Ελληνικό Μεσογειακό Πανεπιστήμιο

Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Πτυχιακή εργασία

Ποσοτικοποίηση σήματος από ιστοπαθολογικές εικόνες φθορισμού και εφαρμογές μηχανικής μάθησης για την κατηγοριοποίηση τους

Στυλιανός Παπαγιαννάκης (4315)

Επιβλέπων εκπαιδευτικός: Κωνσταντίνος Μαριάς

Επιτροπή αξιολόγησης:

Ημερομηνία παρουσίασης:

Περίληψη

Η ποσοτικοποίηση καθώς και η κατηγοριοποίηση των εικόνων αποτελεί πρόκληση ακόμη και σήμερα λόγω της δυσχέρειας των ήδη διαθέσιμων λογισμικών. Η κλινικοί οφείλουν να επιλέγουν και να σχεδιάζουν χειροκίνητα τις περιοχές ενδιαφέροντος σε κάθε εικόνα ξεχωριστά, γεγονός που οδηγεί σε λανθασμένα αποτλέσματα ή ακόμη και σε εσφαλμένα συμπεράσματα. Επίσης, κατά την λήψη μίας απόφασης η οποία αφορά την διάγνωση ή την θεραπεία κάποιου ασθενή προέρχεται αποκλειστικά από την κρίση τους εκάστοτε ειδικού, χωρίς κάποιο εργαλείο επιβεβαίωσης ή υποβοήθειας στην λήψη της απόφασης τους. Λαμβάνοντας υπόψιν τις παραπάνω συνθήκες που επικρατούν ακόμη και σήμερα και εκμεταλλευόμενοι τις δυνατότητες και τα εργαλεία που μας προσφέρει η τεχνολογία, αποφασίστηκε ο σκοπός αυτής της διατριβής να είναι η ανάπτυξη και η αξιολόγηση μίας αυτοματοποιημένης διαδικασίας η οποία στηρίζεται σε τεχνικές επεξεργασίας εικόνας μέσω της οποίας θα πραγματοποιείται ποσοτικοποίηση καρκινικών ή μη κυττάρων καθώς και του κυτταροπλάσματός τους. Για να την υλοποίηση της αυτοματοποιημένης αυτής διαδικασίας έγινε χρήση σε μεγάλο βαθμό των τεχνικών της επεξεργασίας εικόνας καθώς και της στατιστικής . Πιο συγκεκριμένα, έγινε η κατάλληλη προεπεργασία της εικόνας με τον διαχωρισμό αυτής σε κανάλια, την εφαρμογή του φίλτρου μεσαίας τιμής ώστε να εξαλλειφθεί ο πιθανός θόρυβος από κάθε εικόνα ξεχωριστά με στόχο το βέλτιστο αποτέλεσμα, εφαρμόστηκε κατωφλίωση της εικόνας με τη μεθόδου Otsu ώστε να γίνει αυτοματοποιημένος διαχωρισμός των στοχείων της εικόνας χωρίς να χρειάζεται πλέον ο χειροκίνητος σχεδιασμός αυτών εκμηδενίζοντας έτσι το περιθώριο λάθους, και τέλος έγινε χρήση των στατιστικών χαρακτηριστικών όπως είναι για παράδειγμα η μεσαία τιμή και η διακύμανση ώστε να ποσοτικοποιηθεί το σύνολο δεδομένων. Στη συνέχεια, για την στήριξη των κλινικών στην λήψη αποφάσεων, τέθηκε ο στόχος του σχεδιασμού ενός μοντέλου κατηγοριοποίησης σε καρκινικά και φυσιολογικά κύτταρα μέσω του οποίου θα εξάγονται και θα αξιολογούνται τα χαρακτηριστικά από εικόνες ιστοπαθολογίας (εικονες ανοσοφθορισμού) όπως για παράδειγμα, η υφή και το σχήμα των αντικειμένων που απεικονίζονται σε αυτές με βάση την σημαντικότητα τους για την υποβοήθεια των ειδικών στη λήψη των αποφάσεών τους. Για την εξαγωγή των χαρακτηριστικών αυτών, έγινε χρήση της βιβλιοθήκης Pyradiomics. Πιο αναλυτικά, πρώτο βήμα αυτής της υλοποίησης ήταν ο διαχωρισμός των καρκινικών κυττάρων από των μη καρκινικών. Πιο συγκεκριμένα, αφού έγινε η κατάλληλη επεξεργασία του συνόλου δεδομένων, ακολούθησε η διαχωρισμός του συνόλου δεδομένων σε σε σύνολο δεδομένων και σύνολο εκπαίδευσης, έγινε κανονικοποίηση αυτού και εφαρμόστηκε ο αλγόριθμμος Kbest για την επιλογή των K πιο σημαντικών χαρακτηριστικών. Ακόμη, έγινε εφαρμογή των μηχανών διανυσμάτων υποστήριξης (support vector machines) όπως και της λογιστικής παλινδρόμησης (logistic regression) για την ταξινόμηση των δεδομένων σε κατηγορίες με βάση των χαρακτηριστικών που επιλέχθησαν. Τέλος, έγινε εφαρμογή του k-fold cross για τον καθορισμό της ακριβείας του μοντέλου και υπολογίστηκε ο πίνακας σύγχυσης (confusion matrix) που χάρη σε αυτόν υπολογίστηκαν και οι μετρικές για την επίδοση του μοντέλου (area under the curve, accuracy κλπ.).

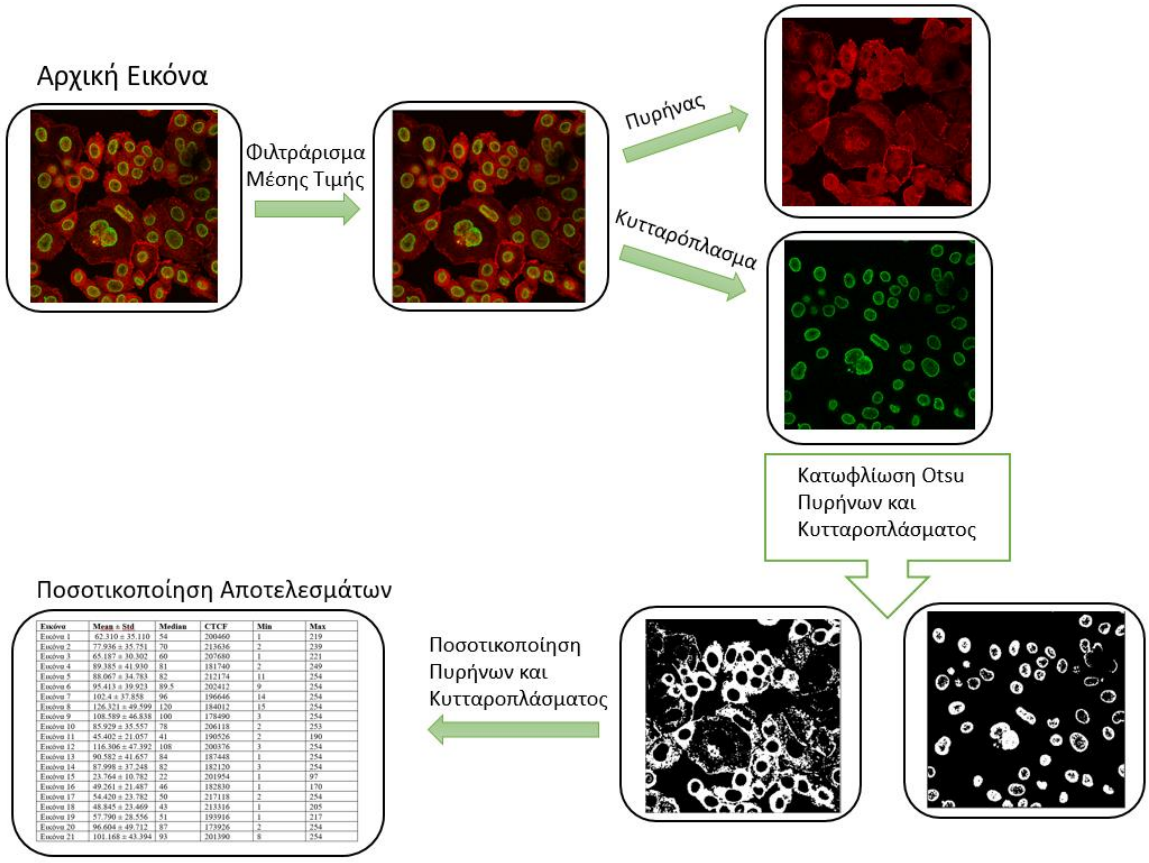
Πίνακας Περιεχομένων

1. Εισαγωγή

Η παρούσα πτυχιακή εργασία πραγματεύεται θέματα σχετικά με την ψηφιακή επεξεργασία εικόνας αλλά και την δημιουργία μοντέλων μηχανικής μάθησης στον τομέα της βιοϊατρικής και πιο συγκεκριμένα της ιστοπαθολογίας. Στο πρώτο τμήμα της εργασίας κατασκευάστηκε ένα μοντέλο προεπεξεργασίας εικόνας σε ένα σύνολο δεδομένων φθορισμού από εικόνες ιστοπαθολογίας με στόχο την ποσοτικοποίηση τους. Πιο συγκεκριμένα, σε αυτήν την εργασία θέλουμε να ποσοτικοποιήσουμε το ποσοστό χρώσης (ένταση σήματος εικόνας) που βρίσκεται στον πυρήνα και στο κυτταρόπλασμα με απώτερο σκοπό την εξαγωγή στατιστικών χαρακτηριστικών για κάθε μια από τις προαναφερθείσες ομάδες (πυρήνες και κυτταρόπλασμα). Για τον διαχωρισμό αυτών των δύο, έχει εφαρμοστεί η τεχνική του ανοσοφθορισμού η οποία σε συνδυασμό με την τεχνική του ανοσοκυτταροχημείας και της ανοσόχρωσης μας δίνουν την δυνατότητα να «χρωματίσουμε» των πυρήνα με πράσινο χρώμα και το κυτταρόπλασμα με κόκκινο. Στο σημείο αυτό αξίζει να σημειωθεί ότι η διαδικασία της ανοσοκυτταροχημείας αποτελεί μία χρονοβόρα διαδικασία για τους ειδικούς καθώς υπάρχει έλλειψη αυτοματοποιημένων λογισμικών με αποτέλεσμα να καλούνται να σχεδιάσουν χειροκίνητα και σε κάθε εικόνα ξεχωριστά την περιοχή ενδιαφέροντος την οποία καλούνται να εξετάσουν δημιουργώντας πλήν των άλλων ένα σημαντικό περιθώριο λάθους. Επίσης, λόγω της πολυπλοκότητας των εικόνων προς εξέταση δημιουργούνται επιπλέον δυσκολίες καθώς δεν είναι δυνατή η ακριβής προσέγγιση των τιμών των εντάσεων τους καθώς εκτός την έλλειψη λογισμικών, τα ήδη υπάρχοντα παρουσιάζουν πολυπλοκότητα ως προς την χρήση και τον χειρισμό τους.

Για τους παραπάνω λόγους δημιουργήθηκε ένα αυτοματοποιημένο εργαλείο που εξάγει στατιστικά χαρακτηριστικά από το ποσοστό της χρώσης που βρίσκεται στον πυρήνα και στο κυτταρόπλασμα με τα παρακάτω βήματα. Αρχικά, σε κάθε εικόνα εφαρμόστηκε το φίλτρο μεσαίας τιμής (media filter) με στόχο την απομάκρυνση θορύβου.

Λαμβάνοντας υπόψιν το γεγονός ότι οι εικόνες είχαν 3 χρωματικά κανάλια (κόκκινο, πράσινο και μπλε) για τον διαχωρισμό του πυρήνα από το κυτταρόπλασμα έγινε κατωφλίωση της εικόνας με τη μεθόδου Otsu σε κάθε χρωματικό κανάλι ξεχωριστά. Έτσι, μειώθηκε σημαντικά το περιθώριο λάθους που παρουσιαζόταν έως τώρα από τους ειδικούς καθώς η διαδικασία αυτή δεν ήταν αυτοματοποιημένη και η σημαντική πληροφορία της περιοχής ενδιαφέροντος μπορεί να παρέμενε εκτός του υπολογισμού ή ακόμη μπορεί να λάμβανε μέρος στον υπολογισμό κάποια πληροφορία εκτός της περιοχής αυτής οδηγώντας σε λάθος αποτελέσματα και συμπεράσματα. Τέλος αφού έγινε διαχωρισμός των στοιχείων αυτών, χρησιμοποιήσαμε στατιστικά χαρακτηριστικά όπως είναι για παράδειγμα η μεσαία τιμή, ο μέσος όρος και η διακύμανση με τα οποία έγινε η ποσοτικοποίηση του συνόλου δεδομένων παρέχοντας έτσι στους ειδικούς όλα τα απαραίτητα στοιχεία και εργαλεία για την εξαγωγή συμπερασμάτων από την έρευνα την οποία πραγματοποιούν τα οποία θα οδηγήσουν στην αποτελεσματική θεραπεία των ασθενών τους καθώς και στην πρόληψη αυτών. Στην εικόνα 1 παρουσιάζεται η ροή των βημάτων που χρησιμοποιήθηκαν για την ποσοτικοποίηση των εικόνων.

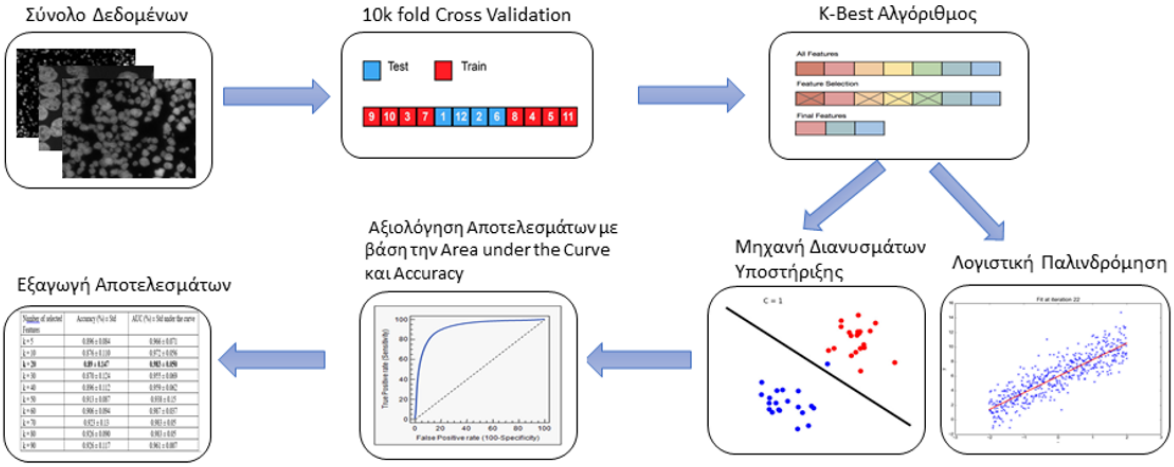


Εικόνα	Mean ± Std	Median	CTCF	Min	Max
Εικόνα 1	62.310 ± 35.110	54	200460	1	219
Εικόνα 2	77.936 ± 35.751	70	213636	2	239
Εικόνα 3	65.187 ± 30.302	60	207680	1	221
Εικόνα 4	89.385 ± 41.930	81	181740	2	249
Εικόνα 5	88.067 ± 34.783	82	212174	11	254
Εικόνα 6	95.413 ± 39.923	89.5	202412	9	254
Εικόνα 7	102.4 ± 37.858	96	196646	14	254
Εικόνα 8	126.321 ± 49.599	120	184012	15	254
Εικόνα 9	108.589 ± 46.838	100	178490	3	254
Εικόνα 10	85.929 ± 35.557	78	206118	2	253
Εικόνα 11	45.402 ± 21.057	41	190526	2	190
Εικόνα 12	116.506 ± 47.392	108	200376	3	254
Εικόνα 13	90.582 ± 41.657	84	187448	1	254
Εικόνα 14	87.998 ± 37.248	82	182120	3	254
Εικόνα 15	23.764 ± 10.782	22	201954	1	97
Εικόνα 16	49.261 ± 21.487	46	182830	1	170
Εικόνα 17	54.420 ± 23.782	50	217118	2	254
Εικόνα 18	48.845 ± 23.469	43	213316	1	205
Εικόνα 19	57.790 ± 28.556	51	193916	1	217
Εικόνα 20	96.604 ± 49.712	87	173926	2	254
Εικόνα 21	101.168 ± 43.394	93	201390	8	254

Εικόνα 1 : Ροή Εργασίας για την Ποσοτικοποίηση.

Όπως προαναφέρθηκε, το δεύτερο σκέλος της εργασίας, επικεντρώνεται στην αυτόματη κατηγοριοποίηση ενός συνόλου εικόνων φθορισμού καθώς και στην εξαγωγή και επιλογή των πιο απαραίτητων χαρακτηριστικών από το σύνολο αυτό με τη χρήση της ευρέως χρησιμοποιούμενης βιβλιοθήκης pyradiomics. Με τη βοήθεια της βιβλιοθήκης αυτής δίνεται η δυνατότητα εξαγωγής πληθώρας χαρακτηριστικών (στατιστικά χαρακτηριστικά πρώτης και δεύτερης τάξης, χαρακτηριστικά σχήματος ακόμη και υφής) από το σύνολο των εικόνων. Με την χρήση της βιβλιοθήκης αυτής καταφέραμε να εξάγουμε χαρακτηριστικά τέτοιου τύπου από τις εικόνες αυτές. Στη συνέχεια, έγινε προ-επεξεργασία του αρχείου csv το οποίο περιείχε όλα τα χαρακτηριστικά που διεξήχθησαν από τις εικόνες. Πρωταρχικός στόχος ήταν ο διαχωρισμός των καρκινικών κυττάρων από των μη καρκινικών έτσι έγινε διαχωρισμός των καρικινικών από των μη καρκινικών υποθέσεων σε 0 και 1. Έπειτα, πραγματοποιήθηκε διαχωρισμός του συνόλου δεδομένων σε σύνολο δεδομένων και σύνολο εκπαίδευσης με ποσοστό 80% - 20% καθώς και κανονικοποίηση του συνόλου των τιμων με τη χρήση εργαλείων της βιβλιοθήκης sklearn για την αποφυγή της υπερπροσαρμογής (overfitting) των δεδομένων. Η περίπτωση δηλαδή που το μονέλο έχει υπερεκπαιδευτεί και εξάγει απόλυτα σωστά συμπεράσματα με αδυναμία να γενικεύσει και να πετύχει ακριβή

αποτελέσματα σε ένα νέο σύνολο δεδομένων. Αφού έγινε η κατάλληλη προεπεξεργασία των δεδομένων, έγινε εφαρμογή του KBest αλγορίθμου ο οποίος είναι υπεύθυνος για την επιλογή των K πιο σημαντικών χαρακτηριστικών για παιρετέρω μελέτη με βάση την αξιολόγηση η οποία προκύπτει από την εφαρμογή του αλγορίθμου. Ακόμη, έγινε εφαρμογή των μηχανών διανυσμάτων υποστήριξης (support vector machine) όπως και της λογιστικής παλινδρόμησης (logistic regression) για την ταξινόμηση των δεδομένων σε κατηγορίες με τη χρήση των χαρακτηριστικών που επιλέχθηκαν από τον αλγόριθμο Kbest. Επίσης, έγινε εφαρμογή του k-fold cross validation (k = 10) για τον αμερόληπτο καθορισμό της ακρίβειας του μοντέλου και υπολογίστηκε ο πίνακας σύγχυσης (confusion matrix) που χάρη σε αυτόν υπολογίστηκαν και οι μετρικές για την επίδοση του μοντέλου (area under the curve, accuracy κλπ.). Σχηματικά, τα βήματα της κατηγοριοποίησης απεικονίζονται στην εικόνα 2. Το μοντέλο μηχανικής μάθησης που κατασκευάστηκε με βάση τα παραπάνω, έχει σκοπό την αυτόματη κατηγοριοποίηση γεγονός το οποίο θα διευκολύνει τους βιολόγους-κλινικούς και θα ελαφρύνει το φόρτο εργασίας τους καθώς θα δουλεύει σαν υποβοηθητικός παράγοντας.

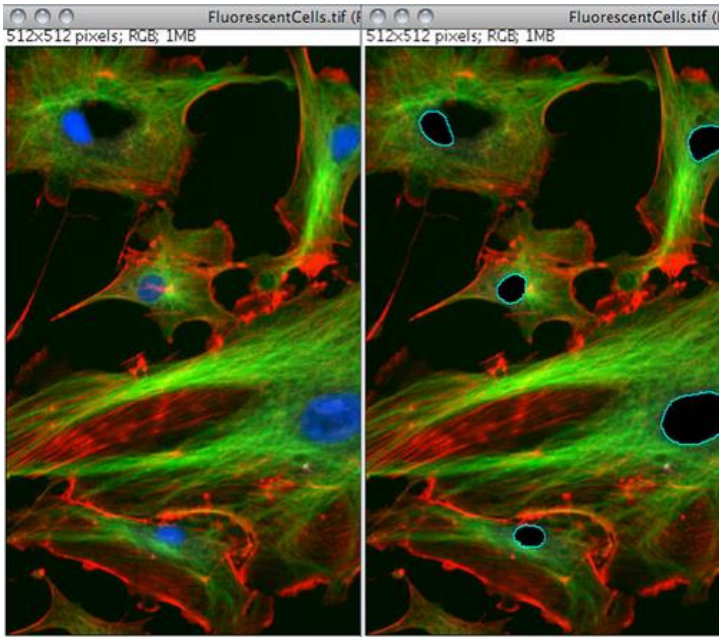


Εικόνα 2 : Ροή Εργασίας για την Κατηγοριοποίηση.

1.1 Κίνητρα για την προτεινόμενη εργασία

Οι ψηφιακές εικόνες αποτελούν ένα κρίσιμο κομμάτι πληροφοριών γι πολλές επιστημονικές εφαρμογές. Η ικανότητα επεξεργασίας και ανάλυσης του μεγάλου όγκου εικόνων που παράγονται από την πληθώρα διαθέσιμων τεχνικών μικροσκοπίας αυξάνει την ανάγκη για εξειδικευμένα εργαλεία λογισμικού. Η επεξεργασία των εικόνων πρέπει να γίνεται με συστηματικό και τυποποιημένο τρόπο, έτσι ώστε τα αποτελέσματα να είναι κατανοητά και να αναπαραχθούν. Η πανταχού παρούσα ψηφιακή τεχνολογία - από τις βασικές ψηφιακές φωτογραφικές μηχανές έως τους εξαιρετικά ειδικούς σαρωτές μικρο-CT έχει κάνει τις εικόνες ένα απαραίτητο «εργαλείο» πολλών ερευνητικών τομέων, από τη νανοτεχνολογία έως την αστρονομία. Είναι κοινή πρακτική για τους κατασκευαστές συσκευών λήψης εικόνων να περιλαμβάνουν ειδικό λογισμικό επεξεργασίας εικόνας, αλλά αυτά τα προγράμματα συνήθως δεν είναι πολύ ευέλικτα και δεν επιτρέπουν πιο

περίπλοκους χειρισμούς εικόνας. Παρά τις αμέτρητες δυνατότητες των λογισμικών αυτών υπάρχουν κάποια ζητήματα τα οποία μπορούν να διεστρεβλώσουν την εξαγωγή συμπερασμάτων και κατά συνέπεια να οδηγήσουν σε μη αναξιόπιστα αποτελέσματα. Για παράδειγμα, στην ανοσοϊστοχημεία όπου οι ιστοπαθολόγοι καλούνται να επιλέξουν μία περιοχή ενδιαφέροντος (ROI) από ένα κύτταρο ή κάποιον πυρήνα δεν υπάρχει αυτόματη αναγνώριση αυτών αλλά ο ειδικός θα πρέπει να σχεδιάσει με κάποιο εργαλείο που παρέχεται από το ImageJ την περιοχή ενδιαφέροντος. Διαπιστώνουμε λοιπόν ότι κάτι τέτοιο δεν θα έχει την ακρίβεια που απαιτείται σε τέτοιες αναλύσεις καθώς δεν είναι δυνατή η ακριβής σχεδίαση των περιοχών αυτών με το χέρι. Τέλος, για την εξέταση ενός συγκεκριμένου τύπου οργάνου, ιστού ή κυττάρου, πρέπει να επιλεγούν μικροσκοπικές τομές για να παρέχουν ένα αντιπροσωπευτικό δείγμα ολόκληρου του δείγματος. Η επιλογή συγκεκριμένων δειγμάτων ή τμημάτων δεν παρέχει ίσες ευκαιρίες για ανάλυση και χάνει την επίδραση της περιφερειακής ετερογένειας. Κάτι τέτοιο δεν οδηγεί σε αντιπροσωπευτικά για το δείγμα συμπεράσματα και τις περισσότερες φορές είναι λανθασμένα πρέπει λοιπόν να χρησιμοποιηθεί ένα σύστημα δειγματοληψίας που περιλαμβάνει όλες τις περιοχές για να αποφευχθεί η προκατάληψη της επιλογής [1]. Με άλλα λόγια χάνεται η έννοια της επαναληψιμότητας του πειράματος.



Εικόνα 3 : Σχεδίαση των περιοχών ενδιαφέροντος (ROI) στο ImageJ [1].

Για τους παραπάνω λόγους η εργασία αυτή εστιάζει στη δημιουργία μίας αυτοματοποιημένης μεθόδου ποσοτικοποίησης και κατηγοριοποίησης ενός συνόλου εικόνων με σκοπό τον περιορισμό λάθους που προκύπτει από την χειροκίνητη τεχνική μέχρι σήμερα και την εξαγωγή πιο αξιόπιστων αποτελεσμάτων.

1.2 Βιβλιογραφική Ανασκόπηση

Στην ενότητα αυτή θα παρουσιαστούν εργασίες οι οποίες σχετίζονται με τα θέματα τα οποία πραγματεύεται η παρούσα εργασία.

α) Ποσοτικοποίηση εικόνων φθορισμού :

Όπως αναφέρθηκε στην προηγούμενη ενότητα, η στατιστική σύγκριση εντάσεων χρώσης από δεδομένα ιστοπαθολογίας είναι σημαντικό εργαλείο για παρακολούθηση ή εξέληση θεραπείας μεταξύ πληθυσμών. Ο μεγαλύτερος περιορισμός όμως για την επαναληψιμότητα των αποτελεσμάτων είναι ο μη αντικειμενικός τρόπος μέτρησης καθώς ο ανθρώπινος παράγοντας παίζει μεγάλο ρόλο στην εφαρμογή της μάσκας για κάθε περιοχή ενδιαφέροντος. Στη βιβλιογραφία, η πλειοψηφία των δημοσιεύσεων που σχετίζονται με την ποσοτικοποίηση κυτταρικών δομών βασίζεται κυρίως σε εμπορικά ή και ελεύθερα λογισμικά [2]–[9] όπως το imageJ (https://imagej.nih.gov/ij/download.html). τα οποία για την ποσοτικοποίηση απαιτούν από τον χρήστη την χειροκίνητη επιλογή περιοχής ενδιαφέροντος όπως φαίνονται παρακάτω . Να επισημανθεί επίσης ότι η χρήση τέτοιων λογισμικών απευθύνεται κυρίως σε βιολόγους και κλινικούς ιστοπαθολόγους ώστε να μπορούν σχετικά εύκολα να διεξάγουν αποτελέσματα [10]. Παραδείγματα τέτοιων ερευνών φαίνονται παρακάτω.

Αρχικά, στο [11] παρουσιάζεται μια συγκριτική μελέτη της πυριτικής φαρμακολογίας, της γονιδιακής έκφρασης και των ανοσοκυτταροχημικών δεδομένων των τριών υποδοχέων σε διάφορες κυτταρικές σειρές καρκίνου του προστάτη και του μαστού. Το συμπέρασμα ήταν ότι ότι η τεστοστερόνη μπορεί να ασκήσει τα αποτελέσματά της μέσω περισσότερων από μία υποδοχέων GPCR μεμβράνης, ενεργώντας τόσο ως αγωνιστής είτε ως ανταγωνιστής, αυξάνοντας την πολυπλοκότητα της απόκρισης στη στοιχειομετρία κάθε υποδοχέα. Στο [12] έγινε διερεύνηση της εξειδίκευσης των υποδοχέων ανδρογόνων των μεμβρανών καρκινικών κυττάρων του προστάτη οι οποίοι ρυθμίζονται με σηματοδότηση Rho / ROCK ακτίνης και ακόμη, έγινε ανάλυση των τελεστών που ελέγχουν την επιβίωση και την απόπτωση σε ανθεκτικά σε ορμόνη καρκινικά κύτταρα DU145-προστάτη τα οποία διεγείρονται με επιλεκτικούς συναγωνιστές μεμβράνης ανδρογόνου. Το συμπέρασμα της έρευνας ήταν ότι οι υποδοχείς αυτοί μπορούν να συντελέσουν ως νέοι αντικαρκινικοί παράγοντες στον καρκίνο του προστάτη. Το [13] ασχολείται με τους υποδοχείς στεροειδών (SR) οι οποίοι χρησιμοποιούνται για να εντοπιστούν στη μεμβράνη του πλάσματος. Με τη χρήση ενός μοτίβου 9 αμινοξέων κατέληξαν στο συμπέρασμα ότι ότι οι μεταλλάξεις που αποτρέπουν επιλεκτικά τον

εντοπισμό της μεμβράνης θα βοηθήσουν στην ταξινόμηση των διακριτών αλλά ολοκληρωμένων λειτουργιών των διαφόρων ομάδων SR. Από την άποψη αυτή, ο αποκλεισμός του εντοπισμού της μεμβράνης ή η σηματοδότηση αποτρέπει μόνο την εξέλιξη του κυτταρικού κύκλου. Αυτό δείχνει τον σημαντικό ρόλο για το πυρηνικό SR ,

και προτείνει τη συνεργασία ομάδων υποκυτταρικών υποδοχέων για την πραγματοποίηση κυτταρικών δράσεων. Στο [14] έγινε διερεύνηση της υπόθεσης ότι το ανθρώπινο ZIP9 λειτουργεί ως υποδοχέας ανδρογόνου μεμβράνης (mAR) και μεσολαβεί σε αποπτωτικές

δράσεις ανδρογόνων. Η τεστοστερόνη δρα στην κυτταρική επιφάνεια του ανθρώπινου καρκίνου του προστάτη και του καρκίνου του μαστού για να προκαλέσει απόπτωση μέσω ενός μη αναγνωρισμένου mAR. Τα αποτελεσμάτα από την έρευνα αύτη παρέχουν την πρώτη απόδειξη για έναν μηχανισμό ο οποίος μεσολαβείται από μία μόνο πρωτεΐνη μέσω της οποίας οι οδούς σηματοδότησης στεροειδών και ψευδαργύρου αλληλεπιδρούν για τη ρύθμιση φυσιολογικών λειτουργιών σε κύτταρα θηλαστικών.

β) Κατηγοριοποίηση εικόνων ανοσοφθορισμού :

Όσον αφορά τις παραδοσιακές προσεγγίσεις μηχανικής μάθησης στο ευρύ πεδίο των εικόνων ανοσοφθορισμού, η βιβλιογραφική μας αναζήτηση είχε ως αποτέλεσμα λίγες σχετικές δημοσιεύσεις που δείχνουν ότι αυτό το πεδίο εφαρμογής τεχνητής νοημοσύνης δεν έχει ακόμη μελετηθεί αρκετά. Ο λόγος που συμβαίνει αυτό είναι ότι στο πεδίο της χειρουργικής παθολογίας χρησιμοποιείται η τεχνική FISH (Fluorescence in situ hybridization) η οποία είναι χρονοβόρα διαδικασία με μεγάλο κόστος [15] και χρησιμοποιείται κυρίως για την διάγνωση σαρκωμάτων, λεμφωμάτων και κάποιων στερεών όγκων. Έτσι, τα παραπάνω δεδομένα κατασκευάζονται μόνο για ερευνητικούς και εκπαιδευτικούς σκοπούς και η εφαρμογή της μηχανικής μάθησης σε δεδομένα ανοσοφθορισμού είναι σποραδικά και περιορισμένα. Παρόλα αυτά σχετικές δημοσιεύσεις με εικόνες ανοσοφθορισμού αναλύονται παρακάτω.

Αρχικά, στο [16] αξιολογήθηκαν διάφορες τεχνικές μηχανικής μάθησης για την ακριβή ανίχνευση μυελίνης σε πολυκαναλικές μικροσκοπικές εικόνες ενός βλαστοκυττάρου ποντικού. Μια άλλη μελέτη παρουσιάζει την εφαρμογή της μηχανικής μάθησης για την οπτικοποίηση σε πραγματικό χρόνο των περιθωρίων όγκου σε αποτμημένα δείγματα μαστού χρησιμοποιώντας απεικόνιση με φθορισμό [17]. Επιπλέον, στην [18] οι συγγραφείς έχουν αναπτύξει μια μέθοδο ταξινόμησης μηχανικής μάθησης για τον σχολιασμό της εξέλιξης μέσω μορφολογικά διακριτών βιολογικών καταστάσεων της απεικόνισης φθορισμού στην πάροδο του χρόνου. Επιπλέον, η παραδοσιακή υφή και τα στατιστικά χαρακτηριστικά εξήχθησαν τόσο σε εικόνες παθολογίας όσο και σε εικόνες ακτινολογίας για τη διερεύνηση των υποκείμενων συσχετίσεων μεταξύ κυτταρικής πυκνότητας και ετερογένειας του όγκου [19].

1.3 Δομή εργασίας

Στο κεφάλαιο 2 γίνεται ανάλυση του θεωρητικού υποβάθρου που χρησιμοποίηθηκε για την εργασία, αναφέροντας και αναλύοντας όλες τις τεχνικές και τις μεθόδους οι οποίες εφαρμόστηκαν για την υλοποίηση της παρούσας εργασίας. Το κεφάλαιο 3 περιέχει την

ανάλυση των μεθόδων που χρησιμοποιήθηκαν για την εκπόνηση της. Πιο αναλυτικά, γίνεται αναφορά στο σύνολο δεδομένων δηλαδή, από που αντλήθηκαν και τι περιέχουν τόσο αυτά για την ποσοτικοποίηση όσο και αυτά για την κατηγηριοποίηση καθώς περιέχονται κι άλλες πληροφορίες οι οποίες αφορούν την ποσότητα αυτών και τη σημαντικότητα τους. Στη συνέχεια, γίνεται παρουσίαση του δείγματος τον εικόνων που χρησιμοποιήθηκαν και ακολουθεί μία σύντομη παράγραφος η οποία περιγράφει το λογισμικό που χρησιμοποιήθηκε για την υλοποίηση της εργασίας καθώς και στους περιορισμούς που ανιχνεύθηκαν. Στο τέλος, γίνεται αναλυτική περιγραφή των βημάτων που εκτελέστηκαν για το πέρας της κάθε μεθοδολογίας ενώ παράλληλα έχουν τοποθετηθεί στιγμιότυπα για την οπτικοποίηση της εφαρμογής του κάθε βήματος ξεχωριστά ώστε να είναι πιο κατανοητή η διαδικασία που τέθηκε σε εφαρμογή. Στο κεφάλαιο 4, παρουσιάζονται τα αποτελέσματα της εργασίας από την ποσοτικοποίηση και την κατηγοριοποίηση των πυρήνων και του κυτταροπλάσματος από τις αρχικές και τις υπολογισμένες μάσκες σε μορφή πινάκων τα οποία θα σχολιαστούν αναλυτικά στο κεφάλαιο 5. Στην μεθοδολογία της ποσοτικοποίησης παρουσιάζονται τα αποτελέσματα για τα στατιστικά χαρακτηριστικά (Μέση τιμή, μέσος όρος, διακύμανση) ενώ στην κατηγοριοποίηση φαίνονται τα αποτελέσματα της μηχανικής μάθησης (ακρίβεια μοντέλου, AUC, Standard Deviation κλπ) με βάση των Κ χαρακτηριστικών τα οποία επιλέχθησαν σε κάθε περίπτωση. Οι πίνακες αυτοί περιέχουν τα αποτελέσματα και από τις 21 εικόνες του συνόλου δεδομένων. Το κεφάλαιο 5 περιλαμβάνει τα αποτελέσματα και από τις δύο μεθόδους, ενώ γίνεται και μία αναφορά εφ' όλης της ύλης σχετικά με τη δομή της εργασίας και τις μεθοδολογίες που εφαρμόστηκαν σε αυτήν ενώ στη συνέχεια πραγματοποιείται αναλυτικός σχολιασμός για τα συμπεράσματα που διεξήχθησαν από αυτές καταλήγοντας στον επίλογο όπου αναφέρονται τα κύρια συμπεράσματα. Στο κεφάλαιο 6 ακολουθεί ο επίλογος στον οποίο σχολιάζονται τα αποτέλεσματα και τα συμπεράσματα ολόκληρης της εργασίας και συζητείται για το πως μπορεί να γίνει εκμετάλλευση αυτών για μελλοντική έρευνα ή προέκταση της συγκεκριμένης. Τέλος το κεφάλαιο 7 περιλαμβάνει όλη την βιβλιογραφία από την οποία αντλίθηκαν όλες οι πληροφορίες που βοήθησαν στην διεξαγωγη της παρούσας πτυχιακής εργασίας.

2.Θεωρητικό Υπόβαθρο

2.1 Εικόνες ανοσοφθορισμού

Ο ανοσοφθορισμός είναι η μέθοδος στην οποία χρησιμοποιούνται φθορίζοντα αντισώματα με σκοπό την ανίχνευση και τον εντοπισμό αντιγόνου ή αντισώματος σε ιστούς ή κύτταρα. Χρησιμοποιείται για τον αξιόπιστο εντοπισμό μορίων σε ένα ευρύ φάσμα σταθερών κυττάρων ή ιστών. Η χρώση με τη μέθοδο αυτή, προσφέρει τη μοναδική δυνατότητα αποκάλυψης των μορίων στην «φυσική» κατάσταση, ελαχιστοποιώντας τις πιθανές διαμορφώσεις της πρωτεΐνης οι οποίες μπορεί να προκύψουν κατά τη σήμανση της με φθοριόχρωμα. Μπορεί επίσης να χρησιμοποιηθεί σε τομές ιστών, καλλιεργημένες κυτταρικές σειρές ή μεμονωμένα κύτταρα και μπορεί να χρησιμοποιηθεί για την ανάλυση

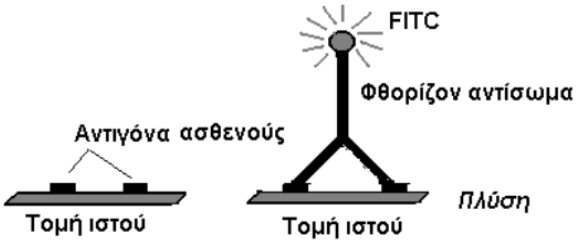
της κατανομής πρωτεϊνών, γλυκανών και μικρών βιολογικών και μη βιολογικών μορίων [20].

2.1.1 Αρχή ανοσοφθορισμού

Όταν το φως απορροφάται από συγκεκριμένα μόρια γνωστά ως φθοροχρώμια η ενέργεια των φωτονίων μπορεί να μεταφερθεί σε ηλεκτρόνια τα οποία χρειάζονται υψηλότερο επίπεδο ενέργειας. Κάποια από την ενέργεια απελευθερώνεται σε 10^-5 δευτερόλεπτα ως θερμότητα όταν το ηλεκτρόνιο επιστρέφει τη χαμηλότερη δονητική ενέργεια της διεγερμένης του κατάστασης. Αυτό το φαινόμενο ονομάζεται φθορισμός. Τα μήκη κύματος που είναι ικανά να προκαλέσουν φθορισμό ενός μορίου είναι γνωστά ως φάσμα διέγερσης και τα μήκη κύματος εκπεμπόμενου φθορισμού φωτός ως φάσμα εκπομπών. Η ένταση του φθορισμού ποικίλλει ανάλογα το φυσικό περιβάλλον. Ορισμένα φθοροχρώματα είναι πολύ φθορίζοντα σε ένα υδάτινο περιβάλλον ενώ άλλα είναι πιο έντονα σε ένα μη πολικό περιβάλλον. Τα φασματικά χαρακτηριστικά εξαρτώνται επίσης από το PH [21].

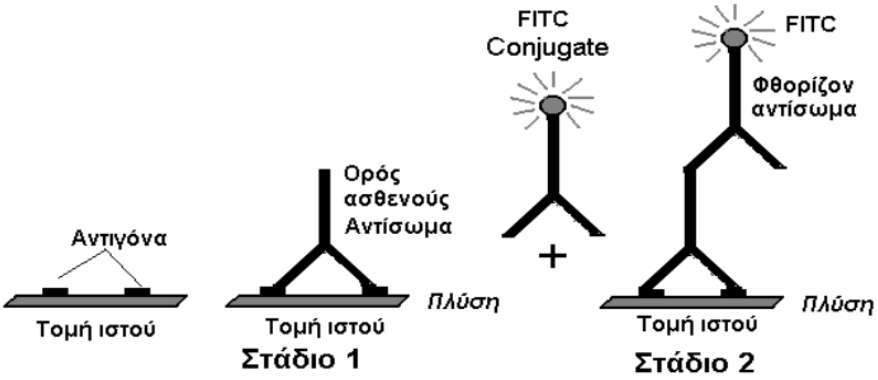
2.1.2 Έμμεσος και Άμεσσος Ανοσοφθορισμός

Ο ανοσοφθορισμός χωρίζεται σε δύο κατηγορίες. Στον άμμεσο (IFA) και στον έμμεσο (DFA). Στον άμεσσο ανοσοφθορισμό, προστίθεται φθορίζον αντίσωμα πάνω στην τομή του ιστού η οποία περιέχει τα αντιγόνα που αναζητούμε. Κατόπιν, ακολουθεί πλύση με φωσφορούχο διάλυμα με σκοπό την απομάκρυνση πλεονάζοντος αντισώματος και τέλος, ακολουθεί μικροσκόπηση σε μικροσκόπιο φθορισμού.



Εικόνα 4: Άμεσσος Ανοσοφθορισμός [22]

Στον **έμμεσο** ανοσοφθορισμό ο ορός του ασθενούς του οποίου τα αντισώματα προσπαθούμε να ανιχνεύσουμε τοποθετείται πάνω σε τομή ιστού που περιέχει τοαντιγόνο. Έπειτα, ακολουθεί έκπλυση με φωσφορούχο διάλυμα για την απομάκρυνση του ασύνδετου αντισώματος και στη συνέχεια, προστίθεται το φθορίζον αντίσωμα. Τέλος, γίνεται νέα πλύση για την απομάκρυνση του μη συνδεδεμένου φθορίζοντος αντισώματος. και ακολουθεί μικροσκόπηση σε μικροσκόπιο φθορισμού.



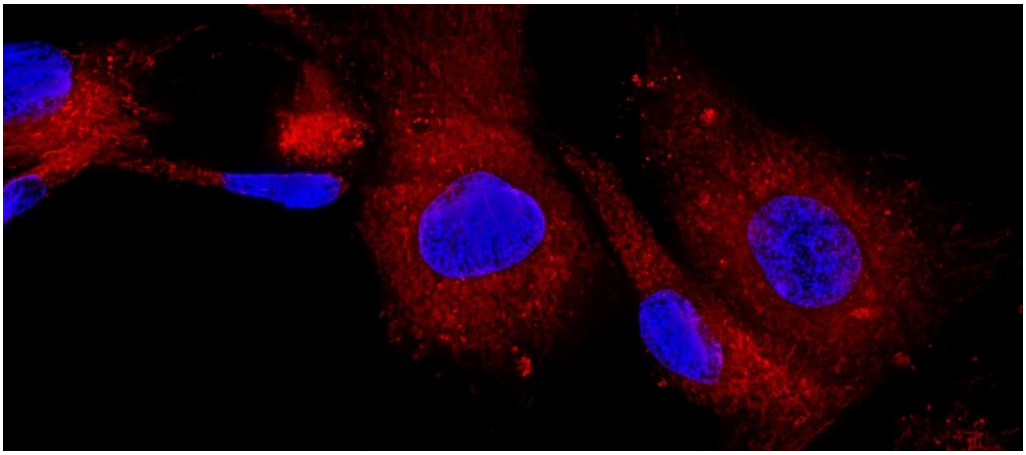
Εικόνα 5: Έμεσσος Ανοσοφθορισμός [22]

Όπως αποδεικνύεται από την περιγραφή των δύο μεθόδων IFA, DFA γίνεται χρήση του άμεσου ανοσοφθορισμού για την ανίχνευση αντιγόνων του ασθενή πάνω σε υλικό βιοψίας.
ενώ του έμμεσου ανοσοφθορισμού για την ανίχνευση αντισωμάτων στον ορό των ασθενών [21].

2.2 Ανοσοχρώση

Η ανοσοχρώση είναι η χρήση μιας μεθόδου που βασίζεται σε αντισώματα για την ανίχνευση μιας συγκεκριμένης πρωτεΐνης σε ένα δείγμα. Η ανοσοχρώση περιλαμβάνει ένα ευρύ φάσμα τεχνικών που χρησιμοποιούνται στην ιστολογία, την κυτταρική βιολογία και τη μοριακή βιολογία που χρησιμοποιούν μεθόδους χρώσης με βάση αντισώματα [23]. Γενικότερα, στις χρώσεις παρατηρείται η ύπαρξη ενός φθοριοχρώματος ή ενός ενζύμου το οποίο με την πραγματοποίηση μίας αντίδρασης θα δημιουργήσει μία χρώση. Το φθοριόχρωμα ή το ένζυμο το οποίο λαμβάνει μέρος στη διαδικασία της χρώσης είναι συζευγμένο με ένα αντίσωμα, το οποίο στη συνέχεια συνδέεται με προς μελέτη αντιγόνο (πχ μία πρωτεΐνη). Ανάλογα με το είδος της χρώσης

μπορούμε να θεωρήσουμε ότι υπάρχει γραμμική σύνδεση της έντασης της με την ύπαρξη του αντιγόνου. Πιο συγκεκριμένα, όσο πιο έντονη είναι η χρώση τόσο περισσότερο είναι το αντιγόνο που υπάρχει. Λόγω αυτού, είναι σημαντικό κατά την επεξεργασία του δείγματος να πραγματοποιηθεί η κατάλληλη μέριμνα ώστε να μην υπάρχει μη συνδεδεμένο φθοριόχρωμα-αντίσωμα, να μην υπάρχει μη ειδική σύνδεση, άρα και χρώση, καθώς και γενικότερα να διασφαλιστεί ότι για τη μέθοδο που χρησιμοποιείται μπορεί να πραγματοποιηθεί η παραπάνω παραδοχή. Είναι σημαντικό ακόμη να εξεταστεί η ύπαρξη της χρώσης ώστε να είναι γνωστό σε ποιο σημείο του κυττάρου εντοπίζεται μια πρωτεΐνη καθώς διαφορετικό σημείο εντόπισης στο κύτταρο αντιστοιχεί σε διαφορετικές δράσεις της συγκεκριμένης πρωτεΐνης. Τέλος, αξίζει να σημειωθεί ότι η υπάρχει η δυνατότητα επέμβασης με διάφορες ουσίες ή και φάρμακα, με σκοπό την αλλαγή του εντοπισμού της πρωτεΐνης και έτσι με αυτόν τον τρόπο αναπτύσσεται η δυνατότητα παρέμβασης στον κύκλο ζωής του κυττάρου.



Εικόνα 6: Παράδειγμα Ανοσοχρώσης

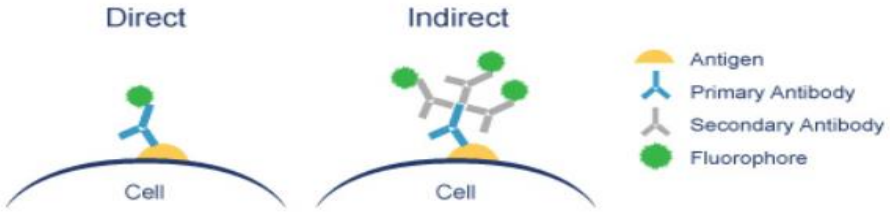
2.3 Ανοσοκυτταροχημεία

Η ανασοκυτταροχημεία είναι μία τεχνική η οποία χρησιμοποιείται για την ανίχνευση και την οπτικοποίηση πρωτεϊνών και πεπιτδίων σε κύτταρα χρησιμοποιώντας βιομόρια ικανά να «δεσμεύσουν» τις πρωτεϊνες ενδιαφέροντος. Συνήθως το βιομόριο είναι ένα αντίσωμα το οποίο είτε έμεσσα είτε άμεσα συνδέεται με τον αντιδραστήρα και είτε ένα ένζυμο είτε ένα φροροφόρο. Ο αντιδραστήρας θα προκαλέσει αύξηση στο σήμα (π.χ ένα χρώμα που θα προκύψει από την ενζυματική αντίδραση το οποίο είναι ανιχνεύσιμο μέσω μικροσκοπίου. Η τεχνική χρωματισμού εφαρμόζεται σε καλλιεργούμενα κύτταρα ή κύτταρα στα οποία έχει αφαιρεθεί η εξωκυτταρική μήτρα [24].

2.3.1 Έμμεση και Άμεση Μέθοδος

Στην άμεση μέθοδο το πρωτογενές αντίσωμα είναι απευθείας συνδεδεμένο με το αντιγόνο. Το πρωτεύον αντίσωμα μπορεί να επισημανθεί είτε με κάποιο ένζυμο με σκοπό να αντιδράσει με το κατάλληλο υπόστρωμα το οποίο είναι υπεύθυνο για την ανάπτυξη χρώματος και το αποτέλεσμα είναι ορατό με τη χρήση μικροσκοπίου φωτός

είτε επισημαίνεται με ένα μόριο φθορισμού με σκοπό τη δημιουργία ενός σήματος όπου φαίνεται η παρουσία των αντιγόνων σε ένα τμήμα ιστού και το αποτέλεσμα μπορεί να φανεί χρησιμοποιώντας μικροσκόπιο φθορισμού. Συμπερασματικά, αποτελεί μία ταχεία μέθοδο με λίγα βήματα και μπορούν να λάβουν μέρος πολλαπλά αντισώματα αλλά το μειονέκτημά της είναι η μειωμένη ευαισθησία. Ένα σημαντικό πλεονέκτημα της έμμεσης μεθόδου είναι η αυξημένη ευαισθησία η οποία οφείλεται στην πολλαπλή σύνδεση δευτερογενών αντισωμάτων με το πρωτογενές αντίσωμα, το οποίο ενισχύει το σήμα όπως επίσης και η ύπαρξη αυξημένης ευκαμψίας λόγω της δυνατότητας μεταβολής του πρωτογενούς και δευτερογενούς συνδυασμού των αντισωμάτων. Τα μειονεκτήματα της έμμεσης μεθόδου είναι ότι αποτελεί μία χρονοβόρα διαδικασία και υπάρχει κίνδυνος μη ακριβής σύνδεσης του δευτερεύοντος αντισώματος [25].



Εικόνα 7: Έμεσση και Άμεση μέθοδος Ανοσοκυτταροχημείας

2.3.2 Αρχή Ανοσοκυτταροχημείας

Ενώ η χρήση των σωστών αντισωμάτων για τη στόχευση των σωστών αντιγόνων και την ενίσχυση του σήματος είναι ζωτικής σημασίας για τη βέλτιστη οπτικοποίηση, η πλήρης προετοιμασία του δείγματος είναι κρίσιμη για τη διατήρηση της μορφολογίας των κυττάρων, της αρχιτεκτονικής των ιστών και της αντιγονικότητας των επιτόπων στόχων. Αρχικά, πρέπει να γίνει καλλιέργεια των κυττάρων. Τα δείγματα τοποθετούνται σε ολισθηρά κομμάτια γυαλιού ώστε να γίνει η επώαση τους (κάθε κύτταρο έχει διαφορετικό χρόνο επώασης, σε κάποια δεν είανι απαραίτητο ενώ άλλα μπορούν να χρειαστούν έως 24 ώρες). Στη συνέχεια, γίνεται στερέωση των κυττάρων ώστε να μην υπάρξει τυχόν μετακίνηση κατά την ανάλυση τους και γίνεται η ανασόχρωση με σκοπό την ανίχνευση συγκεκριμένου δείγματος πρωτεΐνης. Έπετα, χρησιμοποιείται ο αντιδραστήρας στο κομμάτι του γυαλιού και κατόπιν ξεπλένεται το πλεονάζων αντίσωμα. Αφού ολοκληρωθεί η διαδικασία αυτή, είναι δυνατή η χρήση μικροσκοπίου για την οπτικοποίηση των κυττάρων και μπορούν να ληφθούν εικόνες για περαιτέρω ανάλυση των κυτταρικών δομών με τη χρήση καμερών (όπως φαίνεται και στην εικόνα 8) [26] [27].

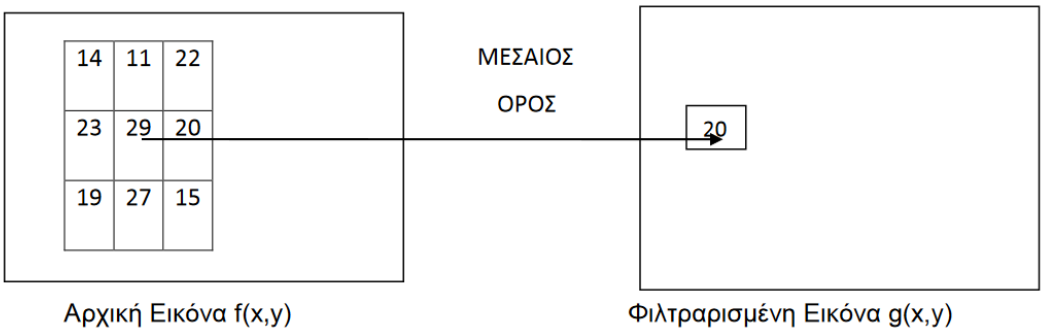
Εικόνα 8 : Αρχή Ανοσοκυτταροχημείας [26]

2.4 Φίλτρο Μεσαίας Τιμής (Median Filter)

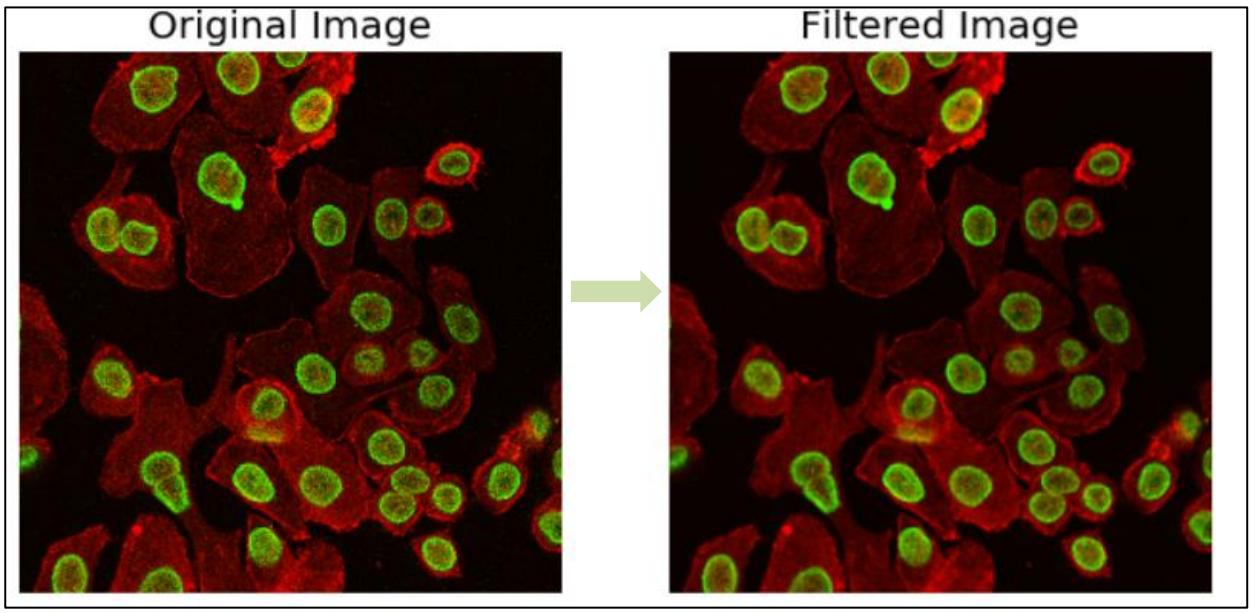
Το φίλτρο μεσαίας τιμής είναι μια μη γραμμική τεχνική ψηφιακού φιλτραρίσματος, που χρησιμοποιείται συχνά για την αφαίρεση θορύβου από μια εικόνα ή ένα σήμα. Αυτή η μείωση θορύβου αποτελεί ένα βήμα προεπεξεργασίας για τη βελτίωση των αποτελεσμάτων της μεταγενέστερης επεξεργασίας (για παράδειγμα, ανίχνευση ακμών σε μια εικόνα). Χρησιμοποιείται ευρέως στην επεξεργασία ψηφιακών εικόνων επειδή, υπό ορισμένες συνθήκες, διατηρεί τις άκρες ενώ αφαιρεί τον θόρυβο (όπως φαίνεται στην εικόνα 10). Σύμφωνα με την μέθοδο αυτή οι

τιμές των pixels μιας γειτονιάς ταξινομούνται και η τιμή κάθε εικονοστοιχείου της νέας εικόνας είναι η μεσαία από τις τιμές των pixel της γειτονιάς του αντίστοιχου εικονοστοιχείου της αρχικής. Συγκεκριμένα, έστω A={α1,α2,α3,..,αn}το σύνολο με στοιχεία α1≤ α2≤...≤αn∈R [28]. Το A είναι ίσο

με median(A) = $f(x) = \begin{cases} a_{\frac{n+1}{2}}, & n, περιττός \\ \frac{1}{2}\left(a_{\frac{n}{2}} + a_{\frac{n}{2}} + 1\right) & n, άρτιος \end{cases}$



Εικόνα 9 : Χρήση Φίλτρου Μεσαίας Τιμής



Εικόνα 10 : Εφαρμογή φίλτρου μεσαίας τιμής

2.5 Κατωφλίωση Εικόνας

Η χρήση κατωφλίων (thresholding) είναι μια από τις επικρατέστερες τεχνικές για την τμηματοποίηση εικόνων που απεικονίζονται με αποχρώσεις του γκρι. Η μέθοδος αυτή

αποτελεί την απλούστερη προσέγγιση στο θέμα της αυτόματης τμηματοποίησης μιας εικόνας και βασίζεται στο γεγονός ότι οι τιμές των εντάσεων των εικονοστοιχείων (pixels) που ανήκουν στα αντικείμενα έχουν αξιοσημείωτη διαφορά από τις τιμές των εντάσεων των pixels που ανήκουν στο φόντο. Συνεπώς, με την επιλογή του κατάλληλου εύρους τιμών που ανήκουν σε κάποιο αντικείμενο υπάρχει η δυνατότητα διαχωρισμού του αντικειμένου από το φόντο του. Η πληροφορία για την επιλογή του κατάλληλου κατωφλιού δίνεται από το ιστόγραμμα μία εικόνας. Η ύπαρξη ενός ή περισσοτέρων αντικειμένων,αντιστοιχεί στην ύπαρξη κορυφών στο ιστόγραμμα, άρα επιλέγοντας τα όρια των κορυφών, δίνεται η δυνατότητα εύρεσης των τιμών των εικονοστοιχείων που ανήκουν σε ένα αντικείμενο [29].

2.5.1 Μέθοδος Otsu

Η μέθοδος κατωφλίωσης Otsu είναι ένας εξαντλητικός αλγόριθμος αναζήτησης του καθολικού βέλτιστου ορίου, μεγιστοποιώντας τη διακύμανση μεταξύ των διάφορων επιπέδων - κλάσεων. Ο αλγόριθμος επιστρέφει ένα όριο κατωφλίωσης το οποίο διαχωρίζει τα εικονοστοιχεία σε δύο κατηγορίες, προσκήνιο και φόντο (όπως φαίνεται στην εικόνα 11). Η μέθοδος επεξεργάζεται το ιστόγραμμα της εικόνας,

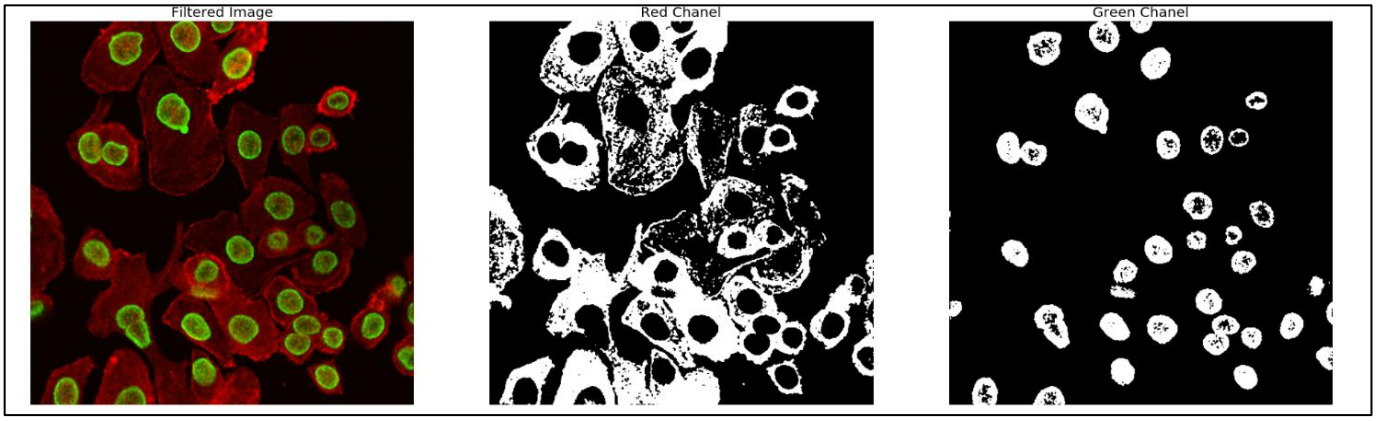
τμηματοποιώντας τα αντικείμενα και ελαχιστοποιώντας τη διακύμανση σε κάθε μία από τις κλάσεις. Το ιστόγραμμα μιας τέτοιας

εικόνας περιέχει δύο σαφώς εκφρασμένες κορυφές, οι οποίες αντιπροσωπεύουν διαφορετικά εύρη τιμών έντασης. Έπειτα, ο αλγόριθμος επεξεργάζεται την εικόνα εισόδου και δημιουργεί το ιστόγραμμα της. Στη συνέχεια υπολογίζει μία τιμή κατωφλιού έστω T και κάνει αντικατάσταση των pixels της εικόνας με λευκό χρώμα για τις περιοχές όπου η τιμή έντασης του pixel είναι ανώτερη του T και με μαύρο χρώμα για τις τιμές που είναι μικρότερες του T.

Έστω ότι τα pixel μιας εικόνας αναπαρίστανται σε L επίπεδα του γκρι με [0, 1, ..., L-1]. Ο αριθμός των pixels σε ένα επίπεδο i συμβολίζεται από το ni και ο συνολικός τους αριθμός με (N = n1+ n2+ ... + nL). Η πιθανότητα i ενός επιπέδου του γκρι υποδηλώνεται από το:

$$pi = \frac{ni}{N}, \qquad pi \geq 0, \sum_{o}^{L-1} pi = 1.$$

Ακολούθως τα εικονοστοιχεία διχοτομούνται σε δύο κλάσεις C1 και C2 (φόντο και αντικείμενα). Η C1 περιλαμβάνει εικονοστοιχεία με επίπεδα [0, 1, ..., t], ενώ η C2 με επίπεδα [t + 1, ..., L-1] και όριο κατωφλίου t [30].



Εικόνα 11 : Κατωφλίωση εικόνας και εφαρμογή μεθόδου otsu.

2.6 Χαρακτηριστικά εικόνας

Χαρακτηριστικά θεωρούνται τα μετρήσιμα μεγέθη τα οποία μπορούν να διεξαχθούν από τις εικόνες και ποικίλλουν ανάλογα με την εικόνα η οποία εξετάζεται. Πιο συγκεκριμένα, ανάλογα με το τι αναπαριστάνει η εικόνα, τα αντικείμενα δηλαδή που απεικονίζονται σε αυτή μπορούν να διεξαχθούν διαφορετικά χαρακτηρισιτκά τα οποία χωρίζονται σε δύο κατηγορίες. Τα στατιστικά χαρακτηριστικά και τα χαρακτηριστικά τα οποία αφορούν την υφή και την μορφολογία της εικόνας.

2.6.1 Στατιστικά Χαρακτηριστικά

Μέση τιμή (mean) : Η τιμή ενός συνόλου ν παρατηρήσεων. Είναι το σπουδαιότερο και το πιο χρήσιμο μέτρο της Στατιστικής και είναι ένα μέτρο θέσης, δείχνει δηλαδή σχετικά τις θέσεις των αριθμών στους οποίους αναφέρεται. Γενικότερα, ορίζεται ως το άθροισμα των παρατηρήσεων δια του πλήθους τους. Αποτελεί δηλαδή τη μαθηματική πράξη ανεύρεσης της «μέσης απόστασης» ανάμεσα σε δύο ή περισσότερους αριθμούς. Η μέση τιμή συμβολίζεται με $\bar{x}$. Ο γενικός τύπος της είναι :

$$\bar{x} = \frac{1}{n}\sum_{k=0}^{n} t_i = \frac{1}{n}(t_1 + \cdots + t_n)$$ όπου t_i η i παρατήρηση και n το πλήθος αυτών.

Μεσαία τιμή (median) : Η διάμεση τιμή ενός δείγματος με παρατηρήσεις n σε αύξουσα σειρά είναι η μέση παρατήρηση εάν το n είναι μονός αριθμός, διαφορετικά ορίζεται ως η μέση τιμή των δύο μεσαίων τιμών.

$$\text{median(A)} = f(x) = \begin{cases} a_{\frac{n+1}{2}}, \ n, περιττός \\ \frac{1}{2}\left(a_{\frac{n}{2}} + a_{\frac{n}{2}} + 1\right) \ n, άρτιος \end{cases}$$

Διακύμναση (Std) : Είναι η αναμενόμενη τιμή της τετραγωνικής απόκλισης της τυχαίας μεταβλητής από τη μέση τιμή, και μη τυπικά μετρά πόσο απέχει ένα σύνολο αριθμών απλώνεται από τη μέση τιμή του.

$$s = \sqrt{\frac{\Sigma(X - \bar{\bar{x}})^2}{n-1}}$$

Ελάχιστη, Μέγιστη τιμη (Min, Max) : Η ελάχιστη και η μέγιστη τιμή αντίστοιχα. Σε ένα όρισμα διανύσματος επιστρέφει την ελάχιστη ή μέγιστη τιμή ενω σε ένα πίνακα, επιστρέφει ένα διάνυσμα γραμμής με την ελάχιστη ή μέγιστη τιμή κάθε στήλης.

Διορθωμένος συνολικός αριθμός φθορισμός κυττάρων (Corrected Total Cell Fluorescence) : Επιστρέφει τον αριθμό των εικονοστοιχείων (pixel) – την πυκνότητα αυτών ανάλογα τα στοιχεία που απεικονίζονται. CTCF = Ολοκληρωμένη πυκνότητα - (Περιοχή επιλεγμένου κυττάρου * Μέσος φθορισμός φόντου). Υπολογίζοντας τη διαφορά μεταξύ του φόντου και παρασκηνίου έχουμε την δυνατότητα να γνωρίζουμε πόσοι πυρήνες ή πόσα κύτταρα απεικονίζονται σε μία εικόνα.

Εκτός από τα στατιστικά χαρακτηριστικά μπορούν να υπολογιστούν και άλλα χαρακτηριστικά τα οποία αφορούν την υφή που υπάρχει στις εικόνες και μπορούν να ανιχνευτούν και να υπολογιστούν με τη χρήση της μεθόδου των Radioimcs.

2.6.2 Radiomics

Τα Radiomics είναι ένα ισχυρό εργαλείο που στόχο έχει την ανάπτυξη και την εξέταση ιατρικών υποθέσεων με την εξαγωγή χρήσιμων πολυδιάστατων δεδομένων από κλινικές εικόνες. Η αρχή των Radiomics είναι ότι τα διακριτικά χαρακτηριστικά των εικόνων μπορεί να είναι χρήσιμα για την πρόβλεψη της πρόγνωσης και της θεραπευτικής απόκρισης για διάφορες ασθένειες, παρέχοντας έτσι πολύτιμες πληροφορίες για μία εξατομικευμένη θεραπεία. Μπορούν να χρησιμοποιηθούν στις περισσότερες μεθόδους απεικόνισης που περιλαμβάνουν ακτινογραφίες, υπερήχους, CT, MRI και PET μελέτες. Χρησιμποιούνται ακόμα για την αύξηση της ακρίβειας στη διάγνωση και την πρόβλεψη της απόκρισης στη θεραπεία, ειδικά σε συνδυασμό με κλινικά, βιοχημικά και γενετικά δεδομένα. Οι βιοϊατρικές εικόνες περιέχουν πληροφορίες όπου ορισμένες από αυτές δεν γίνονται αντιληπτές από το ανθρώπινο μάτι. Μέσω της μαθηματικής εξαγωγής της χωρικής κατανομής των εντάσεων σήματος και των συσχετισμών εικονοστοιχείων, τα radiomics ποσοτικοποιούν τις πληροφορίες υφής που βρίσκονται σε αυτές τις εικόνας χρησιμοποιώντας μεθόδους ανάλυσης από το πεδίο της τεχνητής νοημοσύνης (AI).

Επιπλέον, οι διαφορές στην ένταση, το σχήμα ή την υφή της εικόνας μπορούν να ποσοτικοποιηθούν, ξεπερνώντας έτσι την υποκειμενική φύση της διάγωνσης [31].

2.6.2.1 Χαρακτηριστικά Radiomics

Με την χρήση των Radiomics υπάρχει η δυνατότητα εξαγωγής

Χαρακτηριστικά σχήματος (Shape) : Περιγράφουν το σχήμα της περιοχής ενδιαφέροντος (ROI) σε 2D ή 3D.

Χαρακτηριστικά GLCM (Gray Level Co-occurrence Matrix) : Σχετίζονται με την από κοινού συνάρτηση πιθανότητας δεύτερης τάξης της περιοχής ενδιαφέροντος

Χαρακτηριστικά GLSZM (Gray Level Size Zone Matrix) : Υπολογίζουν τον αριθμό των περιοχών που έχουν pixels/voxels με ίδια ένταση.

Χαρακτηριστικά NGTDM (Neighboring Gray Tone Difference Matrix) : Υπολογίζουν τη διαφορά μεταξύ της έντασης ενός pixel/voxel με τη μέση τιμή έντασης των γειτόνων του σε μια προκαθορισμένη απόσταση.

Χαρακτηριστικά Gray Level Dependence Matrix : Υπολογίζουν τον αριθμό των pixels/voxels που "εξαρτώνται" από καθένα από τα pixels/voxels της εικόνας. Ένα pixel/voxel θεωρείται ότι εξαρτάται από ένα άλλο εάν η απόλυτη τιμή της μεταξύ τους διαφοράς έντασης είναι μικρότερη από μια προκαθορισμένη τιμή.

2.6.2.2 Βήματα για την εξαγωγή χαρακτηριστικών με Radiomics

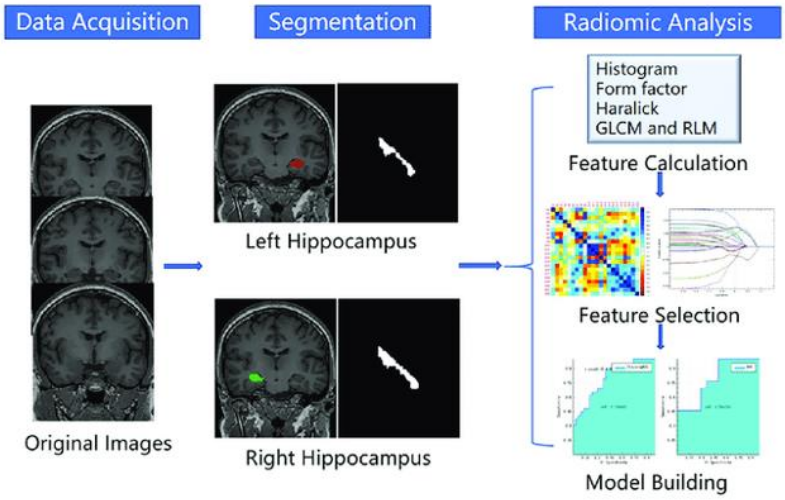
Τα στάδια για την εξαγωγή αποτελεσμάτων είναι τα εξής (όπως παρουσιάζονται στην εικόνα 12) :

α) Λήψη της εικόνας και τμηματοποίηση της: όπου τα δεδομένα της εικόνας που χρησιμοποιούνται για τον χαρακτηρισμό των όγκων παρέχονται από τεχνολογία ιατρικής σάρωσης. Αντί τη λήψη μια φωτογραφίας σαν κάμερα, οι σαρώσεις παράγουν ακατέργαστους όγκους δεδομένων που πρέπει να υποβληθούν σε περαιτέρω επεξεργασία για να μπορούν να χρησιμοποιηθούν σε ιατρικές έρευνες. Για την λήψη εικόνων που μπορούν να ερμηνευθούν, πρέπει να χρησιμοποιηθεί ένα εργαλείο ανακατασκευής. Στη συνέχεια, πρέπει να γίνει τμηματοποίηση της εικόνας με σκοπό να κρατηθούν μόνο οι περιοχές που μας ενδιαφέρουν όπως π.χ όγκοι για την περαιτέρω ανάλυση.

β) Εξαγωγή χαρακτηριστικών και ανάλυση δεδομένων : Μετά την τμηματοποίηση, μπορούν να εξαχθούν πολλά χαρακτηριστικά και μπορεί να υπολογιστεί συσχέτιση μεταξύ

τους. Τα χαρακτηριστικά μπορούν να χωριστούν σε πέντε ομάδες: χαρακτηριστικά βάση μεγέθους και σχήματος, βάση του ιστογράμματος των εντάσεων της εικόνας, συσχετίσεις μεταξύ των voxel της εικόνας (π.χ πίνακας γκρίζων επιπέδων (GLCM)) και με βάση των υφών που αποτυπώνονται. Λόγω της ποικιλίας και του μεγάλου αριθμού των χαρακτηριστικών που μπορούν να διεξαχθούν πρέπει να γίνει η επιλογή των πιο χρήσιμων και κατάλληλων με τη χρήση κάποιου αλγορίθμου και έλεγχος των δεδομένων με επιβλεπόμενο τρόπο ή μη.

γ) αποτύπωση τρισδιάστατης εικόνας και στατιστική ανάλυση αυτής : αφού έχει ολοκληρωθεί η επιλογή των χρήσιμων χαρακτηριστικών και έχουν κριθεί κατάλληλα, γίνεται ανάλυση της εικόνας με σκοπό διεξαγωγή συμπερασμάτων με στόχο την αντιμετώπιση τυχών ασθένειας και τη συνταγογράφηση της πιο κατάλληλης θερπείας για την αντιμετώπισή της [31].



Εικόνα 12 : Βήματα Εφαρμογής των Radiomics [32]

2.7 Μηχανική Μάθηση

Η μηχανική μάθηση είναι μια τεχνική η οποία χρησιμοποιείται για την αναγνώριση προτύπων και μπορεί να εφαρμοστεί σε ιατρικές εικόνες. Η μηχανική εκμάθηση ξεκινά συνήθως με ένα σύστημα αλγορίθμου ο οποίος υπολογίζει τα χαρακτηριστικά τα οποια θεωρουνται ότι είναι σημαντικά για την πρόβλεψη ή τη διάγνωση κάποια ασθένειας. Το σύστημα αυτό προσδιορίζει έπειτα τον καλύτερο συνδυασμό των χαρακτηριστικών της εικόνας με σκοπό την κατηγοριοποίηση της ή τον υπολογισμό κάποιας μετρικής για κάποια δεδομένη περιοχή εικόνας. Στον τομέα της ιατρικής για παράδειγμα, μπορούν να συνεισφέρουν στην ανακάλυψη νέων φαινομένων και την απόκτηση νέων γνώσεων. Πιο συγκεκριμένα, στις περιπτώσεις όπου υπάρχει μεγάλος όγκος δεδομένων τα οποία δεν φαίνεται να είναι εύκολο να κατηγοριοποιηθούν από έναν ειδικό, τα συστήματα αυτά μπορούν και αναλύσουν τα δεδομένα και να βρουν πιο περίπλοκα μοτίβα με απρόσμενες συσχετίσεις μεταξύ τους ενώ ακόμη μπορούν να αναλύσουν προϋπάρχοντα μοντέλα και να δείξουν πώς οι πειραματικές παρατηρήσεις αντιτίθενται στις υπάρχουσες θεωρίες [33].

2.7.1 Μεθοδολογίες Μηχανικής Μάθησης

Κύριες εργασίες της διαδικασίας εξόρυξης γνώσης είναι η εύρεση συσχετίσεων μεταξύ των δεδομένων (κανόνες συσχέτισης), η κατηγοριοποίηση σε προκαθορισμένες κλάσεις (δέντρα απόφασης, νευρωνικά δίκτυα, και η συσταδιοποίηση-ομαδοποίηση (ιεραρχικοί, διαμεριστικοί, με βάση την ποιότητα).

- Ταξινόμηση (classification)

Ανάπτυξη ενός μοντέλου πρόβλεψης της κλάσης των στιγμιοτύπων ενός προβλήματος. Το μοντέλο αυτό χτίζεται με βάση ένα σύνολο δεδομένων εκπαίδευσης .Ηαπόδοσή του αξιολογείται με βάση ένα σύνολο δεδομένων ελέγχου.

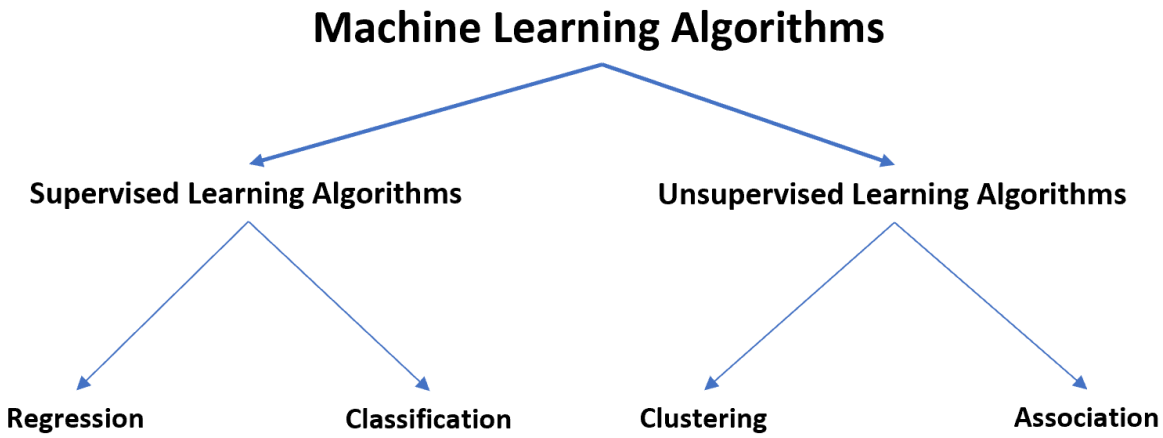
- Ομαδοποίηση (clustering)

Διαχωρισμός των δεδομένων σε ομάδες-συστάδες έτσι ώστε για κάθε εγγραφή που περιλαμβάνει μια συστάδα να είναι μεγαλύτερη από την ομοιότητά της με οποιαδήποτε εγγραφή από άλλες συστάδες. Στην μη εποπτευόμενη μάθηση δεν γνωρίζουμε την κλάση στην οποία ανήκουν τα δεδομένα εκπαίδευσης. Μας δίνεται ένα σύνολο μετρήσεων και παρατηρήσεων με στόχον α ανακαλύψουμε κλάσεις ήομάδες μέσα στα δεδομένα.

Οι τεχνικές μάθησης ανάλογα με το είδος του προβλήματος μπορούν να χωριστούν σε δύο βασικές κατηγορίες (Εικόνα 13) :

- **Μηχανική Μάθηση με επίβλεψη (Supervised Learning**), όπου το σύστημα εκπαιδεύεται με βάση μία έννοια ή συνάρτηση από ένα σύνολο δεδομένων το οποίο αποτελεί την περιγραφή ενός μοντέλου και ο αλγόριθμος επίβλεψης παρέχει στο σύστημα τις σωστές τιμές εξόδου της συνάρτησης για τα δεδομένα τα οποία εισήχθησαν.

- **Μηχανική Μάθηση χωρίς επίβλεψη (Unsupervised Learning)**, όπου το σύστημα καλείται να εκαπιδευτεί αυτόματα, ανακαλύπτοντας ουσιαστικά από μόνο του συσχετίσεις ή ομάδες ανάμεσα σε ένα σύνολο δεδομένων και στη συνέχεια, χωρίς την παροχή εμπειρίας στον αλγόριθμο μάθησης, αναλαμβάνει να βρεί την δομή των δεδομένων εισόδου με σκοπό την ανάπτυξη ορθής πρόβλεψης.



Εικόνα 13 : Μηχανική Μάθηση με επίβλεψη και χωρίς επίβλεψη

Στη μάθηση με επίβλεψη το σύστημα πρέπει να είναι ικανό να μάθει κάποια συνάρτηση στόχο (target function), η οποία συμβολίζεται με c και να αποτελεί έκφραση του μοντέλου που περιγράφει σωστά τα δεδομένα. Συνήθως χρησιμοποιείται για την πρόβλεψη (prediction) της τιμής μιας μεταβλητής οπου ονομάζεται εξαρτημένη μεταβλητή ή μεταβλητή εξόδου $y(i)$ σύμφωνα με ένα σύνολο ανεξάρτητων μεταβλητών ή μεταβλητών εισόδου $x(i)$, όπου το (i) είναι δείκτης. Η συνάρτηση αυτή έχει ως είσοδο ένα σύνολο διαφορετικών μεταβλητών το οποίο είναι το πεδίο ορισμού της και ονομάζεται σύνολο των περιπτώσεων και συμβολίζεται με X. Κάθε υπόθεση περιγράφεται από ένα σύνολο χαρακτηριστικών (features). Το υποσύνολο των περιπτώσεων για το οποίο η τιμή της

μεταβλητής εξόδου είναι γνωστή δηαδή, το σύνολο από τα ζεύγη $(x(i), y(i))$, ονομάζεται σύνολο εκπαίδευσης (training set) και κάθε ζεύγος $(x(i) ,y(i))$ ονομάζεται παράδειγμα εκπαίδευσης (test set).

Αν m είναι ο αριθμός των υποθέσεων που υπάρχουν τότε το σύνολο εκπαίδευσης (training set) θα είναι $\{(x(i),y(i));i=1,...,m\}$ το παράδειγμα εκπαίδευσης (training example). Κατά την προσπάθεια προσέγγισης της συνάρτησης στόχου c, το σύστημα εξετάζει εναλλακτικές συναρτήσεις που ονομάζονται υποθέσεις και συμβολίζονται με h ή $h\theta$. Το σύνολο των υποθέσεων που πρέπει να εξεταστούν ώστε να βρεθεί η συνάρτηση στόχος συμβολίζεται με H. Αν θεωρηθεί ότι η μεταβλητή y είναι γραμμικώς εξαρτημένη από τη x, τότε η αναπαράσταση της υπόθεσης μπορεί να γίνει ως εξής: $h\theta(x)=\theta_0 +\theta_1\chi_1+\theta_2\chi_2$ με θi να είναι

οι παράμετροι (ή βάρη) και $x1$, $x2$ να είναι τα χαρακτηριστικά (features). Για n χαρακτηριστικά και αν θεωρηθεί πως $x0$=1η παράσταση μπορεί να γραφεί ως :

$$h_\theta(x) = \sum_{i=0}^{n} \theta_i \chi_i \; = \; \theta^T \chi$$

Στη μηχανική μάθηση με επίβλεψη διακρίνονται δύο είδη προβλημάτων: i)Ταξινόμηση (Classification), η οποία αναφέραται στη δημιουργία μοντέλων πρόβλεψης διακριτών

τάξεων, όπου η μεταβλητή y θα παίρνει μόνο έναν μικρό αριθμό διακριτών τιμών ii) Παρεμβολή ή Παλινδρόμηση (Regression), που αφορά τη δημιουργία πρόβλεψης αριθμητικών τιμών, δηλαδή η μεταβλητή y θα είναι συνεχής [34].

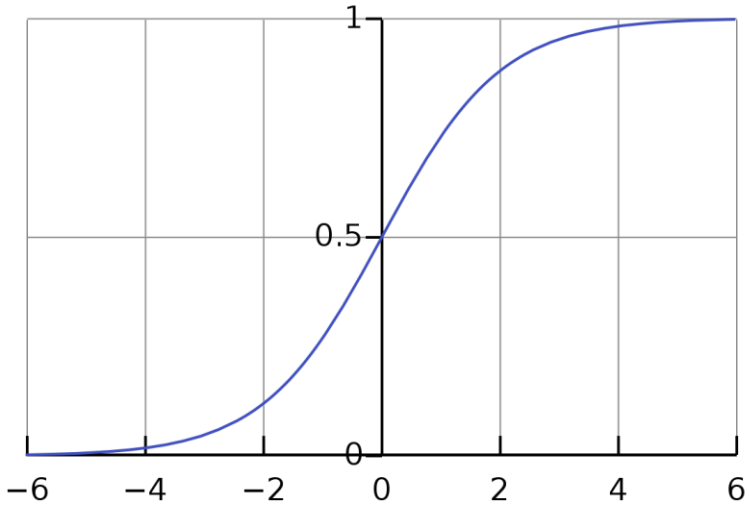
Οι ευρύτερες τεχνικές μηχανικής μάθησης με επίβλεψη είναι :

- Γραμμικη Παλινδρόμηση (Linear Regression)
- Λογιστική Παλινδρόμηση (Logistic Regression) (αποτελεί βελτίωση της μεθόδου της γραμμικής παλινδρόμησης)
- K-NN Algorithm ή K-Nearest Neighbours Algorithm
- Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines (SVMs))
- Νευρωνικά Δίκτυα (Neural Networks)

2.7.2 Λογιστική Παλινδρόμηση (Logistic Regression)

Η λογιστική παλινδρόμηση είναι ένας αλγόριθμος που χρησιμοποιείται για την επίλυση προβλημάτων ταξινόμησης. Πρόκειται για μια προγνωστική ανάλυση που περιγράφει τα δεδομένα και εξηγεί τη σχέση μεταξύ των μεταβλητών. Η λογιστική παλινδρόμηση εφαρμόζεται σε μια μεταβλητή εισόδου (X) όπου η μεταβλητή εξόδου (y) είναι μια διακριτή τιμή που κυμαίνεται μεταξύ 1 και 0 (εικόνα 14). Στόχος της μεθόδου αυτής είναι η δημιουργία ενός μοντέλου πρόβλεψης των τιμών της εξαρτημένης μεταβλητής η οποία εξατάζεται χρησιμοποιώντας κάποιες ποσοτικές και ποιοτικές ανεξάρτητες μεταβλητές. Χρησιμοποιεί τη λογιστική (σιγμοειδής) συνάρτηση για να βρει τη σχέση μεταξύ των μεταβλητών. Η συνάρτηση σιγμοειδούς είναι μια καμπύλη σχήματος S που μπορεί να πάρει οποιονδήποτε πραγματικό αριθμό και να τον αντιστοιχίσει σε τιμή μεταξύ 0 και 1, αλλά ποτέ ακριβώς σε αυτά τα όρια : $0 \leq g(z) \leq 1$ [35].

Η λογιστική συνάρτηση είναι η εξής: $g(z) = \frac{1}{1+ e^{-z}} = \frac{e^{z}}{e^{z+1}}$ (όπως αναπαριστάται στην εικόνα 14)



Εικόνα 14 : Απεικόνιση σιγμοειδούς καμπύλης

2.7.3 Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines, SVMs)

Ο όρος Support Vector Machines, SVMs ή Μηχανές Διανυσμάτων Υποστήριξης, αναφέρεται σε μοντέλα μάθησης με επίβλεψη με συσχετιζόμενους αλγορίθμους μάθησης που αναλύουν δεδομένα τόσο για ταξινόμηση (classification) όσο και για παρεμβολή

ή παλινδρόμηση (regression). Η λειτουργία τους στηρίζεται στη θεωρία στατιστικής μάθησης και στα νευρωνικά δίκτυα τύπου Perceptron.

Για την ταξινόμηση οι μηχανές SVM κάνουν χρήση ταξινομητή μεγίστου περιθωρίου (Maximal Margin Classifier) με σκοπό να βρουν μία υπερεπιφάνεια (hypersurface) η οποία τα αρνητικά από τα θετικά με τρόπο ώστε να διαχωρίζονταιμε το μέγιστο περιθώριο δηλαδή να απέχει η επιφάνεια αυτή όσο γίνεται περισσότερο από τα κοντινότερα θετικά και αρνητικά παραδείγματα. Στην περίπτωση ταξινόμησης με μέγιστο περιθώριο, οι αριθμητικές μεταβλητές εισόδου των δεδομένων σχηματίζουν ένα χώρο nδιαστάσεων. Το υπερεπίπεδο είναι μια γραμμή που διαχωρίζει των χώρο των μεταβλητών εισόδου. Στις μηχανές υποστήριξης ένα υπερεπίπεδο χρησιμοποιείται για τον διαχωρισμό σημείων στον χώρο των μεταβλητών εισόδου ανάλογα με την κλάση τους, κλάση 0 ή κλάση 1. Με δύο διαστάσεις γίνεται να σχεδιαστεί αυτή η γραμμή και αν θεωρηθείότι όλα τα σημεία εισόδουμπορούν να χωριστούν από αυτήν θα υπάρχει : B0 + (B1 * X1) + (B2 * X2) = 0.

Όπου οι συντελεστές B1 και B2 που καθορίζουν την κλίση της γραμμής και η τομή B0 υπολογίζονται από τον αλγόριθμο μάθησης, ενώ τα X1 και X2 είναι μεταβλητές εισόδου. Αυτή η γραμμή μπορεί να χρησιμοποιηθεί γιαταξινόμηση και με την εισαγωγή νέων μεταβλητών μπορεί να υπολογισθεί αν το νέο σημείο είναι πάνω ή κάτω από την γραμμή.

Πιο συγκεκριμένα, πάνω από την γραμμή, η εξίσωση έχει αποτέλεσμα μεγαλύτερο του 0 και το σημείο ανήκει στην κλάση 0, κάτω από την γραμμή, η εξίσωση έχει αποτέλεσμα μικρότερο του 0 και το σημείο ανήκει στην κλάση 1. Επίσης, μια τιμή κοντά στην γραμμή έχει αποτέλεσμα κοντάστο 0 και είναι δύσκολο να ταξινομηθεί ενώ σε περίπτωση που το μέγεθος της τιμής είναι μεγάλο τότε το μοντέλο θα έχει πιο αξιόπιστη πρόβλεψη.

Η πιο ιδανική γραμμή για να διαχωριστούν δύο κλάσεις είναι η γραμμή με το μεγαλύτερο περιθώριο (margin) και ονομάζεται υπερεπίπεδο μεγίστου περιθωρίου. Το περιθώριο υπολογίζεται ως η κάθετη απόσταση που σχηματίζεται από την γραμμή προς τα κοντινότερα σημεία. Τα σημεία αυτά σχετίζονται με τον καθορισμό της γραμμής και την κατασκευή του ταξινομητή. Τα σημεία ονομάζονται διανύσματα υποστήριξης και αυτό που κάνουν είναι να καθορίζουν το υπερεπίπεδο. Το υπερεπίπεδο μαθαίνεται από τη Μηχανή Διανυσμάτων Υποστήριξης μέσα από το σύνολο δεδομένων εισόδου (training set) χρησιμοποιώντας μια διαδικασία βελτιστοποίησης που μεγιστοποιεί το περιθώριο.

Στην πραγματικότητα όμως τα δεδομένα είναι ανακατεμένα και δεν είναι δυνατός ο τέλειος διαχωρισμός με ένα υπερεπίπεδο. Πρέπει λοιπόν να ελλατωθεί ο περιορισμός της μεγιστοποίησης του περιθωρίου της γραμμής που διαχωρίζει τις κλάσεις. Αυτό αποκαλείται Soft Margin Classifier με αυτόν τον τρόπο δίνεται η δυνατότητα σε κάποια σημεία του συνόλου δεδομένων εκπαίδευσης να παραβιάσουν την διαχωριστική γραμμή.

Ένα ακόμη σύνολο συντελεστών δίνουν στο περιθώριο περισσότερο χώρο σε κάθε διάσταση και αποκαλούνται χαλαρές μεταβλητές (slack variables). Η πολυπλοκότητα του

μοντέλου αυξάνεται μιας και υπάρχουν περισσότερες παράμετροι για το μοντέλο που πρέπει να προσασμόσει στα δεδομένα. Επιπλέον γίνεται εισαγωγή μίας ρυθμιστικής παραμέτρου που ονομάζεται C η οποία καθορίζει το μέγεθος της κίνησης που επιτρέπεται σε όλες τις διαστάσεις. Οι παράμετροι C καθορίζουν πόση παραβίαση του περιθωρίου επιτρέπεται. Πιο συγκεκριμένα, αν το C=0 σημαίνει ότι δεν επιτρέπεται καμία παραβίαση και άρα γίνεται λόγος για το άκαμπτο Maximal-Margin Classifier όσο αυξάνεται ο αριθμός του C αυξάνεται και ο αριθμός των παραβιάσεων που επιτρέπονται να γίνουν. Το C επηρεάζει επίσης και τον αριθμό των διανυσμάτων υποστήριξης που χρησιμοποιούνται από το μοντέλο. Όσο μικρότερο είναι το C τόσο πιο ευεπηρέαστος είναι ο αλγόριθμος στα δεδομένα εκπαίδευσης (training data), άρα υπάρχει υψηλότερη διακύμανση (variance) και χαμηλότερη μεροληψία (bias). Ενώ, όσο μεγαλύτερο είναι το C τόσο λιγότερο ευεπηρέαστος είναι ο αλγόριθμος στο training data άρα υπάρχει χαμηλότερη διακύμανση (variance) και υψηλότερη μεροληψία (bias).

Εαν οι κλάσεις δεν είναι γραμμικά διαχωρίσιμες και χρειάζεται να αποφευχθεί η οποιαδήποτε παραβίαση του περιθωρίου είναι δυνατόν να χρησιμοποιηθεί ο κατάλληλος μη-γραμμικός μετασχηματισμός υποθέτονας ότι τα μετασχηματισμένα πρότυπα είναι πλέον γραμμικά διαχωρίσιμα.

Μια σημαντική σημείωση είναι ότι η γραμμική μηχανή SVM μπορεί να αναδιατυπωθεί χρησιμοποιώντας το εσωτερικό γινόμενοδύο οποιωνδήποτε παρατηρήσεων αντί γιατις ίδιες τις παρατηρήσεις . Το εσωτερικό γινόμενο μεταξύ δύο διανυσμάτων είναι το άθροισμα από τον πολλαπλασιασμό του κάθε ζεύγους τιμών εισόδου.Για παράδειγμα, το εσωτερικό γινόμενο των διανυσμάτων [4, 2] και [8 , 5] είναι 4 * 8 + 2 * 5 ή 42.

Αν κάθε μεταβλητή $x^{(i)}$αντικατασταθεί με την μετασχηματισμένη μορφή $\Phi(x^{(i)})$ τότε τόσο στη συνάρτηση κόστους όσο και στη βέλτιστη διαχωριστική επιφάνεια εμφανίζονται εσωτερικά γινόμενα της μορφής $\varphi(x^{(i)})^T\Phi(y^{(i)})$ και έτσι ορίζεται η συνάρτηση k: $k = \varphi(x^{(i)})^T\Phi(y^{(i)})$.
Η συνάρτηση κονομάζεται συνάρτηση πυρήνα (kernel).

Ο αλγόριθμος SVM εφαρμόζεται χρησιμοποιώντας ένα πυρήνα (kernel). Η μάθηση του υπερεπιπέδου στο γραμμικό SVM γίνεται μετατρέποντας το πρόβλημα χρησιμοποιώντας γραμμική άλγεβρα.

Η εξίσωση για την πρόβλεψη μιας νέας εισόδου με τη χρήση του εσωτερικού γινομένου μεταξύ της εισόδου (x) και κάθε διανύσματος υποστήριξης ($x^{(i)}$) υπολογίζεται ως εξής :

$$f(x) = B_0 \sum_{i=1}^{m} a_{(i)(x * x^{(i)}}$$

Αυτή είναι μια εξίσωση που περιλαμβάνει τον υπολογισμό των εσωτερικών γινομένων ενός νέου διανύσματος εισόδου (x) με όλα τα διανύσματα υποστήριξης στο σύνολο δεδομένων εκπαίδευσης (training data). Οι συντελεστές B_0 και $a_{(i)}$ (για κάθε είσοδο) θα πρέπει να εκτιμηθούν από τα δεδομένα εκπαίδευσης μέσω του αλγορίθμου μάθησης.

Το εσωτερικό γινόμενο στην περίπτωση της Μηχανής Διανυσμάτων Υποστήριξης Γραμμικού πυρήνα ονομάζεται πυρήνας και μπορεί να γραφεί ως:

$$K(X, X^{(i)}) = \sum_{i=1}^{m} XX^{(i)}$$

Ο πυρήνας καθορίζει την ομοιότητα ή την απόσταση μεταξύ των νέων δεδομένων και των διανυσμάτων υποστήριξης . Το εσωτερικό γινόμενο είναι το μέτρο ομοιότητας που χρησιμοποιείται για γραμμική μηχανή SVM ή ένα γραμμικό πυρήνα επειδή η απόσταση είναι γραμμικός συνδυασμός των εισόδων. Μπορούν να χρησιμοποιηθούν και άλλοι πυρήνες που μετατρέπουν το χώρο εισόδου σε υψηλότερες διαστάσεις, όπως ένας πυρήνας

πολυωνύμου ή ένας πυρήνας ακτίνας . Αυτό ονομάζεται Kernel trick .Είναι επιθυμητό να χρησιμοποιηθούν πιο πολύπλοκοι πυρήνες καθώς επιτρέπουν στις γραμμές να διαχωρίσουν τις κλάσειςπου είναι καμπυλωτές ή και πιο πολύπλοκες. Αυτό με τη σειρά του μπορεί να

οδηγήσει σε ταξινομητές υψηλότερης ακρίβειας. Σε Μηχανή Διανυσμάτων Υποστήριξης πυρήνα πολυωνύμου, αντί του γινομένου μπορεί να χρησιμοποιηθεί ένας πυρήνας πολυωνύμου για παράδειγμα:

$$K\left(X, X^{(i)}\right) = 1 + \sum_{i=1}^{m} (XX^{(i)})^d$$

Όπου ο βαθμός του πολυωνύμου dπρέπει να καθορίζεται χειροκίνητα για τον αλγόριθμο μάθησης . Όταν d=1, είναι το ίδιο με το γραμμικό πυρήνα. Ο πυρήνας πολυωνύμου επιτρέπει καμπυλωτές γραμμές στο χώρο εισόδου[36].

Προετοιμασία δεδομένων για SVM :

Αριθμητικοί Είσοδοι: Το σύστημα SVM υποθέτει ότι οι είσοδοι είναι αριθμητικοί. Αν υπάρχουν είσοδοι σε κατηγορίες θα πρέπει να μετατραπούν σε δυαδικές ψευδομεταβλητές (μία μεταβλητή για κάθε κατηγορία)

Δυαδική Ταξινόμιση (Binary Classification): Το βασικό SVM προορίζεται για προβλήματα δυαδικής ταξινόμισης. Παρ'όλα αυτά έχουν αναπτυχθεί λύσεις για προβλήματα παρεμβολής (regression) και ταξινόμησης πολλαπλών τάξεων (multi-class classification).

Στις εφαρμογές μηχανών διανυσμάτων υποστήριξης η συνάρτηση κόστους (costfunction) που πρέπει να ελαχιστοποιηθεί είναι συνήθως της μορφής:

$$min_\theta \sum_{i=1}^{m} y^{(i)}\, cost_1(\theta^T X^{(i)}) + \left(1 - y^{(i)}\right) cost_0\,(\theta^T X^{(i)}) + \frac{1}{2} \sum_{j=0}^{n} \theta j^2$$

Όπου $cost_1 = -\log\, h_\theta(x^{(i)})$ και $cost_0 = -\log(1 - h_\theta(x^{(i)}))$.

Και με συνάρτηση υπόθεσης της πιο πολλές φορές να δίνεται από τον τύπο:

$$h_\theta(x) = \begin{cases} 1\ \alpha\nu\ \ (\theta^T x\ \geq 0 \\ 0\ \sigma\tau\iota\varsigma\ \upsilon\pi\acute{o}\lambda o\iota\pi\varepsilon\varsigma\ \pi\varepsilon\rho\iota\pi\tau\acute{\omega}\sigma\varepsilon\iota\varsigma \end{cases}$$

2.7.3.1 Πλεονεκτήματα των SVMs

Κάποια από τα πλεονεκτήματα του SVM αλγορίθμου είναι τα παρακάτω:

- Έχουν την ικανότητα να διαχειρίζονται δεδομένα πολλών διαστάσεων, αφού η πολυπλοκότητα είναι ανεξάρτητη από την διάσταση του χώρου των δεδομένων.

- Παρουσιάζουν ευελιξία στην επιλογή της συνάρτησης του πυρήνα και έχουν σημαντική ικανότητα γενίκευσης σε μη γραμμικά διαχωρίσιμα δεδομένα ενσωματώνοντας το τέχνασμα του πυρήνα (kernel trick). Με αυτό τον τρόπο είναι δυνατή η παραγωγή μη γραμμικών μοντέλων οι οποίοι οδηγούν σε γραμμικότητα σε χώρους μεγαλύτερων διαστάσεων.

- Είναι δυνατός ο έλεγχος της υπερεκπαίδευσης των δεδομένων με την εφαρμογή της προσέγγισης των χαλαρών ορίων (slack variables).

- Έχουν την ικανότητα να διαχωρίσουν μεγάλου πλήθος δεδομένων, αφού αρκούν μόνο οι μηχανές διανυσμάτων για να ορίσουν το επίπεδο διαχωρισμού.

Για την εισαγωγή ενός νέου στοιχείου σε μία κλάση, η διαδικασία ταξινόμησης ελέγχει μόνο την ομοιότητα του άγνωστου προς τον αλγόριθμο στοιχείου και των σημαντικότερων στοιχείων της κάθε κλάσης τα οποία είναι τα διανύσματα υποστήριξης (support vectors). Έτσι ο αλγόριθμος απαλλάσσεται από την σύγκριση μία παρατήρησης με όλα τα γνωστά δεδομένα μειώνοντας το υπολογιστικό του κόστος.

2.7.4 Βασικά Βηματα για Μηχανική Μάθηση

Τα στάδια τα οποία ακολουθούνται στην διαδικασία της μηχανικής μάθησης είναι τα παρακάτω (όπως φαίνονται και στην εικόνα 15) :

Επιλογή Δεδομένων : Δημιουργείται το σύνολο δεδομένων στο οποίο θα εφαρμοστεί η αναζήτηση (training set) με επιλογή στοιχείων (πινάκων, πεδίων) από σχεσιακές βάσεις δεδομένων.

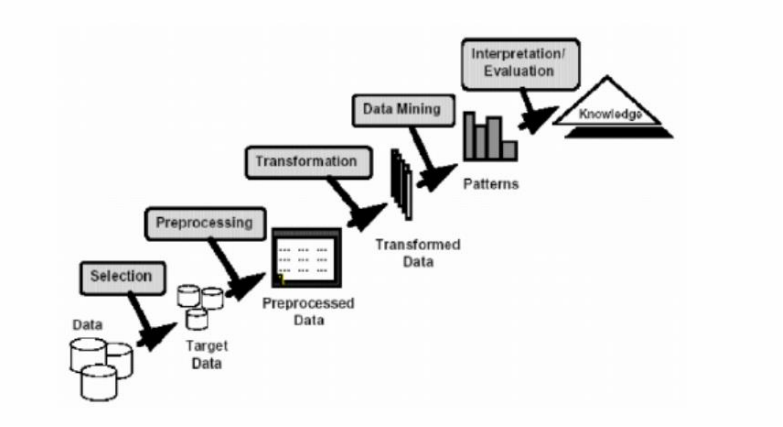
Προεπεξεργασία (preprocessing) Δεδομένων : Αντιμετωπίζονται περιπτώσεις ελλιπών δεδομένων, πεδίων με τιμές που ουσιαστικά τα καθιστούν κενά και μηδενικές τιμές.

Μετασχηματισμός δεδομένων : Τα δεδομένα μετασχηματίζονται με τέτοιο τρόπο ώστε να διευκολύνουν την εξόρυξη γνώσης. Οι μετασχηματισμοί αυτοί περιλαμβάνουν για

παράδειγμα: τη μείωση του αριθμού των υπό-εξέταση χαρακτηριστικών με επιλογή μερικών από αυτά, την ομοιόμορφη κωδικοποίηση της ποιοτικά ίδιας πληροφορίας καθώς και την την μετατροπή των συνεχόμενων αριθμητικών τιμών σε διακριτές.

Επιλογή αλγόριθμου και εφαρμογή του : Καθορίζεται τι είδους γνώση θα αναζητηθεί, κάτι που έμμεσα προσδιορίζει και την κατηγορία αλγορίθμου που θα χρησιμοποιηθεί. Πρόκετια για ένα υπολογιστικό στάδιο, στο οποίο γίνεται η ουσιαστική αναζήτηση της γνώσης στα δεδομένα το οποίο περιγράφεται με τον όρο εξόρυξη σε δεδομένα (data mining).

Ερμηνεία και αξιολόγηση: Πραγματοποιείται ερμηνεία και αξιολόγηση των αποτελεσμάτων, πιθανώς με την βοήθεια απεικονιστεικών μεθόδων ή και των δεδομένων [37]. Επίσης για τη διασφάλιση της ορθότητας των αποτελεσμάτων χρεισημοποιούνται οι τεχνικές επικύρωσης (cross validation) οι οποίες αναφέρονται στη υποενότητα (1.8).



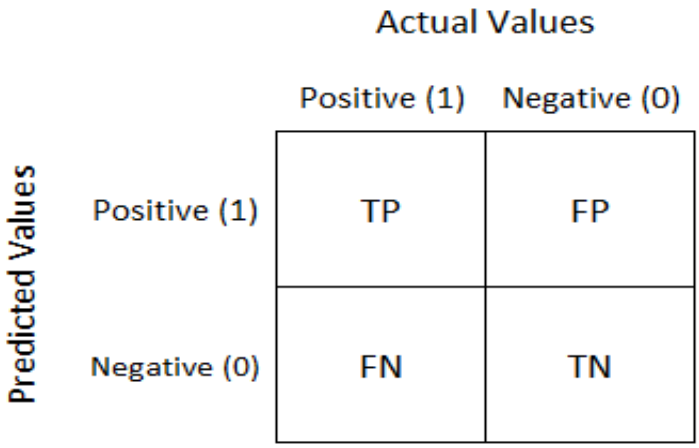
Εικόνα 15 : Διαδικασία της Μηχανικής Μάθησης [37].

2.8 Πίνακας Σύγχυσης (Confusion Matrix)

Απαραίτητη είναι επίσης η εκτίμηση της απόδοσης ενός μοντέλου η οποία γίνεται με διαφορετικές μεθόδους. Η αποτελεσματικότητα ενός ταξινομητή εξαρτάται από την ικανότητά του να μπορεί διαχωρίσει σωστά τα δεδομένα εισόδου στις διάφορες κλάσεις εξόδου. Στην προκειμένη μελέτη οι κατηγορίες εξόδου είναι δύο και αντιπροσωπεύουν την ύπαρξη καρκίνου. Για την κατανόηση των κριτηρίων αξιολόγησης πρέπει να δοθούν οι ορισμοί των παρακάτω μεταβλητών. Στη δυαδική ταξινόμηση (ταξινόμηση των δεδομένων σε δύο κατηγορίες) χρησιμοποιούνται συχνά οι έννοιες True Positive, True Negative, False Positive και False Negative. Σε περίπτωση που θεωρηθεί ότι οι κατηγορίες στι οποίες τοποθετούνται τα στοιχεία είναι θετικά (positive) και αρνητικά (negative), τότε μπορεί να οριστεί ως True Positive (TP, Αληθή Θετικά) ο αριθμός των εξόδων που

κατηγοριοποιήθηκαν ως θετικά από το σύστημα και ήταν όντως θετικά, ως True Negative (TN, Αληθή Αρνητικά) ο αριθμός των εξόδων που κατηγοριοποιήθηκαν ως αρνητικά και ήταν όντως αρνητικά, FalsePositive (FP, Ψευδή Θετικά) ο αριθμός των εξόδων που κατηγοριοποήθηκαν ως θετικά ενώ ήταν κανονικά αρνητικά και ως False Negative (FN, Ψευδή Αρνητικά) ο αριθμός των εξόδων που κατηγοριοποιήθηκαν ως αρνητικά ενώ στην πραγματικότητα ήταν θετικά.

Αυτές οι τέσσερις έννοιες μπορούν να τοποθετηθούν σε ένα πίνακα Σύγχυσης (Confusion Matrix) ώς εξής :



Εικόνα 16 : Πίνακας Σύγχυσης

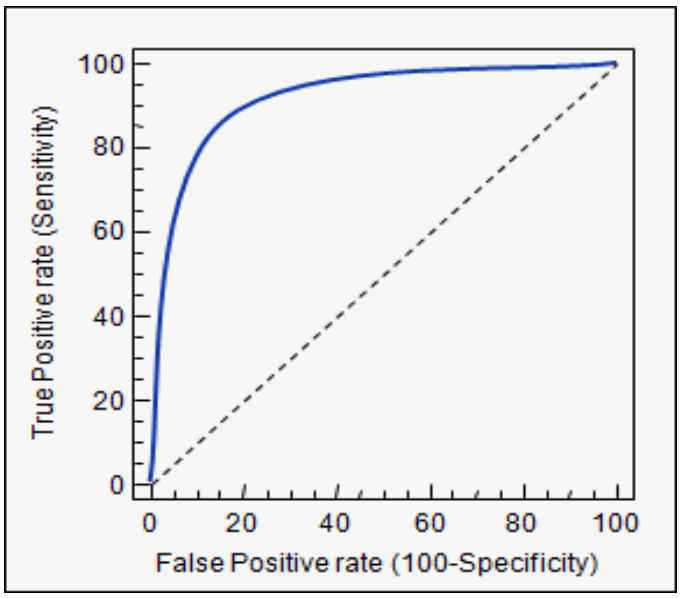
Το άθροισμα των τιμών που βρίσκονται στη διαγώνιο του πίνακα (TP+ TN) αποτελεί το σύνολο των στοιχείων που κατηγοριοποιήθηκαν σωστά. Έτσι αποτελεί ένα εργαλείο που εύκολα δείχνει την ικανότητα ταξινόμησης των αποτελεσμάτων του συστήματος. Ένα

θεωρητικό σύστημα που θα είχε κατηγοριοποιήσει σωστά όλες τις τιμές, θα έδινε ένα διαγώνιο πίνακα, δηλαδή θα υπήρχε 0 σε κάθε άλλη θέση πλην της διαγωνίου (αij=0$\forall i \neq j$)[38].

Μέσω του πίνακα είναι εύκολο να υπολογιστούν τα παρακάτω: [39]

- Ορθότητα (Aqquracy) : $\frac{TP+TN}{Σύνολο} = \frac{TP+TN}{TP+TN+FP+FN}$
- Ακρίβεια (Precision) : $\frac{TP}{Προβλεφθέντα\ θετικά} = \frac{TP}{TP+FP}$
- Ευαισθησία (Sensitivity) : $\frac{TP}{Προβλεφθέντα\ θετικά} = \frac{TP}{TP+FN}$
- Ειδικότητα (Specificity) ή True Negative Rate (TNR) : $\frac{TN}{TN+FP}$
- Ψευδής Θετικός Ρυθμός (False Positive Rate (FPR)) : $\frac{FP}{Πραγματικά\ Αρνητικά} = \frac{FP}{FP+TN}$
- F1 score ή F-measure : $2\ \frac{Precision*Recall}{Precision+Recall} = \frac{2*TP}{2*TP+FP+FN}$
- Matthews Correlation Coefficient (MCC): $\frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)+(TN+FP)(TN+FN)}}$

- Καμπύλη ROC : είναι ένα γράφημα το οποίο αναπαριστά την απόδοση ενός μοντέλου ταξινόμησης σε όλα τα κατώφλια (εικόνα 17). Για την υλοποίηση της καμπύλης αρκεί η γραφική παράσταση δύο παραμέτρων, του TPR (άξονας y) και του FPR (άξονας x). Η καμπύλη αυτή αντιπροσωπεύει τον βαθμό ή το μέτρο της διαχωριστικότητας. Για παράδειγμα, ένα μοντέλο με υψηλότερο AUC (area under the curve) είναι καλύτερο στην πρόβλεψη των True Positives και True Negatives από ένα με χαμηλότερο. Επιπλέον, η βαθμολογία της AUC μετρά το συνολικό εμβαδόν κάτω από την καμπύλη ROC, η οποία μπορεί να υπολογιστεί με απλές τεχνικές αριθμητικές προσεγγίσεις (Μέθοδος τραπεζίου, Simspon κ.α). Τέλος, η AUC είναι αναλλοίωτη κλίμακα και επίσης αναλλοίωτη τιμή κατωφλίου[40].



Εικόνα 17 : Καμπύλη ROC[40].

2.9 Τεχνικές Επικύρωσης

2.9.1 K-Fold Cross Validation

Η μέθοδος Cross Validation είναι μια διαδικασία που μπορεί να χρησιμοποιηθεί για την εκτίμηση της ποιότητας ενός αλγορίθμου μηχανικής μάθησης. Όταν εφαρμόζεται σε διάφορα πλήθη δεδομένων με διαφορετικές ελεύθερες τιμές παραμέτρων τα αποτελέσματα της διασταυρούμενης επικύρωσης (Cross-validation) μπορούν να χρησιμοποιηθούν για την επιλογή του καλύτερου συνόλου τιμών παραμέτρων.

Ένας τύπος cross validation που αποτελεί μια μέθοδο αξιολόγησης και επικύρωσης των αλγορίθμων μάθησης είναι η μέθοδος k-fold Cross Validation. Η βασική ιδέα είναι ο

διαχωρισμός των δεδομένων μας σε k υποσύνολα τα οποία είναι ίσα ή σχεδόν ίσα μεταξύ τους (το k συνήθως είναι ίσο με 10 και αυτό χρησιμοποιήθηκε και σε αυτή την υλοποίηση). Στη συνέχεια, διατρέχουμε τον αλγόριθμο εκπαίδευσης k φορές. Την πρώτη φορά που τρέχουμε τον αλγόριθμο, χρησιμοποιούμε τα k-1 από τα k σύνολα δεδομένων για εκπαίδευση και έπειτα υπολογίζουμε την ακρίβεια του αλγορίθμου μηχανικής μάθησης χρησιμοποιώντας το σύνολο που περισσεύει. Αυτή η διαδικασία επαναλαμβάνεται k φορές έτσι ώστε κάθε ένα από τα k υποσύνολα να χρησιμοποιηθεί ως σύνολο επικύρωσης (εικόνα 18). Με το πέρας της διαδικασίας, το τελικό αποτέλεσμα είναι ο μέσος όρος όλως των τμημάτων ,τον οποίο θα χρησιμοποιήσουμε ως συνολική εκτίμηση της ακρίβειας του αλγορίθμο.

Σύμφωνα με τη διαδικασία αυτή το σύνολο των δεδομένων X χωρίζεται σε K υποσύνολα ίσου μεγέθους. Κάθε φορά, το k-οστό υποσύνολο σχηματίζει ένα σύνολο ελέγχουν X_{val},

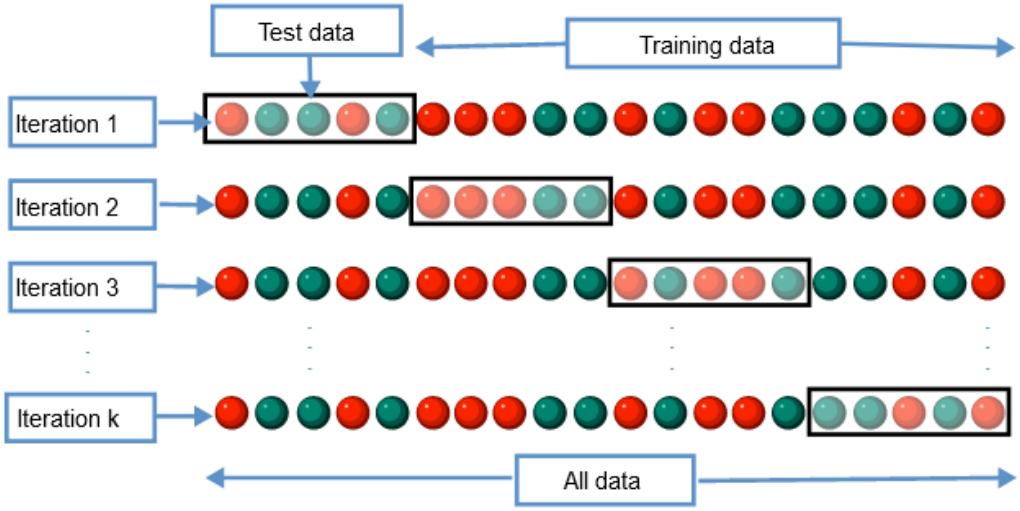
ενώ τα υπόλοιπα αποτελούν το σύνολο εκπαίδευσης X_{learn}που χρησιμοποιείται για την ανάπτυξη του μοντέλου. Το σφάλμα υπολογίζεται από τον τύπο :

$$\mathrm{Err}^{cvk} = \frac{\varepsilon_{i=1}^{m/k} (g(x_i{}^{val}) - ((y_i{}^{val})^2}{m/K}$$

Με $(x_i{}^{val}, y_i{}^{val})$ τα στοιχεία του X_{val} και $g(x_i{}^{val})$ η προσέγγιση του $y_i{}^{val}$ απ' το μοντέλο g. Η παραπάνω διαδικασία επαναλαμβάνεται για όλα τα K υποσύνολα, δηλαδή από το 1 μέχρι το K [41].

2.9.2 Leave one out Cross Validation

Μια πιο συγκεκριμένη περίπτωση της μεθόδου k-fold Cross-Validation, είναι η Leave-One-Out όπου εκεί το K είναι ίσο με το m. Δηλαδή, κάθε φορά μένει έξω από το σύνολο εκπαίδευσης ένα μόνο σημείο, αυτό για το οποίο πρόκειται να γίνει η πρόβλεψη. Πιο συγκεκριμένα, περιλαμβάνει τη χρήση ενος εκ των δεδομένων ως σύνολο επικύρωσης και τις υπόλοιπες παρατηρήσεις ως σύνολο εκπαίδευσης και επαναλαμβάνεται για όλα τα δεδομένα [41].



Εικόνα 18 : Μέθοδος k-fold cross validation [41].

2.10 Επιλογή Χαρακτηριστικών

Η επιλογή χαρακτηριστικών χρησιμοποιείται στους τομείς όπως : η αναγνώριση προτύπων, η στατιστική, και κυρίως στην εξόρυξη δεδομένων. Η κύρια ιδέα της επιλογής χαρακτηριστικών είναι να επιλέγει ένα υποσύνολο μεταβλητών εξαλείφοντας στοιχεία με μικρή ή περιττή πληροφορία. Η επιλογή χαρακτηριστικών μπορεί να βελτιώσει σημαντικά την απόδοση των ταξινομητών (classifiers) και να βοηθήσει στην δημιουργία ενός μοντέλου που επεξεργάζεται καλύτερα περίπλοκα δεδομένα. Κατά την προσέγγιση επιλογής χαρακτηριστικών, βασικός σκοπός είναι να εντοπιστούν και απορριφθούν εκείνα τα χαρακτηριστικά τα οποία παρέχουν περιττή πληροφορία. Με τον τρόπο αυτό μειώνεται η διάσταση των δεδομένων αφού τα περιττά διανύσματα δεν συμμετέχουν στις περαιτέρω διαδικασίες.

2.10.1 Μέθοδοι επιλογής χαρακτηριστικών

Η επιλογή χαρακτηριστικών περιλαμβάνει μεθόδους οι οποίες αναφέρονται είτε στο σύνολο των δεδομένων του κάθε διανύσματος (μη επιβλεπόμενες μέθοδοι) είτε στη γραφική και στατιστική ανάλυση προκειμένου να καθοριστεί ο βαθμός της διαχωριστικότητας μεταξύ των κλάσεων (επιβλεπόμενες μέθοδοι).Οι πιο γνωστοί μέθοδοι επιλογής υποσυνόλου χαρακτηριστικών είναι οι παρακάτω δυο:

Μέθοδος φίλτρου (Filter): Η ανεξάρτητη επιλογή, που είναι βασισμένη στα γενικά χαρακτηριστικά των δεδομένων .

Μέθοδος περιτυλίγματος (Wrapper): Για να βρεθεί το διάστημα όλων των υποσυνόλων χαρακτηριστικών γνωρισμάτων, ένας αλγόριθμος αναζήτησης είναι συνέχεια τυλιγμένος (wrapped) γύρω από το πρότυπο ταξινόμησης.

2.10.1.1 Μέθοδος φίλτρου (Filter)

Οι μέθοδοι χρησιμοποιούνται ως στάδιο προεξεπεργασίας Η επιλογή των χαρακτηριστικών είναι ανεξάρτητη από οποιονδήποτε αλγόριθμο μηχανικής μάθησης. Τα χαρακτηριστικά επιλέγονται βάση των βαθμολογιών τους σε διάφορες στατιστικές δοκιμές για τη συσχέτιση τους με τη μεταβλητή εξόδου. Τα χαρακτηριστικά γνωρίσματα τα οποία δεν έχουν υψηλή σχετικότητα αφαιρούνται. Όσα απομένουν χρησιμοποιούνται ως εισαγωγή στον αλγόριθμο ταξινόμησης . Η τεχνική αυτή έχει το πλεονέκτημα ότι είναι απλή και γρήγορη σε περιπτώσεις υπολογιστικής πολυπλοκότητας που συναντάται κυρίως στα δεδομένα υψηλής διάστασης (όπως δεδομένα DNA, κείμενο κτλ) και είναι ανεξάρτητη από τον αλγόριθμο ταξινόμησης ενώ ως μειονέκτημα αξίζει να σημειωθεί ότι δεν αφαιρεί την την πολυγραμμικότητα. Επομένως, πρέπει να εξεταστεί η πολυγραμμικότητα των χαρακτηριστικών πριν από την εκπαίδευση του μοντέλου. Τεχνικές οι οποίες χρησιμοποιούν την μέθοδο φίλτρου για παράδειγμα είναι η Γραμμική ανάλυση διακρίσεων (LDA) και η εύρεση του Chi-Square. Η ανάλυση γραμμικής διάκρισης χρησιμοποιείται για την εύρεση ενός γραμμικού συνδυασμού χαρακτηριστικών που χαρακτηρίζει ή διαχωρίζει δύο ή περισσότερες κατηγορίες (ή επίπεδα) μιας κατηγορηματικής μεταβλητής ενώ όσον αφορά Chi-Square πρόκειται για ένα στατιστικό τεστ που εφαρμόζεται στις ομάδες κατηγορηματικών χαρακτηριστικών για την αξιολόγηση της πιθανότητας συσχέτισης ή συσχέτισης μεταξύ τους χρησιμοποιώντας την κατανομή συχνότητας.

Εικόνα 19 : Μέθοδος Φίλτρου

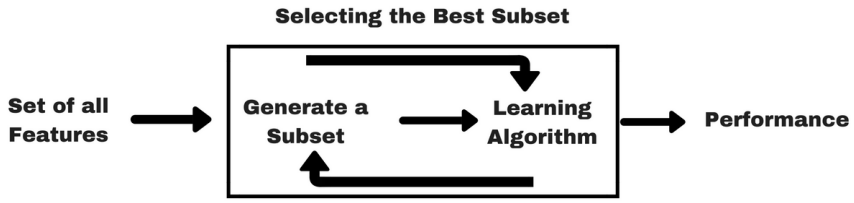
2.10.1.2 Μέθοδος περιτυλίγματος (Wrapper)

Η μέθοδος αυτή ενσωματώνει τους ταξινομητές στην αναζήτηση υποσυνόλων χαρακτηριστικών γνωρισμάτων. Τα χαρακτηριστικά γνωρίσματα συνήθως αξιολογούνται σε ομάδες και όχι μεμονωμένα. Τα ήδη υπάρχονται διαθέσιμα χαρακτηριστικά γνωρίσματα χρησιμοποιούνται με στόχο την παραγωγή των υποσύνολων χαρακτηριστικών γνωρισμάτων που τίθενται σε αξιολόγηση. Η αξιολόγηση αυτή πραγματοποιείται με την κατάρτιση και τη δοκιμή ενός συγκεκριμένου προτύπου ταξινόμησης.

Με σκοπό να βρεθεί το διάστημα όλων γνωρισμάτων, ένας αλγόριθμος αναζήτησης είναι τυλιγμένος (wrapped) γύρω από το πρότυπο ταξινόμησης . Με αυτόν τον τρόπο, καθώς το υποσυνόλο των χαρακτηριστικών γνωρισμάτων αυξάνεται εκθετικά σε σχέση με τον

αριθμό χαρακτηριστικών γνωρισμάτων, οι ευρετικές μέθοδοι αναζήτησης χρησιμοποιούνται ώστε να καθοδηγήσουν την αναζήτηση ενός βέλτιστου υποσυνόλου. Στα πλεονεκτήματα της μεθόδου αυτής περιλαμβάνεται η δυνατότητα να ληφθούν υπόψη οι εξαρτήσεις μεταξύ των χαρακτηριστικών γνωρισμάτων. Αντίθετα, μειονέκτημα είναι ότι υπάρχει μεγαλύτερο ρίσκο για υπερφόρτωση (overfiting) σε σχέση με τις μεθόδους Wrapper, και συνήθως υπολογιστικά η μέθοδος αυτή είναι πολλή ακριβή.

Μερικά παραδείγματα μεθόδων περιτύλιξης είναι η επιλογή χαρακτηριστικών προς τα εμπρός (Forward Selection), η εξάλειψη χαρακτηριστικών προς τα πίσω (Backward Elimination), η αναδρομική εξάλειψη χαρακτηριστικών (Recursive Feature elimination) κ.λπ. Η Forward Selection τεχνική είναι μια επαναληπτική μέθοδος η οποία αρχίζει την διαδικασία χωρίς να υπάρχει δυνατότητα στο μοντέλο. Σε κάθε επανάληψη, προσθέτει εκείνο το χαρακτηριστικό που βελτιώνει καλύτερα το μοντέλο έως ότου η προσθήκη ενός νέου δεν βελτιώνει την απόδοση του. Αντίθετα, η Backward Elimination ξεκινάει με όλα τα χαρακτηριστικά και καταργεί το λιγότερο σημαντικό σε κάθε επανάληψη με σκοπό την βελτίωση της απόδοσης του μοντέλουκαι η επανάληψη συνεχίζεται μέχρι ώστε να μην παρατηρηθεί βελτίωση κατά την κατάργηση των χαρακτηριστικών. Τέλος, όσον αφορά την τεχνική Recursive Feature elimination πρόκειται για ένας αλγόριθμο βελτιστοποίησης που στοχεύει στην εύρεση του υποσυνόλου χαρακτηριστικών με την καλύτερη απόδοση. Δημιουργεί επανειλημμένα μοντέλα και διατηρεί κατά μέρος την καλύτερη ή τη χειρότερη απόδοση σε κάθε επανάληψη. Επίσης, κατασκευάζει το επόμενο μοντέλο με τα χαρακτηριστικά τα οποία έχουν απομείνει μέχρι να εξαντληθούν όλα και στη συνέχεια κατατάσσει τα χαρακτηριστικά με βάση τη σειρά της εξάλειψής τους [42].



Εικόνα 20 : Μέθοδος περιτυλίγματος

2.10.2 Επιλογή Χαρακτηριστικών με τον αλγόριθμο K-Best

Ο αλγόριθμος παίρνει ως παράμετρο μια συνάρτηση η οποία υπολογίζει κάποιο σκορ το οποίο ουσιαστικά εκτελεί μία αξιολόγηση όπου πρέπει να ισχύει για ένα ζεύγος (X, y) όπου είναι τα εκπαιδευόμενα δεδομένα (Xtrain, ytrain) ενώ το K ως όρισμα αντιπροσωπεύει τον αριθμό των χαρακτηριστικών τα οποία θα επιλεγούν. Η συνάρτηση σκορ επιστρέφει μια σειρά από βαθμολογίες (σκορ), μία για κάθε χαρακτηριστικό X [:, i] του X. Έχει την δυνατότητα επιπλέον να επιστρέψει p-values, αλλά αυτές δεν είναι απαραίτητες. Στη συνέχεια, ο K-Best διατηρεί τα πρώτα χαρακτηριστικά του X με τις υψηλότερες βαθμολογίες. Η χρήση της default συνάρτησης του KBest με την κλήση έπειτα της

select.fit_transform (Xtrain, ytrain) χρησιμοποιεί τη συνάρτηση f_classif σκορ. Αυτό ερμηνεύει τις τιμές του y ως ετικέτες κλάσης. Θα μπορούσε επίσης να γίνει χρήση της chi-square . Έτσι για παράδειγμα, εάν αντί του f_classif χρησιμοποιηθεί το chi2 ως συνάρτηση σκορ, θα υπολογιστεί το chi2 μεταξύ κάθε χαρακτηριστικού του X και y (υποτίθεται ότι

είναι ετικέτες τάξης). Μια μικρή τιμή chi-square σημαίνει ότι το χαρακτηριστικό αυτό είναι ανεξάρτητο από το y. Μια μεγάλη τιμή θα σημαίνει ότι το χαρακτηριστικό αυτό δεν σχετίζεται τυχαία με το y και είναι πιθανό να παρέχει σημαντικές πληροφορίες.

3. Μεθοδολογία

Σε συνεργασία με το παθολογικό τμήμα του Πανεπιστημιακού Γενικού Νοσοκομείου Ηρακλείου (Πα.Γ.Ν.Η.) έγινε η απόκτηση των εικόνων ανοσοφθορισμού στις οποίες το κόκκινο χρώμα αντιπροσωπεύει το κυτταρόπλασμα και το πράσινο τον πυρήνα του κυττάρου. Οι εφαρμογές της ποσοτικοποίησης των κυττάρων είναι ολοένα αυξανόμενες και τα αποτελέσματα της μεθόδου είναι γρήγορα και ακριβή. Η χρήση της σε κλινικά και σε ερευνητικά πρωτόκολλα είναι η χρώση και ανάλυση επιφανειακών και ενδοκυτταρικών πρωτεϊνών αλλά διάφορων πληθυσμών κυττάρων. Αυτό έχει ως αποτέλεσμα την ταυτοποίηση επιφανειακών πρωτεϊνών η οποία είναι ιδιαίτερα χρήσιμη στην αναγνώριση διάφορων κυτταρικών υποτύπων καθώς και στη μελέτη της διακύμανσης του αριθμού τους ή στη μεταβολή του φαινοτύπου τους σε παθολογικές και μη καταστάσεις. Επιπλέον, κάποιες ακόμη εφαρμογές είναι η διάκριση των κυτταρικών συστατικών του αίματος, η χρώση και ανάλυση νουκλεϊκών οξέων, η ανάλυση του κυτταρικού κύκλου, η ανάλυση χρωμοσωμάτων, η σχετική και απόλυτη ποσοτικοποίηση των κυτταρικών διαιρέσεων, η κινητική αύξησης κυτταρικού ασβεστίου και η κυτταρομετρία ροής σε υδάτινα οικοσυστήματα.

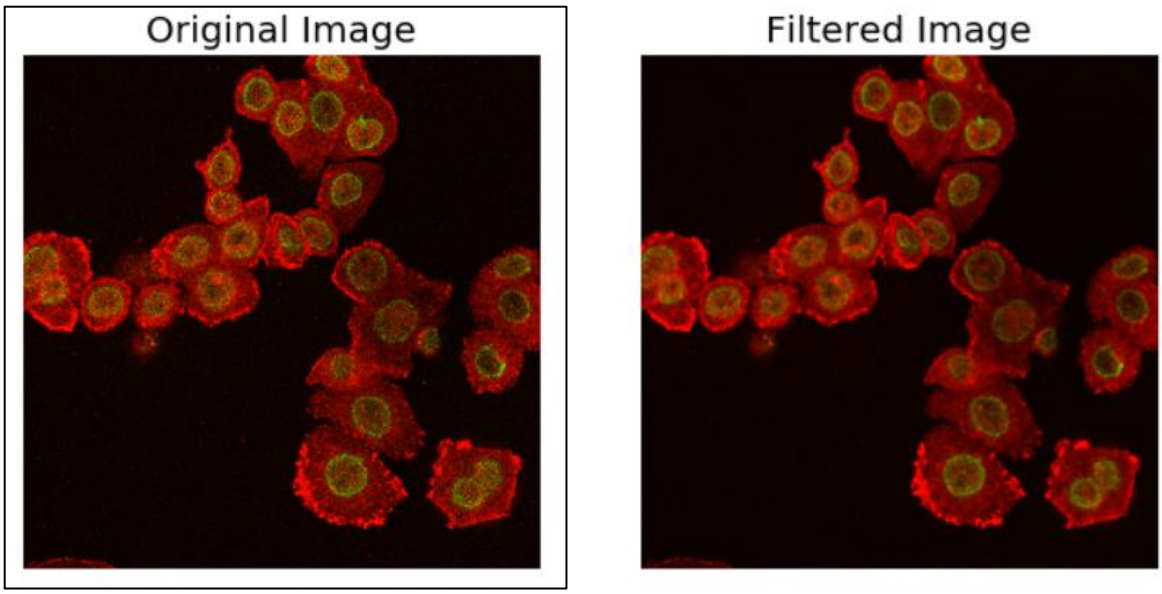
Ακόμη και σήμερα, η ποσοτικοποίηση των κυττάρων γίνεται με ιστολογικά παρασκευάσματα και καταμέτρηση κυττάρων. Η συγκεκριμένη μέθοδος είναι ημιποσοτική και χρονοβόρα. Η διαδικασία απαιτεί τουλάχιστον τρεις ημέρες για να ολοκληρωθεί και τα αποτελέσματα προκύπτουν ανά τομή ή ανά εικόνα τομής και τα συμπεράσματα δεν μπορούν να είναι ακριβή αφού είναι αδύνατον να καταμετρηθούν όλες οι τομές. Ο σκοπός της παρούσας εργασίας είναι η εφαρμογή και η διερεύνηση μίας μεθόδου ποσοτικοποίησης των κυττάρων η οποία να παρέχει ακριβή, αντικειμενικά και άμεσα αποτελέσματα.

Το λογισμικό που χρησιμοποιήθηκε για την υλοποίηση της μεθοδολογίας αυτής είναι το Spyder το οποίο είναι μία πλατφόρμα ανοιχτού κώδικα όπου περιέχει ένα ολοκληρωμένο περιβάλλον ανάπτυξης (IDE) για προγραμματισμό σε Python 3.5.

Έγινε επιλογή αυτής της γλώσσας λόγω της εύκολης αναγνωσιμότητας και συγγραφής του κώδικα καθώς και η ύπαρξη της μεγάλης κοινότητας προγραμματισμού σε αυτήν. Κάτι τέτοιο, καθιστά την λύση τεχνικών δυσκολιών σχετικών με τη γλώσσα μία αρκετά εύκολη και μη χρονοβόρα διαδικασία καθώς υπάρχει πληθώρα υποστήριξη από χιλιάδες χρήστες παγκοσμίως.

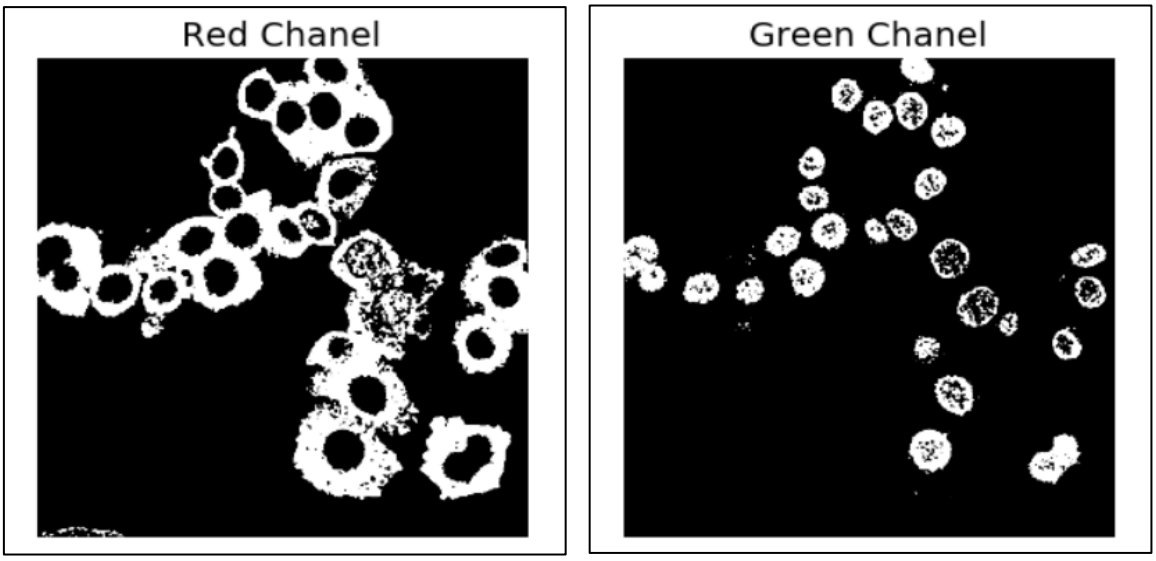
3.1 Ποσοτικοποίηση σήματος από εικόνες ανοσοφθορισμού

- Η μέθοδος η οποία εφαρμόστηκε περιλαμβάνει αρχικά το φιλτράρισμα της αρχικής εικόνας σε κάθε χρωματικό της κανάλι ξεχωριστά με φίλτρο μεσαίας τιμής για την εξάλλειψη πιθανού θορύβου (εικόνα 21).



Εικόνα 21 : Αποτέλεσμα Φιλτραρίσματος με Φίλτρο Μεσαίας Τιμής

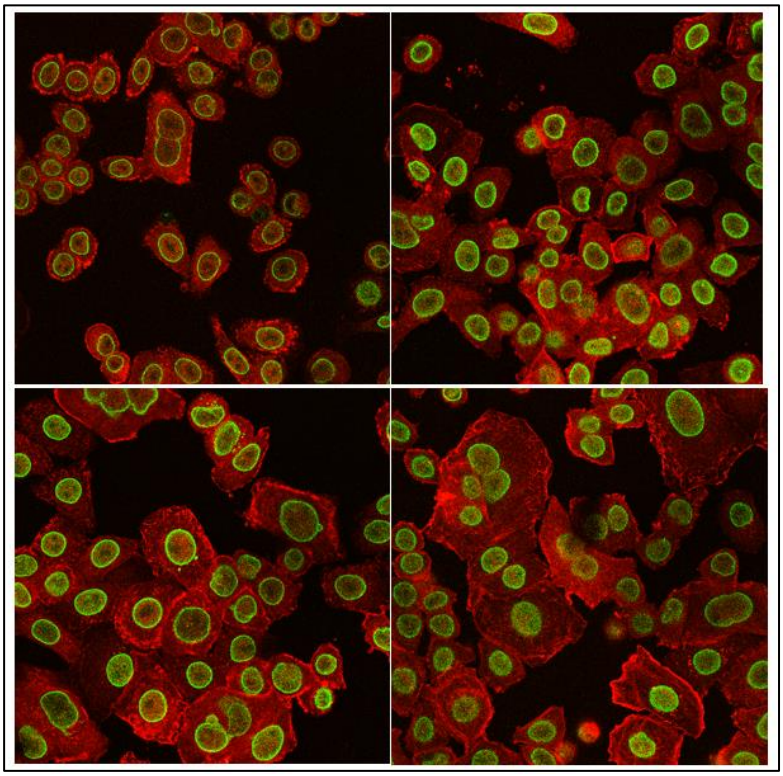
- Στη συνεχεια, αφού όπως αναφέρθηκε παρπάνω ο πυρήνας του κυττάρου αποτυπώνεται με πράσινο χρώμα ενώ το κυτταρόπλασμα με κόκκινο κρίθηκε απαραίτητος ο διαχωρισμός της εικόνας σε 2 κανάλια : το κόκκινο και το πράσινο εφαρμόζοντας ταυτόχρονα σε καθένα από αυτά την τεχνική OTSU το κατώφλι της οποίας θα μας δώσει την δυνατότητα να κρατήσουμε αποκλειστικά τα στοιχεία που αποτυπώνοται σε κάθε ένα από τα κανάλια (εικόνα 22).



Εικόνα 22 : Αποτέλεσμα Διαχωρισμού Καναλιών και Τεχνικής Otsu.

- Στη συνέχεια αφού έχει γίνει διαχωρισμός των πυρήνων από το κυτταρόπλασμα υπάρχει η δυνατότητα να γίνει ποσοτικοποίηση τους ξεχωριστά και να γίνει η περαιτέρω ανάλυση των αποτελεσμάτων.

Παρακάτω (εικόνα 23) παρουσιάζονται περισσότερα δείγματα εικόνων τα οποία χρησιμοποιήθηκαν για την εφαρμογή της παραπάνω μεθοδολογίας.



Εικόνα 23 : Δείγματα εικόνων ανοσοφθορισμού

3.2 Κατηγοριοποίηση

Ο σχεδιασμός των μεθόδων τμηματοποίησης που λειτουργούν ανεξάρτητα από τον τύπο ή το παρασκεύασμα ιστού είναι πολύπλοκος, λόγω των διακυμάνσεων στην μορφολογία των πυρήνων, την ένταση χρώσης, την πυκνότητα κυττάρων και τις συσσωματώσεις πυρήνων. Οι μέθοδοι τμηματοποίησης που βασίζονται στη μηχανική μάθηση μπορούν να ξεπεράσουν αυτές τις προκλήσεις, ωστόσο απαιτούνται εικόνες υψηλής ποιότητας. Στην ενότητα αυτή εξετάζεται πως ο ηλεκτρονικός υπολογιστής μπορεί να κατηγοριοποιήσει, με την βοήθεια της τεχνητής νοημοσύνης, μια εικόνα με φυσιολογικά και καρκινικά κύτταρα.

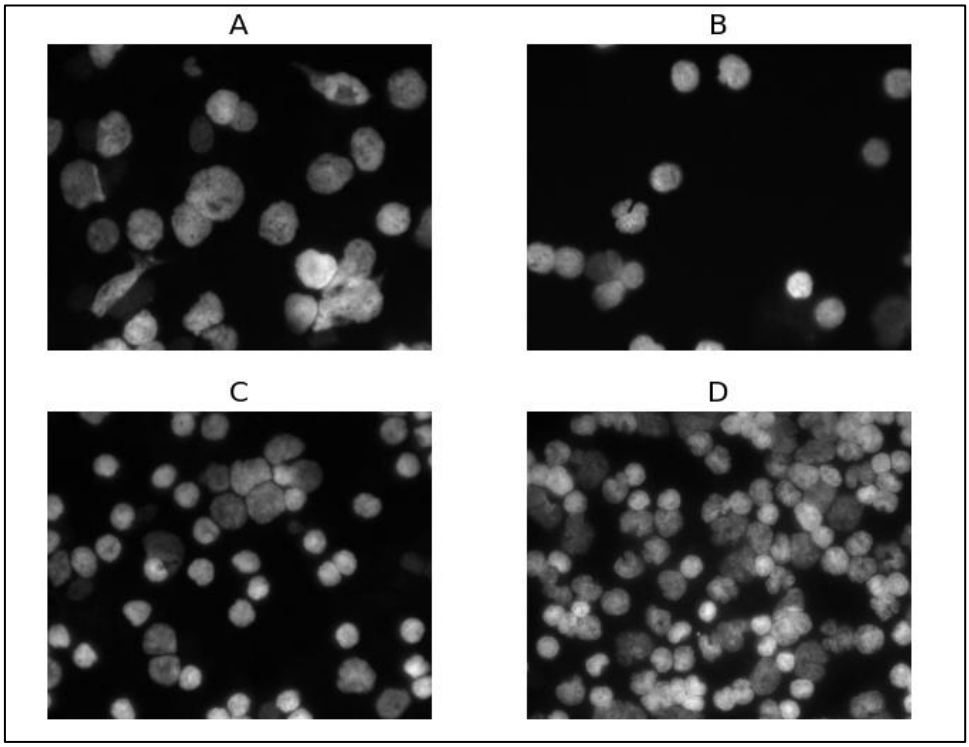
3.2.1 Σύνολο Δεδομένων (Dataset)

Για την έρευνα αυτή, έγινε χρήση ενός δημόσιου συνόλου δεδομένων. Το σύνολο δεδομένων αποτελείται από 79 εικόνες ανοσοφθορισμού (IF) διαφορετικών βιολογικών ιστών και από κύτταρα παθολογικής και μη παθολογικής προέλευσης ώστε να καλύπτει και να συμβαδίζει με αυτές που χρησιμοποιούνται σε βιοϊατρικές έρευνες. Από το σύνολο το δεδομένων φθορισμού, χρησιμοποιήθηκαν οι εικόνες ασθενών με νευροβλάστωμα (Neuroblastoma), γαγγλινονευροβλάστωμα (Ganglioneuroblastoma), καθώς και φυσιολογικών κυττάρων [43] . Επίσης μαζί με τις εικόνες δίνονται και οι μάσκες με τα κύτταρα ενδιαφέροντος όπως είχαν υπολογιστεί από την έρευνα των συγγραφέων στο [43] καθώς και ένα αρχείο excel το οποίο περιέχει τεχνικές πληροφορίες σχετικά με τις παραμέτρους απόκτησης της κάθε εικόνας.

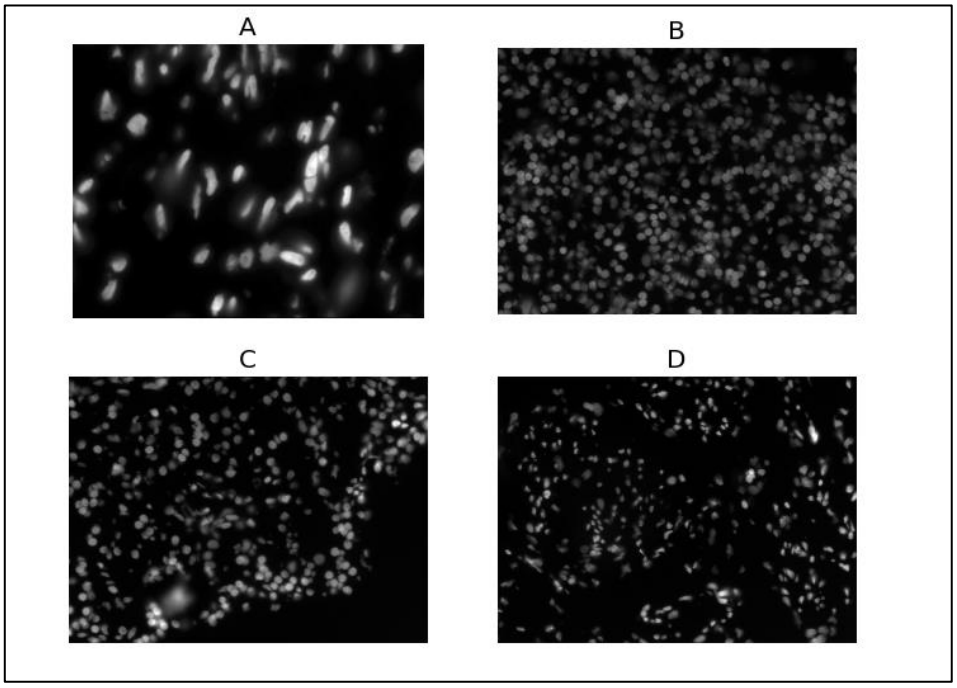
Νευροβλάστωμα : Πρόκειται για μια ασθένεια, κατά την οποία σχηματίζονται καρκινικά κύτταρα στον νευρικό ιστό των επινεφριδίων, του θώρακος, του λαιμού ή του νωτιαίου μυελού και δημιουργείται συχνά στον νευρικό ιστό των επινεφριδίων. Αρχίζει να εμφανίζεται κατά την πρώιμη παιδική ηλικία, συνήθως πριν τα πέντε ηλικιακά έτη ενώ καμιά φορά σχηματίζεται πριν από την γέννηση σε σπάνιες περιπτώσεις.

Γαγγλιονευροβλάστωμα : Είναι μία παραλλαγή του νευροβλαστώματος που περιβάλλεται απο γαγγλιοκύτταρα. Το γάγγλιο είναι μια ομάδα κυττάρων νευρώνων στο περιφερικό νευρικό σύστημα. Τα γάγγλια παρέχουν σημεία αναμετάδοσης και ενδιάμεσων συνδέσεων μεταξύ διαφορετικών νευρολογικών δομών στο σώμα, όπως το περιφερειακό και το κεντρικό νευρικό σύστημα.

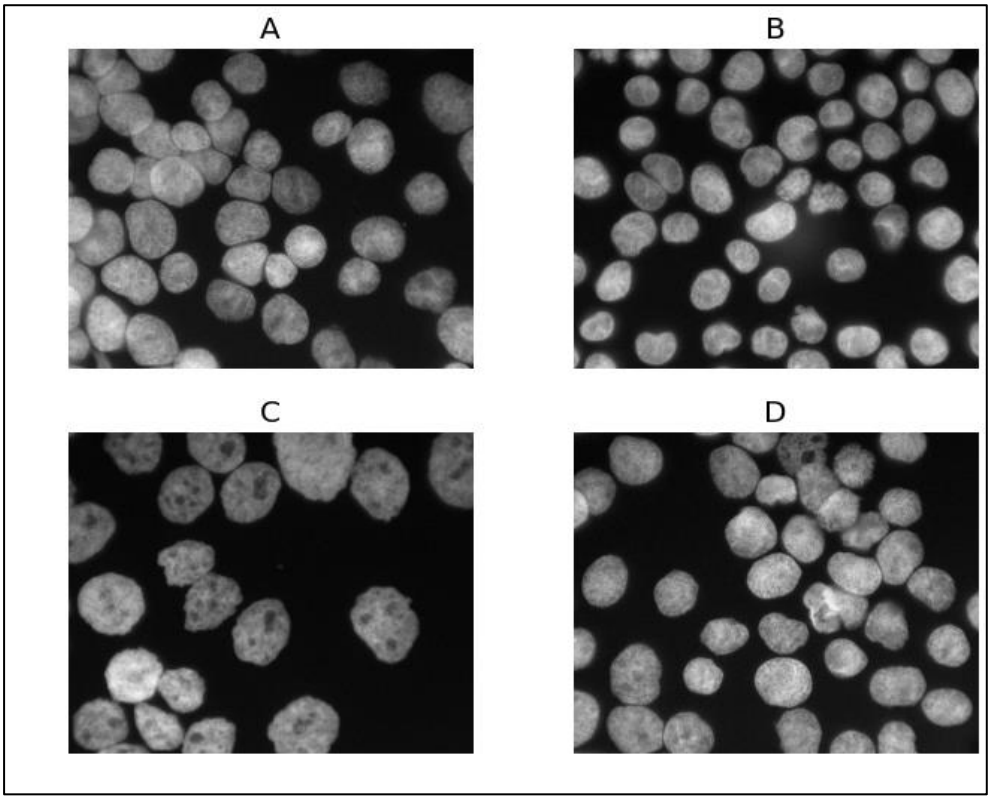
Παρακάτω (εικόνα 24,25,26) φαίνονται δείγματα μερικών εικόνων οι οποίες χρησιμοποιήθηκαν στην μέθοδο της κατηγοριοποίησης.



Εικόνα 24 : Δείγματα *εικόνων από το σύνολο δεδομένων (**A, B, C, D** :* Νευροβλάστωμα*)*



Εικόνα 25 : Δείγματα *εικόνων από το σύνολο δεδομένων (**A, B, C, D** :* Γαγγλινονευροβλάστωμα*)*



Εικόνα 26 : Δείγματα εικόνων από το σύνολο δεδομένων (A, B, C, D :φυσιολογικά κύτταρα)

3.3 Περιορισμοί (Limitations)

Το σύνολο δεδομένων παρουσιάζει μεγάλη ανομοιογένεια σε σχέση με τις απεικονιστικές παραμέτρους απόκτησης των εικόνων. Αναλυτικότερα οι εικόνες αποκτήθηκαν με διαφορετική μεγέθυνση, διαφορετικό οπτικό φακό και διαστάσεις εικόνας, οδηγώντας και σε διαφορετικό signal to noise ratio. Για το λόγο αυτό από τα μεταδεδομένα των εικόνων (excel file [43]) λήφθηκαν υπ όψη αυτά με σταθερή μεγέθυνση x63 καταλήγοντας σε 63 εικονες (35 με φυσιολογικά κύτταρα και 28 με παθολογικα κύτταρα).

3.4 Εξαγωγή χαρακτηριστικών Radiomics

Όπως αναφέρθηκε στο κεφάλαιο 2.6, αναπόσπαστο κομμάτι της ανάλυσης εικόνας με Radiomics είναι η χρήση μάσκας ή αλλιώς περιοχή ενδιαφέροντος. Στην εργασία αυτή χρησιμοποιήθηκαν δύο ειδών μάσκες για την εξαγωγή χαρακτηριστικών και μετέπειτα για την κατηγοριοποίηση.

Αρχικά, για την εξαγωγή χαρακτηριστικών έγινε η χρήση των μασκών που δόθηκαν από το σύνολο δεδομένων όπως προέκυψαν από την έρευνα στο [43]. Σε δεύτερη φάση χρησιμοποιήθηκε η μεθοδολογία τμηματοποίησης όπως αναφέρθηκε στο κεφάλαιο 3.1 (median filtering, Otsu thresholding) για την απόκτηση καινούριων μασκών. Για την ευκολία του αναγνώστη οι περιοχές ενδιαφέροντος θα αποκαλούνται ως **αρχικές μάσκες** και **υπολογισμένες μάσκες** αντίστοιχα**.**

Τα χαρακτηριστικά τα οποία εξήχθησαν από την βιβλιοθήκη pyradiomics με σταθερό bin width = 25 περιείχαν στατιστικά χαρακτηριστικά όπως στατιστικά χαρακτηριστικά πρώτης τάξης (mean, min, max, media, standard deviation…) καθώς και υψηλότερης τάξης χαρακτηριστικά υφής όπως Grey-Level Run Length Matrix (GLRLM), shape-based 2D features, texture features such as Grey-Level Co-Occurrence Matrix (GLCM), Grey Level Size Zone Matrix (GLSZM) and Grey Level Difference Matrix (GLDM). Επιπρόσθετα, χρησιμοποιήθηκαν και χαρακτηριστικά που αφορούν γειτονικά εικονοστοιχεία όπως είναι το local binary patterns 2D (LBP) καθώς και τεχνικές μετασχηματισμού εικόνας όπως Logarithmic, Exponential, Gradient, and wavelet. Τελικά το σύνολο των χαρακτηριστικών που προέκυψαν ήταν 946 χαρακτηριστικά.

3.5 Μεθοδολογία Κατηγοριοποίησης

Πρωταρχικός στόχος ήταν ο διαχωρισμός των καρκινικών κυττάρων από των μη καρκινικών. Έτσι στην περιγραφή της εικόνας όπου υπήρχε η λέξη “normal” έγινε η δημιουργία στήλης με 0 για την αναπαράσταση απουσίας του καρικίνου ενώ στις υπόλοιπες περιγραφές έγινε η δημιουργία στήλης με τιμή 1. Στη συνέχεια, έγινε διαχωρισμός των δεδομένων σε train και test set με ποσοστό 80% - 20% αντιστοίχως έπειτα από πολλές δοκιμές για την μέγιστη δυνατή απόδοση και της αποφυγής υπερπροσαρμογής (overfitting) όπου το μοντέλο έχει υπερκεπευδευτεί στην εξαγωγή απόλυτα σωστών αποτελεσμάτων για το σύνολο εκπαίδευσης (training test), αλλά αδυνατεί να γενικεύσει και να πετύχει εξίσου καλά αποτελέσματα για νέα δεδομένα. Επίσης, έγινε κανονικοποίηση των δεδομένων με τον αλγόριθμο standard scaler της βιβλιοθήκης sklearn [44] έτσι ώστε οι μικρές τιμές των παραμέτρων να οδηγηθούν σε απλούστερες υποθέσεις και, κατά συνέπεια, να μειωθούν οι πιθανότητες να εμφανιστεί το φαινόμενο της υπερπροσαρμογής.

Μετά το πέρας της προεπεξεργασίας των δεδομένων έγινε εφαρμογή του KBest αλγορίθμου για την επιλογή των N = [5,10,20,30,40,50,60,70,80,90] πιο χρήσιμων για περαιτέρω ανάλυση και μελέτη χαρακτηριστικών με βάση την αξιολόγηση – βαθμολογία (score) του αλγορίθμου. Ως ταξινομητές χρησημοποιήθηκαν οι: μηχανές διανυσμάτων υποστήριξης (Support Vector Machines, SVM) και λογιστική παλινδρόμηση (logistic regression) καθώς έχουν χρησιμοποιηθεί σε ανάλογες δημοσιευμένες εργασίες [15], [45].

Οι μηχανές αυτές είναι υπεύθυνες για την ταξινόμηση των δεδομένων σε κατηγορίες καθώς και για την αναζήτηση σχέσεων ομοιοτήτων - διαφορών μεταξύ αυτών.
Για την αποφυγή εσφαλμένων μετρικών που επηρεάζονται από την επιλογή του δείγματος (training και testing set) εφαρμόστηκε η επικύρωση k-fold cross validation με k=10 και υπολογίστηκε ο πίνακας σύγχυσης (Confusion Matrix) και οι επαγόμενεςμετρικές. Οι κύριες μετρικές που χρησιμοποιήθηκαν ήταν η ακρίβεια (Acuuracy, ACC) και το AUC (Area under the Curve) καθώς είναι τα πιο ευρέως χρησιμοποιούμενα σε τέτοια προβλήματα. Επίσης η παραπάνω μεθοδολογία επαναλήφθηκε και τα τα χαρακτηριστικά Radiomics τα οποία υπολογίστηκαν από τις μάσκες που παρήχθησαν με την μεθοδολογία του κεφαλαίου 3.1 (median filtering και Otsu thresholding).

4. Αποτελέσματα

4.1 Αποτελέσματα Ποσοτικοποίησης

Παρακάτω παρουσιάζονται τα αποτελέσματα από τις μετρήσεις των στατιστικών μεταβλητών που πραγματοποιήθηκαν στις 21 εικόνες δεδομένων ανοσοφθορισμού. Τα αποτελέσματα είναι χωρισμένα στην ποσοτικοποίηση του πυρήνα του κυττάρου (Πίνακας 1) και στην ποσοτικοποίηση του κυτταροπλάσματος του (Πίνακας 2) για κάθε μία από τις εικόνες ξεχωριστά. Παρακάτω παρουσιάζονται δείγματα των εικόνων που αναλύθηκαν. Οι μετρήσεις των πυρήνων αφορούν τις περιοχές οι οποίες απεικονίζονται με πράσινο χρώμα ενώ του κυτταροπλάσματος αυτών με κόκκινο χρώμα.

4.1.1 Ποσοτικοποίηση Πυρήνων

Εικόνα	**Mean ± Std**	**Median**	**CTCF**	**Min**	**Max**
Εικόνα 1	62.310 ± 35.110	54	200460	1	219
Εικόνα 2	77.936 ± 35.751	70	213636	2	239
Εικόνα 3	65.187 ± 30.302	60	207680	1	221
Εικόνα 4	89.385 ± 41.930	81	181740	2	249
Εικόνα 5	88.067 ± 34.783	82	212174	11	254
Εικόνα 6	95.413 ± 39.923	89.5	202412	9	254
Εικόνα 7	102.4 ± 37.858	96	196646	14	254
Εικόνα 8	126.321± 49.599	120	184012	15	254
Εικόνα 9	108.589± 46.838	100	178490	3	254
Εικόνα 10	85.929 ± 35.557	78	206118	2	253
Εικόνα 11	45.402 ± 21.057	41	190526	2	190
Εικόνα 12	116.306 ± 47.392	108	200376	3	254
Εικόνα 13	90.582 ± 41.657	84	187448	1	254
Εικόνα 14	87.998 ± 37.248	82	182120	3	254
Εικόνα 15	23.764 ± 10.782	22	201954	1	97
Εικόνα 16	49.261 ± 21.487	46	182830	1	170
Εικόνα 17	54.420 ± 23.782	50	217118	2	254
Εικόνα 18	48.845 ± 23.469	43	213316	1	205
Εικόνα 19	57.790 ± 28.556	51	193916	1	217
Εικόνα 20	96.604 ± 49.712	87	173926	2	254
Εικόνα 21	101.168± 43.394	93	201390	8	254

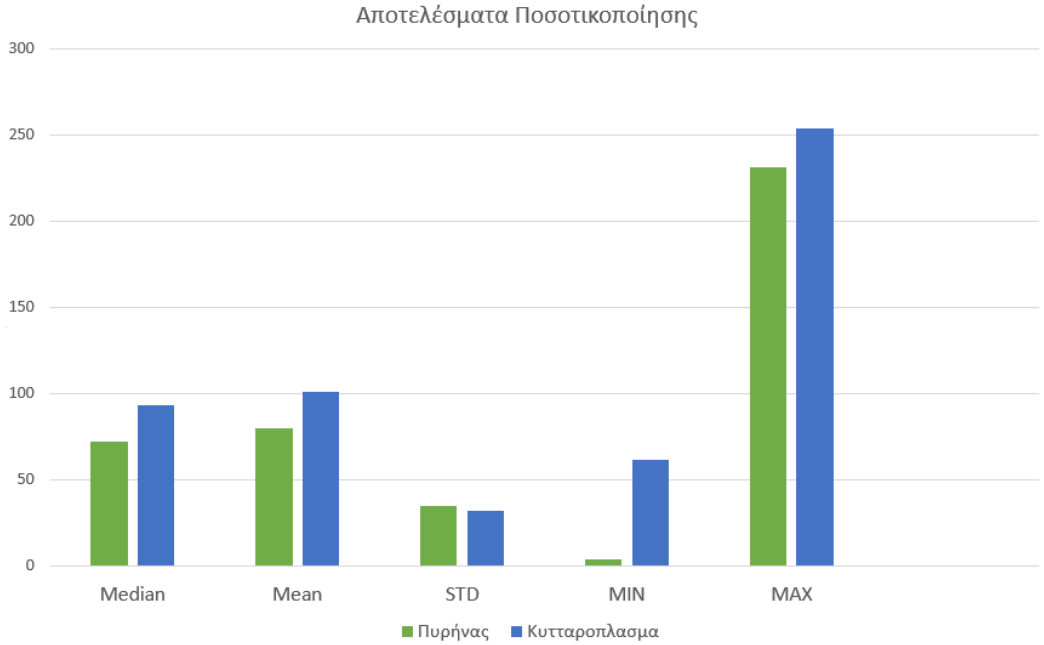
(Πίνακας 1 : Ποσοτικοποίηση Πυρήνων)

4.1.2 Ποσοτικοποίηση Κυτταροπλάσματος

Εικόνα	Mean ± Std	Median	CTCF	Min	Max
Εικόνα 1	99.28 ± 30.426	94	103432	58	254
Εικόνα 2	117.361 ± 35.165	112	98342	67	254
Εικόνα 3	115.319 ± 37.115	107	118856	67	254
Εικόνα 4	115.475 ± 32.469	108	57316	76	254
Εικόνα 5	121.353 ± 39.895	113	67762	70	254
Εικόνα 6	102.791 ± 33.869	95	24916	60	254
Εικόνα 7	86.601 ±27.238	81	50842	52	254
Εικόνα 8	104.630 ± 32.528	97	16832	65	254
Εικόνα 9	100.885 ± 32.059	93	21838	62	254
Εικόνα 10	127.428 ± 46.518	113	76226	77	254
Εικόνα 11	83.664 ± 23.832	77	9856	59	254
Εικόνα 12	92.270 ± 31.099	85	50432	56	254
Εικόνα 13	95.377 ± 29.751	89	34098	59	254
Εικόνα 14	105.500 ± 34.910	95	32968	70	254
Εικόνα 15	56.427 ± 16.429	53	35886	37	254
Εικόνα 16	96.117 ± 27.143	91	18652	62	254
Εικόνα 17	119.299 ± 37.224	113	113304	68	254
Εικόνα 18	96.041 ± 35.564	86	59494	60	254
Εικόνα 19	90.313 ± 30.682	82	14662	57	254
Εικόνα 20	101.986 ± 32.597	95	8250	62	254
Εικόνα 21	88.763 ± 30.795	81	28990	54	254

(Πίνακας 2 : Ποσοτικοποίηση Κυτταροπλάσματος)

Παρακάτω αναπαριστάνονται τα παραπάνω αποτελέσματα σε μορφή ραβδοδιαγράμματος.



Εικόνα 27 : Αποτελέσματα Ποσοτικοποίησης

4.2 Αποτελέσματα Κατηγοριοποίησης

Σε αυτήν την ενότητα παρουσιάζονται τα αποτελέσματα του δεύτερου μέρους της εργασίας τα οποία αφορούν την κατηγοριοποίηση των εικόνων ανοσοφθορισμού από το ελεύθερης πρόσβασης σύνολο δεδομένων σε καρκινικά και φυσιολογικά.

4.2.1 Αποτελέσματα Κατηγοριοποίησης (αρχικές μάσκες)

Στους παρακάτω πίνακες παρουσιάζονται τα αποτελέσματα τα οποία αφορούν τα χαρακτηριστικά Radiomics τα οποία υπολογίστηκαν με την χρήση των αρχικών μασκών (κεφάλαιο 3.4). Αναλυτικότερα, στους πίνακες 3 και 4 παρουσιάζονται η ακρίβεια με την τυπική απόκλιση για κάθε k-fold (k=10) του μοντέλου κατηγοριοποίησης καθώς και το εμβαδόν της περιοχής κάτω από την καμπύλη (Area under the curve) για διαφορετικό αύξοντα αριθμό επιλεγμένων χαρακτηριστικών με τη μέθοδο k-best για τους ταξινομητές SVM και logistic regression αντίστοιχα.

Number of selected Features	Accuracy ± Std	AUC ± Std
N = 5	0.896 ± 0.084	0.966 ± 0.071
N = 10	0.876 ± 0.110	0.972 ± 0.056
N = 20	**0.89 ± 0.147**	**0.983 ± 0.050**
N = 30	0.870 ± 0.124	0.955 ± 0.069
N = 40	0.896 ± 0.112	0.959 ± 0.062
N = 50	0.913 ± 0.087	0.938 ± 0.15
N = 60	0.906 ± 0.094	0.987 ± 0.037
N = 70	0.923 ± 0.13	0.983 ± 0.05
N = 80	0.926 ± 0.090	0.983 ± 0.05
N = 90	0.926 ± 0.117	0.961 ± 0.007

(Πίνακας 3 : Αποτελέσματα Κατηγοριοποίησης με SVM)

Number of selected Features	Accuracy ± Std	AUC ± Std under the curve
N = 5	0.896 ± 0.084	0.955 ± 0.073
N = 10	0.896 ± 0.112	0.972 ± 0.056
N = 20	**0.910 ± 0.117**	**0.988 ± 0.033**
N = 30	0.906 ± 0.094	0.938 ± 0.076
N = 40	0.896 ± 0.112	0.959 ± 0.062
N = 50	0.93 ± 0.086	0.95 ± 0.15
N = 60	0.890 ± 0.090	0.987 ± 0.037
N = 70	0.943 ± 0.124	0.983 ± 0.05
N = 80	0.960 ± 0.079	0.983 ± 0.05
N = 90	0.926 ± 0.117	0.977 ± 0.066

(Πίνακας 4 : Αποτελέσματα Κατηγοριοποίησης με Λογιστική Παλινδρόμηση)

4.2.2 Αποτελέσματα Κατηγοριοποίησης (υπολογισμένες μάσκες)

Στους παρακάτω πίνακες παρουσιάζονται τα αποτελέσματα τα οποία αφορούν τα χαρακτηριστικά Radiomics τα οποία υπολογίστηκαν με την χρήση των υπολογισμένων μασκών (κεφάλαιο 3.4). Αναλυτικότερα, στους πίνακες 5 και 6 παρουσιάζονται η ακρίβεια με την τυπική απόκλιση για κάθε k-fold (k=10) του μοντέλου κατηγοριοποίησης καθώς και το εμβαδόν της περιοχής κάτω από την καμπύλη (Area under the curve) για διαφορετικό αύξοντα αριθμό επιλεγμένων χαρακτηριστικών με τη μέθοδο k-best για τους ταξινομητές SVM και logistic regression αντίστοιχα.

Number of selected Features	Accuracy ± Std	AUC ± Std
N = 5	0.805 ± 0.182	0.941 ± 0.118
N = 10	0.865 ± 0.119	0.95 ± 0.106
N = 20	0.85 ± 0.196	0.908 ± 0.120
N = 30	0.885 ± 0.161	0.9 ± 0.2
N = 40	0.865 ± 0.158	0.958 ± 0.08
N = 50	0.89 ± 0.111	0.95 ± 0.15
N = 60	0.915 ± 0.104	0.91 ± 0.17
N = 70	0.955 ± 0.090	0.95 ± 0.15
N = 80	0.89 ± 0.157	0.941 ± 0.118
N = 90	**0.894 ± 0.138**	**0.966 ± 0.1**

(Πίνακας 5 : Αποτελέσματα Κατηγοριοποίησης με SVM)

Number of selected Features	Accuracy ± Std	AUC ± Std
N = 5	0.784 ± 0.212	0.941 ± 0.118
N = 10	0.845 ± 0.136	0.966 ± 0.1
N = 20	0.85 ± 0.196	0.891 ± 0.139
N = 30	0.885 ± 0.161	0.9 ± 0.2
N = 40	0.915 ± 0.104	0.975 ± 0.075
N = 50	0.915 ± 0.104	0.965 ± 0.13
N = 60	**0.915 ± 0.104**	**0.983 ± 0.050**
N = 70	0.96 ± 0.079	0.95 ± 0.15
N = 80	0.894 ± 0.105	0.941 ± 0.118
N = 90	0.934 ± 0.100	0.94 ± 0.21

(Πίνακας 6 : Αποτελέσματα Κατηγοριοποίησης με Λογιστική Παλινδρόμηση)

5. Συμπεράσματα

5.1 Ποσοτικοποίηση

Το πρώτο σκέλος της εργασίας αφορούσε την ποσοτικοποίηση εικόνων ανοσοφθορισμού. Αφού προηγήθηκε ανάλυση των εννοιών που σχετίζονται με τα στοιχεία τα οποία απεικονίζονται στις εικόνες π.χ ανοσοκυτταροχημεία, φθορισμός κλπ έγινε εισαγωγή στην μεθοδολογία που εφαμόστηκε για τον αυτοματοποιημένο εντοπισμό των πυρήνων και του κυτταροπλάσματος στις εικόνες αυτές και την εξαγωγή των στατιστικών χαρακτηριστικών από αυτές με τέτοιο τρόπο ώστε να μειωθεί το ρίσκο του κινδύνου το οποίο εντοπίζεται με τη χρήση του λογισμικού ImageJ στο οποίο πρέπει να σχεδιαστούν οι περιοχές ενδιαφέροντος δια χειρός. Το συμπέρασμα μετά την ολοκλήρωση αυτής της υλοποίησης είναι οτι παρατηρήθηκε πολύ χαμηλότερο σήμα στους πυρήνες σε αντίθεση με αυτό του κυτταροπλάσματος γεγονός που αποδεικνύει ότι η χρώση είναι πιο έντονη στο κυτταρόπλασμα γεγονός που σημαίνει ότι υπάρχει περισσότερη ύπαρξη αντιγόνου σε αυτό.

5.2 Κατηγοριοποίηση

Στόχος της κατηγοριοποίησης ήταν η δημιουργία ενός μοντέλου το οποίο θα ήταν ικανό να κατηγοριοποιήσει κάθε εικόνα σύμφωνα με τον τύπο των κυττάρων που περιέχει σε φυσιολογικά και μη κύτταρα. Αφού έγινε εισαγωγή στις έννοιες της μηχανικής μάθησης και τις τεχνικές Radiomics οι οποίες εφαρμόζονται για την ανίχνευση και τον υπολογισμό χαρακτηριστικών σχετικών με την υφή για παράδειγμα που απεικονίζεται σε μία εικόνα επετεύχθη η κατασκευή του μοντέλου και έγινε καταγραφή των αποτελεσμάτων τα οποία διεξήχθησαν από αυτό όπως φαίνεται στους πίνακες 3,4. Τα αποτελέσματα είναι στηριζόμενα στο γεγονός ότι η Area Under the Curve (AUC) ως μετρική είναι πιο αντιπροσωπευτική για τον διαχωρισμό μεταξύ δύο κλάσεων. Με βάση τα παραπάνω, προκύπτει ότι μετά το k = 20 (k = αριθμός σημαντικότερων επιλεγμένων χαρακτηριστικών) και οι δύο αλγόριθμοι (Support Vector Machines & Logistic Regression) παρουσιάζουν σταθερή και παρόμοια συμπεριφορά με αυτόν του Logistic Regression όμως να σημειώνει καλύτερη απόδοση σε σχέση με αυτόν του SVM. Γενικότερα, η περιοχή της τιμής κάτω από την καμπύλη (Area Under the Curve) και η ακρίβεια (accuracy) του παρουσιάζουν μία σταθερότητα γεγονός το οποίο σημαίνει ότι ο αριθμός των ελάχιστων χαρακτηριστικών που είναι απαραίτητα προς διεξαγωγή είναι 20. Επίσης, παρόλο τους περιορισμούς με τα τα διαφορετικά μεγέθη και τις μεγενθύνσεις του φακού που υπήρχαν στο σύνολο δεδομένων (όπως προαναφέρθηκε στην ενότητα 3.3 Περιορισμοί) οι ταξινομητές ανταποκρίθηκαν με αρκετή ακρίβεια στον διαχωρισμό των δύο κλάσεων μεταξύ παθολογικών και φυσιολογικών κυττάρων. Τέλος, αξίζει να σημειωθούν τα 20 αυτά πιο χρήσιμα χαρακτηριστικά όπως φαίνονται στον Πίνακα 7 με βάση τον αλγόριθμο K-best.

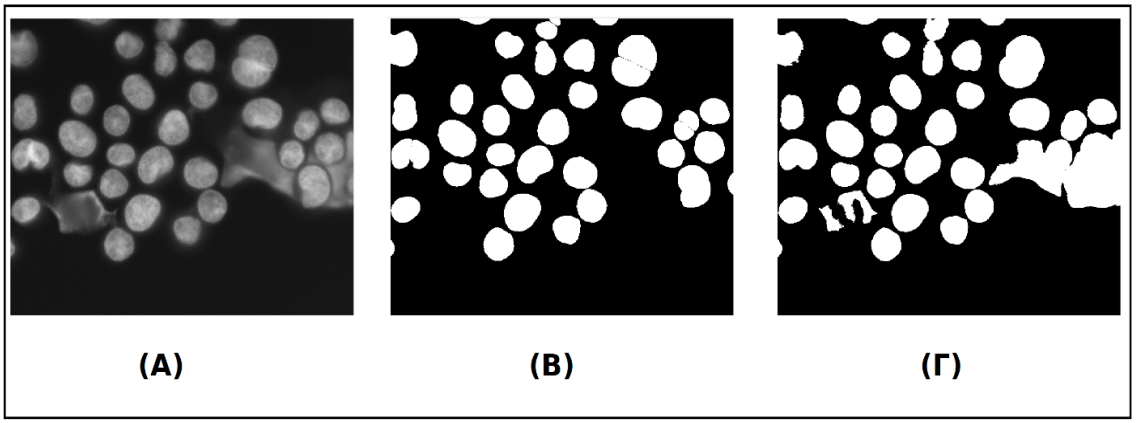
Όνομα Χαρακτηριστικού	**Βαθμολογία Σημαντικότητας**
Original firstorder 10Percentile	71.405
Wavelet-LL firstorder 10 Percentile	70.861
Square firstorder 10 Percentile	67.842
Logarithm firstorder Robust Mean Absolute Deviation	57.011
Logarithm firstorder Interquartile Range	56.003
Exponential firstorder 10 Percentile	53.894
Logarithm firstorder Robust Mean Absolute Deviation	50.639
Square gldm Low Gray Level Emphasis	48.644
Squareroot firstorder 10 Percentile	44.100
Square gldm Large Dependence Low Gray Level Emphasis	40.746
Wavelet LL firstorder Minimum	39.283
Original firstorder Minimum	39.165
Squareroot firstorder Minimum	38.987
Square firstorder Median	38.617
Logarithm firstorder Variance	35.803
Logarithm firstorder Entropy	35.591
Squareroot firstorder InterquartileRange	34.947
Wavelet - LL_firstorder Median	34.301
Squareroot firstorder Robust Mean Absolute Deviation	34.285
Exponential glrlm Short Run Low Gray Level Emphasis	33.980

(Πίνακας 7 : Ονόματα των 20 πιο απαραίτητων χαρακτηριστικών σε συνδυασμό με τη βαθμολογία τους)

Επίσης ανάλογη δουλειά είχε γίνει και στο [15] με τα ίδια δεδομένα που χρησιμοποιήθηκαν στην εργασία αυτή. Στην έρευνα αυτή παρουσιάζεται ένα πρωτόκολλο προ-επεξεργασίας εικόνας με στόχο την ταξινόμηση ενός ετερογενούς συνόλου δεδομένων εικόνων φθορισμού με τη χρήση μηχανικής μάθησης καθώς και βαθιάς μάθησης (Deep learning) για τη διαφοροποίηση τριών κλάσεων. Η μία τεχνική αφορά την παθολογική ανάλυση των εικόνων ενώ η δεύτερη εισάγει μοντέλα βαθιάς μηχανικής μάθησης για την ταξινόμηση των εικόνων.

Το σύνολο των εικόνων που χρησιμοποιήθηκαν ήταν 79 (ολόκληρο το σύνολο δεδομένων) ενώ στην παρούσα εργασία έγινε χρήση των 63 λόγω των περιορισμών μεγέθυνσης που προέκυψαν όπως αναφέρθηκε παραπάνω και η κατηγοριοποίηση έγινε σε δύο κλάσεις έναντι τριών (φυσιολογικά, καλοήθη και κακοήθη κύτταρα). Παρόλο τις διαφορές στο σύνολο των δεδομένων που χρησιμοποιήθηκαν και στον αριθμό των κλάσεων τα αποτελέσματα όσον αφορά την απόδοση των αλγορίθμων μηχανικής μάθησης όπου εφαρμόστηκαν ήταν παραπλήσια . Είναι άξιο λόγου ότι τα αποτελέσματα της παρούσας εργασίας (πίνακες 3,4) καθώς και αυτά του [15] δείξανε την μέγιστη απόδοση σύμφωνα με την μετρική AUC κατά την επιλογή των 20 καλύτερων χαρακτηριστικών με τη χρήση διαφορετικών μεθόδων επιλογής χαρακτηριστικών (K-best έναντι pymrmr [46]).

Η απόδοση του μοντέλου (πίνακες 5,6) κατηγοριοποίησης με την χρήση των χαρακτηριστικών Radiomics που προέκυψαν από τη χρήση των υπολογισμένων μασκών παρουσιάζει εφάμιλλες τιμές με αυτές των πινάκων 3 και 4. Η μόνη διαφορά είναι ότι στην περίπτωση με τις υπολογισμένες μάσκες για να βγει η μεγαλύτερη ακρίβεια χρειάστηκαν περισσότερα χαρακτηριστικά (90 με SVM και 60 με λογιστική παλινδρόμηση). Αυτό είναι λογικό διότι οι αρχικές μάσκες ήταν πιο προσεγμένες καθώς έχουν αφαιρεθεί περιοχές θορύβου/ψευδενδείξεων (artifacts) από την εικόνα σε αντίθεση με τις υπολογισμένες. Παράδειγμα τέτοιων μασκών αναδεικνύεται στην εικόνα 28 παρακάτω.



Εικόνα 28 : Παράδειγμα εικόνας ανοσοφθορισμού (Α) με τις μάσκες του. (Β) Αρχική μάσκα, (Γ) υπολογισμένη μάσκα.

6. Επιλογος

α) Ανακεφαλαιώνοντας, στο πρώτο σκέλος της εργασίας διαπιστώθηκε ότι με τη χρήση μικροσκοπίου και την τεχνικής τμηματοποίησης, δίνεται η δυνατότητα ποσοτικοποίησης της χρώσης των εικόνων ιστοπαθολογίας με αποτέλεσμα να είναι γνωστά τα σημεία στα οποία αυτή εμφανίζεται πιο έντονη. Στα δεδομένα που χρησιμοποιήθηκαν στην υλοποίηση της παρούσας μελέτης παρουσιάστηκε πιο έντονο ποσοστό χρώσης στο κυτταρόπλασμα γεγονός που σημαίνει ότι υπήρχε περισσότερο αντιγόνο σε αυτό το σημείο του κυττάρου. Αξίζει να σημειωθεί ότι θα ήταν ενδιαφέρον ως μελλοντική εργασία η κατασκευή μίας πλατφόρμας ανοιχτού κώδικα με σκοπό την υποβοήθηση των κλινικών για πιο αξιόπιστα ακόμη αποτελέσματα καθώς τα λογισμικά τα οποία χρησιμοποιούνται ακόμη και σήμερα δεν είναι ακόμη πλήρως αυτοματοποιημένα εμφανίζοντας έτσι σημαντικό περιθώριο λάθους λόγω της αναγκαίας επέμβασης των ειδικών οι οποίοι καλούνται να σχεδιάσουν χειροκίνητα τις περιοχές ενδιαφέροντος όπως αναφέρθηκε παραπάνω στην εργασία.

β) Όσον αφορά το δεύτερο σκέλος της εργασίας το οποίο αφορούσε την κατηγοριοποίηση των εικόνων ανοσοφθορισμού σε καρκινικά και φυσιολογικά κύτταρα, η ταξινόμηση σημείωσε σημαντική και αξιόπιστη απόδοση (ACC έως: 0.910 ± 0.117 με AUC 0.988 ± 0.033). Τα αποτελέσματα αυτά οδηγούν στο συμπέρασμα ότι η τεχνική κατηγοριοποίησης είναι ένα πολλά υποσχόμενο πλαίσιο για την ενίσχυση της ανάλυσης και ερμηνείας των εικόνων παθολογίας οι οποίες βασίζονται στην χρήση φθορισμού. Επιπλέον, για την επιτάχυνση της μεταφοράς στην κλινική πράξη τέτοιων εργαλείων απαιτείται στενότερη συνεργασία μεταξύ ερευνητών και ιατρών οι οποίοι ειδικεύονται στην τεχνητή νοημοσύνη ενώ τέλος, η ανάπτυξη μιας μεγαλύτερης βάσης δεδομένων φθορισμού είναι απαραίτητη προϋπόθεση για την γενίκευση και την βελτιστοποίηση τέτοιων μοντέλων και την αύξηση της ευρωστίας.

Βιβλιογραφία

[1] E. C. Jensen, "Quantitative Analysis of Histological Staining and Fluorescence Using ImageJ," *Anat. Rec.*, vol. 296, no. 3, pp. 378–381, Mar. 2013, doi: 10.1002/ar.22641.

[2] D. Gao *et al.*, "FLIMJ: An open-source ImageJ toolkit for fluorescence lifetime image data analysis," *PLoS One*, vol. 15, no. 12, p. e0238327, Dec. 2020, doi: 10.1371/journal.pone.0238327.

[3] S. Fontenete *et al.*, "FISHji: New ImageJ macros for the quantification of fluorescence in epifluorescence images," *Biochem. Eng. J.*, vol. 112, pp. 61–69, Aug. 2016, doi: 10.1016/j.bej.2016.04.001.

[4] T. J. Collins, "ImageJ for microscopy.," *BioTechniques*, vol. 43, no. 1 Suppl. pp. 25–30, 2007, doi: 10.2144/000112505.

[5] S. M. Hartig, "Basic Image Analysis and Manipulation in ImageJ," *Curr. Protoc. Mol. Biol.*, vol. 102, no. 1, Apr. 2013, doi: 10.1002/0471142727.mb1415s102.

[6] A. Shivanandan, A. Radenovic, and I. F. Sbalzarini, "MosaicIA: An ImageJ/Fiji plugin for spatial pattern and interaction analysis," *BMC Bioinformatics*, vol. 14, no. 1, Dec. 2013, doi: 10.1186/1471-2105-14-349.

[7] J. F. Dorn, G. Danuser, and G. Yang, "Computational Processing and Analysis of Dynamic Fluorescence Image Data," *Methods in Cell Biology*, vol. 85. pp. 497–538, 2008, doi: 10.1016/S0091-679X(08)85022-4.

[8] A. R. Cohen, "Extracting meaning from biological imaging data," *Molecular Biology of the Cell*, vol. 25, no. 22. American Society for Cell Biology, pp. 3470–3473, Nov. 05, 2014, doi: 10.1091/mbc.E14-04-0946.

[9] S. van Teeffelen, J. W. Shaevitz, and Z. Gitai, "Image analysis in fluorescence microscopy: Bacterial dynamics as a case study," *BioEssays*, vol. 34, no. 5, pp. 427–436, May 2012, doi: 10.1002/bies.201100148.

[10] P. Bankhead, "Analyzing fluorescence microscopy images with ImageJ," 2014. Accessed: Oct. 19, 2021. [Online]. Available: http://imagej.nih.gov/ij/images/.

[11] K. Kalyvianaki *et al.*, "Membrane androgen receptors (OXER1, GPRC6A AND ZIP9) in prostate and breast cancer: A comparative study of their expression," *Steroids*, vol. 142, pp. 100–108, Feb. 2019, doi: 10.1016/J.STEROIDS.2019.01.006.

[12] N. Papadopoulou *et al.*, "Membrane androgen receptor activation triggers down-regulation of PI-3K/ Akt/NF-kappaB activity and induces apoptotic responses via Bad, FasL and caspase-3 in DU145 prostate cancer cells," *Mol. Cancer*, vol. 7, no. 1, pp. 1–13, Dec. 2008, doi: 10.1186/1476-4598-7-88.

[13] A. Pedram, M. Razandi, R. C. A. Sainson, J. K. Kim, C. C. Hughes, and E. R. Levin, "A conserved mechanism for steroid receptor translocation to the plasma membrane," *J. Biol. Chem.*, vol. 282, no. 31, pp. 22278–22288, Aug. 2007, doi: 10.1074/jbc.M611877200.

[14] P. Thomas, Y. Pang, J. Dong, and A. H. Berg, "Identification and characterization of membrane androgen receptors in the ZIP9 zinc transporter subfamily: II. Role of human ZIP9 in testosterone-induced prostate and breast cancer cell apoptosis," *Endocrinology*, vol. 155, no. 11, pp. 4250–4265, Nov. 2014, doi: 10.1210/en.2014-1201.

[15] G. S. Ioannidis, E. Trivizakis, I. Metzakis, S. Papagiannakis, E. Lagoudaki, and K. Marias, "Pathomics and Deep Learning Classification of a Heterogeneous Fluorescence Histology Image Dataset," *Appl. Sci.*, vol. 11, no. 9, p. 3796, Apr. 2021, doi: 10.3390/app11093796.

[16] S. Çimen Yetiş, A. Çapar, D. A. Ekinci, U. E. Ayten, B. E. Kerman, and B. U. Töreyin, "Myelin detection in fluorescence microscopy images using machine learning," *J. Neurosci. Methods*, vol. 346, p. 108946, Dec. 2020, doi: 10.1016/j.jneumeth.2020.108946.

[17] J. Unger *et al.*, "Real-time diagnosis and visualization of tumor margins in excised breast

specimens using fluorescence lifetime imaging and machine learning," *Biomed. Opt. Express*, vol. 11, no. 3, p. 1216, Mar. 2020, doi: 10.1364/boe.381358.

[18] M. Held *et al.*, "CellCognition: time-resolved phenotype annotation in high-throughput live cell imaging," *Nat. Methods*, vol. 7, no. 9, pp. 747–754, Sep. 2010, doi: 10.1038/nmeth.1486.

[19] C. Alvarez-Jimenez, A. A. Sandino, P. Prasanna, A. Gupta, S. E. Viswanath, and E. Romero, "Identifying Cross-Scale Associations between Radiomic and Pathomic Signatures of Non-Small Cell Lung Cancer Subtypes: Preliminary Results," *Cancers (Basel).*, vol. 12, no. 12, p. 3663, Dec. 2020, doi: 10.3390/cancers12123663.

[20] A. U. Acuña, F. Amat-Guerri, P. Morcillo, M. Liras, and B. Rodríguez, "Structure and formation of the fluorescent compound of lignum nephriticum," *Org. Lett.*, vol. 11, no. 14, pp. 3020–3023, Jul. 2009, doi: 10.1021/ol901022g.

[21] A. Kawamura and Y. Aoyama, *Immunofluorescence in medical science*. University of Tokyo Press, 1983.

[22] "Ανοσοφθορισμός, Εργαστηριακές διαγνώσεις, Σημειώσεις Πέτρος Καρκαλούσος, Πανεπιστήμιο Δυτικής Αττικής,Τμήμα Βιοϊατρικών Επιστημών."

[23] J. A. Ramos-Vara, "Technical aspects of immunohistochemistry," *Veterinary Pathology*, vol. 42, no. 4. pp. 405–426, Jul. 2005, doi: 10.1354/vp.42-4-405.

[24] S. Renshaw, *Immunohistochemistry and Immunocytochemistry: Essential Methods*. Chichester, UK: John Wiley & Sons, Ltd, 2017.

[25] A. H. Coons, H. J. Creech, and R. N. Jones, "Immunological Properties of an Antibody Containing a Fluorescent Group," *Proc. Soc. Exp. Biol. Med.*, vol. 47, no. 2, pp. 200–202, 1941, doi: 10.3181/00379727-47-13084P.

[26] F. Edfors *et al.*, "Immunoproteomics using polyclonal antibodies and stable isotope-labeled affinity-purified recombinant proteins," *Mol. Cell. Proteomics*, vol. 13, no. 6, pp. 1611–1624, 2014, doi: 10.1074/mcp.M113.034140.

[27] "- Methods | Antibodypedia." .

[28] O. Appiah, M. Asante, and J. B. Hayfron-Acquah, "Improved approximated median filter algorithm for real-time computer vision applications," *J. King Saud Univ. - Comput. Inf. Sci.*, Apr. 2020, doi: 10.1016/j.jksuci.2020.04.005.

[29] "A straightforward introduction to Image Thresholding using python | by Sagar Kumar | Spinor | Medium." .

[30] W. Liu, H. Shi, X. He, S. Pan, Z. Ye, and Y. Wang, "An application of optimized Otsu multi-threshold segmentation based on fireworks algorithm in cement SEM image," *J. Algorithm. Comput. Technol.*, vol. 13, p. 174830181879702, Jan. 2019, doi: 10.1177/1748301818797025.

[31] P. Lambin *et al.*, "Radiomics: The bridge between medical imaging and personalized medicine," *Nature Reviews Clinical Oncology*, vol. 14, no. 12. Nature Publishing Group, pp. 749–762, Dec. 01, 2017, doi: 10.1038/nrclinonc.2017.141.

[32] Q. Feng *et al.*, "Hippocampus Radiomic Biomarkers for the Diagnosis of Amnestic Mild Cognitive Impairment: A Machine Learning Method," *Front. Aging Neurosci.*, vol. 11, Nov. 2019, doi: 10.3389/FNAGI.2019.00323.

[33] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, "Machine learning for medical imaging," *Radiographics*, vol. 37, no. 2, pp. 505–515, Mar. 2017, doi: 10.1148/rg.2017160130.

[34] M. W. Berry, A. Mohamed, and B. W. Yap, Eds., *Supervised and Unsupervised Learning for Data Science*. Cham: Springer International Publishing, 2020.

[35] J. S. Cramer, "The Origins of Logistic Regression," *SSRN Electron. J.*, Nov. 2005, doi: 10.2139/ssrn.360300.

[36] M. Awad and R. Khanna, "Support Vector Machines for Classification," in *Efficient*

Learning Machines, Berkeley, CA: Apress, 2015, pp. 39–66.

[37] "'Ταξινόμηση με χρήση αλγορίθμων Data mining και ασαφούς λογικής: Μια εφαρμογή στο μετρό του Παρισιού'Χρήστος Παλαιολόγος, Διπλωματικη Εργασία, Πολυτεχνέιο Κρήτης, τμήμα ηλεκτρονικών μηχανικών & μηχανικών υπολογιστών, Χανιά, 2009."

[38] R. Susmaga, "Confusion Matrix Visualization," *Intelligent Information Processing and Web Mining*, Springer, Berlin, Heidelberg, pp. 107–116, 2004.

[39] D. Chicco, N. Tötsch, and G. Jurman, "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation," *BioData Min.*, vol. 14, pp. 1–22, 2021, doi: 10.1186/S13040-021-00244-Z.

[40] K. Hajian-Tilaki, "Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation," *Casp. J. Intern. Med.*, vol. 4, no. 2, p. 627, 2013.

[41] D. Berrar, "Cross-validation," in *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, vol. 1–3, Elsevier, 2018, pp. 542–545.

[42] K. M. Jablonka, D. Ongari, S. M. Moosavi, and B. Smit, "Big-Data Sciencein Porous Materials: Materials Genomicsand Machine Learning," *Chem. Rev.*, vol. 120, no. 16, p. 8066, Aug. 2020, doi: 10.1021/ACS.CHEMREV.0C00004.

[43] F. Kromp *et al.*, "An annotated fluorescence image dataset for training nuclear segmentation methods," *Sci. Data*, vol. 7, no. 1, p. 262, Dec. 2020, doi: 10.1038/s41597-020-00608-w.

[44] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Jan. 2012, [Online]. Available: http://arxiv.org/abs/1201.0490.

[45] E. Trivizakis, G. S. Ioannidis, I. Souglakos, A. H. Karantanas, M. Tzardi, and K. Marias, "A neural pathomics framework for classifying colorectal cancer histopathology images based on wavelet multi-scale texture analysis," *Sci. Rep.*, vol. 11, no. 1, p. 15546, Dec. 2021, doi: 10.1038/s41598-021-94781-6.

[46] Hanchuan Peng, Fuhui Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005, doi: 10.1109/TPAMI.2005.159.