρησιμοποιείται η δοκιμασία Wilcoxon και η σταδιακή διακριτική ανάλυση σε ένα υποσύνολο των δεδομένων που συνίσταται από 80 ασθενείς με καρκίνο και 45 υγιείς. Η επιλογή έγινε τυχαία από το αρχικό σύνολο δεδομένων για την ανάπτυξη κανόνων ταξινόμησης του συνόλου δοκιμών που περιέχει τα υπολειπόμενα υποκείμενα ελέγχου. Οι ταξινομητές ασθένειας μεγάλης ευαισθησίας θα μπορούσαν να κατασκευαστούν με οπτική επιθεώρηση και με συσσωμάτωση των τιμών M/Z με βάση τις τιμές P-Value της δοκιμής Wilcoxon σε συνδυασμό με τη διακριτική ανάλυση. Η μελέτη καταλήγει στο συμπέρασμα ότι η πρωτεϊνική διαμόρφωση ορού είναι μια νέα προσέγγιση στη διάγνωση του καρκίνου. Βέβαια, αντιμετωπίζει ένα δύσκολο περιβάλλον με τεχνολογίες μέτρησης οι οποίες είναι νέες στο κλινικό περιβάλλον με νέες προσεγγίσεις για την επεξεργασία δεδομένων υψηλών διαστάσεων. Τα αποτελέσματα που παρουσιάζονται σε αυτή τη μελέτη είναι χρήσιμα διότι οι προηγμένες τεχνικές εξόρυξης δεδομένων πρέπει να συγκρίνονται με τις παραδοσιακές μεθόδους. Επίσης πρέπει να οριστούν και άλλα παρόμοια σύνολα εκπαίδευσης για μία τέτοια σύγκριση διότι τα αποτελέσματα μπορεί να εξαρτώνται σε μεγάλο βαθμό από τα δείγματα που επιλέχθηκαν. Η ανάπτυξη ταξινομητών που

κάνουν χρήση ανάλυσης ρουτίνας αποδείχθηκε απλή και ταυτόχρονα πολύ αποδοτική. Επίσης στην μελέτη αυτή επιβεβαιώνεται ότι ένας ταξινομητής μπορεί να αναπτυχθεί με τιμές M/Z μεγαλύτερες από 2000. Υποδεικνύεται ότι υπάρχουν πληροφορίες σχετικά με τη διαφορά μεταξύ των καρκινικών και των υγιών δειγμάτων σε όλη την περιοχή M/Z που μελετήθηκε. Ακόμη η ανάλυση ρουτίνας επιτρέπει στους ερευνητές να αναθεωρήσουν τα δεδομένα τους και να συσχετίσουν τα ευρήματά τους με άλλες πληροφορίες. Οι χαμηλές τιμές M/Z δείχνουν ότι η πρέπει να δοθεί ιδιαίτερη προσοχή σε αυτές τις περιοχές, διότι αυτές οι περιοχές προσφέρουν το βέλτιστο για την ανάπτυξη διαγνωστικών δοκιμών του καρκίνου των ωοθηκών. Προκειμένου να αντιμετωπιστούν θέματα θορύβου και bias εξετάζεται η χρήση μεθόδων φασματομετρίας μάζας με αυξημένη ευαισθησία στην χαμηλή περιοχή M/Z . Οι μεταβολές στην χαμηλή περιοχή M/Z αντιπροσωπεύουν ένα αποτέλεσμα το οποίο προκαλείται από διαφοροποιημένα περιβάλλοντα πρωτεΐνης μεταξύ καρκίνου και υγιή. Σημαντική όμως είναι και οι ικανότητα διάκρισης υγιών και μη δειγμάτων με βάση τις τιμές M/Z 2,79 και 245,5. Όλα αυτά τα ευρήματα δείχνουν ότι οι κανόνες που προέρχονται από το συγκεκριμένο σύνολο δεδομένων, μπορεί να ανιχνεύουν διαφορές που οφείλονται στην πειραματική διαδικασία. Για αυτό τον λόγο οι ερευνητές προσπαθούν να ελαχιστοποιήσουν τις πιθανότητες λανθασμένων συμπερασμάτων χρησιμοποιώντας στατιστικές μεθόδους ρουτίνας.

3. Περιγραφή Δεδομένων, Αλγορίθμων και Μεθοδολογίας

3.1 Μηχανική Μάθηση

Η μηχανική μάθηση διερευνά τη μελέτη και την κατασκευή αλγορίθμων που είναι ικανοί να μαθαίνουν από σύνολα δεδομένων και να κάνουν προβλέψεις. Οι αλγόριθμοι αυτοί κατασκευάζουν μοντέλα με την χρήση των πειραματικών δεδομένων, προκειμένου να πραγματοποιούν προβλέψεις ή να εξάγουν αποφάσεις που εκφράζονται ως αποτελέσματα. Στο πεδίο της ανάλυσης δεδομένων η μηχανική μάθηση είναι μια

μέθοδος που χρησιμοποιείται για την επιλογή μοντέλων και αλγορίθμων που οδηγούν στην πρόβλεψη.

Στον τομέα της ιατρικής γίνεται μεγάλη προσπάθεια ανάπτυξης μοντέλων πρόβλεψης με την χρήση αλγορίθμων μηχανικής μάθησης και γενετικών δεδομένων. Για τον λόγο αυτό συλλέγονται κλινικά δεδομένα ασθενών με συγκεκριμένο τύπο καρκίνου, τα οποία αναλύονται μέσω αλγορίθμων μηχανικής μάθησης. Η έγκαιρη ανίχνευση του καρκίνου μπορεί να μειώσει την θνησιμότητα αυτού. Η χρήση αλγορίθμων μηχανικής μάθησης και δεδομένων από μετρήσεις φασματομετρίας μάζας, δίνει μία νέα πνοή στον τομέα της έγκαιρης διάγνωσης του καρκίνου.

3.2 Εργαλείο Επεξεργασίας

Το εργαλείο επεξεργασίας που χρησιμοποιήθηκε είναι το RapidMiner. Το εργαλείο αυτό είναι μια open source πλατφόρμα λογισμικού για Data Analytics που υλοποιήθηκε το 2001 από τους Ralf Klinkenberg, Ingo Mierswa και Simon Fischer στο τμήμα τεχνητής νοημοσύνης του Dortmund. Το RapidMiner παρέχει ένα ολοκληρωμένο περιβάλλον για την προετοιμασία δεδομένων, τη μηχανική μάθηση, την εξόρυξη δεδομένων και την ανάλυση προγνωστικών. Χρησιμοποιείται για έρευνα, εκπαίδευση, ταχεία ανάπτυξη πρωτοτύπων καθώς υποστηρίζει όλα τα στάδια της διαδικασίας της μηχανικής μάθησης.

Εκτός από CSV το RapidMiner δέχεται ένα σύνολο αρχείων όπως PDF, ARFF, XML, JSON και XLS. Τα δεδομένα μας έχουν την μορφή ARFF που καθιστά εύκολη την εισαγωγή τους και την επεξεργασία τους από το εργαλείο.

Σημαντικό είναι να αναφερθεί πως το RapidMiner διαθέτει τρεις εκδόσεις ανάλογα με το μέγεθος των δεδομένων που θέλουμε να επεξεργαστούμε αλλά και τον αριθμό των χρηστών που συμμετέχουν στο project. Οι τρεις εκδόσεις αναφέρονται παρακάτω :

➤ **Rapid Miner Studio :**

- Μπορεί να εγκατασταθεί σε προσωπικό υπολογιστή.
- Παρέχει ένα φιλικό προς τον χρήστη περιβάλλον για την σχεδίαση και την εκτέλεση αναλυτικών ροών εργασίας οι οποίες καλούνται Processes και αποτελούνται από Operators.

➤ **Rapid Miner Server :**

- Εγκαθίσταται και λειτουργεί σε Server και είναι ικανή να δεχθεί παράλληλη επεξεργασία δεδομένων από πολλούς χρήστες.

➤ **Rapid Miner Radoop :**

- Επεξεργάζεται δεδομένα ενός Cluster υπολογιστών. Στην περίπτωση αυτή το Rapid Miner εγκαθίσταται σε ένα σύστημα Hadoop και είναι ικανό να εκτελέσει ανάλυση μεγάλου όγκου δεδομένων.

3.3 Περιγραφή Δεδομένων

Το σύνολο των δεδομένων που αναλύθηκε και επεξεργάστηκε στην παρούσα εργασία ονομάζεται Ovarian DataSet 8-7-02 και είναι διαθέσιμο στο διαδίκτυο στην ιστοσελίδα Βάσης Δεδομένων του προγράμματος Κλινικών Πρωτεϊνωμάτων του Εθνικού Ινστιτούτου Καρκίνου των Ηνωμένων Πολιτειών (United States National Cancer Institute's Center for Cancer Research's Clinical Proteomics Program Databank – <http://home.ccr.cancer.gov/ncifdaproteomics>).

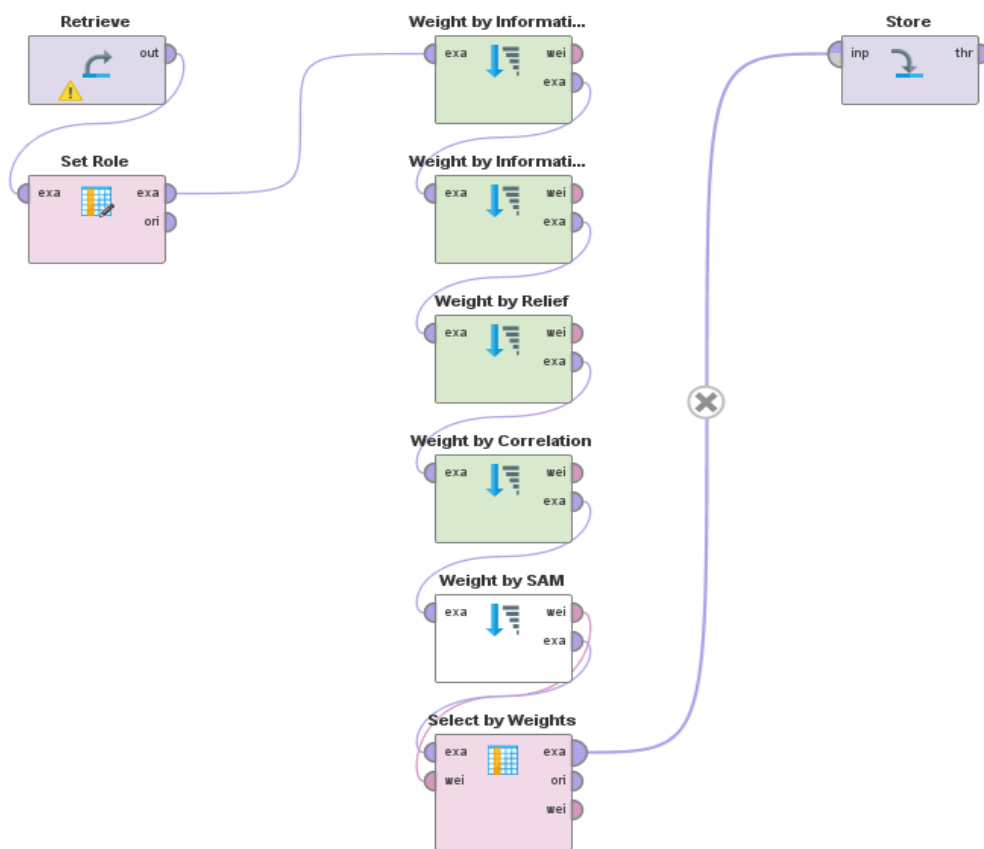
Τα δεδομένα περιέχουν μετρήσεις φασματομετρίας μάζας ορού και προέρχονται από 162 ασθενείς με καρκίνο των ωοθηκών και 91 φυσιολογικούς ανθρώπους. Τα δείγματα προετοιμάστηκαν με ρομποτικό όργανο και τα ακατέργαστα δεδομένα δημοσιεύθηκαν για δωρεάν λήψη. Το προφίλ των περιπτώσεων περιλαμβάνει 15.154 διαφορετικές αναλογίες μάζας προς φορτίο (M/Z) που κυμαίνονται από 0,0000786 έως 19995,513. Για την αποφυγή ανισορροπίας, λόγω της μεγάλης κλίμακας τιμών τα δεδομένα κανονικοποιήθηκαν σύμφωνα με τον παρακάτω τύπο :

$$KE = \frac{AE - Min}{Max - Min}$$

Όπου KE είναι η κανονικοποιημένη ένταση, AE είναι η αρχική ένταση και Min,Max είναι η ελάχιστη και η μέγιστη τιμή έντασης του συνόλου. Οι τιμές που εξετάζονται μετά την διαδικασία της κανονικοποίησης κυμαίνονται μεταξύ του 0 και του 1. Το σύνολο των δεδομένων συλλέχθηκε από δείγματα αίματος ανθρώπων με τη χρήση της συστοιχίας Ciphergen WCX2 ProteinChip. Η συγκεκριμένη συστοιχία είναι μια αδύναμη συστοιχία ανταλλαγής κατιόντων με καρβοξυλική επιφάνεια για να δεσμεύει κατιονικές πρωτεΐνες. Οι αρνητικά φορτισμένες καρβοξυλικές ομάδες στην επιφάνεια του τσιπ αλληλοεπιδρούν με τα θετικά φορτία που εκτίθενται στις πρωτεΐνες στόχους.

3.4 Αλγόριθμοι μείωσης χαρακτηριστικών

Όπως αναφέρθηκε παραπάνω, το αρχείο που μελετήθηκε περιέχει 15,154 διαφορετικούς δείκτες για 253 διαφορετικές περιπτώσεις. Για αυτό στο στάδιο της προ επεξεργασίας των δεδομένων, κύριο μέλημα ήταν η ελάττωση των χαρακτηριστικών. Αυτό πραγματοποιήθηκε μέσω μιας σειράς αλγορίθμων που βασίζονται στο βάρος των χαρακτηριστικών και εφαρμόστηκαν στα αρχικά δεδομένα μέσω του RapidMiner. Παρακάτω παρουσιάζεται η εικόνα με τους αλγορίθμους που χρησιμοποιήθηκαν για την προ επεξεργασία των δεδομένων.



Σχήμα 5 Διαδικασία επεξεργασίας δεδομένων

Αναλυτικά περιγράφονται παρακάτω οι τελεστές :

- **Weight By Information Gain** : Αυτός ο τελεστής υπολογίζει τη συνάφεια των χαρακτηριστικών με βάση το κέρδος πληροφοριών και τους αντιστοιχεί στα βάρη.

- **Weight By Information Gain Ratio** : Αυτός ο τελεστής υπολογίζει τη συνάφεια των χαρακτηριστικών με βάση την αναλογία κέρδους πληροφοριών και τους αντιστοιχεί στα βάρη.
- **Weight By Relief** : Αυτός ο τελεστής υπολογίζει τη συνάφεια των χαρακτηριστικών από το Relief. Η βασική ιδέα του Relief είναι να εκτιμηθεί η ποιότητα των χαρακτηριστικών ανάλογα με το πόσο καλά οι τιμές τους διακρίνουν μεταξύ των παρουσιών των ίδιων και διαφορετικών τάξεων που βρίσκονται κοντά στο άλλο.
- **Weight By Correlation** : Αυτός ο τελεστής υπολογίζει τη συνάφεια των χαρακτηριστικών υπολογίζοντας την τιμή συσχέτισης για κάθε χαρακτηριστικό του συνόλου δεδομένων σε σχέση με το χαρακτηριστικό label. Αυτό το σχήμα στάθμισης βασίζεται στη συσχέτιση και επιστρέφει την απόλυτη ή τετραγωνική τιμή της συσχέτισης ως βάρος χαρακτηριστικού.
- **Weight By SAM** : Αυτός ο τελεστής υπολογίζει το βάρος για αριθμητικά χαρακτηριστικά σύμφωνα με την "Ανάλυση Σημασίας για Μικροσυστοιχίες".
- **Select By Weights** : Αυτός ο τελεστής επιλέγει μόνο εκείνα τα χαρακτηριστικά ενός παραδείγματος εισαγωγής του οποίου τα βάρη ικανοποιούν το καθορισμένο κριτήριο σε σχέση με τα βάρη εισόδου. Το κριτήριο εισόδου που επιλέχθηκε είναι για βάρη μεγαλύτερα του 2,8.

Μετά το πέρας της παραπάνω διαδικασίας απομονώθηκε ένα υποσύνολο τριάντα δύο (32) χαρακτηριστικών, το οποίο χρησιμοποιήθηκε στην συνέχεια για την κατασκευή ενός μοντέλου ταξινόμησης. Παρακάτω παρατίθενται ο Πίνακας με τα χαρακτηριστικά του υποσυνόλου που επιλέχθηκαν.

MZ25.49556	MZ25.589892	MZ221.86191	MZ222.97536
MZ223.53313	MZ223.81227	MZ224.09159	MZ244.36855
MZ244.66041	MZ244.95245	MZ245.24466	MZ245.53704
MZ245.8296	MZ246.41524	MZ246.70832	MZ261.58446
MZ261.88643	MZ262.18857	MZ262.49088	MZ417.73207
MZ434.68588	MZ435.07512	MZ7999.9616	MZ8003.3008
MZ8004.9707	MZ8006.6407	MZ8008.3109	MZ8009.9813
MZ8011.6517	MZ223.25415	MZ246.12233	MZ8001.6311

Πίνακας 9 Επιλεγμένα χαρακτηριστικά

3.5 Αλγόριθμοι Ταξινόμησης

Στο επόμενο στάδιο γίνεται χρήση αλγορίθμων ταξινόμησης με σκοπό την δημιουργία ενός μοντέλου πρόβλεψης. Χρησιμοποιήθηκαν τέσσερις (4) μέθοδοι ταξινόμησης για την αξιολόγηση των αποτελεσμάτων. Παρακάτω παρουσιάζονται συνοπτικά οι μέθοδοι που χρησιμοποιήθηκαν.

- Ταξινομητής **Naïve Bayes**
- Δέντρο απόφασης **Random Forest**
- Νευρωνικό δίκτυο **Neural Net**
- Μηχανή διανυσματικής υποστήριξης **SVM (Support Vector Machine)**

Ταξινομητής Naïve Bayes

Ο ταξινομητής Naïve Bayes προσφέρει μια πιθανοθεωρητική προσέγγιση στα προβλήματα μάθησης, όπου ο στόχος είναι να προβλέψουμε ακριβώς την κατηγορία των στιγμιότυπων δοκιμής χρησιμοποιώντας ταξινομημένα στιγμιότυπα εκπαίδευσης. Συγκεκριμένα υποθέτει ότι το κάθε χαρακτηριστικό των στιγμιότυπων είναι στοχαστικά ανεξάρτητο των υπολοίπων και ότι δεν υπάρχουν άλλα κρυφά χαρακτηριστικά που επηρεάζουν την διαδικασία της πρόβλεψης. Η υπόθεση της ανεξαρτησίας των χαρακτηριστικών απλοποιεί σε μεγάλο βαθμό τους υπολογισμούς που απαιτούνται για την δημιουργία του μοντέλου πρόβλεψης.

Έχοντας τώρα ένα στιγμιότυπο δοκιμής $\mathbf{x}$ για ταξινόμηση, χρησιμοποιούμε τον νόμο του Bayes για να υπολογίσουμε το εκ των υστέρων γινόμενο πιθανοτήτων όπως φαίνεται στην παρακάτω εικόνα :

$$\arg \max_{\forall c} \{p(C = c | \mathbf{X} = \mathbf{x})\} = \arg \max_{\forall c} \left\{ p(C = c) \prod_i p(X_i = x_i | C = c) \right\}$$

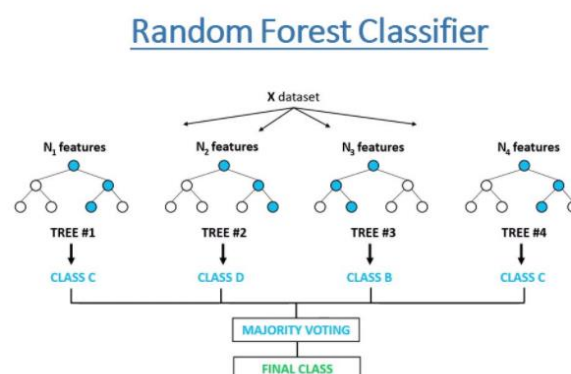
Κάθε στιγμιότυπο του προβλήματος αποτελείται από ένα σύνολο γνωρισμάτων. Εάν το πρόβλημα που μελετάται έχει n κλάσεις τότε ο ταξινομητής προβλέπει την στην οποία ανήκει το στιγμιότυπο με βάση την μεγαλύτερη εκ των υστέρων πιθανότητα. Τα μειονεκτήματα του συγκεκριμένου αλγορίθμου είναι ότι απαιτείται μεγάλο σύνολο

δεδομένων για την μεγιστοποίηση της απόδοσης του, καθώς και σε κάποιες περιπτώσεις τα αποτελέσματά του είναι μεροληπτικά.

Random Forest

Ο αλγόριθμος Random Forest αποτελεί ένα είδος μάθησης που λειτουργεί με την κατασκευή πολλών δέντρων απόφασης. Κάθε δέντρο απόφασης που κατασκευάζεται δίνει μια κατηγοριοποιημένη τιμή της ανεξάρτητης μεταβλητής. Αυτά τα δέντρα δημιουργούνται και εκπαιδεύονται σε ένα υποσύνστημα bootstrapped του συνόλου δεδομένων. Κάθε κόμβος ενός δέντρου αντιπροσωπεύει έναν κανόνα διαχωρισμού για ένα συγκεκριμένο χαρακτηριστικό. Μόνο ένα υποσύνολο χαρακτηριστικών λαμβάνεται υπόψη για την επιλογή των κανόνων διαχωρισμού. Για την ταξινόμηση ο κανόνας είναι ο διαχωρισμός των τιμών που ανήκουν σε διαφορετικές κατηγορίες, ενώ για την παλινδρόμηση οι τιμές διαχωρίζονται για να μειωθεί το ποσοστό σφάλματος που δημιουργήθηκε κατά την διαδικασία της εκτίμησης. Η δημιουργία νέων κόμβων επαναλαμβάνεται έως ότου πληρούνται τα κριτήρια διακοπής.

Κάθε τυχαίο δέντρο δημιουργεί μία πρόβλεψη για το κάθε παράδειγμα ακολουθώντας τα κλαδιά του δέντρου σύμφωνα με τους κανόνες διαχωρισμού αλλά και αξιολογώντας το κάθε φύλλο. Οι προβλέψεις βασίζονται στην πλειονότητα των παραδειγμάτων, ενώ οι εκτιμήσεις λαμβάνονται μέσω του μέσου όρου των τιμών που φτάνουν σε ένα φύλλο. Το τελικό μοντέλο που προκύπτει είναι ένα μοντέλο ψηφοφορίας όλων των κατασκευασμένων δέντρων. Δεδομένου τώρα ότι όλες οι μεμονωμένες προβλέψεις θεωρούνται εξίσου σημαντικές και βασίζονται σε υποσύνολα παραδειγμάτων, η πρόβλεψη που προκύπτει τείνει να διαφέρει λιγότερο από τις μεμονωμένες προβλέψεις.



Σχήμα 6 Δομή Random Forest

Μια ιδέα που ονομάζεται κλάδεμα (pruning) μπορεί να αξιοποιηθεί για να μειώσει την πολυπλοκότητα του μοντέλου αντικαθιστώντας τα υπό-δέντρα, τα οποία παρέχουν μικρή προγνωστική ισχύ. Σημαντικές παράμετροι για τη ρύθμιση αυτής της μεθόδου είναι το ελάχιστο μέγεθος του φύλλου και ο λόγος διαχωρισμού. Αυτό μπορεί να αλλάξει μετά την απενεργοποίηση του λόγου διάσπασης. Οι καλύτερες επιλογές για το ελάχιστο μέγεθος φύλλων είναι δυο (2) για την ταξινόμηση και πέντε (5) για τα προβλήματα παλινδρόμησης.

Neural Net

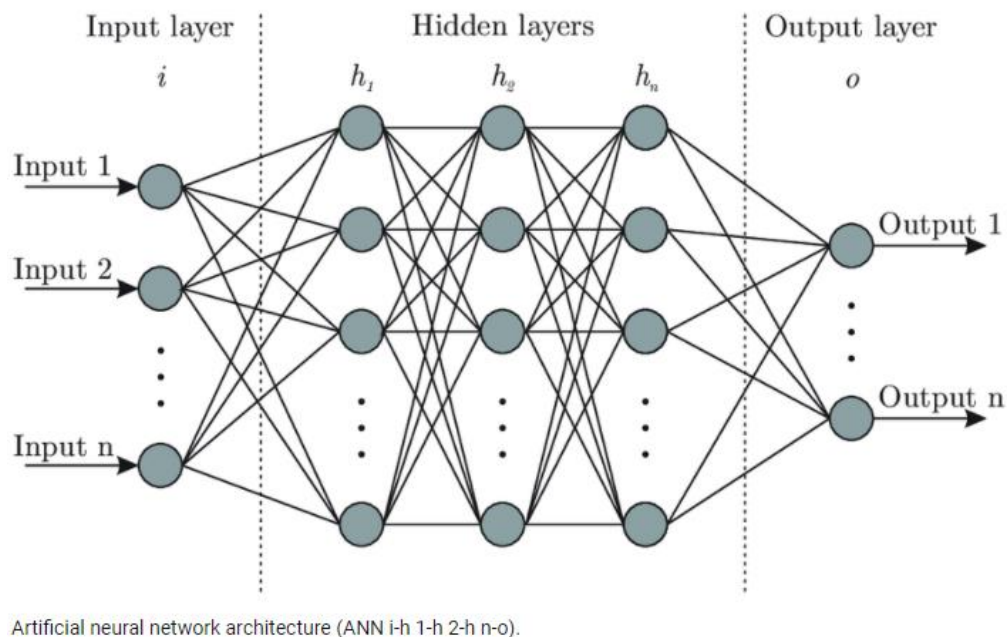
Τα νευρωνικά δίκτυα (Neural Nets) αποτελούν μια νέα περιοχή στις φυσικές επιστήμες, αφού είναι γνωστά στον επιστημονικό κόσμο τα τελευταία περίπου σαράντα (40) χρόνια. Ένα τεχνητό νευρωνικό δίκτυο (AAN), που συνήθως ονομάζεται νευρωνικό δίκτυο (NN), είναι ένα μαθηματικό μοντέλο που εμπνέεται από τη δομή και τις λειτουργικές πτυχές των βιολογικών νευρωνικών δικτύων.

Ένα νευρωνικό δίκτυο αποτελείται από μία διασυνδεδεμένη ομάδα τεχνητών νευρώνων και επεξεργάζεται πληροφορίες χρησιμοποιώντας μία προσέγγιση σύνδεσης με τον υπολογισμό. Στις περισσότερες περιπτώσεις, το ANN είναι ένα προσαρμοστικό σύστημα που αλλάζει τη δομή του με βάση εξωτερικές ή εσωτερικές πληροφορίες που ρέουν μέσω του δικτύου κατά τη φάση μάθησης. Τα σύγχρονα νευρωνικά δίκτυα χρησιμοποιούνται συνήθως για τη μοντελοποίηση πολύπλοκων σχέσεων μεταξύ εισόδων και εξόδων ή για την εύρεση μοτίβων στα δεδομένα που αναλύονται.

Ένα νευρωνικό δίκτυο τροφοδοσίας είναι ένα τεχνητό νευρωνικό δίκτυο όπου οι συνδέσεις μεταξύ των μονάδων δεν σχηματίζουν κατευθυνόμενο κύκλο. Σε αυτό το δίκτυο οι πληροφορίες μετακινούνται σε μία μόνο κατεύθυνση προς τα εμπρός, από τους κόμβους εισόδου, μέσω των κρυφών κόμβων στους κόμβους εξόδου. Αξιοσημείωτο είναι ότι δεν υπάρχουν κύκλοι ή βρόχοι στο δίκτυο.

Έχουμε και το multilayer perceptron (MLP) μοντέλο τεχνητού νευρωνικού δικτύου τροφοδοσίας που χαρτογραφεί σύνολα δεδομένων εισόδου σε ένα σύνολο εξόδου. Ένα MLP αποτελείται από πολλαπλά στρώματα κόμβων σε κατευθυνόμενο γράφημα, με κάθε επίπεδο να είναι πλήρως συνδεδεμένο με το επόμενο. Εκτός από τους κόμβους εισόδου, κάθε κόμβος είναι ένας νευρώνας με λειτουργία μη γραμμικής ενεργοποίησης. Το MLP χρησιμοποιεί πολλαπλασιασμό για την εκπαίδευση του

δικτύου και αυτή η κατηγορία δικτύων αποτελείται από πολλαπλά επίπεδα υπολογιστικών μονάδων.

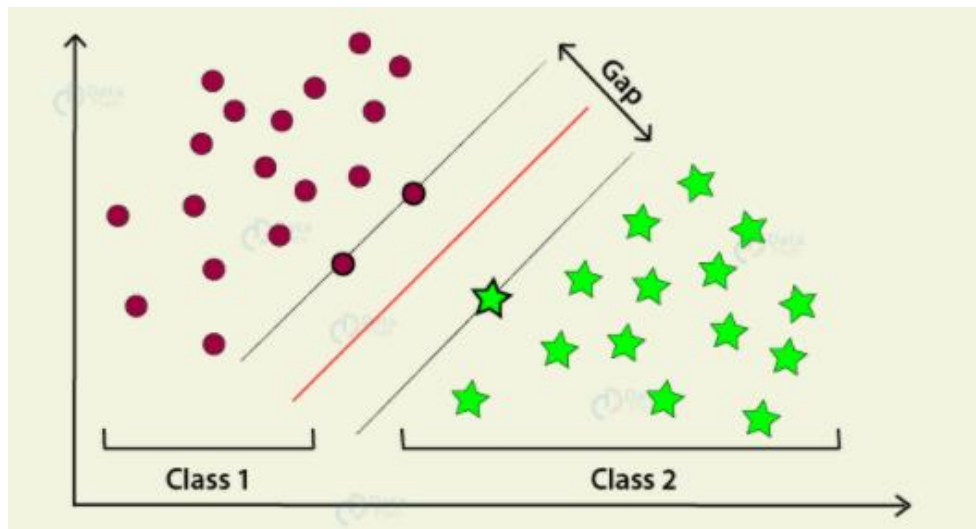


Σχήμα 7 Δομή νευρωνικού δικτύου

Τέλος είναι απαραίτητο να ειπωθεί πως το εύρος των τιμών των χαρακτηριστικών πρέπει να κλιμακωθεί σε -1 και +1. Αυτό μπορεί να γίνει μέσω της παραμέτρου κανονικοποίησης. Ο τύπος του κόμβου εξόδου είναι στιγμοειδής εάν τα δεδομένα μάθησης περιγράφουν μια εργασία ταξινόμησης και γραμμική εάν τα δεδομένα εκμάθησης περιγράφουν μια εργασία αριθμητικής παλινδρόμησης.

Support Vector Machine (SVM)

Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machine) αποτελούν ένα εργαλείο επεξεργασίας βασιζόμενο στη στατιστική θεωρία. Τα SVMs ανήκουν στην κατηγορία των μηχανών εκμάθησης και η κύρια λειτουργία τους είναι να μάθουν την αντιστοιχία εισόδων εξόδων με σκοπό την αντιστοίχιση μιας νέας εισόδου με την σωστή έξοδο. Η κατάσταση λειτουργίας μιας μηχανής, καθορίζει την τιμή που παίρνουν ορισμένα μεγέθη οπότε ανάλογα με αυτή την τιμή βγαίνουν και τα αντίστοιχα συμπεράσματα.



Σχήμα 8 Διαχωρισμός κλάσεων SVM

Αναλυτικότερα το πρότυπο SVM παίρνει ένα σύνολο δεδομένων εισόδου και προβλέπει για κάθε δεδομένη είσοδο ποια από τις δύο πιθανές κατηγορίες περιλαμβάνει την είσοδο. Λαμβάνοντας υπόψη ένα σύνολο παραδειγμάτων εκπαίδευσης, το καθένα επισημαίνεται ότι ανήκει σε μία από τις δύο κατηγορίες. Ένα μοντέλο SVM είναι μία αναπαράσταση των παραδειγμάτων ως σημεία χαρτογραφημένα έτσι ώστε τα παραδείγματα των ξεχωριστών κατηγοριών να διαιρούνται με ένα σαφές κενό όσο το δυνατόν μεγαλύτερο. Στην συνέχεια νέα παραδείγματα χαρτογραφούνται στον ίδιο χώρο και προβλέπεται σε ποια κατηγορία ανήκουν με βάση την πλευρά του κενού στην οποία βρίσκονται. Ένας αποτελεσματικός διαχωρισμός επιτυγχάνεται από το υπερπλάνο που έχει τη μεγαλύτερη απόσταση από τα πλησιέστερα σημεία δεδομένων εκπαίδευσης οποιασδήποτε κλάσης. Όσο μεγαλύτερο είναι το περιθώριο τόσο μικρότερο είναι το σφάλμα γενίκευσης του ταξινομητή. Για να διατηρηθεί το υπολογιστικό φορτίο λογικό, η αντιστοίχιση που χρησιμοποιείται από τα σχήματα SVM έχει σχεδιαστεί έτσι ώστε να διασφαλίζει ότι τα προϊόντα κουκκίδων μπορούν να υπολογιστούν εύκολα από την άποψη των μεταβλητών στον αρχικό χώρο, ορίζοντάς τα με όρους μιας λειτουργίας πυρήνα $K(x, y)$ που επιλέχθηκε να ταιριάζει με το πρόβλημα. Τα υπερπλάνα στο διάστημα υψηλότερης διάστασης ορίζονται ως το σύνολο σημείων των οποίων το εσωτερικό προϊόν με ένα φορέα σε αυτόν τον χώρο είναι σταθερό.

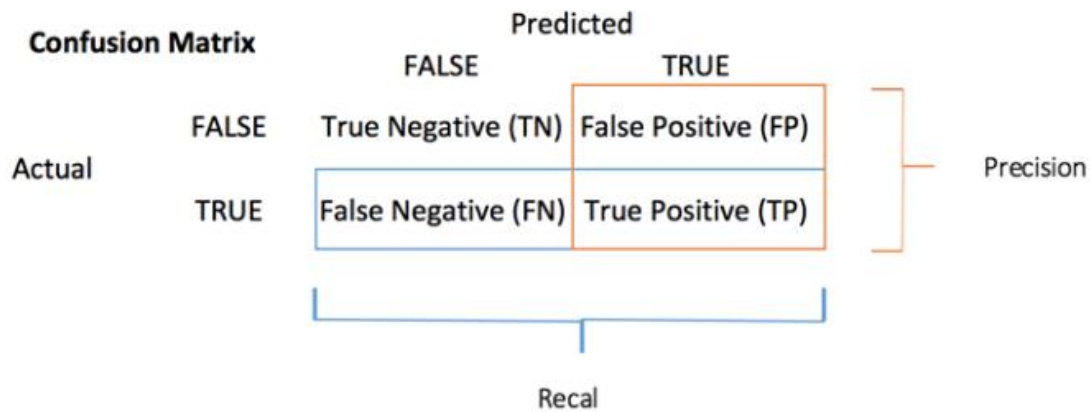
4. Πειραματικά Αποτελέσματα

4.1 Κριτήρια αξιολόγησης (Ακρίβεια, Ευαισθησία, Εξειδίκευση, Πίνακας Σύγχυσης)

Προκειμένου να διαπιστωθεί η αξιοπιστία μίας μεθόδου που έχει χρησιμοποιηθεί σε μία μελέτη ελέγχονται κάποια κριτήρια για την αποτελεσματικότητά της. Τα κριτήρια αυτά είναι η ακρίβεια, η ευαισθησία και η εξειδίκευση. Αναλυτικά :

- **Ακρίβεια (Accuracy)** : Αναφέρεται στην απόκλιση του αποτελέσματος από την πραγματική τιμή που υπάρχει στο δείγμα. Όσο μικρότερη είναι η τιμή της απόκλισης τόσο μεγαλύτερη είναι η τιμή της ακρίβειας.
- **Ευαισθησία (Sensitivity)** : Αναφέρεται στη μικρότερη συγκέντρωση ενός στοιχείου, όπου η μέθοδος που εφαρμόστηκε μπορεί να αντιδιαστείλει από το μηδέν.
- **Εξειδίκευση (Specificity)** : Αναφέρεται στη μέθοδο που εφαρμόζεται στο δείγμα, η οποία θα πρέπει να προσδιορίζει αποκλειστικά και μόνο το στοιχείο της καταμέτρησης.

Εκτός όμως από τα παραπάνω κριτήρια αξιολόγησης, εξίσου σημαντικός είναι και ο πίνακας σύγχυσης (Confusion Matrix) ο οποίος αποτυπώνει σε έναν πίνακα, το σύνολο των σωστών και το σύνολο των λανθασμένων ταξινομήσεων. Οι βέλτιστες λύσεις του αποτυπώνονται με μηδενικά περιμετρικά από την κύρια διαγώνιο του πίνακα. Η ταξινόμηση των στοιχείων συμβαίνει, είτε τα στοιχεία είναι αληθώς θετικά (True Positive - TP), είτε είναι ψευδώς αρνητικά (True Negative - TN). Επίσης πρέπει να αναφερθεί, ότι τα ψευδώς αρνητικά (False Negative – FN), καθώς και τα ψευδώς θετικά (False Positive – FP) αντιπροσωπεύουν τις λανθασμένες ταξινομήσεις για τον υπολογισμό του συνολικού σφάλματος. Παρακάτω ακολουθεί ένα παράδειγμα πίνακα σύγχυσης.



Σχήμα 9 Πίνακας σύγχυσης

Σύμφωνα με τις παραπάνω τιμές υπολογίζονται οι παρακάτω στατιστικές μετρήσεις. Οι μετρήσεις αυτές αποτελούν σημαντικό στοιχείο καθώς καθορίζουν τα αποτελέσματα του μοντέλου ταξινόμησης.

- **True Positive Rate ή Recall ή Sensitivity:** $\frac{TP}{TP + FN}$
- **True Negative Rate ή Specificity:** $\frac{TN}{TN + FP}$
- **Accuracy:** $\frac{TP + TN}{TP + TN + FP + FN}$
- **Precision:** $\frac{TP}{TP + FP}$
- **False Positive Rate:** $\frac{FP}{FP + TN}$

4.2 Αποτελέσματα σχήματος Naïve Bayes

Στο περιβάλλον του Rapid Miner επιλέχθηκε το σχήμα του Naïve Bayes το οποίο εφαρμόστηκε στα επεξεργασμένα δεδομένα μετά την επιλογή των χαρακτηριστικών. Η μέθοδος δοκιμής που επιλέχθηκε είναι η δεκαπλή διασταυρούμενη επικύρωση (10 Fold Cross Validation).

Στον ακόλουθο πίνακα φαίνεται το μοντέλο που παράχθηκε από αυτή την επεξεργασία:

Attribute	Parameter	Normal	Cancer
MZ25.49556	mean	0.668	0.354
MZ25.49556	standard deviation	0.165	0.134
MZ25.589892	mean	0.619	0.299
MZ25.589892	standard deviation	0.174	0.129
MZ221.86191	mean	0.671	0.320
MZ221.86191	standard deviation	0.183	0.182
MZ222.97536	mean	0.921	0.592
MZ222.97536	standard deviation	0.096	0.220
MZ223.25415	mean	0.825	0.481
MZ223.25415	standard deviation	0.154	0.204
MZ223.53313	mean	0.720	0.378
MZ223.53313	standard deviation	0.195	0.177
MZ223.81227	mean	0.639	0.297
MZ223.81227	standard deviation	0.225	0.157
MZ224.09159	mean	0.564	0.234
MZ224.09159	standard deviation	0.243	0.136
MZ244.36855	mean	0.494	0.123
MZ244.36855	standard deviation	0.193	0.118
MZ244.66041	mean	0.556	0.154
MZ244.66041	standard deviation	0.160	0.091
MZ244.95245	mean	0.694	0.210
MZ244.95245	standard deviation	0.159	0.102
MZ245.24466	mean	0.731	0.229
MZ245.24466	standard deviation	0.158	0.106
MZ245.53704	mean	0.671	0.206
MZ245.53704	standard deviation	0.158	0.097
MZ245.8296	mean	0.591	0.176
MZ245.8296	standard deviation	0.160	0.087
MZ246.12233	mean	0.518	0.146
MZ246.12233	standard deviation	0.163	0.077
MZ246.41524	mean	0.452	0.120
MZ246.41524	standard deviation	0.163	0.069
MZ246.70832	mean	0.413	0.107
MZ246.70832	standard deviation	0.164	0.063
MZ261.58446	mean	0.585	0.284
MZ261.58446	standard deviation	0.157	0.096
MZ261.88643	mean	0.589	0.286

MZ261.88643	standard deviation	0.143	0.091
MZ262.18857	mean	0.602	0.292
MZ262.18857	standard deviation	0.141	0.108
MZ262.49088	mean	0.623	0.314
MZ262.49088	standard deviation	0.155	0.135
MZ417.73207	mean	0.566	0.263
MZ417.73207	standard deviation	0.154	0.084
MZ434.68588	mean	0.292	0.609
MZ434.68588	standard deviation	0.096	0.136
MZ435.07512	mean	0.240	0.542
MZ435.07512	standard deviation	0.082	0.129
MZ7999.9616	mean	0.668	0.340
MZ7999.9616	standard deviation	0.221	0.182
MZ8001.6311	mean	0.660	0.332
MZ8001.6311	standard deviation	0.222	0.179
MZ8003.3008	mean	0.657	0.326
MZ8003.3008	standard deviation	0.225	0.177
MZ8004.9707	mean	0.655	0.322
MZ8004.9707	standard deviation	0.228	0.174
MZ8006.6407	mean	0.644	0.313
MZ8006.6407	standard deviation	0.228	0.170
MZ8008.3109	mean	0.637	0.305
MZ8008.3109	standard deviation	0.230	0.166
MZ8009.9813	mean	0.631	0.298
MZ8009.9813	standard deviation	0.233	0.162
MZ8011.6519	mean	0.619	0.289
MZ8011.6519	standard deviation	0.235	0.157

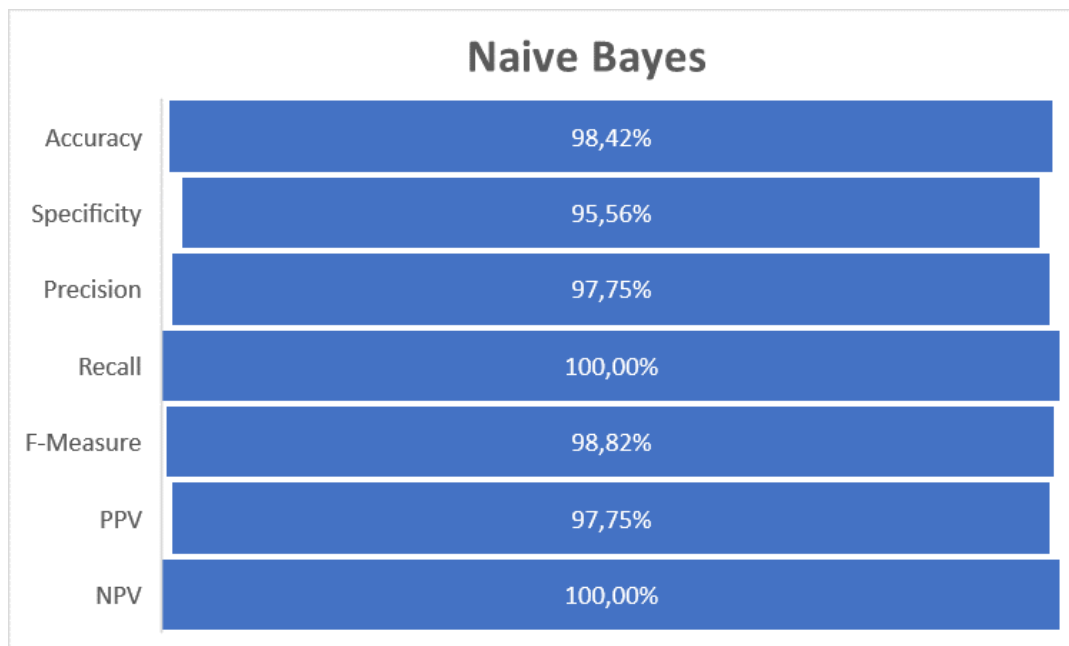
Πίνακας 10 Επιλεγμένες ιδιότητες απο το μοντέλο Naive Bayes

	True Normal	True Cancer
Normal	87	0
Cancer	4	162

Πίνακας 11 Πίνακας σύγχυσης

Recall	Precision	Class
95,50%	100%	Normal
100%	97,59%	Cancer

Πίνακας 12 Τιμές ανά κλάση



Σχήμα 10 Μετρικές Naive Bayes

4.3 Αποτελέσματα Random Forest

Στο περιβάλλον του Rapid Miner επιλέχθηκε ο αλγόριθμος Random Forest, ο οποίος εφαρμόστηκε στα επεξεργασμένα δεδομένα μετά την επιλογή των χαρακτηριστικών. Η μέθοδος δοκιμής που επιλέχθηκε είναι η δεκαπλή διασταυρούμενη επικύρωση (10 Fold Cross Validation).

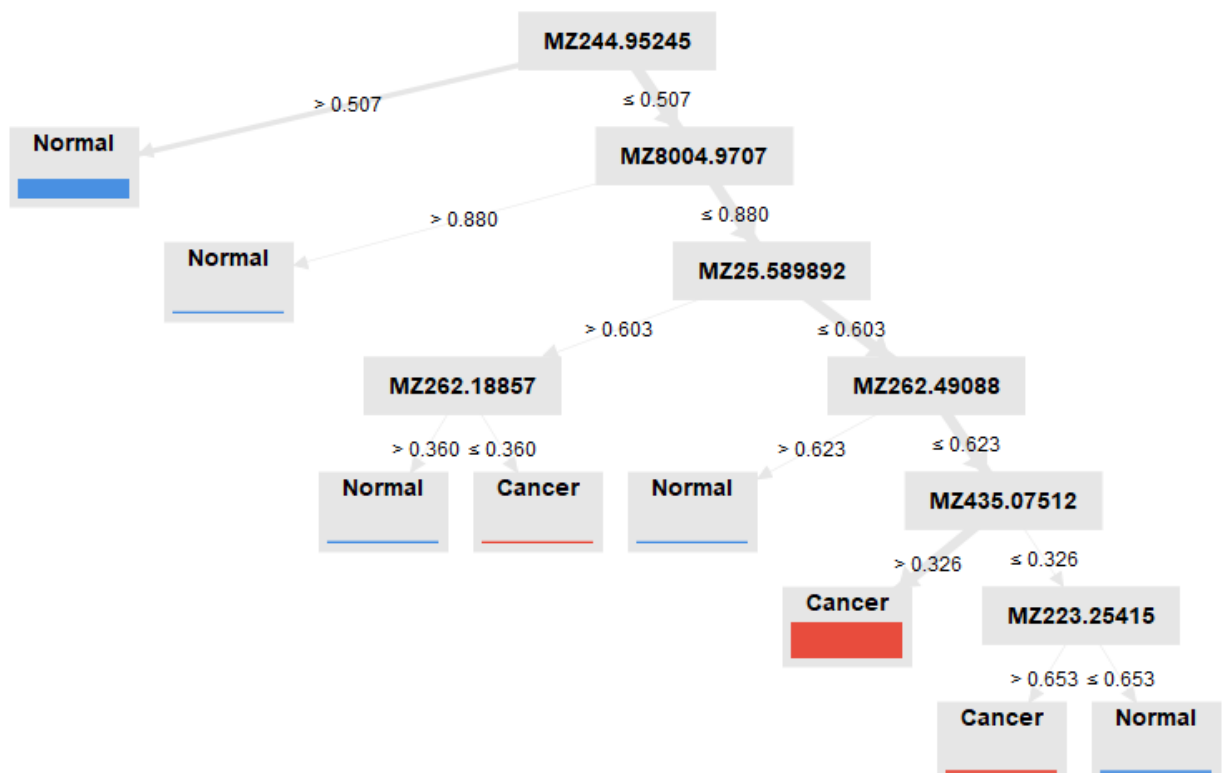
Παρακάτω φαίνονται τα αποτελέσματα του αλγορίθμου καθώς και το επικρατέστερο δέντρο απόφασης που δημιουργήθηκε.

	True Normal	True Cancer
Normal	88	2
Cancer	3	160

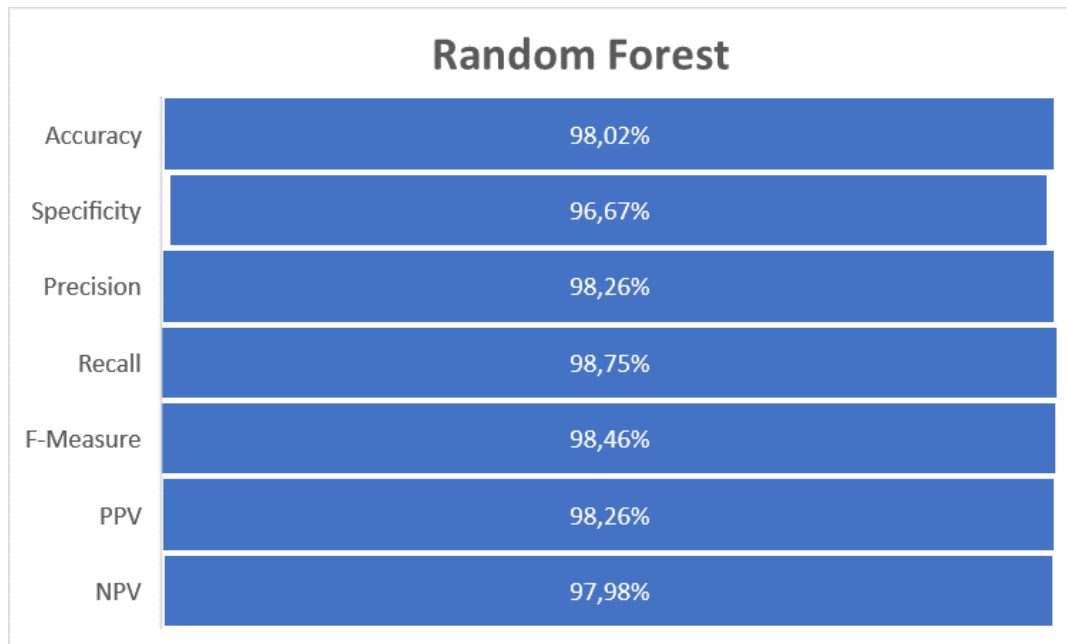
Πίνακας 13 Πίνακας σύγχυσης

Recall	Precision	Class
96,70%	98%	Normal
99%	98,16%	Cancer

Πίνακας 14 Τιμές ανά κλάση



Σχήμα 11 Ενδεικτικό δέντρο απόφασης Random Forest



Σχήμα 12 Μετρικές Random Forest

4.4 Αποτελέσματα Neural Net

Στο περιβάλλον του Rapid Miner επιλέχθηκε το νευρωνικό δίκτυο (Neural Net – MLP), το οποίο εφαρμόστηκε στα επεξεργασμένα δεδομένα μετά την επιλογή των χαρακτηριστικών. Η μέθοδος δοκιμής που επιλέχθηκε είναι η δεκαπλή διασταυρούμενη επικύρωση (10 Fold Cross Validation).

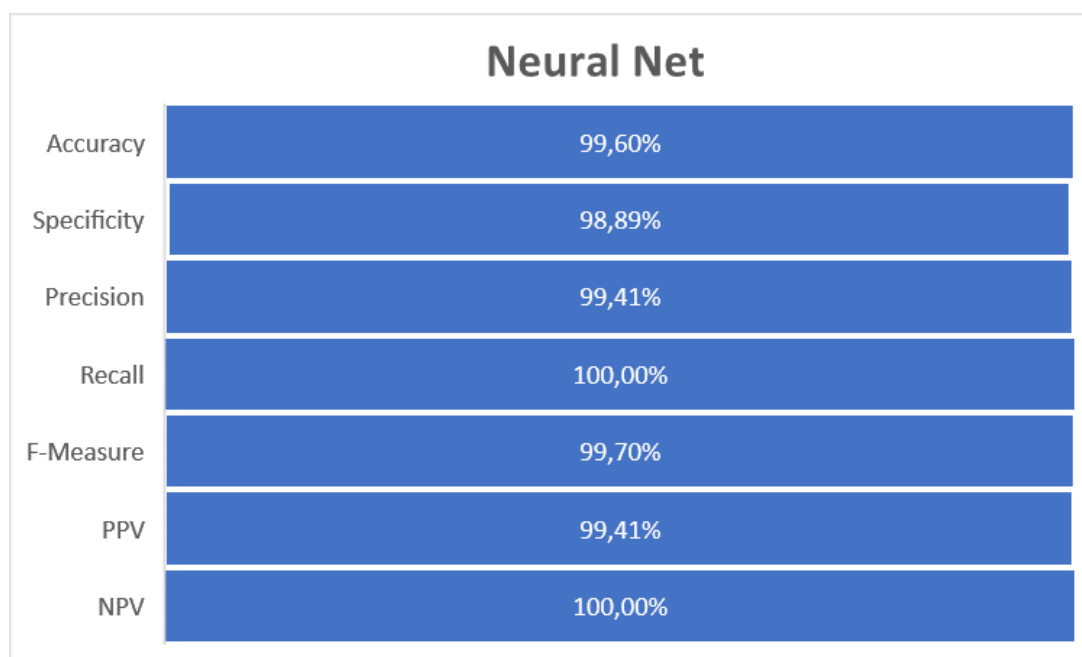
Παρακάτω φαίνονται τα αποτελέσματα του νευρωνικού δικτύου καθώς και η απεικόνιση των layers όπως αυτή κατασκευάστηκε κατά την δοκιμή του.

	True Normal	True Cancer
Normal	90	0
Cancer	1	162

Πίνακας 15 Πίνακας σύγχυσης

Recall	Precision	Class
98,90%	100%	Normal
100%	99,39%	Cancer

Πίνακας 16 Τιμές ανά κλάση



Σχήμα 13 Μετρικές Neural Net

4.5 Αποτελέσματα Support Vector Machine

Στο περιβάλλον του Rapid Miner επιλέχθηκε η μηχανή διανυσμάτων υποστήριξης (Support Vector Machine) η οποία εφαρμόστηκε στα επεξεργασμένα δεδομένα μετά την επιλογή των χαρακτηριστικών. Η μέθοδος δοκιμής που επιλέχθηκε είναι η δεκαπλή διασταυρούμενη επικύρωση (10 Fold Cross Validation).

Παρακάτω φαίνονται τα αποτελέσματα του Support Vector Machine καθώς και το διάγραμμα που κατασκευάστηκε με τον διαχωρισμό των χαρακτηριστικών.

	True Normal	True Cancer
Normal	89	0
Cancer	2	162

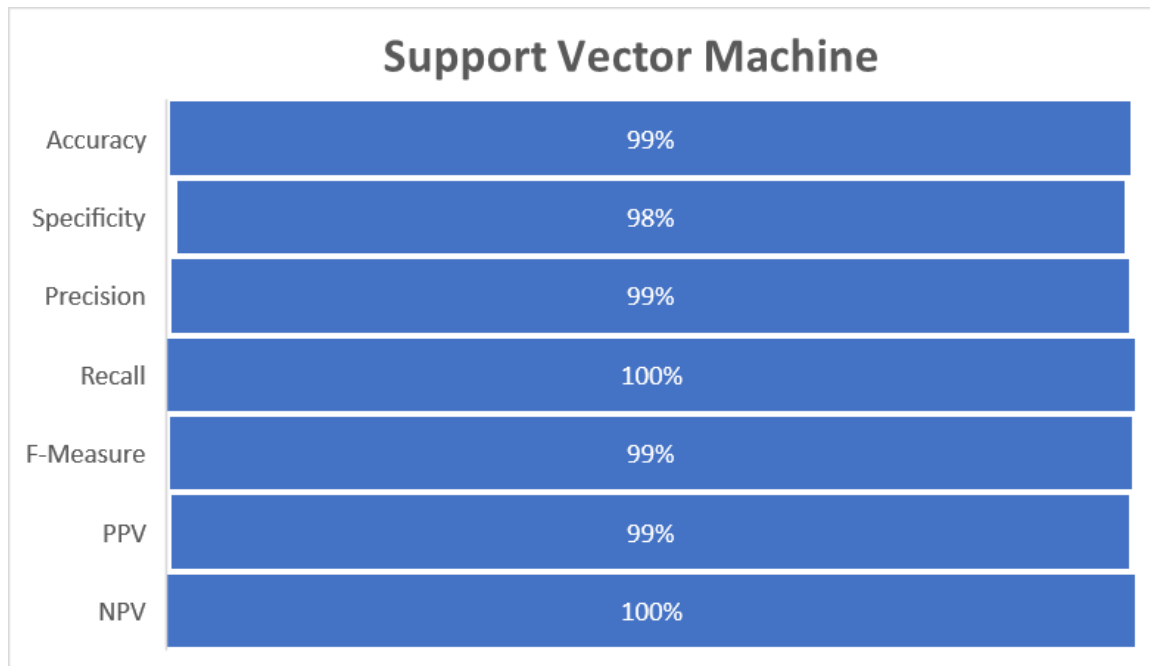
Πίνακας 17 Πίνακας σύγχυσης

Recall	Precision	Class
97,80%	100%	Normal
100%	98,78%	Cancer

Πίνακας 18 Τιμές ανά κλάση



Σχήμα 14 Διαχωρισμός δειγμάτων SVM



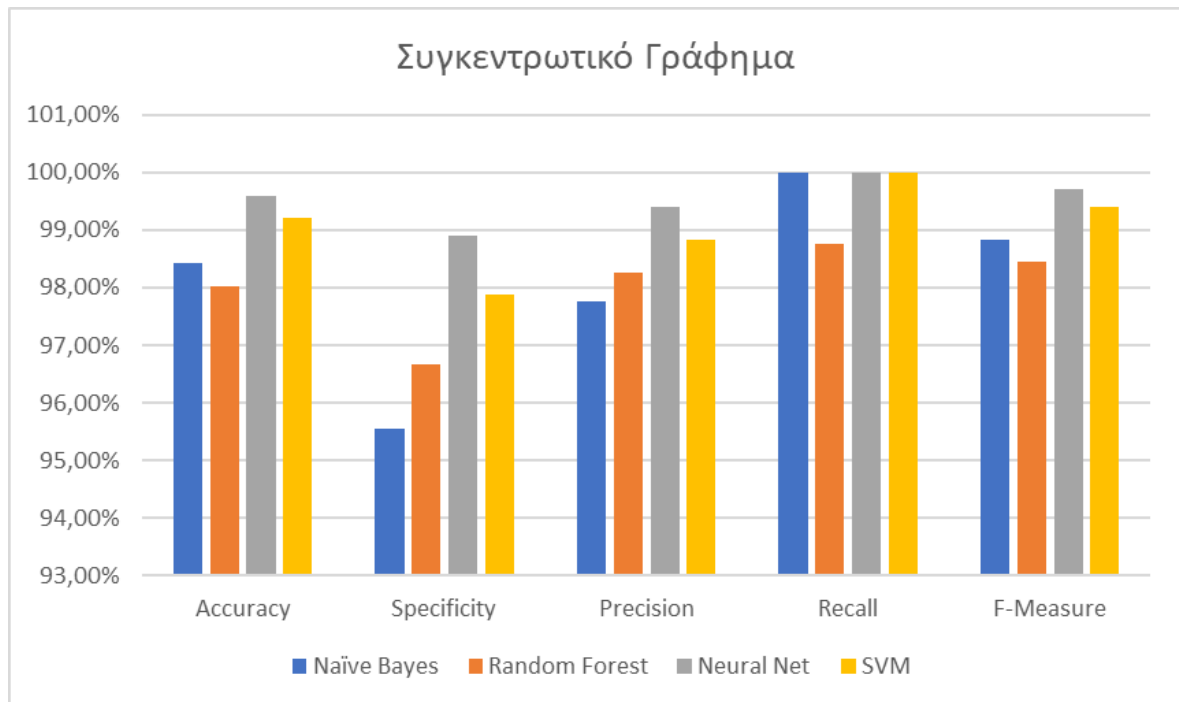
Σχήμα 15 Μετρικές SVM

5. Συμπεράσματα

Ανακεφαλαιώνοντας σκόπιμο είναι να δούμε συγκεντρωτικά τις αποδόσεις όλων των μεθόδων που χρησιμοποιήθηκαν στην μελέτη, ώστε να μπορέσει να γίνει η σύγκριση των αποτελεσμάτων. Ακολουθεί ένας πίνακας με τις αποδόσεις όλων των αλγορίθμων καθώς και ένα γράφημα για την οπτικοποίηση των αποτελεσμάτων.

	Accuracy	Sensitivity	Specificity	Precision	F-Measure
Naïve Bayes	98,42%	100%	95,56%	97,75%	98,82%
Random Forest	98,02%	98,75%	96,67%	98,26%	98,46%
Neural Net	99,60%	100%	98,89%	99,41%	99,70%
SVM	99%	100%	97,89%	98,82%	99,39%

Πίνακας 19 Συγκεντρωτικός πίνακας αποδόσεων των αλγορίθμων



Σχήμα 16 Συγκεντρωτικό γράφημα αποδόσεων

Παρατηρώντας τον πίνακα των αποτελεσμάτων σε συνδυασμό με το γράφημα μπορούμε να συμπεράνουμε τα εξής. Το νευρωνικό δίκτυο πετυχαίνει την μεγαλύτερη ακρίβεια (Accuracy) 99.6% με δεύτερο τον SVM 99%. Οι Random Forest και Naïve Bayes ακολουθούν με εξίσου υψηλά ποσοστά 98,02% και 98,42 % αντίστοιχα. Επίσης το Sensitivity κυμαίνεται σε υψηλά ποσοστά όπως βλέπουμε με τις τρεις από τις τέσσερις μεθόδους να πετυχαίνουν το μέγιστο 100%. Ακόμη το Specificity όλων των μεθόδων κυμαίνεται σε ποσοστά μεγαλύτερα του 95 %. Τα παραπάνω ποσοστά επιβεβαιώνονται αν δούμε τον συγκεντρωτικό πίνακα σύγχυσης (Confusion Matrix) όλων των αλγορίθμων που μελετήθηκαν.

	Naïve Bayes		Random Forest		Neural Net		SVM	
	True Normal	True Cancer	True Normal	True Cancer	True Normal	True Cancer	True Normal	True Cancer
Normal	87	0	88	2	90	0	89	0
Cancer	4	162	3	160	1	162	2	162

Πίνακας 20 Συγκεντρωτικός πίνακας σύγχυσης

Στην τρέχουσα εργασία, η επεξεργασία των βιολογικών δεδομένων ανέδειξε ένα περιορισμένο υποσύνολο δεικτών. Τόσο σε μελέτες που μελετήθηκαν όσο και στην συγκεκριμένη οι δείκτες που επεξεργάζονται και μελετώνται είναι περιορισμένου μεγέθους, παρά τους διαφορετικούς τρόπους επεξεργασίας και ανάλυσης των δεδομένων. Τα δεδομένα που εξετάστηκαν περιέχουν μετρήσεις M/Z και ο αριθμός των μετρήσεων που διαχωρίζει τις περιπτώσεις ασθενών με καρκίνο των ωοθηκών είναι πολύ μικρός. Η παρατήρηση αυτή μας οδηγεί στο συμπέρασμα ότι πρέπει να τηρούνται αυστηροί κανόνες κατά τη διαδικασία λήψης των δειγμάτων.

Σαφές πρέπει να γίνει ότι κρίνεται απαραίτητη η συνεργασία με βιολόγους και άλλους επιστήμονες της υγείας, ώστε να αντιστοιχηθούν οι μετρήσεις φασματομετρίας με γονίδια. Καθώς χρειάζεται και η χρήση ειδικού αλγορίθμου αντιστοιχίας των μετρήσεων ώστε να βρεθούν τα γονίδια από τα οποία προέρχονται.

Τέλος, μελλοντικός σκοπός της παρούσας μελέτης είναι να βρεθεί ένα αξιόπιστο μοντέλο για ασθενείς με καρκίνο των ωοθηκών, το οποίο θα περιγράφει με εξαιρετική ευαισθησία τις αλληλεπιδράσεις των γονιδίων του καρκίνου. Επίσης, το μελλοντικό μοντέλο θα είναι ικανό να αναγνωρίσει τα μοτίβα των γονιδίων που εμπλέκονται στην έγκαιρη διάγνωση του καρκίνου των ωοθηκών στον γυναικείο πληθυσμό. Αυτό αποτελεί σημαντικό επιστημονικό άλμα καθώς η έγκαιρη διάγνωση αλλά και θεραπεία των ασθενών μπορεί να μειώσει δραματικά την θνησιμότητα του συγκεκριμένου καρκίνου.

6. Βιβλιογραφία

- [1] "(PDF) Learning Causal Models for Noisy Biological Data Mining: An Application to Ovarian Cancer Detection." https://www.researchgate.net/publication/221603093_Learning_Causal_Models_for_Noisy_Biological_Data_Mining_An_Application_to_Ovarian_Cancer_Detection (accessed Dec. 14, 2020).
- [2] "A comparative study of different machine learning methods on microarray gene expression data | SpringerLink." <https://link.springer.com/article/10.1186/1471-2164-9-S1-S13> (accessed Dec. 04, 2020).
- [3] R. Tabares-Soto, S. Orozco-Arias, V. Romero-Cano, V. S. Bucheli, J. L. Rodríguez-Sotelo, and C. F. Jiménez-Varón, "A comparative study of machine learning and deep

learning algorithms to classify cancer types based on microarray gene expression data,” *PeerJ Comput. Sci.*, vol. 6, p. e270, Apr. 2020, doi: [10.7717/peerj-cs.270](https://doi.org/10.7717/peerj-cs.270).

[4] J. M. Sorace and M. Zhan, “A data review and re-assessment of ovarian cancer serum proteomic profiling,” *BMC Bioinformatics*, vol. 4, no. 1, p. 24, Jun. 2003, doi: [10.1186/1471-2105-4-24](https://doi.org/10.1186/1471-2105-4-24).

[5] Z.-J. Lee, “An integrated algorithm for gene selection and classification applied to microarray data of ovarian cancer,” *Artificial Intelligence in Medicine*, vol. 42, no. 1, pp. 81–93, Jan. 2008, doi: [10.1016/j.artmed.2007.09.004](https://doi.org/10.1016/j.artmed.2007.09.004).

[6] “An integrated network of microRNA and gene expression in ovarian cancer | SpringerLink.” <https://link.springer.com/article/10.1186/1471-2105-16-S5-S5> (accessed Dec. 04, 2020).

[7] P. Yasodha and N. R. Ananthanarayanan, “Analysing Big Data to Build Knowledge Based System for Early Detection of Ovarian Cancer,” *Indian Journal of Science and Technology*, vol. 8, no. 14, Jul. 2015, doi: [10.17485/ijst/2015/v8i14/65745](https://doi.org/10.17485/ijst/2015/v8i14/65745).

[8] Jin-Tsong Jeng, Tsu-Tian Lee, and Yung-Cheng Lee, “Classification of Ovarian Cancer based on Intelligent Systems with Microarray Data,” in *2005 IEEE International Conference on Systems, Man and Cybernetics*, Oct. 2005, vol. 2, pp. 1053–1058, doi: [10.1109/ICSMC.2005.1571285](https://doi.org/10.1109/ICSMC.2005.1571285).

[9] C. Coveney, D. J. Boockock, R. C. Rees, S. Deen, and G. R. Ball, “Data Mining of Gene Arrays for Biomarkers of Survival in Ovarian Cancer,” *Microarrays (Basel)*, vol. 4, no. 3, pp. 324–338, Jul. 2015, doi: [10.3390/microarrays4030324](https://doi.org/10.3390/microarrays4030324).

[10] Md. A. Hossain, S. M. Saiful Islam, J. M. W. Quinn, F. Huq, and M. A. Moni, “Machine learning and bioinformatics models to identify gene expression patterns of ovarian cancer associated with disease progression and mortality,” *Journal of Biomedical Informatics*, vol. 100, p. 103313, Dec. 2019, doi: [10.1016/j.jbi.2019.103313](https://doi.org/10.1016/j.jbi.2019.103313).

[11] O. Klein *et al.*, “MALDI-Imaging for Classification of Epithelial Ovarian Cancer Histotypes from a Tissue Microarray Using Machine Learning Methods,” *PROTEOMICS – Clinical Applications*, vol. 13, no. 1, p. 1700181, 2019, doi: <https://doi.org/10.1002/prca.201700181>.

[12] S. H. Bouazza, K. Auhmani, A. Zeroual, and N. Hamdi, “Selecting significant marker genes from microarray data by filter approach for cancer diagnosis,” *Procedia Computer Science*, vol. 127, pp. 300–309, Jan. 2018, doi: [10.1016/j.procs.2018.01.126](https://doi.org/10.1016/j.procs.2018.01.126).

[13] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, “Support vector machine classification and validation of cancer tissue samples using microarray expression data,” *Bioinformatics*, vol. 16, no. 10, pp. 906–914, Oct. 2000, doi: [10.1093/bioinformatics/16.10.906](https://doi.org/10.1093/bioinformatics/16.10.906).

[14] J.-H. Hong and S.-B. Cho, "The classification of cancer based on DNA microarray data that uses diverse ensemble genetic programming," *Artificial Intelligence in Medicine*, vol. 36, no. 1, pp. 43–58, Jan. 2006, doi: [10.1016/j.artmed.2005.06.002](https://doi.org/10.1016/j.artmed.2005.06.002).