



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

**ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ
ΥΛΙΚΩΝ**

**Εξαγωγή χαρακτηριστικών μαστογραφικών μαζών και σύγκριση της
διακριτικής τους ικανότητας κατά την ταξινόμηση με χρήση τεχνικών
Μηχανικής Μάθησης**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΚΩΝΣΤΑΝΤΙΝΑΣ Ι. ΤΖΑΒΕΛΛΑ

Επιβλέπουσα: Κ. Σπ. Νικήτα
Καθηγήτρια Ε.Μ.Π

ΑΘΗΝΑ, ΙΟΥΛΙΟΣ 2015



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

**ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ
ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ**

Εξαγωγή χαρακτηριστικών μαστογραφικών μαζών και σύγκριση της διακριτικής τους ικανότητας κατά την ταξινόμηση με χρήση τεχνικών Μηχανικής Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΚΩΝΣΤΑΝΤΙΝΑΣ Ι. ΤΖΑΒΕΛΛΑ

Επιβλέπουσα: Κωνσταντίνα Σ. Νικήτα
Καθηγήτρια Ε.Μ.Π

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 24/07/2015.

.....
Κ. Σ. Νικήτα
Καθηγήτρια Ε.Μ.Π

.....
Γ-Α Σταφυλοπάτης
Καθηγητής Ε.Μ.Π

.....
Δ-Δ Κουτσούρης
Καθηγητής Ε.Μ.Π

ΑΘΗΝΑ, ΙΟΥΛΙΟΣ 2015

.....
Κωνσταντίνα Τζαβέλλα

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Copyright © Κωνσταντίνα Ι. Τζαβέλλα, 2015

Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Περίληψη

Τα τελευταία χρόνια ο καρκίνος του μαστού εξελίσσεται στην κυριότερη αιτία θανάτου για τις γυναίκες παγκοσμίως. Είναι ήδη ο πιο κοινός τύπος καρκίνου για τις γυναίκες. Παρόλα αυτά η έγκαιρη και έγκυρη διάγνωσή του μπορεί να οδηγήσει στην αποτελεσματική του θεραπεία. Η πιο σημαντική ως τώρα μέθοδος διάγνωσης είναι η Μαστογραφία λόγω της αποτελεσματικότητας και αξιοπιστίας της. Ωστόσο, ο εντοπισμός της βλάβης, ο υπολογισμός της πιθανότητας κακοήθειας και η κατηγοριοποίησή της σε καλοήθεια ή κακοήθεια είναι ένα πολύ δύσκολο εγχείρημα ακόμα και για έμπειρους ακτινολόγους. Για τον λόγο αυτό, τα τελευταία χρόνια πολλά συστήματα υποβοηθούμενης διάγνωσης (CAD_x) αναπτύχθηκαν με σκοπό να βοηθήσουν τους ακτινολόγους στην διάγνωση της ασθένειας.

Ο σκοπός της παρούσας διπλωματικής εργασίας είναι η ανάπτυξη μεθοδολογιών που μπορούν να χρησιμοποιηθούν σε ένα αυτόματο σύστημα υποβοηθούμενης διάγνωσης (CAD_x) για μάζες που εντοπίζονται σε μαστογραφικές εικόνες. Οι μεθοδολογίες ελέγχονται σε μεγάλο πλήθος μαστογραφιών από τη βάση μαστογραφιών DDSM ώστε να έχουμε όσο το δυνατόν πιο αξιόπιστα αποτελέσματα. Συγκεκριμένα, συγκρίναμε τρεις ομάδες χαρακτηριστικών μαστογραφικών μαζών ως προς τη διακριτική τους ικανότητα στην ταξινόμηση μαζών διαφορετικής πυκνότητας και διακριτικότητας. Παράλληλα, συγκρίναμε τις επιδόσεις τριών από τους πιο δημοφιλείς ταξινομητές, SVM, Νευρωνικό Δίκτυο και ταξινομητή k-NN. Τα αποτελέσματα ήταν ενθαρρυντικά αφού το συνολικό ποσοστό ακρίβειας για τις 1214 Περιοχές Ενδιαφέροντος έφτασε το 72.82% ενώ σε κάποια υποσύνολα το 87.57%.

Λέξεις κλειδιά: σύστημα υποβοηθούμενης διάγνωσης, μαστογραφία, μάζα, μηχανές υποστήριξης διανυσμάτων, νευρωνικά δίκτυα, ταξινομητής k-κοντινότερων γειτόνων χαρακτηριστικά υφής, μορφολογικά χαρακτηριστικά

Abstract

In recent years breast cancer is turning into the leading cause of death for women worldwide. It is already the most common type of cancer for women. Nevertheless, the early and accurate diagnosis may lead to an effective treatment. Mammography constitutes the most efficient method of diagnosis yet, due to its effectiveness and credibility. However, the detection of the lesion, the calculation of the probability of malignancy and the categorization as a benign or malignant tumour is a very difficult task even for experienced radiologists. For this reason, in recent years many computer aided diagnosis systems (CADx) have been developed in order to improve radiologists' efficiency in the diagnostic interpretation of mammograms.

The aim of this work is the development of methodologies that can be used in an automatic computer aided diagnosis system (CADx) for masses detected in mammographic images. These methodologies were tested in a large amount of cases from the database DDSM in order to obtain the most reliable results possible. Specifically, we compared three sets of features from mammographic masses based on their resolution capability in the classification of masses of different density and subtlety. Furthermore, we compared the performance of the three of the most popular classifiers, SVM, Neural Network and classifier kNN. The results were encouraging since the overall accuracy rate for the 1214 Regions of Interest was 72.82% and in some subsets reached 87.57%.

Keywords: computer aided diagnosis system, mammography, mass, support vector machines, neural networks, k-nearest neighbors algorithm, texture features, morphological features

ΕΥΧΑΡΙΣΤΙΕΣ

Με τη διπλωματική μου εργασία ολοκληρώνονται οι σπουδές μου στη Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου, κατά τη διάρκεια των οποίων είχα την τύχη να συναναστραφώ με πολύ αξιόλογους καθηγητές. Από όλους θα ήθελα να ξεχωρίσω την Καθηγήτρια Κωνσταντίνα Νικήτα, επιβλέπουσα της εργασίας, για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα θέμα εξαιρετικά ενδιαφέρον για μένα.

Θα ήθελα να εκφράσω επίσης ένα μεγάλο ευχαριστώ στον μεταδιδακτορικό ερευνητή Γιάννη Ανδρεάδη για το χρόνο, τις ιδέες και την έμπρακτη βοήθεια που μου προσέφερε από την αρχή της ενασχόλησής μου με την εργασία.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου που ήταν πάντα δίπλα μου και με την στήριξη της οποίας κατάφερα να ολοκληρώσω τον κύκλο των φοιτητικών μου χρόνων.

Περιεχόμενα

Περίληψη.....	4
Abstract.....	5
Κατάλογος Εικόνων.....	10
Κατάλογος Πινάκων.....	12
1 Εισαγωγή.....	14
1.1 Εισαγωγή, Κίνητρο και Δομή Εργασίας.....	14
1.2 Μαστογραφία και Μαστογραφικά Ευρήματα.....	16
1.2.1 Μάζες.....	17
1.2.2 Ασβεστώσεις.....	19
1.2.3 Αρχιτεκτονικές Διαστρεβλώσεις.....	20
1.3 Συστήματα Υποβοηθούμενης Διάγνωσης.....	21
2 Εξαγωγή Χαρακτηριστικών για Μαστογραφικές Μάζες.....	24
2.1 Χαρακτηριστικά Σχήματος.....	24
2.1.1 Κυκλικότητα.....	24
2.1.2 Ορθογωνικότητα.....	25
2.1.3 Εκκεντρότητα.....	25
2.1.4 Κυρτότητα.....	26
2.1.5 Στερεότητα.....	26
2.2 Μορφολογικά Χαρακτηριστικά.....	27
2.2.1 Μέσος Όρος Ακτινικής Απόστασης.....	27
2.2.2 Δείκτης Τραχύτητας.....	28
2.2.3 Εντροπία Ακτινικού Μήκους.....	28
2.2.4 Δείκτης Κυρτότητας.....	28
2.2.5 Zero-crossings Ακτινικής Απόστασης.....	30
2.3 Χαρακτηριστικά Υφής.....	30
2.3.1 Laws' Texture Energy Measures.....	31
2.3.2 Πίνακες Grey Level Run-length.....	32

2.4	Αξιολόγηση.....	32
3	Επιλογή Χαρακτηριστικών και Επισκόπηση Ταξινομητών.....	34
3.1	Επιλογή Χαρακτηριστικών.....	34
3.2	Ταξινομητές.....	36
3.2.1	Κανονικοποίηση των Εισόδων.....	37
3.2.2	Γενίκευση και Χρήση της Μεθόδου k-fold Cross-Validation.....	37
3.2.3	Νευρωνικό Δίκτυο.....	38
3.2.4	Support Vector Machine.....	41
3.2.5	Ταξινομητής K-κοντινότερων Γειτόνων.....	43
4	Μετρήσεις και Αποτελέσματα.....	45
4.1	Σύνολο Δεδομένων.....	45
4.2	Αξιολόγηση Ταξινομητών.....	47
4.3	Αξιολόγηση και Σύγκριση Ομάδων Χαρακτηριστικών.....	49
4.3.1	Υλοποίηση σε Μηχανή Υποστήριξης Διανυσμάτων.....	49
4.3.2	Υλοποίηση σε Νευρωνικό Δίκτυο.....	57
4.3.3	Υλοποίηση σε Ταξινομητή K-κοντικότερων Γειτόνων.....	61
4.4	Συμπεράσματα.....	63
5	Συμπεράσματα και Μελλοντικές Προοπτικές.....	64
5.1	Γενικά Συμπεράσματα.....	64
5.2	Κατάτμηση με Χρήση του Αλγορίθμου K-means.....	65
5.3	Μελλοντικές Προεκτάσεις.....	71
6	Βιβλιογραφία.....	73

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

Εικόνα 1: Μαστογραφική ανατομία του μαστού.....	18
Εικόνα 2: Κακοήθης μάζα (cancer _08_1 468 RMLO).....	19
Εικόνα 3: Κατηγοριοποίηση με βάση το σχήμα και το περίγραμμα της μάζας σύμφωνα με το BI-RADS.....	19
Εικόνα 4: Κατηγοριοποίηση με βάση την πυκνότητα της μάζας σύμφωνα με το BI-RADS.....	20
Εικόνα 5: Περιοχή ενδιαφέροντος με σύμπλεγμα μικροασβεστώσεων.....	21
Εικόνα 6: Αναπαράσταση αρχιτεκτονικής διαστρέβλωσης.....	22
Εικόνα 7: Μοντέλο Υποβοηθούμενης Διάγνωσης CADx.....	24
Εικόνα 8: Παράδειγμα σύνθεσης χαρακτηριστικών Laws.....	33
Εικόνα 9: Μη γραμμικό μοντέλο νευρώνα.....	40
Εικόνα 10: Συνάρτηση Κατωφλίου.....	41
Εικόνα 11: Οι τρεις πιο κοινές συναρτήσεις ενεργοποίησης των κρυφών επιπέδων: (a) the tanh-sigmoid (tansig), (b) log-sigmoid (logsig) and (c) the radial basis function (RBF), ενώ η πιο κοινή συνάρτηση για το επίπεδο εξόδου είναι η (d) (purelin).....	41
Εικόνα 12: Τυπική αρχιτεκτονική Artificial Neural Network.....	42
Εικόνα 13: Σχηματική αναπαράσταση βέλτιστου υπερεπίπεδου για γραμμικώς διαχωρίσιμα πρότυπα	44
Εικόνα 14: Εξέλιξη ακρίβειας του SVM συναρτήσει του g.....	51
Εικόνα 15: Εξέλιξη ακρίβειας του SVM συναρτήσει του C.....	51
Εικόνα 16: Σύγκριση Επιδόσεων του SVM για χρήση όλων των χαρακτηριστικών, των μορφολογικών, των Laws και Run-Length (Ομάδες 1,2,3 και 4 αντίστοιχα).....	54
Εικόνα 17: Σύγκριση Ακρίβειας του SVM για πυκνότητα 1 και χρήση όλων των χαρακτηριστικών, των μορφολογικών, των Laws και Run-Length (Ομάδες 1,2,3 και 4 αντίστοιχα).....	57
Εικόνα 18: Σύγκριση Ακρίβειας του SVM για πυκνότητα 2 και χρήση όλων των χαρακτηριστικών, των μορφολογικών, των Laws και Run-Length (Ομάδες 1,2,3 και 4 αντίστοιχα).....	57

Εικόνα 19: Σύγκριση Ακρίβειας του SVM για πυκνότητα 3 και χρήση όλων των χαρακτηριστικών, των μορφολογικών, των Laws και Run-Length (Ομάδες 1,2,3 και 4 αντίστοιχα).....	58
Εικόνα 20: Σύγκριση Ακρίβειας του SVM για πυκνότητα 4 και χρήση όλων των χαρακτηριστικών, των μορφολογικών, των Laws και Run-Length (Ομάδες 1,2,3 και 4 αντίστοιχα).....	58
Εικόνα 21: Σύγκριση Ακρίβειας του NN για πυκνότητα 1 και χρήση όλων των χαρακτηριστικών, των μορφολογικών, των Laws και Run-Length (Ομάδες 1,2,3 και 4 αντίστοιχα).....	62
Εικόνα 22: Σύγκριση Ακρίβειας του NN για πυκνότητα 2 και χρήση όλων των χαρακτηριστικών, των μορφολογικών, των Laws και Run-Length (Ομάδες 1,2,3 και 4 αντίστοιχα).....	62
Εικόνα 23: Σύγκριση Ακρίβειας του NN για πυκνότητα 3 και χρήση όλων των χαρακτηριστικών, των μορφολογικών, των Laws και Run-Length (Ομάδες 1,2,3 και 4 αντίστοιχα).....	62
Εικόνα 24: Σύγκριση Ακρίβειας του NN για πυκνότητα 4 και χρήση όλων των χαρακτηριστικών, των μορφολογικών, των Laws και Run-Length (Ομάδες 1,2,3 και 4 αντίστοιχα).....	62
Εικόνα 25: Σύγκριση Ακρίβειας του kNN για πυκνότητα 1 και χρήση όλων των χαρακτηριστικών, των μορφολογικών, των Laws και Run-Length (Ομάδες 1,2,3 και 4 αντίστοιχα).....	64
Εικόνα 26: Σύγκριση Ακρίβειας του kNN για πυκνότητα 2 και χρήση όλων των χαρακτηριστικών, των μορφολογικών, των Laws και Run-Length (Ομάδες 1,2,3 και 4 αντίστοιχα).....	64
Εικόνα 27: Σύγκριση Ακρίβειας του kNN για πυκνότητα 3 και χρήση όλων των χαρακτηριστικών, των μορφολογικών, των Laws και Run-Length (Ομάδες 1,2,3 και 4 αντίστοιχα).....	64
Εικόνα 28: Σύγκριση Ακρίβειας του kNN για πυκνότητα 4 και χρήση όλων των χαρακτηριστικών, των μορφολογικών, των Laws και Run-Length (Ομάδες 1,2,3 και 4 αντίστοιχα).....	64
Εικόνα 29: Σύγκριση Επιδόσεων του kNN για χρήση όλων των χαρακτηριστικών, των μορφολογικών, των Laws και Run-Length (Ομάδες 1,2,3 και 4 αντίστοιχα).....	66
Εικόνα 30: ROC καμπύλη για τους τρεις ταξινομητές (με μπλε χρώμα το SVM, πράσινο το NN και κόκκινο ο kNN).....	68
Εικόνα 31: Γραφική αναπαράσταση ολοκληρωμένου CADx.....	70
Εικόνα 32: Μετασχηματισμός Fourier Λογαριθμικού Μετασχηματισμού.....	72
Εικόνα 33: Αρχική Εικόνα (1594_LMLO_ROI).....	73
Εικόνα 34: Εικόνα μετά από τον Λογαριθμικό Μετασχηματισμό.....	73
Εικόνα 35: Εικόνα μετά από την Εξίσωση Ιστογράμματος.....	74

Εικόνα 36: Περίγραμμα Τελικής Εικόνας.....	74
Εικόνα 37: Μαστογραφική Μάζα (4016_LMLO_ROI).....	75
Εικόνα 38: Εικόνα μετά από το Λογαριθμικό Αλγόριθμο	75
Εικόνα 39: Εικόνα μετά από την Εξίσωση Ιστογράμματος.....	75
Εικόνα 40: Μαστογραφική Μάζα μετά από κατάτμηση με τον Αλγόριθμο K-means.....	76

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Πίνακας 1: Πυρήνες Εσωτερικού Γινομένου για SVMs.....	43
Πίνακας 2: Καταμερισμός Μαστογραφιών στο σύνολο.....	47
Πίνακας 3: Επιδόσεις SVM για ισομεγέθη καρκινικά-καλοήγη σύνολα εκπαίδευσης.....	52
Πίνακας 4: Επιδόσεις SVM με εκπαίδευση σε όλη βάση.....	53
Πίνακας 5: Ακρίβεια SVM με χρήση όλων των διαθέσιμων χαρακτηριστικών.....	55
Πίνακας 6: Ευαισθησία SVM με χρήση όλων των διαθέσιμων χαρακτηριστικών.....	55
Πίνακας 7: Ειδικότητα SVM με χρήση όλων των διαθέσιμων χαρακτηριστικών.....	55
Πίνακας 8: Ακρίβεια SVM με χρήση μόνο Μορφολογικών χαρακτηριστικών.....	56
Πίνακας 9: Ευαισθησία SVM με χρήση μόνο Μορφολογικών χαρακτηριστικών.....	56
Πίνακας 10: Ειδικότητα SVM με χρήση μόνο Μορφολογικών χαρακτηριστικών.....	56
Πίνακας 11: Ακρίβεια SVM με χρήση μόνο των χαρακτηριστικών Laws.....	57
Πίνακας 12: Ευαισθησία SVM με χρήση μόνο των χαρακτηριστικών Laws.....	57
Πίνακας 13: Ειδικότητα SVM με χρήση μόνο των χαρακτηριστικών Laws.....	57
Πίνακας 14: Ακρίβεια SVM με χρήση μόνο Run-length χαρακτηριστικών.....	57
Πίνακας 15: Ευαισθησία SVM με χρήση μόνο Run-length χαρακτηριστικών.....	58
Πίνακας 16: Ειδικότητα SVM με χρήση μόνο των Run-length χαρακτηριστικών.....	58
Πίνακας 17: Ακρίβεια Νευρωνικού Δικτύου με χρήση όλων των χαρακτηριστικών.....	60
Πίνακας 18: Ευαισθησία Νευρωνικού Δικτύου με χρήση όλων των χαρακτηριστικών.....	60
Πίνακας 19: Ειδικότητα Νευρωνικού Δικτύου με χρήση όλων των χαρακτηριστικών.....	60
Πίνακας 20: Ακρίβεια Νευρωνικού Δικτύου με χρήση μόνο Μορφολογικών χαρακτηριστικών.....	61
Πίνακας 21: Ευαισθησία Νευρωνικού Δικτύου με χρήση μόνο Μορφολογικών χαρακτηριστικών.....	61
Πίνακας 22: Ειδικότητα Νευρωνικού Δικτύου με χρήση μόνο Μορφολογικών χαρακτηριστικών.....	61
Πίνακας 23: Ακρίβεια Νευρωνικού Δικτύου με χρήση μόνο των χαρακτηριστικών Laws.....	61

Πίνακας 24: Ευαισθησία Νευρωνικού Δικτύου με χρήση μόνο των χαρακτηριστικών Laws.....	62
Πίνακας 25: Ειδικότητα Νευρωνικού Δικτύου με χρήση μόνο των χαρακτηριστικών Laws.....	62
Πίνακας 26: Ακρίβεια Νευρωνικού Δικτύου με χρήση μόνο Run-length χαρακτηριστικών.....	62
Πίνακας 27: Ευαισθησία Νευρωνικού Δικτύου με χρήση μόνο Run-length χαρακτηριστικών.....	62
Πίνακας 28: Ειδικότητα Νευρωνικού Δικτύου με χρήση μόνο Run-length χαρακτηριστικών.....	63
Πίνακας 29: Ακρίβεια kNN με χρήση όλων των διαθέσιμων χαρακτηριστικών.....	64
Πίνακας 30: Ευαισθησία kNN με χρήση όλων των διαθέσιμων χαρακτηριστικών.....	64
Πίνακας 31: Ειδικότητα kNN με χρήση όλων των διαθέσιμων χαρακτηριστικών.....,	64

Κεφάλαιο 1

Εισαγωγή

1.1 Εισαγωγή, Κίνητρο και Δομή Εργασίας

Σύμφωνα με τα Κέντρα Ελέγχου Ασθενειών και Πρόληψης (Centers for Disease Control and Prevention), του κυριότερου ιδρύματος Δημόσιας Υγείας των ΗΠΑ, ο καρκίνος του μαστού αποτελεί την πιο συχνά απαντώμενη μορφή καρκίνου στις γυναίκες. Υπολογίζεται ότι μια στις οκτώ γυναίκες θα εμφανίσει την συγκεκριμένη μορφή καρκίνου κατά την διάρκεια της ζωής της. Στη δεύτερη θέση ακολουθεί ο καρκίνος των πνευμόνων. Όμως η κατάσταση αντιστρέφεται όταν κανείς μελετήσει τους δείκτες θνησιμότητας των δύο αυτών μορφών καρκίνου. Ο καρκίνος των πνευμόνων βρίσκεται στην πρώτη θέση ως η κυριότερη αιτία θανάτου στις γυναίκες, δίνοντας τη δεύτερη θέση στον καρκίνο του μαστού. Αυτή είναι και η ισχυρότερη απόδειξη της σημασίας της έγκαιρης διάγνωσης του καρκίνου του μαστού που μπορεί να αλλάξει σε μεγάλο βαθμό την αποτελεσματικότητα της αντιμετώπισής του.

Το πιο αξιόπιστο και ευρέως αναγνωρισμένο εργαλείο που είναι διαθέσιμο σήμερα για τον έγκαιρο εντοπισμό του καρκίνου του μαστού και τη διάγνωσή του θεωρείται η μαστογραφία. Με τη χρήση μαστογραφίας είναι δυνατό να ανιχνεύεται κατά μέσο όρο το 80%-90% των καρκίνων σε γυναίκες χωρίς συμπτώματα [1]. Οι ακτινολόγοι χρησιμοποιούν το σύστημα Bi-RADS (Breast Imaging-Reporting and Data System) το οποίο εκδίδεται σε μορφή άτλαντα από το Αμερικάνικο Κολλέγιο Ραδιολογίας. Το τυποποιημένο αυτό σύστημα παρέχει έναν αντικειμενικό τρόπο αξιολόγησης της Περιοχής Ενδιαφέροντος (ROI) και ταξινόμησής της σε μια από τις επτά κατηγορίες του Bi-RADS. Αυτές είναι οι:

- Κατηγορία 0: Ασαφές
- Κατηγορία 1: Αρνητικό, καμία εμφάνιση ανωμαλίας στον μαστό
- Κατηγορία 2: Καλοήγη ευρήματα, μηδεμική πιθανότητα κακοήθειας.

- Κατηγορία 3: Πιθανόν καλοήθης, με πιθανότητα άνω του 98%
- Κατηγορία 4: Ύποπτη ανωμαλία
- Κατηγορία 5: Ιδιαίτερα ενδεικτική κακοήθεια με πιθανότητα τουλάχιστον 95%
- Κατηγορία 6: Γνωστή βιοψία – Αποδεδειγμένη κακοήθεια

Όμως η μαστογραφία είναι μια επαναλαμβανόμενη και επιρρεπής σε λάθη διαδικασία. Αυτό οδηγεί στον μη εντοπισμό από τους ακτινολόγους του 10% - 30% όλων των καρκίνων. Από την άλλη μεριά, η μαστογραφία είναι μια δισδιάστατη απεικόνιση μιας τρισδιάστατης οντότητας που οδηγεί έτσι στην υπέρθεση του μαστογραφικού ιστού και στη δημιουργία περιοχών που μοιάζουν με ύποπτες μάζες. Το γεγονός αυτό έχει ως αποτέλεσμα λιγότερο του 30% όλων των βιοψιών στήθους που πραγματοποιούνται δείχνουν εν τέλει κακοήγη παθογένεια.

Είναι λοιπόν σαφές ότι για την εξάλειψη των παραπάνω φαινομένων αλλά και για την μείωση της ψυχικής και σωματικής δυσφορίας των ασθενών αλλά των δαπανών, είναι αναγκαίο ένα σύστημα Υποβοηθούμενης Διάγνωσης στις υπάρχουσες ιατρικές πρακτικές. Ως Υποβοηθούμενη Διάγνωση (CAD) μπορεί να οριστεί η πρόταση διάγνωσης που πραγματοποιείται από ένα υπολογιστικό σύστημα για την ερμηνεία της μαστογραφίας. Το σύστημα πρέπει να παρέχει στον ακτινολόγο μία αξιόπιστη δεύτερη γνώμη, την οποία αυτός συνυπολογίζει προτού βγάλει την τελική του απόφαση για την ύπαρξη ή όχι καρκίνου. Ασφαλώς, η τελική διάγνωση γίνεται πάντοτε από τον ακτινολόγο και η διάγνωση του συστήματος λειτουργεί μόνο επικουρικά. Τα τελευταία χρόνια έχουν αναπτυχθεί πολλά συστήματα υποβοηθούμενης διάγνωσης από ιδιωτικές εταιρείες, Πανεπιστήμια, νοσοκομεία και ακαδημαϊκούς φορείς. Παρά τον αρχικό σκεπτικισμό, τα συστήματα CAD έχουν εισχωρήσει πλέον ουσιαστικά στις καθημερινή κλινική διάγνωση με ευεργετικά αποτελέσματα.

Ωστόσο, στη βιβλιογραφία επανέρχεται συχνά η απουσία μιας άμεσης σύγκρισης των ομάδων χαρακτηριστικών που έχουν αναπτυχθεί από διάφορες ερευνητικές ομάδες στα ίδια ακριβώς δεδομένα [18]. Βασικός σκοπός της παρούσας διπλωματικής εργασίας, λοιπόν, είναι η υλοποίηση ενός συστήματος ταξινόμησης μαστογραφικών μαζών και η σύγκριση των μεθόδων σε μια μεγάλη και δημόσια βάση δεδομένων. Ως βάση με αυτές τις ιδιότητες χρησιμοποιήθηκε η Digital Database for Screening Mammography (DDSM) από την οποία λήφθηκαν 1214 διαφορετικές περιπτώσεις ευρημάτων μαζών. Οι μάζες αυτές έχουν ήδη χωριστεί σε τέσσερις κατηγορίες με βάση την πυκνότητα (density) και σε πέντε κατηγορίες με βάση την ευκολία

διάκρισής τους (subtlety). Η ευκολία διάκρισης δεν είναι ένα Bi-RADS κριτήριο. Είναι η υποκειμενική εντύπωση ενός έμπειρου ακτινολόγου για το πόσο ευδιάκριτη ή όχι είναι η παρουσία μιας μάζας. Παίρνει τιμές από 1 έως 5, όπου το 1 είναι η "δυσδιάκριτη" και 5 είναι η "προφανής". Η τιμή της ευκολίας διάκρισης μπορεί να καταδείξει πόσο δύσκολο είναι να βρεθεί μια βλάβη. Και κατά συνέπεια μπορούμε να ελέγξουμε την αποτελεσματικότητα κάθε μεθόδου ξεκινώντας από τις πιο εύκολες και καταλήγοντας στις πιο δυσδιάκριτες μάζες.

1.2 Μαστογραφία και Μαστογραφικά Ευρήματα

Στο κομμάτι αυτό της εργασίας γίνεται μια συνολική επισκόπηση στην είσοδο του συστήματος Υποβοηθούμενης Διάγνωσης, την απεικόνιση της εσωτερικής δομής του στήθους. Όπως έχει ήδη αναφερθεί η μαστογραφία έχει αναδειχτεί στην αποτελεσματικότερη απεικονιστική μέθοδος διάγνωσης του καρκίνου. Η εξέταση πραγματοποιείται με χρήση ενός ειδικού μηχανήματος, του μαστογράφου, ο οποίος χρησιμοποιεί ακτίνες X χαμηλής ενέργειας, για να απεικονίσει σε ένα ειδικό φιλμ την παρουσία ανεξέλεγκτου πολλαπλασιασμού των κυττάρων στο στήθος. Αρχικά ο μαστός συμπιέζεται για την καλύτερη δυνατή απεικόνισή του, και στη συνέχεια στοχεύεται με μια δέσμη ακτίνων X. Η δέσμη αυτή φιλτράρεται προτού φτάσει στο μαστό, με σκοπό την ευθυγράμμισή της και την απομάκρυνση των φωτονίων χαμηλής ενέργειας. Τα φωτόνια καθώς διαπερνούν τον μαστό εξασθενούν. Καθώς τα φωτόνια X απορροφώνται από το φιλμ, φωτόνια φωτός εκπέμπονται και σχηματίζουν την εικόνα μαστογραφίας. Διαφορετικά επίπεδα εξασθένισης των ακτίνων καθώς διέρχονται από το μαστό είναι ενδεικτικά των διαφορετικών ιστών του μαστού και των πιθανών ευρημάτων. Οι βασικές κατηγορίες στις οποίες αυτές κατατάσσονται είναι: (i) οι μάζες, (ii) οι αρχιτεκτονικές διαστρεβλώσεις και (iii) οι ασβεστώσεις, καθώς βέβαια και συνδυασμός των κατηγοριών αυτών. Η αξιολόγηση του συνόλου των ευρημάτων αυτών οδηγούν στην διάγνωση της ασθένειας. Βέβαια, πρέπει να τονίσουμε ότι οι πρώτες δύο κατηγορίες, δηλαδή οι μάζες και οι ασβεστώσεις, είναι οι συνηθέστερες και οι πλέον πιο ανησυχητικές για την εκδήλωση του καρκίνου. Όλες οι προαναφερθείσες κατηγορίες αναλύονται στις παρακάτω υποενότητες του κεφαλαίου.

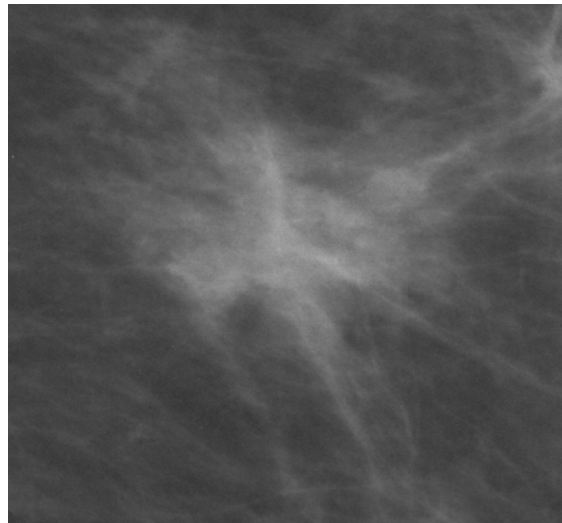


Εικόνα 1: Μαστογραφική ανατομία του μαστού

1.2.1 Μάζες

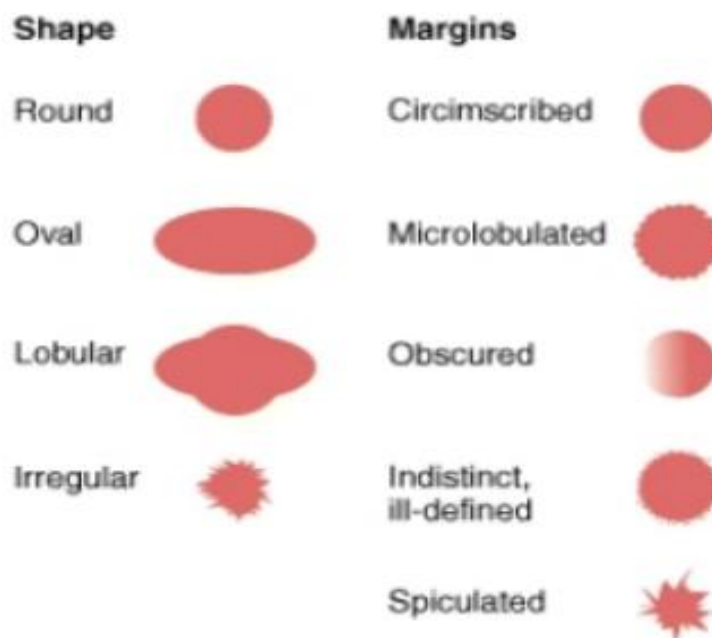
Μάζα θεωρείται η τρισδιάστατη αλλοίωση της δομής του εσωτερικού του μαστού η οποία μπορεί να επαληθευτεί σε δύο διαφορετικές προβολές της μαστογραφίας [3]. Είναι περιοχές με μεγαλύτερη πυκνότητα από τον περιβάλλοντα ιστό και απεικονίζονται ως περιοχές με μεγαλύτερη ένταση στη μαστογραφία. Επιπρόσθετα η κακοήθης μάζα συχνά διεισδύει στον περιβάλλοντα ιστό. Σαν αποτέλεσμα, η επιφάνειά της μπορεί να αποτελείται από λεπτά γραμμικά νήματα που εκτείνονται ακανόνιστα προς τα έξω από την κεντρική μάζα [4]. Μια αντιπροσωπευτική απεικόνιση καρκινικής μάζας φαίνεται στην Εικόνα 2 η οποία έχει εξαχθεί από τη βάση DDSM.

Οι ακτινολόγοι ταξινομούν τις μάζες με βάση το σχήμα, τα όριά τους καθώς και την πυκνότητά τους. Όσον αφορά το σχήμα των μαζών, ανώμαλες μάζες με τραχύ σχήμα χαρακτηρίζονται ως μάλλον κακοήθεις, ενώ στρογγυλές ή ωοειδείς αθρώνονται ως καλοήθεις. Κάτι ανάλογο ισχύει με τα όρια των μαζών τα οποία μπορούν να ταξινομηθούν ως συνεχή, ημιτελή ή κακώς ορισμένα, με τα τελευταία να μαρτυρούν συνήθως κακοήθεια.

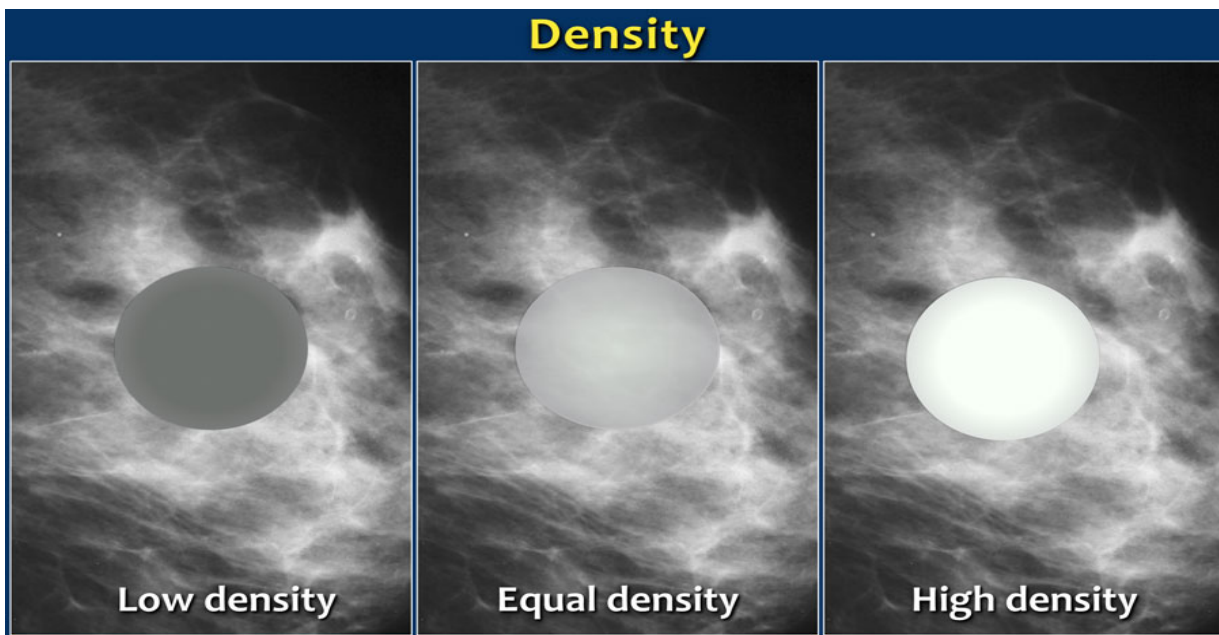


Εικόνα 2: Κακοήθης μάζα (cancer_08_1468 RMLO)

Οι κατηγοριοποιήσεις αυτές σύμφωνα με το American College of Cancer φαίνονται στην Εικόνα 3. Τέλος, όσον αφορά στην πυκνότητα της μάζας, όσο μεγαλύτερη είναι η διαφοροποίηση από την πυκνότητα του περιβάλλοντα ιστού, τόσο μεγαλύτερη είναι η πιθανότητα η μάζα να είναι κακοήθης. Στις περισσότερες των περιπτώσεων οι καλοήθεις μάζες μπορούν να διακριθούν σχετικά εύκολα, σε αντίθεση με τις κακοήθεις [5]. Μια σχηματική αναπαράσταση των παραπάνω τύπων ακολουθεί στην Εικόνα 4.



Εικόνα 3: Κατηγοριοποίηση με βάση το σχήμα και το περίγραμμα της μάζας σύμφωνα με το BI-RADS



Εικόνα 4: Κατηγοριοποίηση με βάση την πυκνότητα της μάζας σύμφωνα με το BI-RADS

1.2.2 Ασβεστώσεις

Οι ασβεστώσεις είναι μικρά συμπλέγματα ασβεστίου στον μαστό. Αποτελούν το πιο διαδεδομένο αλλά και το πιο μικρό εύρημα που εμφανίζεται στις μαστογραφίες. Πρόκειται εν γένει για καλοήγη ευρήματα, που μπορεί να εντοπιστούν σε οποιοδήποτε σημείο του μαστού. Διακρίνονται σε δύο κατηγορίες, τις μακροασβεστώσεις και τις μικροασβεστώσεις, προφανώς ανάλογα με το μέγεθός τους.

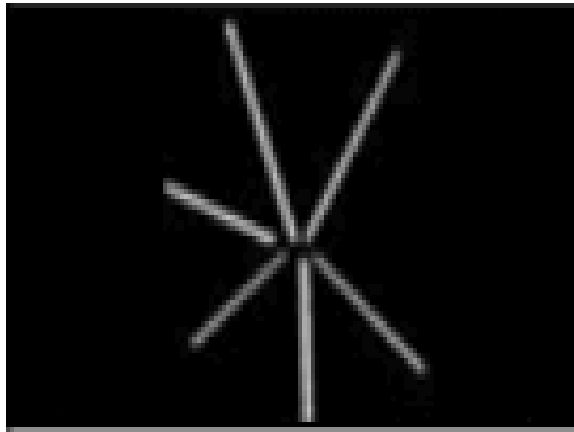
Οι μακροασβεστώσεις εντοπίζονται σε περίπου το μισό γυναικείο πληθυσμό άνω των 5-ετών. Είναι συνήθως μεγαλύτερες από 0.5 mm και στρογγυλές, είναι πιο σπάνιες από τις μικροασβεστώσεις, ενώ η εμφάνισή τους αποδίδεται σε καλοήθεις συνθήκες. Οι μικροασβεστώσεις χρίζουν ιδιαίτερης ανάλυσης, αφού ενώ δεν είναι άμεσα σημάδια καρκίνου, μπορεί να εκδηλώνονται ως προ-καρκινικές αλλαγές. Οι ακτινολόγοι θεωρούν ότι η αξιολόγηση τους βοηθάει σημαντικά στην σωστή κατηγοριοποίηση των βλαβών και στην έγκαιρη διάγνωση του καρκίνου του μαστού. Μάλιστα, έχει αποδειχθεί ότι ένα ποσοστό 30-50% περιπτώσεων ασθενών που πάσχουν από την συγκεκριμένη ασθένεια έχει διαγνωστεί χάρη στην παρουσία μικροασβεστώσεων.



Εικόνα 5: Περιοχή ενδιαφέροντος με σύμπλεγμα μικροασβεστώσεων

1.2.3 Αρχιτεκτονικές Διαστρεβλώσεις

Οι αρχιτεκτονικές διαστρεβλώσεις είναι η τρίτη πιο κοινή ένδειξη εμφάνισης μη-ψηλαφητού καρκίνου του στήθους. Ως αρχιτεκτονική διαστρέβλωση μπορεί να οριστεί μία σημαντική αλλαγή στη φυσιολογική δομή του μαστού, χωρίς όμως την ύπαρξη συγκεκριμένης ορατής μάζας. Η συνηθέστερη μορφή μιας διαστρέβλωσης είναι λεπτές γραμμές που ακτινοβολούνται από ένα σημείο στο μαστό, όπως φαίνεται στην Εικόνα 7. Οι διαστρεβλώσεις αυτού του είδους εμφανίζονται συνήθως στην περιοχή του παρεγχύματος του μαστού, όπου και παρατηρείται ανώμαλη διάταξη των ιστών. Οι διαστρεβλώσεις αντιπροσωπεύουν το 6% των ανωμαλιών που ανιχνεύονται σε μια μαστογραφία. Παρόλα αυτά όμως είναι πιθανόν να αποτελούν την πρωιμότερη ένδειξη καρκίνου του μαστού. Είναι ύποπτες για κακοήθεια και απαιτείται βιοψία, ιδιαίτερα μάλιστα αν δεν υπάρχει ιστορικό τραύματος ή προηγούμενης χειρουργικής επέμβασης. Ακόμα, ορισμένες μελέτες καταδεικνύουν ότι ένας πρώιμος εντοπισμός Αρχιτεκτονικής Διαστρέβλωσης μπορεί να οδηγήσει σε βελτιωμένη πρόγνωση από ότι ο πρώιμος εντοπισμός ασβεστώσεων [2].



Εικόνα 6: Αναπαράσταση αρχιτεκτονικής διαστρέβλωσης

1.3 Υποβοηθούμενη Διάγνωση και Δομή Εργασίας

Οι πρώτες προσπάθειες για την ανάπτυξη ενός συστήματος Υποβοηθούμενης Διάγνωσης του καρκίνου του μαστού έγιναν περίπου πενήντα χρόνια πριν, όταν το 1967 οι Winsberg et al χρησιμοποίησαν κατανομές της οπτικής πυκνότητας της μαστογραφίας για διαφοροποιήσουν φυσιολογικούς από μη φυσιολογικούς ιστούς [8]. Η επιτακτική αυτή ανάγκη για επικουρία των γιατρών οδήγησε κατά τη διάρκεια των δύο τελευταίων δεκαετιών στην υλοποίηση και ανάπτυξη συστημάτων υποβοηθούμενης διάγνωσης καθώς και την έγκρισή τους από τον FDA το 1998 [6]. Τα παραπάνω συστήματα διακρίνονται σε δύο κατηγορίες:

- Συστήματα Υποβοηθούμενης Ανίχνευσης Ευρημάτων (CADE)

Το σύστημα τροφοδοτείται με μια μαστογραφία και εντοπίζει και επισημαίνει στον ακτινολόγο τις ύποπτες περιοχές. Στόχος του συγκεκριμένου συστήματος είναι η ελαχιστοποίηση των απαρατήρητων ευρημάτων ενώ αφήνουν την ανάλυση και διάγνωσή τους αποκλειστικά στον ακτινολόγο.

- Συστήματα Υποβοηθούμενης Κατηγοριοποίησης Ευρημάτων (CADx)

Τα συστήματα αυτά δημιουργήθηκαν ώστε να καταστήσουν εφικτή την αυτόματη διάγνωση ενός μαστογραφικού ευρήματος, την οποία ο ακτινολόγος θα συνυπολογίσει στη δική του διάγνωση. Αρχικά το σύστημα δέχεται μια Περιοχή Ενδιαφέροντος (Region of Interest) του ευρήματος και

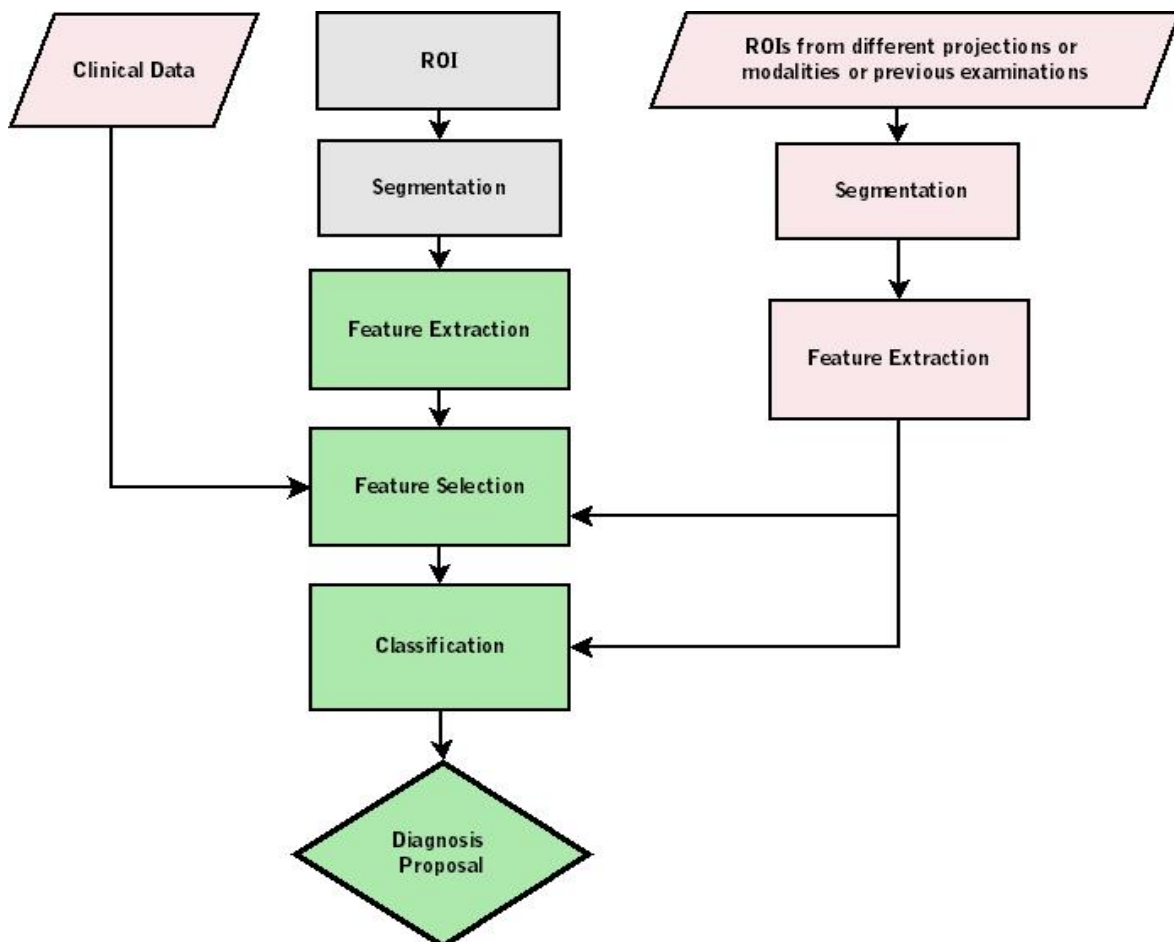
αφού το επεξεργαστεί μαθηματικά καταλήγει στην απόφαση.

Ένα σχηματικό μοντέλο για το Σύστημα Υποβοηθούμενης Κατηγοριοποίησης Ευρημάτων (CAD_x) ακολουθεί στην Εικόνα 7. Αποτελείται από αυτοτελείς μονάδες καθεμία από τις οποίες πρέπει να σχεδιαστεί με ανεξάρτητο τρόπο από τις υπόλοιπες, μιας και υλοποιεί αυτόνομο έργο. Το σύστημα αρχικά δέχεται την περιοχή ενδιαφέροντος και πραγματοποιεί κατάτμηση του ευρήματος, ώστε να αναλυθούν οι ιδιότητες του ίδιο αλλά και της γύρω περιοχής. Στη συνέχεια, το σύστημα πραγματοποιεί την εξαγωγή των χαρακτηριστικών, χρησιμοποιώντας τεχνικές της Όρασης Υπολογιστών, και τα οποία θα καθορίσουν σε μεγάλο βαθμό την κατάληξη της μετέπειτα απόφασης. Από τα χαρακτηριστικά αυτού του σταδίου θα επιλεγούν εκείνα με τη μεγαλύτερη διακριτική ικανότητα και τα οποία θα οδηγηθούν στο επόμενο στάδιο, στο στάδιο της ταξινόμησης. Στο συγκεκριμένο στάδιο χρησιμοποιούνται μέθοδοι ενός άλλου επιστημονικού πεδίου, αυτού της Τεχνητής Νοημοσύνης και συγκεκριμένα της Μηχανικής Μάθησης. Μέσω αυτών των διαδικασιών πραγματοποιείται αυτόματα η υπόδειξη του συστήματος. Όπως φαίνεται και στο σχήμα, σε αυτή την αλληλουχία βημάτων μπορούν να προστεθούν και άλλα όπως παλιότερες απεικονίσεις του ευρήματος ή και από διαφορετικές απεικονιστικές μεθόδους. Τέλος, σε πολλές περιπτώσεις γίνεται χρήση και άλλων χαρακτηριστικών που δεν προκύπτουν από την ανάλυση της εικόνας, π.χ. ιστορικό ασθενούς ή ιατρικά δεδομένα, στοιχεία που μπορούν να αποδειχθούν κρίσιμα για την σωστή ταξινόμηση του ευρήματος.

Συνοπτικά, κατά τη διαδικασία της υλοποίησης της παρούσας διπλωματικής εργασίας μας εξάγονται και συγκρίνονται οι τρεις κυριότερες ομάδες χαρακτηριστικών μαζών που έχουν χρησιμοποιηθεί τα τελευταία χρόνια ως προς τη διακριτική τους ικανότητα καθώς και τρεις από τις πιο δημοφιλείς υλοποιήσεις Μηχανικής Μάθησης. Πραγματοποιείται μελέτη ευρείας κλίμακας σε μεγάλο πλήθος μαστογραφιών ώστε να διερευνηθούν επαρκώς οι παράγοντες που επηρεάζουν την υπολογιστική διάγνωση. Πιο συγκεκριμένα, το μοντέλο δέχεται σαν είσοδο μια περιοχή ενδιαφέροντος στη μαστογραφία που στη συγκεκριμένη περίπτωση απεικονίζει, σχεδιασμένο από τον ακτινολόγο, το περίγραμμα της μάζας προς εξέταση. Μετά από την εξαγωγή διαφόρων ομάδων διακριτικών χαρακτηριστικών και την επιλογή των σημαντικότερων εξ αυτών πραγματοποιείται η εκπαίδευση με τεχνικές Μηχανικής Μάθησης (Machine Learning) η οποία και καταλήγει στην κατηγοριοποίησή της για το αν σχετίζεται με καλοήθειες ή κακοήθειες συνθήκες.

Η εργασία πραγματοποιήθηκε εξ ολοκλήρου με χρήση της γλώσσας προγραμματισμού Matlab. Η δομή της υπόλοιπης εργασίας είναι η εξής: στο Κεφάλαιο 2 αναλύεται το θεωρητικό

υπόβαθρο στο οποίο βασίστηκε η υλοποίηση των αλγορίθμων για την εξαγωγή των χαρακτηριστικών των μαζών. Στο Κεφάλαιο 3 περιγράφεται η διαδικασία της επιλογής των πιο κρίσιμων χαρακτηριστικών καθώς και το μαθηματικό υπόβαθρο στο οποίο βασίζονται τα μοντέλα ταξινόμησης Μηχανικής Μάθησης. Στο Κεφάλαιο 4 παρουσιάζονται και αναλύονται τα αποτελέσματα της σύγκρισης των διαφόρων μοντέλων και πραγματοποιείται η σύγκριση με τα αποτελέσματα προηγούμενων μελετών ερευνητικών ομάδων με δημοσιεύσεις στο συγκεκριμένο θέμα ταξινόμησης μαστογραφικών μαζών. Τέλος, στο Κεφάλαιο 5 αναλύονται οι μελλοντικές προοπτικές με βάση τα συμπεράσματα της εργασίας μας. Διερευνάται η δυνατότητα βελτίωσης της απόδοσης του συστήματος εάν σε αυτό ενταχθούν η αυτόματη κατάτμηση από τον υπολογιστή καθώς και διαφορετικές μέθοδοι απεικόνισης του εσωτερικού του μαστού.



Εικόνα 7: Μοντέλο Υποβοηθούμενης Διάγνωσης CADx

Κεφάλαιο 2

Εξαγωγή Χαρακτηριστικών για Μαστογραφικές Μάζες

Οι ακτινολόγοι χαρακτηρίζουν τις μαστογραφικές μάζες βασιζόμενοι στο σχήμα τους, στην ιδιομορφία του περιγράμματός τους και στην οπτική τους πυκνότητα. Με τη βοήθεια της Υπολογιστικά Υποβοηθούμενης Διάγνωσης, εισάγονται στη διαδικασία διάγνωσης χαρακτηριστικά παραπλήσια εκείνων που χρησιμοποιούνται από τους ακτινολόγους, αλλά πολλά ακόμα υψηλότερης ανάλυσης. Τα χαρακτηριστικά μπορούν να κατηγοριοποιηθούν σε τέσσερις γενικές κατηγορίες: σχήματος, μορφολογίας περιγράμματος, οπτικής πυκνότητας και υφής. [18] Σε αυτό το κεφάλαιο θα αναφερθούν αναλυτικά οι ιδιαιτερότητες των χαρακτηριστικών κάθε κατηγορίας καθώς και η διακριτική τους ικανότητα στην κατηγοριοποίηση μαζών σύμφωνα με μελέτες προηγούμενων ετών.

2.1 Χαρακτηριστικά Σχήματος

Στα χαρακτηριστικά αυτά συμπεριλαμβάνεται η περίμετρος, το εμβαδόν, η κυκλικότητα, η ορθογωνικότητα, ο προσανατολισμός, η κυρτότητα και η εκκεντρότητα της μάζας. Στη βιβλιογραφία μπορούν να βρεθούν παραδείγματα που τα ταυτίζουν με τα μορφολογικά χαρακτηριστικά όμως στην πλειονότητα των περιπτώσεων αναφέρονται ως χαρακτηριστικά σχήματος. Τα χαρακτηριστικά που χρησιμοποιήθηκαν στη συγκεκριμένη εργασία για την υλοποίηση του αλγορίθμου αναφέρονται παρακάτω.

2.1.1 Κυκλικότητα

Η κυκλικότητα προσφέρει μια περιγραφή του σχήματος της μάζας συγκρίνοντας την περιφέρεια και το εμβαδόν της μέσω της σχέσης:

$$C = \frac{P^2}{A}$$

όπου P η περίμετρος και A το εμβαδόν της. Παρατηρούμε ότι όσο πιο ακτινωτό και τραχύ το περίγραμμα της μάζας, τόσο μεγαλύτερη η τιμή του C [19].

2.1.2 Ορθογωνικότητα

Η ορθογωνικότητα εκφράζει το μέτρο επικάλυψης του εμβαδού της μάζας με το εμβαδόν του μικρότερου δυνατού περιγεγραμμένου ορθογωνίου. Ορίζεται δηλαδή ως:

$$rect = \frac{A}{AreaBoundingBox}$$

όπου A το εμβαδόν της μάζας και AreaBoundingBox το εμβαδόν του μικρότερου ορθογωνίου που μπορεί να την περιέχει [20].

2.1.3 Εκκεντρότητα

Η εκκεντρότητα χρησιμοποιείται ως μέτρο της επιμήκυνσης ενός σχήματος και πιο συγκεκριμένα για την αναγνώριση ελλειπτικών χαρακτηριστικών του. Για τον λόγο αυτό, ένας συμμετρικός πίνακας H ορίζεται ως:

$$H_{1,1} = \sum_{i=1}^N (x_i - X_0)^2$$

$$H_{2,2} = \sum_{i=1}^N (y_i - Y_0)^2$$

$$H_{1,2} = H_{2,1} = \sum_{i=1}^N (x_i - X_0) \cdot (y_i - Y_0)$$

όπου x_i και y_i είναι οι συντεταγμένες ενός τυχαίου pixel του περιγράμματος της μάζας, N ο αριθμός των pixels του περιγράμματος της μάζας και X_0 και Y_0 οι συντεταγμένες του γεωμετρικού κέντρου του.

Εάν τώρα λ_1 και λ_2 είναι οι ιδιοτιμές του πίνακα H , στην ελλειπτική προσέγγιση του ROI, οι άξονες μπορούν να αναπαρασταθούν ως:

$$S_1 = \sqrt{\left| \frac{\lambda_1}{2} \right|} \quad \text{και} \quad S_2 = \sqrt{\left| \frac{\lambda_2}{2} \right|}$$

και η εκκεντρότητα ορίζεται ως:

$$Eccent = \frac{S_1}{S_2}$$

με $S_1 < S_2$. Εκκεντρότητα κοντά στο 1 φανερώνει ότι η περιοχή ενδιαφέροντος προσεγγίζει τον κύκλο, ενώ τιμές κοντά στο μηδέν μια περιοχή πιο επιμηκυσμένη [19].

2.1.4 Κυρτότητα

Ορίζεται ως ο λόγος της περιμέτρου της περιοχής ενδιαφέροντος προς την περίμετρο του κυρτού σχήματος που την περιέχει.

$$Conv = \frac{Perimeter(Region)}{Perimeter(ConvexEnvelop)}$$

Εάν η μάζα έχει κανονικό σχήμα, με ομαλό περίγραμμα, ο λόγος αυτός θα ισούται με ένα. Σε οποιαδήποτε άλλη περίπτωση, δηλαδή όταν το περίγραμμα της μάζας είναι αστεροειδές, ο λόγος θα είναι μικρότερος [20].

2.1.5 Στερεότητα (Solidity)

Προσδιορίζει το ποσοστό των pixels του κυρτού περιγεγραμμένου σχήματος της μάζας που ανήκουν στη μάζα:

$$Solid = \frac{Area}{ConvexArea}$$

Εάν η μάζα έχει κανονικό σχήμα, το οποίο συνεπάγεται ότι είναι κυρτό, ο λόγος αυτός θα ισούται με ένα. Σε οποιαδήποτε άλλη περίπτωση ο λόγος θα είναι μικρότερος αφού μια μάζα με

αστεροειδές σχήμα θα καλύπτει μικρότερο ποσοστό του εμβαδού του κυρτού σχήματος που την περικλείει [20].

2.2 Μορφολογικά Χαρακτηριστικά

Η μελέτη των Μορφολογικών Χαρακτηριστικών είναι ίσως η πιο σημαντική στο να καθοριστεί το αν μια μάζα είναι πιο πιθανόν να είναι καλοήθης ή κακοήθης. Τα χαρακτηριστικά που συναντώνται πιο συχνά στη βιβλιογραφία είναι ο μέσος όρος και η τυπική απόκλιση της Κανονικοποιημένης Ακτινικής Απόστασης, η Εντροπία του Ακτινικού Μήκους, τα Zero-crossings της Ακτινικής Απόστασης, ο Δείκτης Τραχύτητας, η Γωνία Στροφής (Turning Angle) και πιο συγκεκριμένα ο Δείκτης Κυρτότητας που εξάγεται από αυτήν. Επιπλέον, μια άλλη κατηγορία χαρακτηριστικών σχήματος βασίζεται στους Τελεστές Fourier του περιγράμματος της μάζας. Τα χαρακτηριστικά που χρησιμοποιήθηκαν στην παρούσα εργασία είναι τα παρακάτω:

2.2.1 Μέσος Όρος Ακτινικής Απόστασης

Ο μέσος όρος ακτινικής απόστασης είναι το άθροισμα των κανονικοποιημένων Ευκλείδειων αποστάσεων ανάμεσα στο κέντρο μάζας και κάθε ένα pixel περιγράμματος [19]. Η κανονικοποίηση είναι απαραίτητη ώστε ο δείκτης να μην επηρεάζεται από το μέγεθος της κάθε μάζας και να έχουμε συγκρίσιμα αποτελέσματα. Οι ακτινικές Αποστάσεις κανονικοποιήθηκαν με τη χρήση της μέγιστης και ελάχιστης τιμής στη βάση με βάση τον παρακάτω τύπο:

$$\hat{d} = \frac{d(i) - d_{min}}{d_{max} - d_{min}}$$

Στη συνέχεια ο Μέσος Όρος Ακτινικής Απόστασης ορίστηκε ως:

$$\frac{1}{N} \sum_{i=1}^N d(i)$$

όπου:

$$d(i) = \sqrt{(x_i - X_0)^2 + (y_i - Y_0)^2}$$

2.2.2 Δείκτης Τραχύτητας

Για τον υπολογισμό του δείκτη τραχύτητας το περίγραμμα χωρίζεται σε μικρά τμήματα ίσου μήκους. Ο δείκτης υπολογίζεται για κάθε ένα τμήμα χρησιμοποιώντας τα εξής:

$$R(j) = \sum_{i=j}^{L+j} |d(i) - d(i+1)|, \text{ όπου } j = 1, \dots, \left[\frac{N}{L} \right]$$

και

$$R = \left[\frac{L}{N} \right] \sum_{j=1}^{N/L} R(j)$$

όπου $R(j)$ είναι ο δείκτης τραχύτητας για το j τμήμα, L είναι ο αριθμός των σημείων του περιγράμματος στο συγκεκριμένο τμήμα και N ο συνολικός αριθμός σημείων του περιγράμματος. Για τη συγκεκριμένη εργασία ορίσαμε $L = 16$ [21].

2.2.3 Εντροπία Ακτινικού Μήκους

Η Εντροπία Ακτινικού Μήκους είναι ένα πιθανολογικό μέτρο το οποίο υπολογίζεται με τη χρήση του ιστογράμματος του κανονικοποιημένου ακτινικού μήκους ως εξής:

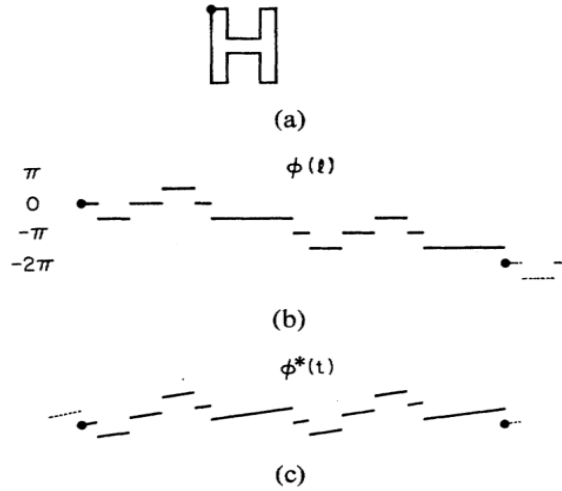
$$E = - \sum_{k=1}^{N_{bins}} P_k \log P_k$$

Η παράμετρος P_k είναι η πιθανότητα του να βρεθεί το κανονικοποιημένο ακτιμικό μήκος ανάμεσα στις αποστάσεις $d(i)$ και $d(i) + 1/N_{bins}$, N_{bins} είναι ο αριθμός των περιοχών στις οποίες χωρίζεται ο άξονας των x στο ιστόγραμμα που εκτείνεται στο $[0, 1]$. Στη δική μας εργασία επιλέχθηκε $N_{bins} = 100$ [21].

2.2.4 Δείκτης Κυρτότητας

Η έννοια της **γωνίας στροφής** (turning angle) έχει εισαχθεί εδώ και αρκετές δεκαετίες στον τομέα του pattern recognition [22]. Θεωρείται ως ένα εργαλείο για τον προσδιορισμό του σχήματος μιας κλειστής καμπύλης, συγκρίνοντας πρακτικά την γωνία κάθε σημείου μιας καμπύλης με τη γωνία ενός αρχικού σημείου αναφοράς. Ειδικά τα τελευταία χρόνια έχουν αναπτυχθεί μέθοδοι

ταξινόμησης που βασίζονται σε αυτήν με πολύ ελπιδοφόρα αποτελέσματα [23]. Στην εργασία μας, γωνία αναφοράς ορίστηκε η γωνία 0^0 . Παράδειγμα υπολογισμού της γωνίας στροφής φαίνεται παρακάτω:



Ο Δείκτης Κυρτότητας συνδυάζει πληροφορία σχετικά με την παρουσία κυρτών και κοίλων περιοχών στο περίγραμμα της μάζας. Ορίζεται ως:

$$CX_{TA} = 1 - \left(\frac{VR_{TA}}{2} + \frac{1 - XR_{TA}}{2} \right)$$

όπου

$$VR_{TA} = \frac{\sum_{i=1}^{Nd} (1 + \cos(\theta(i))) L_a(i)}{2 \sum_{i=1}^{Nd} L_a(i)}$$

όπου $L_a(i)$ είναι το άθροισμα των μηκών δύο διπλανών τμημάτων, ενωμένα από μια πτώση της turning angle $\theta(i)$ και Nd ο αριθμός των μειώσεων της turning angle. Ακόμα:

$$XR_{TA} = 1 - \frac{\sum_{j=1}^{Nb} (1 + \cos(\varphi(j))) L_b(j)}{2 \sum_{j=1}^{Nb} L_b(j)}$$

όπου $L_b(j)$ είναι το άθροισμα των μηκών δύο διπλανών τμημάτων, ενωμένα από μια αύξηση της

turning angle $\varphi(i)$ και N_b ο αριθμός των αυξήσεων της turning angle. Ο Δείκτης Κυρτότητας είναι κανονικοποιημένος στο διάστημα $[0,1]$. Σε μια κυρτή καμπύλη παίρνει την τιμή 1 ενώ μικραίνει όσο μεγαλώνει η παρουσία κοίλων περιοχών.

2.2.5 Zero-crossings Ακτινικής Απόστασης

Τα Zero-crossings Ακτινικής Απόστασης είναι τα σημεία στα οποία η Ακτινική Απόσταση τέμνει τον Μέσο Όρο Ακτινικής Απόστασης. Περισσότερες αυξομειώσεις στην Ακτινική Απόσταση δείχνουν μεγαλύτερη ανωμαλία στο περίγραμμα της μάζας [24].

2.3 Χαρακτηριστικά Υφής

Σημαντικό κομμάτι της διαγνωστικής διαδικασίας από τους ακτινολόγους είναι ο έλεγχος της υφής της περιοχής ενδιαφέροντος. Αν και δύσκολο να οριστεί επακριβώς, ως υφή μπορούμε να ορίσουμε διαισθητικά τις ιδιότητες μιας επιφάνειας οι οποίες μπορούν να την χαρακτηρίσουν ως ομαλή ή ανώμαλη, απαλή ή σκληρή, τραχιά ή λεία. Στην ασάφεια του ακριβούς ορισμού της διαφαίνεται και η δυσκολία που αντιμετωπίζουν οι ακτινολόγοι την οποία μπορούν να υπερνικήσουν μέσω της εμπειρίας και της εξάσκησης [2]. Τα συστήματα Υποβοηθούμενης Διάγνωσης μπορούν να εκμεταλλευτούν χαρακτηριστικά υφής υψηλότερης τάξης τα οποία δεν χρησιμοποιούνται από τους ακτινολόγους κατά τη διαδικασία διάγνωσης.

Όπως τις καταγράφουν οι Golzalez και Woods, οι μέθοδοι ανάλυσης υφής χωρίζονται σε τρεις βασικές κατηγορίες [41]:

- (α) Στατιστικές
- (β) Φασματικές/ενεργειακές
- (γ) Δομικές (Model – Based)

Οι στατιστικές μέθοδοι ασχολούνται με στατιστικά μεγέθη των περιοχών της εικόνας, όπως αυτά προκύπτουν από το ιστόγραμμα, τις ροπές ή τους πίνακες αυτοσυσχέτισης της αρχικής εικόνας. Οι πιο διαδεδομένες μέθοδοι που ανήκουν στη συγκεκριμένη κατηγορία είναι τα στατιστικά 1ης τάξης, τα στατιστικά 2ης τάξης, όπως μητρώα συνεμφάνισης και μητρώα χωρικής εξάρτησης επιπέδων του γκρι. Οι ενεργειακές/φασματικές τεχνικές βασίζονται στην ανάλυση του

φάσματος Fourier και χρησιμοποιούνται για να εντοπίσουν καθολική περιοδικότητα σε μια εικόνα ανιχνεύοντας κορυφές του φάσματος υψηλής ενέργειας. Εν γένει, οι συγκεκριμένες μέθοδοι βασίζονται στην εφαρμογή κάποιου μετασχηματισμού ή φίλτρου στην αρχική εικόνα. Τυπικά παραδείγματα μεθόδων της συγκεκριμένης κατηγορίας αποτελούν ο μετασχηματισμός Fourier και ο μετασχηματισμός κυματιδίου, ο διακριτός μετασχηματισμός συνημίτονου και τα φίλτρα Gabor. Τέλος, οι δομικές μέθοδοι χαρακτηρίζουν μια εικόνα υφής με βάση ένα γενικότερο μαθηματικό μοντέλο, το οποίο είναι είτε δομικό που χρησιμοποιεί τη γεωμετρική πολυπλοκότητα για την περιγραφή των μοτίβων της υφής (π.χ. η μέθοδος εκτίμησης κλασματικής διάστασης-fractals), είτε μοντέλο πιθανότητας το οποίο χαρακτηρίζει τη στοχαστική διαδικασία δημιουργίας υφής από την κατανομή των μοτίβων (π.χ. η μέθοδος Τυχαίων Μαρκοβιανών Πεδίων- Markov Random Field).

Τα πιο διαδεδομένα χαρακτηριστικά υφής είναι οι πίνακες Grey Level Co-Occurrence, η Wavelet ανάλυση, οι πίνακες Grey Level Run-Length, τα Completed Local Binary Pattern, Laws' Texture Energy Measures κ.ά. Στην παρούσα εργασία χρησιμοποιήθηκαν αλγόριθμοι υλοποίησης των πινάκων Grey Level Run-Length και Laws' Texture Energy Measures. Ακολουθεί μια σύντομη αναφορά στο θεωρητικό υπόβαθρο στο οποίο βασίστηκε η υλοποίηση των αλγορίθμων.

2.3.1. Laws' Energy Texture Measures

Η μέθοδος εξαγωγής αυτών των χαρακτηριστικών υφής προτάθηκε από τον Laws το 1979. Σύμφωνα με αυτήν, αρχικά κατασκευάζονται είκοσι πέντε μάσκες Laws, συνελίσνοντας πέντε πίνακες μιας διάστασης. Οι πυρήνες αυτοί είναι οι:

- $L5 = [1 \ 4 \ 6 \ 4 \ 1]$
- $E5 = [-1 \ -2 \ 0 \ 2 \ 1]$
- $S5 = [-1 \ 0 \ 2 \ 0 \ -1]$
- $W5 = [-1 \ 2 \ 0 \ -2 \ 1]$
- $R5 = [1 \ -4 \ 6 \ -4 \ 1]$

Στη συνέχεια οι είκοσι πέντε αυτές μάσκες φιλτράρουν την εικόνα, δημιουργώντας είκοσι πέντε Πίνακες Ενέργειας Υφής (Texture Energy Matrices). Λαμβάνοντας υπόψιν ότι από αυτές οι είκοσι είναι συμμετρικές, παράγονται δεκαπέντε TR εικόνες (αμετάβλητες περιστροφής) από τις οποίες εξάγονται πέντε στατιστικά μεγέθη πρώτης τάξης (μέσος όρος, τυπική απόκλιση, εύρος,

ασυμμετρία και κύρτωση) [25].

2.3.2. Πίνακες Grey Level Run-Length

Μια διαδρομή γκριζου επιπέδου είναι μια δομή διαδοχικών pixels προς μια κατεύθυνση, τα οποία έχουν το ίδιο επίπεδο έντασης (grey level value). Με αυτόν τον τρόπο δίνουν πληροφορία σχετικά με την ομοιογένεια της εικόνας. Στη συγκεκριμένη εργασία παράχθηκαν τέσσερις πίνακες Grey Level Run-Length, που αντιστοιχούν στις γωνίες 0° , 45° , 90° και 135° [25].

<i>Picture:</i>		0	1	2	3
		0	2	3	3
		2	1	1	1
		3	0	3	0

<i>Run length:</i>						
<i>G r a y l e v e l</i>	0°	1	2	3	4	
	0	4	0	0	0	
	1	1	0	1	0	
	2	3	0	0	0	
	3	3	1	0	0	

		45°	1	2	3	4
		0	4	0	0	0
		1	4	0	0	0
		2	0	0	1	0
		3	3	1	0	0

<i>G r a y l e v e l</i>	90°	1	2	3	4	
	0	2	1	0	0	
	1	4	0	0	0	
	2	3	0	0	0	
	3	3	1	0	0	

		135°	1	2	3	4
		0	4	0	0	0
		1	4	0	0	0
		2	3	0	0	0
		3	5	0	0	0

Εικόνα 8: Παράδειγμα σύνθεσης χαρακτηριστικών Laws

2.4 Αποτελέσματα Προηγούμενων Μελετών

Αρκετές μελέτες έχουν πραγματοποιηθεί πάνω στην ταξινόμηση μαστογραφικών μαζών, χρησιμοποιώντας τις προαναφερθείσες κατηγορίες χαρακτηριστικών. Οι μελέτες αυτές όμως είχαν συνήθως μικρή κλίμακα. Παράδειγμα είναι η μελέτη των Tang και Liu [15] στην οποία

χρησιμοποιήθηκαν 309 μαστογραφίες από την DDSM για σύγκριση μορφολογικών χαρακτηριστικών και χαρακτηριστικών συνόλου επιπέδων. Για τα μορφολογικά χαρακτηριστικά με χρήση LDA και SVM τα ποσοστά ακρίβειας ήταν 73% και 76% αντίστοιχα, ενώ για τα χαρακτηριστικά υφής 65% και για τους δύο ταξινομητές. Με τη χρησιμοποίηση και των δύο ομάδων χαρακτηριστικών η απόδοση ανέβηκε και έφτασε το 77% για ταξινομητή LDA και 78% για SVM. Το 2012 σε μια άλλη μελέτη [16] οι Wong et al εξήγαγαν 50 Περιοχές Ενδιαφέροντος από τη βάση MIAS και πραγματοποιώντας εκπαίδευση με 12 Grey Level Co-occurrence Matrices και ταξινόμηση με Νευρωνικό Δίκτυο, πέτυχαν τιμή ακριβείας 86%. Οι Choi και Ro [17] χρησιμοποιώντας Linear Binary Patterns στη βάση MIAS επέλεξαν 89 μαστογραφίες και κατάφεραν να τις ταξινομήσουν με ακρίβεια 0.96. Αντίστοιχα, επέλεξαν 303 μαστογραφίες από τη βάση DDSM τις οποίες ταξινόμησαν με ακρίβεια 0.87. Οι Mohanty et al το 2011 συνέλεξαν 254 ROIs από τη βάση DDSM και χρησιμοποιώντας τεχνικές GLCM έφτασαν ποσοστό ακρίβειας 94,9%. Τέλος, ο Cascio et al [19] συνέλεξαν 3762 ανεξάρτητες μαστογραφίες και χρησιμοποιώντας μορφολογικά χαρακτηριστικά κατάφεραν να έχουν επίδοση στην ακρίβεια 0.86 και ποσοστό ευαισθησίας 82%

Κεφάλαιο 3

Επιλογή Χαρακτηριστικών και Επισκόπηση Ταξινομητών

3.1 Επιλογή Χαρακτηριστικών

Ένα σύνηθες πρόβλημα στη στατιστική αναγνώριση προτύπων είναι η επιλογή χαρακτηριστικών. Ο όρος αυτός αναφέρεται σε μια διεργασία κατά την οποία ένας χώρος δεδομένων μετασχηματίζεται σε έναν χώρο χαρακτηριστικών. Ο μετασχηματισμός αυτός όμως σχεδιάζεται έτσι ώστε το σύνολο των δεδομένων να μπορεί να αναπαρασταθεί μέσω ενός μειωμένου αριθμού χαρακτηριστικών τα οποία διατηρούν το μεγαλύτερο μέρος της εγγενούς πληροφορίας που περιέχεται στα αρχικά δεδομένα. Το σύνολο δεδομένων, δηλαδή, υφίσταται μείωση της διαστασιμότητάς του. Στην περίπτωση που αυτός ο μετασχηματισμός δεν πραγματοποιηθεί, ο χώρος των χαρακτηριστικών έχει πολλές διαστάσεις, με αποτέλεσμα τα σημεία-δείγματα να είναι πολύ αραιά στο χώρο και ο διαχωρισμός τους πολύ δύσκολος. Το φαινόμενο αυτό ονομάζεται “κατάρρα της διαστασιμότητας”. Επιπρόσθετα, η επεξεργασία των δεδομένων μας πρέπει να είναι ταχύς ώστε οι ακτινολόγοι να μπορούν άμεσα να συμβουλευτούν τα αποτελέσματα των υπολογισμών του CAD. Τέλος, μειώνοντας τα χαρακτηριστικά που χρησιμοποιεί το σύστημα κατά την ταξινόμηση, μειώνεται και η πιθανότητα υπερπροσαρμογής του ταξινομητή (over-fitting).

Όλοι οι παραπάνω λόγοι καθιστούν σαφή τη σπουδαιότητα της εφαρμογής μεθόδων επιλογής χαρακτηριστικών έτσι ώστε να αναζητηθεί το κατάλληλο υποσύνολο των απαραίτητων αρχικών χαρακτηριστικών της βάσης δεδομένων. Υπάρχουν δύο κατηγορίες πρακτικών προσεγγίσεων για την επιλογή των χαρακτηριστικών. Οι προσεγγίσεις της πρώτης κατηγορίας ονομάζονται προσεγγίσεις φίλτρου (filter approaches) ή στατιστικές μέθοδοι και της δεύτερης προσεγγίσεις ενσωματωμένου ταξινομητή (wrapper approaches). Οι μέθοδοι της πρώτης κατηγορίας είναι γενικά ανεξάρτητες του ταξινομητή που χρησιμοποιείται και επιλέγουν

χαρακτηριστικά με βάση ένα μέτρο που συνήθως αντικατοπτρίζει τη διακριτική τους ικανότητα. Τα πλεονεκτήματά της είναι η ταχύτητα και η απλότητα των υπολογισμών που απαιτούν ενώ λόγω της ανεξαρτησίας τους από το σχήμα ταξινόμησης μπορούν να χρησιμοποιηθούν με πληθώρα ταξινομητών. Μειονέκτημά τους είναι το γεγονός ότι εξετάζουν κάθε χαρακτηριστικό μεμονωμένα χωρίς να υπολογίζουν τις εξαρτήσεις και τους συσχετισμούς μεταξύ των χαρακτηριστικών. Στις μεθόδους της δεύτερης προσέγγισης, ο ίδιος ο ταξινομητής χρησιμοποιείται ώστε να βρει το κατάλληλο υποσύνολο χαρακτηριστικών χρησιμοποιώντας στρατηγική μη-εξαντλητικής αναζήτησης. Έχουν το πλεονέκτημα ότι πλέον συνυπολογίζονται στην επιλογή του τελικού συνόλου οι συσχετισμοί και εξαρτήσεις μεταξύ διαφόρων χαρακτηριστικών. Τα μεγάλα μειονεκτήματά τους είναι η μεγάλη πολυπλοκότητα, η καθυστέρηση στην εύρεση του βέλτιστου υποσυνόλου ενώ επίσης υπάρχει και αυξημένος κίνδυνος για υπερπροσαρμογή στα δεδομένα.

Με βάση τις μεθόδους που αναλύθηκαν στο προηγούμενο κεφάλαιο αναπτύξαμε αλγόριθμους για την εξαγωγή 12 μορφολογικών χαρακτηριστικών και 85 χαρακτηριστικών υφής. Για την υλοποίηση της επιλογής χαρακτηριστικών στην παρούσα εργασία επιλέχθηκε ο αλγόριθμος T-test. Η μέθοδος αυτή είναι μια στατιστική μέθοδος η οποία δέχεται ανά δύο τα χαρακτηριστικά ως ανεξάρτητα τυχαία δείγματα και τα συγκρίνει ως προς μέσο όρο τους. Η αρχική υπόθεση της συνάρτησης είναι ότι τα δύο δείγματα που προέρχονται από κανονική κατανομή και έχουν ίδιο μέσο όρο και ίδια, αλλά άγνωστη, τυπική απόκλιση. Η μέθοδος εξάγει το test-στατιστικό ως εξής:

$$t = \frac{\hat{x} - \hat{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$$

όπου $\hat{x}$ και $\hat{y}$ είναι οι μέσοι όροι των δύο κατανομών x και y , s_x και s_y οι τυπικές αποκλίσεις τους και n και m ο αριθμός των δειγμάτων σε κάθε κατανομή ο οποίος στην περίπτωση μας είναι ίδιος.

Στη συνέχεια εξάγεται η εκ συνένωσης τυπική απόκλιση s ως εξής:

$$s = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}}$$

Τέλος, με βάση τους παραπάνω υπολογισμούς, η μέθοδος εξάγει p-τιμές που δείχνουν τη σημασία του αποτελέσματος για τη απόρριψη ή την αποδοχή της αρχικής υπόθεσης για ομοιότητα

των δύο κατανομών ενός χαρακτηριστικού. Για παράδειγμα, αν η p -τιμή για ένα χαρακτηριστικό βρεθεί ίση με 0.01 αυτό σημαίνει ότι κατά 99% μπορούμε να απορρίψουμε την αρχική υπόθεση περί ομοιότητας των δύο κατανομών. Προφανώς, όσο μικρότερες οι p -τιμές που υπολογίζονται για τα χαρακτηριστικά, τόσο μεγαλύτερη η σημαντικότητα του χαρακτηριστικού για το διαχωρισμό καλόηθων-κακόηθων περιπτώσεων.

Η ίδια διαδικασία επαναλαμβάνεται για όλα τα χαρακτηριστικά του δείγματος ώστε να υπολογίσουμε για το καθένα την p -τιμή μεταξύ της κατανομής του για τις καλοήθειες και κακοήθειες περιπτώσεις αντίστοιχα. Συνηθίζεται να θεωρούνται ως στατιστικά σημαντικά χαρακτηριστικά εκείνα που παρουσιάζουν p -τιμές μικρότερες του 0.05, οπότε και υπάρχει 95% διάστημα εμπιστοσύνης στη διαφορά των μέσων τιμών. Παρά το γεγονός ότι το συγκεκριμένο στατιστικό τεστ είναι ευρέως χρησιμοποιούμενο, παρουσιάζει το μειονέκτημα ότι προϋποθέτει κανονικές κατανομές των μετρήσεων και ίσες διακυμάνσεις μεταξύ των δειγμάτων.

3.2 Ταξινομητές

Στη συγκεκριμένη ενότητα παρουσιάζουμε τους ταξινομητές που χρησιμοποιήθηκαν στην υλοποίηση του σχήματος ταξινόμησης των μαστογραφικών μαζών. Όπως έχει ήδη αναφερθεί, ένας από τους στόχους της συγκεκριμένης εργασίας είναι η σύγκρισή τους ώστε να καταλήξουμε στον καταλληλότερο ταξινομητή για την κατηγοριοποίηση των μαστογραφικών μαζών εν γένει. Κάποιοι από τους πιο δημοφιλείς ταξινομητές είναι οι Μηχανές Υποστήριξης Διανυσμάτων (SVMs), τα πολυστρωματικά νευρωνικά δίκτυα (Multi-Layer Perceptrons), ο ταξινομητής k -κοντινότερων γειτόνων (k -NN), δίκτυα Ακτινικών Συναρτήσεων Βάσης (RDF), τα πιθανοτικά νευρωνικά δίκτυα (Probabilistic Neural Networks PNN), ο ταξινομητής Bayes, τα γενικευμένα δυναμικά fuzzy νευρωνικά δίκτυα καθώς και συστήματα μη επιβλεπόμενης ταξινόμησης όπως οι Χάρτες Αυτο-Οργάνωσης (SOMs). Στην παρούσα εργασία υλοποιούμε τους τρεις πιο συχνά απαντώμενους στη βιβλιογραφία ταξινομητές, τα SVMs, Πολυστρωματικά Νευρωνικά Δίκτυα και τον ταξινομητή k -NN ώστε να μπορούμε να συγκρίνουμε τα αποτελέσματά μας με όσο το δυνατόν περισσότερες υπάρχουσες εργασίες. Στις επόμενες ενότητες αναλύεται στο θεωρητικό υπόβαθρο πίσω από την υλοποίηση των ταξινομητών αυτών.

3.2.1 Κανονικοποίηση των Εισόδων

Πριν από την ανάλυση του τρόπου λειτουργίας των διαφόρων σχημάτων ταξινόμησης κρίνουμε απαραίτητο να αναφέρουμε έναν σημαντικό παράγοντα για την σωστή λειτουργία του ταξινομητών, την κανονικοποίηση των εισόδων. Εάν ένας ταξινομητής συνδυάζει τα δεδομένα μέσω μιας συνάρτησης απόστασης (όπως η Ευκλείδεια απόσταση που χρησιμοποιούμε στην παρούσα εργασία), η κανονικοποίηση των εισόδων μπορεί να είναι μείζονος σημασίας. Η συνεισφορά μιας εισόδου-χαρακτηριστικού θα εξαρτηθεί σε μεγάλο βαθμό από τη διαφοροποίησή της σε σχέση με τα υπόλοιπα χαρακτηριστικά. Μια είσοδος η οποία έχει ένα μεγάλο εύρος τιμών, θα συνεισφέρει περισσότερο στο μέτρο της απόστασης από ότι μια είσοδος με μικρότερο εύρος. Ένας ακόμη παράγοντας είναι η γρηγορότερη σύγκλιση του αλγορίθμου Κλίσης Καθόδου (Gradient Descent) κατά την ελαχιστοποίηση της Συνάρτησης Κόστους. Μετά την κανονικοποίηση των χαρακτηριστικών η καμπύλη λάθους παίρνει μια πιο κυκλική μορφή σε σχέση με την ελλειπτική πριν την κανονικοποίηση, κι άρα επιταχύνεται η σύγκλιση. Συνεπώς καταλήγουμε στο ότι το εύρος όλων των χαρακτηριστικών πρέπει να κανονικοποιείται ώστε να βελτιστοποιείται η μάθηση. Συνήθως επιλέγουμε να κλιμακώσουμε τα δεδομένα στο διάστημα $[-1,1]$, ανάλογα με τις τιμές της συνάρτησης ενεργοποίησης. Για τον ίδιο λόγο στην υλοποίησή μας κλιμακώνουμε τις εισόδους ώστε να έχουν μέσο όρο μηδέν και τυπική απόκλιση ένα, καθώς χρησιμοποιήσαμε σαν συνάρτηση ενεργοποίησης την σιγμοειδή “tansig”.

3.2.2 Γενίκευση και χρήση Μεθόδου k-fold Cross-Validation

Η ουσία της μάθησης με οπισθοδιάδοση σφάλματος είναι να κωδικοποιήσει την αντιστοίχιση εισόδων-εξόδων στα συναπτικά βάρη και τα κατώφλια ενός perceptron πολλών επιπέδων. Με αυτό τον τρόπο στοχεύουμε στο να καταλήξει το δίκτυο καλά εκπαιδευμένο, ώστε μαθαίνοντας αρκετά για το παρελθόν να μπορεί να γενικεύει για το μέλλον. Η διαδικασία της μάθησης, δηλαδή, ισοδυναμεί με το πρόβλημα επιλογής του βέλτιστου μοντέλου για το δίκτυο. Αυτό πραγματοποιείται επιλέγοντας το μοντέλο που ελαχιστοποιεί τη συνάρτηση κόστους σε ένα σύνολο. Εάν αυτό που θα χρησιμοποιηθεί θα είναι το σύνολο στο οποίο έχει εκπαιδευτεί το μοντέλο, υπάρχει σαφής πιθανότητα το μοντέλο που επιλέγεται με αυτόν τον τρόπο να καταλήξει

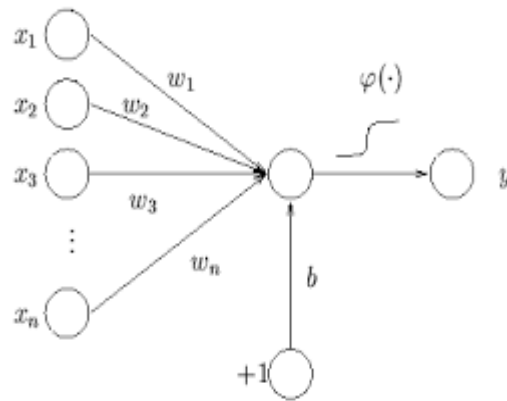
σε υπερπροσαρμογή. Αντίθετα, αν συγκρίνουμε τα μοντέλα σε ένα σύνολο στο οποίο δεν έχουν εκπαιδευτεί, αποκτούμε μια καλύτερη εκτίμηση για το λάθος γενίκευσης κι άρα για το καλύτερο μοντέλο. Σε αυτό το πλαίσιο, χρησιμοποιείται ένα καθιερωμένο εργαλείο της στατιστικής γνωστό ως διασταυρωμένη επικύρωση (cross-validation). Το σύνολο δεδομένων χωρίζεται αρχικά σε ένα δείγμα εκπαίδευσης και ένα σύνολο ελέγχου. Με τη σειρά του το δείγμα εκπαίδευσης διαμερίζεται περαιτέρω σε δύο ξένα υποσύνολα:

- ένα υποσύνολο εκτίμησης, το οποίο χρησιμοποιείται για την επιλογή του μοντέλου
- ένα σύνολο επικύρωσης, το οποίο χρησιμοποιείται για τον έλεγχο (την επικύρωση) του μοντέλου.

Η χρήση της μεθόδου διασταυρωμένης επικύρωσης είναι ιδιαίτερα ελκυστική όταν καλούμαστε να σχεδιάσουμε ένα μεγάλο νευρωνικό δίκτυο με στόχο την καλή γενίκευση. Το μειονέκτημα της μεθόδου είναι ότι στην ουσία “χαραμίζει” μέρος των αρχικών δεδομένων. Στην εργασία μας χρησιμοποιήσαμε τη μέθοδο χωρίζοντας στα αρχικά δεδομένα μας σε 80% σύνολο εκπαίδευσης και 20% σύνολο ελέγχου.

3.2.3 Νευρωνικό Δίκτυο

Ένα νευρωνικό δίκτυο είναι ένα σύστημα σχεδιασμένο να μοντελοποιεί τον τρόπο με τον οποίο ο εγκέφαλος εκτελεί μια συγκεκριμένη λειτουργία. Προσλαμβάνει τη γνώση από το περιβάλλον του, μέσω μιας διαδικασίας μάθησης, και αποθηκεύει τη γνώση αυτή στις συνδέσεις μεταξύ των νευρώνων του, με τη χρήση των συναπτικών βαρών. Ο νευρώνας είναι η θεμελιώδης μονάδα επεξεργασίας πληροφορίας η οποία παίζει καθοριστικό ρόλο στη λειτουργία ενός νευρωνικού δικτύου. Το σχηματικό διάγραμμα (& παρουσιάζει το μοντέλο ενός νευρώνα που αποτελεί τη βάση για τη σχεδίαση μιας μεγάλης οικογένειας νευρωνικών δικτύων. Συγκεκριμένα αποτελείται από ένα σύνολο n κλάδων διασύνδεσης (συνάψεων), έναν κόμβο άθροισης και μια συνάρτηση ενεργοποίησης. Αναλυτικότερα, το σήμα x_j στην είσοδο της σύναψης j πολλαπλασιάζεται με το συναπτικό βάρος w_j και στη συνέχεια προστίθεται μαζί με τα υπόλοιπα στον αθροιστή. Το αποτέλεσμα του αθροιστή, ύστερα από μια πόλωση b , οδηγείται σε μια συνάρτηση ενεργοποίησης (activation function) για τον περιορισμό του πλάτους του σήματος εξόδου του νευρώνα.



Εικόνα 9: Μη γραμμικό μοντέλο νευρώνα

Με μαθηματικούς όρους, μπορούμε να περιγράψουμε το νευρώνα με το ακόλουθο ζεύγος εξισώσεων:

$$u = \sum_{j=1}^m w_j x_j$$

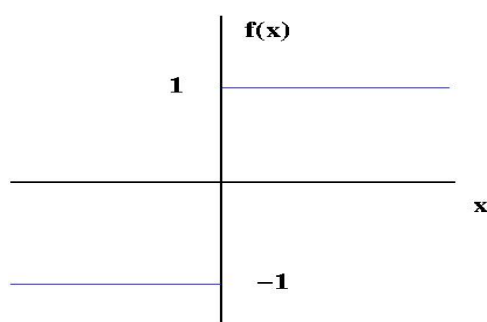
και

$$y = \phi(u + b)$$

Η συνάρτηση ενεργοποίησης $\phi(u)$ μπορεί να είναι είτε η συνάρτηση Κατωφλίου (ή Heaviside), είτε η σιγμοειδής, είτε η γραμμική, είτε η συνάρτηση ακτινικής βάσης. Η σιγμοειδής συνάρτηση είναι η πλέον κοινή μορφή συνάρτησης ενεργοποίησης που χρησιμοποιείται στην κατασκευή νευρωνικών δικτύων. Ορίζεται ως:

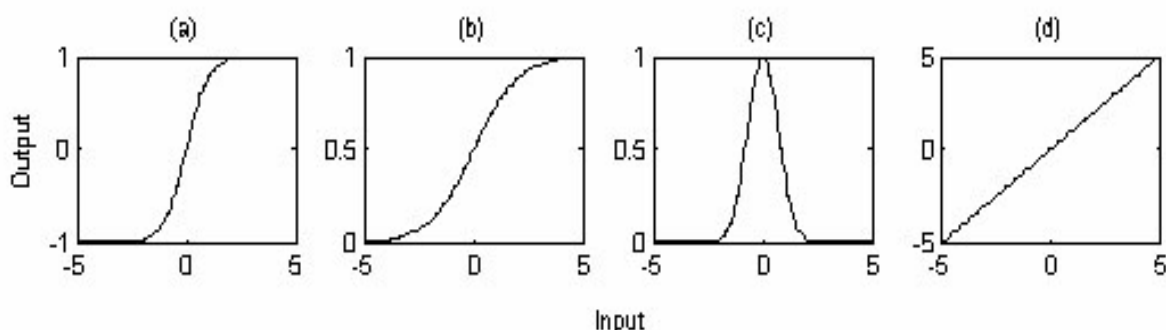
$$\phi(v) = \frac{1}{1 + e^{-\alpha v}}$$

όπου α είναι η παράμετρος κλίσης της σιγμοειδούς συνάρτησης. Η γραμμική συνάρτηση είναι αυτή που χρησιμοποιείται κατά κόρον στο τελευταίο επίπεδο του νευρωνικού δικτύου (output layer). Έπειτα από δοκιμές καταλήξαμε να θεωρήσουμε ως ιδανική συνάρτηση ενεργοποίησης για την υλοποίησή μας την 'tansig' για όλα τα επίπεδα, συμπεριλαμβανομένου και του επιπέδου εξόδου.



Εικόνα 10: Συνάρτηση Κατωφλίου

Στις παρακάτω εικόνες φαίνεται η μορφή των υπόλοιπων συναρτήσεων ενεργοποίησης:



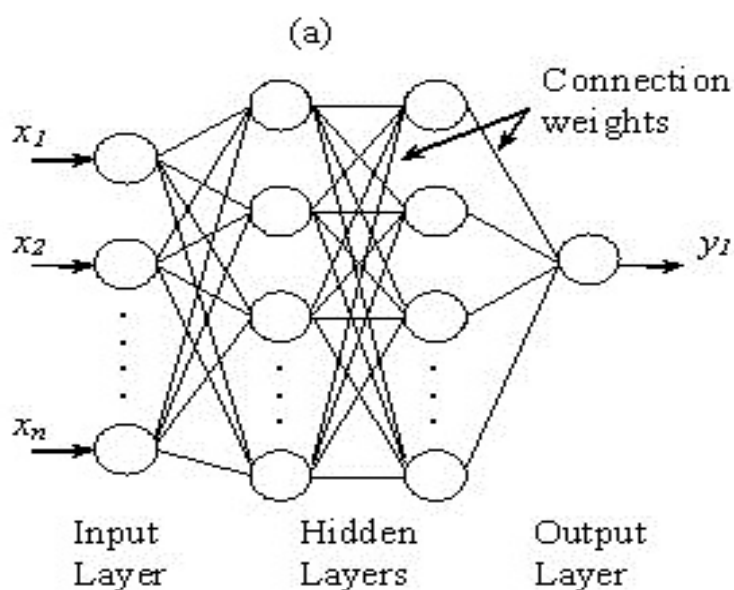
Εικόνα 11: Οι τρεις πιο κοινές συναρτήσεις ενεργοποίησης των κρυφών επιπέδων: (a) the *tanh-sigmoid* (*tansig*), (b) *log-sigmoid* (*logsig*) and (c) the *radial basis function* (*RBF*), ενώ η πιο κοινή συνάρτηση για το επίπεδο εξόδου είναι η (d) (*purelin*)

Οι διαφορετικές αρχιτεκτονικές νευρωνικών δικτύων προκύπτουν ανάλογα με τον τρόπο που οι νευρώνες συνδέονται μεταξύ τους. Κατά την εκπαίδευση του νευρωνικού δικτύου γίνεται χρήση των δεδομένων εκπαίδευσης ώστε να προσαρμοστούν κατάλληλα οι τιμές των βαρών των συνάψεων και οι τιμές πόλωσης των συναρτήσεων ενεργοποίησης. Στην παρούσα εργασία χρησιμοποιούμε το δίκτυο πρόσθιας τροφοδότησης (*feed forward*) με αρχιτεκτονική:

- ένα επίπεδο εισόδου
- ένα ή περισσότερα κρυμμένα επίπεδα
- ένα επίπεδο εξόδου

το οποίο εκπαιδεύεται με τον αλγόριθμο ανάστροφης διάδοσης σφάλματος (*Back Propagation*).

Στην εργασία μας αποτυπώσαμε τις επιδόσεις του νευρωνικού συστήματος με ένα και δύο κρυφά επίπεδα, οι οποίες και ακολουθούν σε επόμενο κεφάλαιο.



Εικόνα 12: Τυπική αρχιτεκτονική Artificial Neural Network

3.2.4 Μηχανές Διανυσμάτων Υποστήριξης

Η Μηχανή Διανυσμάτων Υποστήριξης είναι μια δυαδική μηχανή με δυνατότητα μάθησης δοθέντος ενός δείγματος εκπαίδευσης. Προτάθηκε το 1995 από τον Vapnik για δυαδικά προβλήματα ταξινόμησης, δηλαδή ταξινόμηση προτύπων σε δύο κατηγορίες. Ουσιαστικά κατασκευάζει ένα υπερεπίπεδο ως επιφάνεια απόφασης με τρόπο ώστε το περιθώριο διαχωρισμού μεταξύ των θετικών και αρνητικών παραδειγμάτων να μεγιστοποιείται. Τα διανύσματα υποστήριξης αποτελούνται από ένα μικρό υποσύνολο σημείων δεδομένων τα οποία εξάγει ο αλγόριθμος μάθησης από το ίδιο το δείγμα εκπαίδευσης. Θεωρώντας x το διάνυσμα εισόδου, w ένα προσαρμόσιμο διάνυσμα βαρών και b μια πόλωση ορίζεται ως επιφάνεια απόφασης το υπερεπίπεδο:

$$w^T \cdot x + b = 0$$

και αν ορίσουμε ως d_i την αντίστοιχη επιθυμητή απόκριση μπορούμε να γράψουμε:

$$\begin{aligned} w^T \cdot x + b &\geq 0 \quad , \text{για } d_i = +1 \\ w^T \cdot x + b &< 0 \quad , \text{για } d_i = -1 \end{aligned}$$

Τα συγκεκριμένα σημεία δεδομένων (x_i, d_i) , για τα οποία ή η πρώτη γραμμή ή η δεύτερη

ικανοποιείται με το σύμβολο ισότητας, αποκαλούνται διανύσματα υποστήριξης (support vectors) και είναι τα πιο δύσκολα στο να ταξινομηθούν.

Ως συναρτήσεις πυρήνα χρησιμοποιήθηκαν η ακτινικής βάσης συνάρτηση (radial basis function, RBF), η σιγμοειδής (sigmoid) συνάρτηση, η πολυωνυμική (polynomial) συνάρτηση και η γραμμική (linear) συνάρτηση, οι τύποι των οποίων δίνονται αντίστοιχα στον παρακάτω πίνακα:

Πίνακας 1: Πυρήνες Εσωτερικού Γινομένου για SVMs

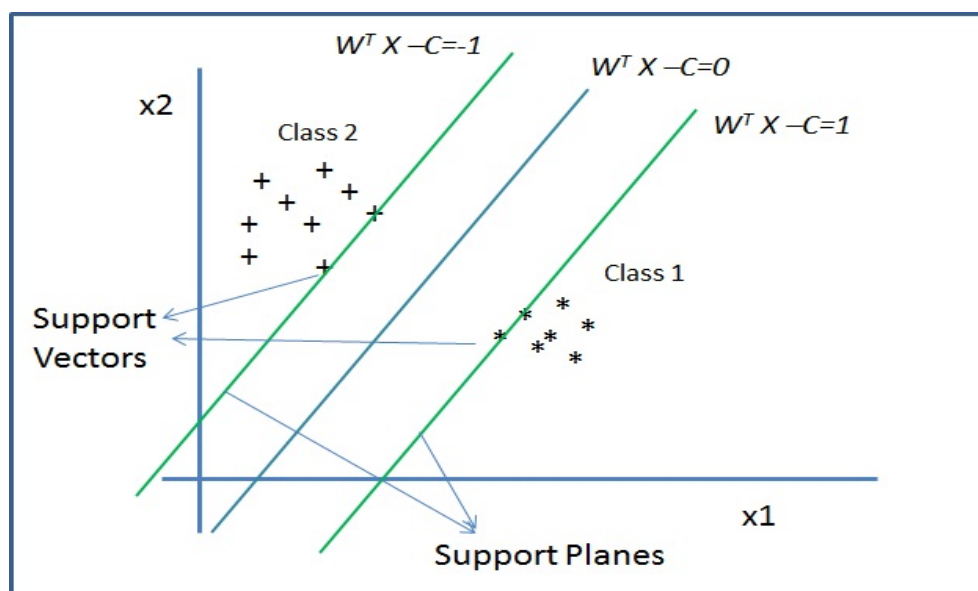
Τύπος Μηχανής	Πυρήνας Εσωτερικού Γινομένου $K(x, x_i)$, $i=1,...,N$	Σχόλια
Πολυωνυμική Μηχανή Μάθησης	$(x^T x_i + 1)^p$	Η δύναμη p καθορίζεται εκ των προτέρων από τον χρήστη
Γκαουσιανή Συνάρτηση Ακτινικής Βάσης	$e^{-\frac{\ x - x_i\ ^2}{2\sigma^2}}$	Το εύρος σ^2 καθορίζεται εκ των προτέρων από τον χρήστη
Perceptron δύο επιπέδων	$\tanh(\beta_0 x^T x + \beta_1)$	Το θεώρημα Mercer ικανοποιείται μόνο για κάποιες τιμές των β_0 και β_1

Για την εκπαίδευση του ταξινομητή απαιτείται ο καθορισμός των παραμέτρων εκπαίδευσης που εμφανίζονται στις εξισώσεις των πυρήνων εσωτερικού γινομένου $K(x, x_i)$. Για την περίπτωση της πολυωνυμικής συνάρτησης πρέπει να καθοριστεί η παράμετρος p που δείχνει το βαθμό του πολυωνύμου, Για την συνάρτηση ακτινικής βάσης ορίζουμε την παράμετρο g η οποία είναι αντιστρόφως ανάλογη του τετραγώνου της διακύμανσης σ^2 . Σε όλες τις περιπτώσεις, πρέπει να καθοριστεί η παράμετρος C , που αποτελεί ένα συντελεστή βαρύτητας των σφαλμάτων εκπαίδευσης κατά τον προσδιορισμό του υπερεπιπέδου [Burges 1998]. Είναι διαφορετικά ο συμβιβασμός μεταξύ του σφάλματος εκπαίδευσης και της ομαλότητας της λύσης. Με την αύξηση του C μικραίνει το περιθώριο διαχωρισμού στο υπερεπίπεδο κι άρα περισσότερα σημεία ταξινομούνται σωστά. Με αυτό τον τρόπο όμως δημιουργείται ο κίνδυνος να χαθούν ιδιότητες γενίκευσης του ταξινομητή καθώς θα επιδιώξει να ταιριάζει όσο καλύτερα γίνεται τα παραδείγματα του συνόλου εκπαίδευσης, τα οποία όμως μπορεί να εμπεριέχουν και πιθανά λάθη της βάσης. Προφανώς, διαφορετικά ζευγάρια τιμών αυτών των παραμέτρων οδηγούν σε διαφορετικά αποτελέσματα εκπαίδευσης και

κατ' επέκταση σε διαφορετικές επιδόσεις ταξινόμησης.

Στην συγκεκριμένη εργασία χρησιμοποιήθηκε η συνάρτηση ακτινικής βάσης (rbf) για τον πυρήνα του SVM και επιλέχθηκαν, έπειτα από δοκιμές στο σύστημα ταξινόμησης, τιμές παραμέτρων για το $g \in \{2^{-5}, \dots, 2^9\}$ και το $C \in \{2^{-3}, \dots, 2^{15}\}$.

Οι μηχανές διανυσμάτων υποστήριξης έχει καθιερωθεί ως ο πιο συχνά χρησιμοποιούμενος αλγόριθμος βασιζόμενος σε πυρήνα μάθησης, λόγω της καλής τους απόδοσης στη γενίκευση, της σχετικής ευκολίας χρήσης και της αυστηρής θεωρητικής θεμελίωσή τους. Ωστόσο, το βασικό μειονέκτημα των μηχανών διανυσμάτων υποστήριξης είναι η ταχεία αυξανόμενη απαίτησή τους σε υπολογιστικούς πόρους και αποθήκευση, λόγω του πλήθους των παραδειγμάτων εκπαίδευσης.



Εικόνα 13: Σχηματική αναπαράσταση βέλτιστου υπερεπίπεδου για γραμμικός διαχωρίσιμα πρότυπα

3.2.5 Ταξινομητής K-κοντινότερων Γειτόνων

Η λειτουργία του ταξινομητή k-κοντινότερων γειτόνων (k-NN) στηρίζεται στον υπολογισμό της ομοιότητας μεταξύ ενός διανύσματος προς ταξινόμηση και των υπόλοιπων διανυσμάτων του συνόλου εκπαίδευσης. Ως μέτρο ομοιότητας χρησιμοποιείται η Ευκλείδεια απόσταση. Λόγω της απλότητας και του μικρού υπολογιστικού του κόστους επιλέγεται συχνά σε εφαρμογές πάνω σε πολυποίκιλα προβλήματα.

Πιο συγκεκριμένα, αρχικά ο k-NN ταξινομητής υπολογίζει την ευκλείδεια απόσταση μεταξύ του διανύσματος που θέλουμε να ταξινομήσουμε και κάθε διάνυσμα του συνόλου

εκπαίδευσης. Η κατηγοριοποίησή του γίνεται εξετάζοντας μόνο τα k πιο κοντινά διανύσματα από το σύνολο εκπαίδευσης, δηλαδή τα k διανύσματα με τις χαμηλότερες τιμές ευκλείδειας απόστασης. Η απόσταση του καθενός από τα k αυτά διανύσματα από το άγνωστο διάνυσμα είναι ο κύριος παράγοντας που καθορίζει την συνεισφορά του στην τελική κατηγοριοποίηση. Στη συνέχεια, μεταξύ των γειτόνων αυτών πραγματοποιείται ψηφοφορία, όπου η ψήφος του κάθε γείτονα προκύπτει από την ακόλουθη σχέση:

$$Y_i = \frac{1}{D_i + 1}$$

όπου Y_i η ψήφος του i -γειτόνα και D_i η ευκλείδεια απόστασή του από το άγνωστο διάνυσμα [Dudani, 1976]. Οι ψήφοι όλων των γειτόνων αθροίζονται και το άγνωστο διάνυσμα κατηγοριοποιείται στην κλάση που συγκεντρώνει το υψηλότερο άθροισμα ψήφων. Συγκεκριμένα, ο k -NN ταξινομητής χρησιμοποιεί την ακόλουθη συνάρτηση απόφασης:

$$F = \text{sgn}(\sum_{i=1}^b Y_i - \sum_{j=1}^m Y_j)$$

όπου b είναι ο αριθμός των διανυσμάτων-γειτόνων που ανήκουν στην κλάση καλοήθειας και m ο αντίστοιχος αριθμός των διανυσμάτων-γειτόνων που είναι κακοήθεις. Προφανώς ισχύει $b+m=k$.

Στην ακραία περίπτωση όπου $k=1$, δεν διεξάγεται ψηφοφορία, μιας και αναζητείται μόνο ο κοντινότερος γείτονας, δηλαδή το διάνυσμα με την μικρότερη απόσταση. Ο ταξινομητής στην περίπτωση αυτή δεν πραγματοποιεί ψηφοφορία, αλλά κατατάσσει απευθείας το άγνωστο διάνυσμα στην κλάση που ανήκει ο κοντινότερος γείτονας.

Είναι φανερό ότι ο k NN αλγόριθμος είναι μια μη-παραμετροποιήσιμη μέθοδος. Η μόνη παράμετρος την οποία μπορούμε να προσδιορίσουμε είναι η επιλογή του αριθμού των γειτόνων k και τον τρόπο υπολογισμού της απόστασης. Κατά τη διάρκεια της υλοποίησής μας δοκιμάστηκαν: η Ευκλείδεια απόσταση, το άθροισμα των απολύτων διαφορών και η συσχέτιση, με καλύτερα αποτελέσματα να δίνει η ευκλείδεια απόσταση.

Κεφάλαιο 4

Μετρήσεις και Αποτελέσματα

Όπως έχουμε ήδη επισημάνει παρά το γεγονός ότι τα τελευταία χρόνια πολλές μελέτες έχουν δημοσιευτεί σχετικά με την ταξινόμηση των μαστογραφικών ευρημάτων, δεν έχουμε ακόμα στα χέρια μας μια καθολική εικόνα για την αποτελεσματικότητα της κάθε μεθόδου ταξινόμησης. Το γεγονός αυτό οφείλεται στο ότι οι περισσότερες από αυτές τις μελέτες έχουν πραγματοποιηθεί σε μικρές βάσεις μαστογραφιών, οι οποίες επιπρόσθετα είναι διαφορετικές για κάθε έρευνα που πραγματοποιείται. Λόγω του μικρού μεγέθους του συνόλου δεδομένων το σύστημα οδηγείται σε υπερεκπαίδευση, δηλαδή σε πολύ καλή επίδοση διαχωρισμού για τα δεδομένα εκπαίδευσης όμως με πολύ χαμηλή ικανότητα γενίκευσης. Επιπρόσθετα, συνήθως δεν γίνεται καμία αναφορά στην ποιότητα των μαστογραφικών εικόνων κι έτσι δεν διερευνάται η επίδοση του συστήματος υποβοηθούμενης διάγνωσης σε ευδιάκριτες και δυσδιάκριτες περιπτώσεις. Στο κεφάλαιο αυτό παρουσιάζουμε τα αποτελέσματα της ταξινόμησης σε μια μεγάλη και δημόσια βάση, εντοπίζοντας την αποδοτικότερη ομάδα χαρακτηριστικών σε συνδυασμό με την επιλογή του καταλληλότερου ταξινομητή. Παράλληλα πραγματοποιούμε μια διερεύνηση των παραμέτρων που ενσωματώνονται σε κάθε ταξινομητή.

4.1 Σύνολο Δεδομένων

Για να στοχεύσουμε στα προβλήματα που αναφέραμε επιλέγουμε μια βάση μαστογραφιών με πληθώρα διαφορετικών περιπτώσεων μαστογραφίας η οποία είναι ελεύθερα προσβάσιμη από την επιστημονική κοινότητα. Η βάση που χρησιμοποιήθηκε είναι η DDSM (Digital Database for Screening Mammography) του Πανεπιστημίου της Νότιας Φλόριντα.

Η βάση μαστογραφιών DDSM περιέχει μαστογραφίες με ευρήματα, είτε καλοήγη είτε κακοήγη, καθώς και μαστογραφίες χωρίς ευρήματα. Σχεδόν για όλες τις περιπτώσεις περιέχονται εικόνες μαστογραφίας από δύο διαφορετικές λήψεις : (α) MedioLateral (MLO - μεσοπλάγια λοξή

όψη) και (β) Cranio-caudal (CC – κρανιο-ουραία όψη). Για κάθε εικόνα υπάρχουν συνοδευτικές πληροφορίες για την ακριβή θέση του ευρήματος καθώς και το αποτέλεσμα της βιοψίας που έχει προηγηθεί, ώστε να είναι γνωστή η ιστοπαθολογική του κατάσταση (καλοήθεια/κακοήθεια). Παράλληλα, παρέχονται στοιχεία από τον ιατρικό φάκελο της ασθενούς (π.χ. ηλικία) και πληροφορίες για την ποιότητα των μαστογραφικών εικόνων και τη διαδικασία ψηφιοποίησής τους. Όλες οι εικόνες της βάσης είναι ψηφιοποιημένες και έχουν ψηφιοποιηθεί με χρήση ενός από τους ακόλουθους ψηφιοποιητές: (α) LUMISYS 2000 σε 12bits με χωρική ανάλυση 50μm, (β) HOWTEK MultiRad 850 και HOWTEK 960 με ανάλυση 43.5μm και (γ) ψηφιοποιητής DBA (M2100 ImageCLear) με ανάλυση 42μm.

Από τη βάση απομονώσαμε 1214 μαστογραφικές απεικονίσεις σε μεσοπλάγια λοξή όψη (MLO) που περιέχουν ευρήματα μαζών. Επιπρόσθετα στη συγκεκριμένη εργασία διαχωρίζουμε τις μαστογραφικές μάζες με βάση δύο κριτήρια, την πυκνότητα (density) και τη διακριτικότητα τους (subtlety) και επιχειρούμε να αποφανθούμε για το πως οι δυο αυτές ιδιότητες επηρεάζουν την απόδοση στην ταξινόμηση. Στον παρακάτω πίνακα αποτυπώνεται εποπτικά το σύνολο των περιπτώσεων χωρισμένες ανά κατηγορία πυκνότητας και διακριτικότητας που χρησιμοποιήθηκαν ως είσοδος στο σύστημα.

Πίνακας 2: Καταμερισμός Μαστογραφιών στο σύνολο

		Διακριτικότητα					
		1	2	3	4	5	Άθροισμα
Πυκνότητα	1	3	21	26	38	112	200
	2	22	45	89	155	221	533
	3	17	53	85	96	86	337
	4	17	25	51	33	18	144
	Άθροισμα	59	144	253	322	437	1214

Χρησιμοποιώντας τα διανύσματα των χαρακτηριστικών που έχουμε εξάγει από τις παραπάνω μαστογραφίες με βάση τις μεθόδους του Κεφαλαίου 2, δημιουργούμε μια ενιαία δομή την οποία στη συνέχεια προωθούμε στους ταξινομητές.

4.2 Αξιολόγηση Ταξινομητών

Έχουμε ήδη εξηγήσει σε προηγούμενο κεφάλαιο τη σημασία της διασταυρωμένης επικύρωσης για την ικανοποιητική γενίκευση του συστήματος ταξινόμησης. Στην εργασία μας χρησιμοποιούμε την 10-fold cross-validation η οποία έχει χρησιμοποιηθεί σε προηγούμενες μελέτες με πολύ καλά αποτελέσματα. [5,7]. Πριν την εκπαίδευση των ταξινομητών το διαθέσιμο σύνολο των 1214 μαστογραφιών χωρίζεται σε 10 ισομεγέθη και μεταξύ τους ξένα υποσύνολα, ώστε να μην υπάρχει κανένα κοινό διάνυσμα μεταξύ των υποσυνόλων. Η εκπαίδευση και αξιολόγηση των ταξινομητών γίνεται συνολικά 10 φορές. Σε κάθε επανάληψη, ως σύνολο εκπαίδευσης χρησιμοποιείται η ένωση των 9 υποσυνόλων, ενώ το εναπομένον υποσύνολο χρησιμοποιείται ως σύνολο ελέγχου των επιδόσεων του ταξινομητή. Έτσι, ο κάθε ταξινομητής ελέγχεται σε ένα σύνολο ελέγχου με άγνωστα δεδομένα τα οποία δεν χρησιμοποιήθηκαν καθόλου στην εκπαίδευσή του. Σε κάθε επανάληψη, χρησιμοποιείται επίσης διαφορετικός συνδυασμός υποσυνόλων ώστε κάθε ένα από τα 10 ξένα υποσύνολα να χρησιμοποιείται μόνο μια φορά ως σύνολο ελέγχου. Σημειώνουμε εδώ ότι για την εκπαίδευση και ταξινόμηση των απεικονίσεων συγκεκριμένων πολύ μικρότερων υποκατηγοριών χρησιμοποιήθηκε διαχωρισμός σε λιγότερα ξένα υποσύνολα. Επιπρόσθετα, για την καλύτερη δυνατή στατιστική εικόνα της απόδοσης των ταξινομητών ολόκληρη η διαδικασία επαναλήφθηκε 20 φορές από τις οποίες εξάγαμε τελικά έναν μέσο όρο απόδοσης.

Για την εκτίμηση της απόδοσης αυτής απαιτείται η μέτρηση των επιδόσεων με σαφώς ορισμένα μεγέθη, έτσι ώστε να είναι απολύτως συγκρίσιμη η αποτελεσματικότητα ανάμεσα σε όλα τα σχήματα ταξινόμησης. Το πρόβλημα της διάγνωσης μιας μαστογραφικής μάζας αποτελεί ένα δυαδικό πρόβλημα. Η μάζα μπορεί να είναι είτε κακοήθης (θετικό δείγμα) είτε καλοήθης (αρνητικό δείγμα) σύμφωνα με τη βιοψία που γίνεται στην περιοχή για τον προσδιορισμό της ιστοπαθολογικής της κατάστασης. Επιπλέον, η πρόβλεψη του χρησιμοποιούμενου ταξινομητή μπορεί να είναι είτε αληθής (επιτυχής πρόβλεψη) είτε ψευδής (ανεπιτυχής πρόβλεψη). Έχοντας καθορίσει αυτές τις τέσσερις εκβάσεις μπορούμε να ορίσουμε τέσσερις μετρικές που τις συνδυάζουν ως εξής:

- Αληθώς θετική ένδειξη (True Positive - TP), σε περίπτωση που η περιοχή είναι κακοήθης και ο ταξινομητής πραγματοποιήσει σωστή πρόβλεψη.
- Αληθώς αρνητική ένδειξη (True Negative - TN), στην περίπτωση κατά την οποία ο

ταξινομητής προβλέψει επιτυχώς ότι η μάζα είναι καλοήθης.

- Ψευδώς θετική ένδειξη (False Positive - FP), όταν μια καλοήθης μάζα κατηγοριοποιείται ως κακοήθης από τον ταξινομητή.
- Ψευδώς αρνητική ένδειξη (False Negative - FN), όταν μια κακοήθης μάζα κατηγοριοποιείται από τον ταξινομητή ως καλοήθης.

Είναι φανερό ότι στις πρώτες δύο περιπτώσεις έχουμε μια επιτυχημένη ταξινόμηση της μάζας ενώ στις δύο τελευταίες μια λανθασμένη. Στόχος όλων των μεθόδων ταξινόμησης είναι ακριβώς η ελαχιστοποίηση αυτών των λαθών η οποία έχει δυστυχώς μεγάλες συνέπειες. Ίσως να μπορούμε να θεωρήσουμε ως την χειρότερη από αυτές τις δύο αποτυχίες την τελευταία (Ψευδώς αρνητική ένδειξη) αφού προτείνει ως καλοήθη μια καρκινική μάζα και την αθώνει από την περαιτέρω εξέτασή της. Για το λόγο αυτό όλα τα συστήματα υποβοηθούμενης διάγνωσης εστιάζονται στο να κρατήσουν την FN ένδειξη χαμηλή, γεγονός που έρχεται με ένα τίμημα. Αυτό δεν είναι άλλο από τη μεγάλη ευαισθησία που αναπτύσσουν τα συστήματα στην προσπάθειά τους να μην παραβλέψουν καμία περίπτωση καρκίνου. Κατά συνέπεια πολλά καλοήθη ευρήματα κατηγοριοποιούνται ως ύποπτα για κακοήθεια γεγονός που αποτελεί μεγάλη πρόκληση για όλα τα συστήματα υποβοηθούμενης διάγνωσης.

Σαν στατιστικά μεγέθη επίδοσης σε δυαδικά προβλήματα, χρησιμοποιώντας τις προηγούμενες ενδείξεις, ορίζονται τα ακόλουθα μεγέθη:

- Ακρίβεια

Η Ακρίβεια (Accuracy) σχετίζεται με την ικανότητα του ταξινομητή να εντοπίσει σωστά όλες τις περιπτώσεις του συνόλου ελέγχου.

$$\text{Ακρίβεια} = \frac{\text{Πλήθος Αληθινών Ενδείξεων}}{\text{Πλήθος Συνόλου Περιπτώσεων}} = \frac{TPs + TNs}{TPs + TNs + FPs + FNs}$$

Παρά το γεγονός ότι αποτελεί ένα άμεσο τρόπο μέτρησης του ποσοστού επιτυχίας, η λειτουργικότητά του σε συγκεκριμένες συνθήκες μπορεί να αποδειχθεί περιορισμένη. Συγκεκριμένα, εξαρτάται σε μεγάλο βαθμό από το ποσοστό εμφάνισης κακοήθων περιπτώσεων, με αποτέλεσμα να μην είναι αντιπροσωπευτικό μέγεθος σε δείγματα χωρίς ισορροπία, όπου ο αριθμός των θετικών ή αρνητικών περιπτώσεων υπερτερεί κατά πολύ σε σχέση με τις περιπτώσεις της άλλης κατηγορίας.

- Ευαισθησία

Η Ευαισθησία (Sensitivity) δείχνει τον ικανότητα του ταξινομητή να καταδεικνύει ως θετικές τις περιπτώσεις στις οποίες όντως υπάρχει καρκινική παθογένεια. Ορίζεται ως:

$$\text{Ευαισθησία} = \frac{\text{Πλήθος Αληθινών Θετικών Ενδείξεων}}{\text{Πλήθος Συνόλου Θετικών Περιπτώσεων}} = \frac{TPs}{TPs + FNs}$$

- Ειδικότητα

Η Ειδικότητα (Specificity) σχετίζεται με την ικανότητα του ταξινομητή να εντοπίζει ως αρνητικές περιπτώσεις μαζών που δεν έχουν καρκινική παθογένεια.

$$\text{Ειδικότητα} = \frac{\text{Πλήθος Αληθινών Αρνητικών Ενδείξεων}}{\text{Πλήθος Συνόλου Αρνητικών Περιπτώσεων}} = \frac{TNs}{TNs + FPs}$$

Τα παραπάνω μεγέθη αποτέλεσαν τα σημεία σύγκρισης των μεθόδων που υλοποιήθηκαν κατά τη διάρκεια της εργασίας και παραθέτονται στις επόμενες ενότητες.

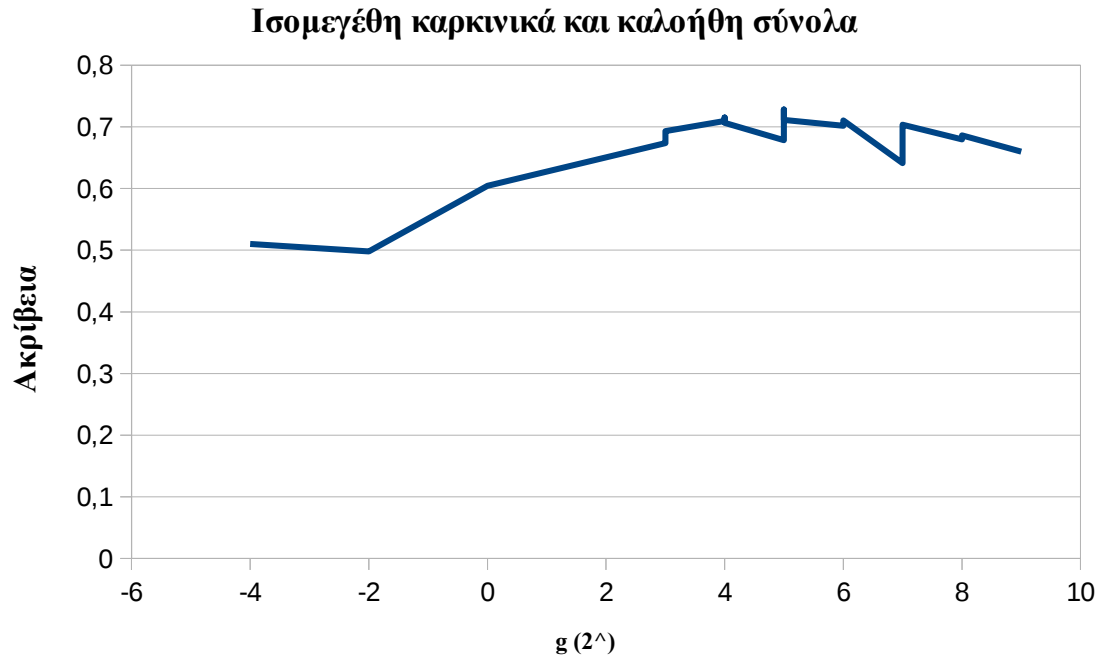
4.3 Σύγκριση Ομάδων Χαρακτηριστικών ανά Ταξινομητή

4.3.1 Υλοποίηση σε Μηχανές Υποστήριξης Διανυσμάτων

Στην ενότητα αυτή παρουσιάζονται τα αποτελέσματα της κατηγοριοποίησης με χρήση ταξινομητή SVM. Αρχικά χρησιμοποιούνται όλα τα διαθέσιμα χαρακτηριστικά που έχουμε εξάγει και στη συνέχεια απομονώνονται και συγκρίνονται τα μορφολογικά χαρακτηριστικά με αυτά της υφής. Όταν ολοκληρή η βάση με τις 1214 μαστογραφίες χρησιμοποιείται, το SVM παρουσιάζει ακρίβεια 72,82%. Για την απόδοση αυτή χρησιμοποιούνται οι παράμετροι στις τιμές $C = 2^9$ και $g = 2^5$. Στις τιμές αυτές καταλήξαμε έπειτα από έλεγχο της παραμετροποίησης για τις τιμές που φαίνονται στις Εικόνες 13 και 14.

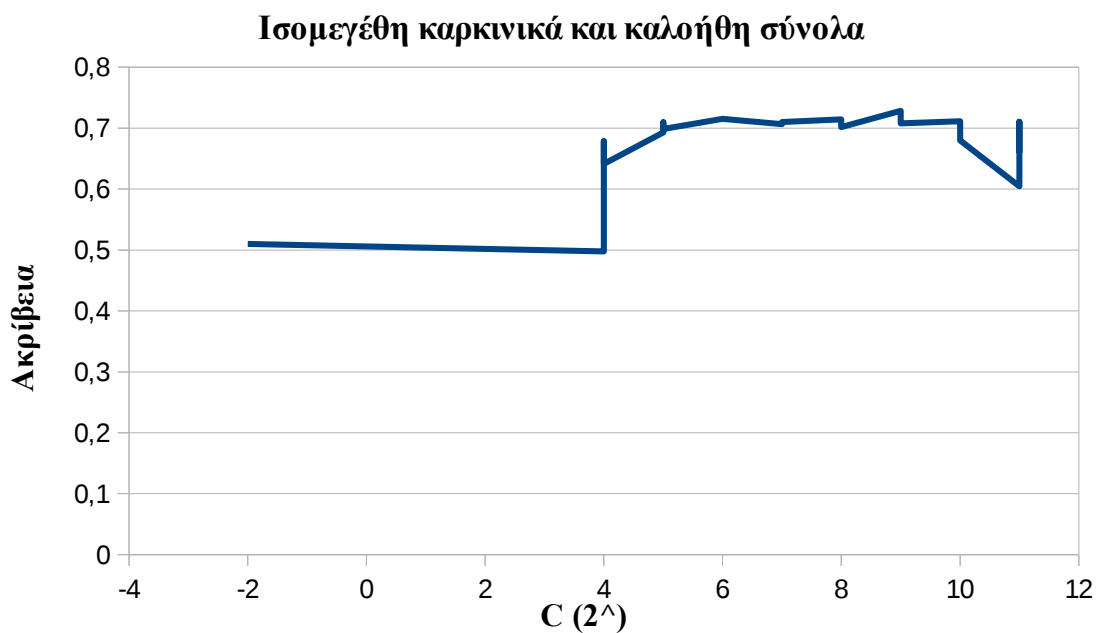
Αρχικά για την περίοδο εκπαίδευσης του SVM επιλέχθηκε ένα σύνολο εκπαίδευσης με ίσο αριθμό καρκινικών και καλοήθων μαζών. Στη συνέχεια συγκρίναμε τις επιδόσεις του SVM όταν στη βάση εκπαίδευσης έχουμε άνισα μέλη από τις δύο αυτές ομάδες.

Η μεταβολή της ακρίβειας σε συνάρτηση με τη μεταβολή του C :



Εικόνα 14: Εξέλιξη ακρίβειας του SVM συναρτήσει του g

Η μεταβολή των επιδόσεων του SVM σε συνάρτηση με το C δείχνει ότι σε μεγαλύτερες τιμές της



Εικόνα 15: Εξέλιξη ακρίβειας του SVM συναρτήσει του C

παραμέτρου C παρατηρούνται μεγαλύτερες τιμές ακριβείας για το σύστημα. Το γεγονός αυτό αποτυπώνεται στην Εικόνα 14.

Παρατηρούμε ότι, σύμφωνα και με τη θεωρία, η εκπαίδευση του SVM σε ισομεγέθη σύνολα καρκινικών και καλοηθών μαζών έχει σαν αποτέλεσμα βελτιωμένες επιδόσεις ακρίβειας. Αντίθετα, όταν χρησιμοποιείται όλο το δείγμα των 1214 μαστογραφιών έχουμε βελτιωμένες επιδόσεις στην ευαισθησία του συστήματος αφού σε όλη τη βάση υπάρχουν περισσότερες περιπτώσεις καρκινικών μαζών από ότι καλοήθεις (700 και 514 αντίστοιχα). Στον Πίνακα 3 παραθέτονται ενδεικτικά 20 τιμές των παραμέτρων και τα αντίστοιχα αποτελέσματα και για τις τρεις ενδείξεις επίδοσης του συστήματος.

Πίνακας 3: Επιδόσεις SVM για ισομεγέθη καρκινικά-καλοήθη σύνολα εκπαίδευσης

Παράμετρος C (2 [^])	Παράμετρος g (2 [^])	Ακρίβεια	Ευαισθησία	Ειδικότητα
-2	-4	0,51007	0,23365	0,78695
4	-2	0,49776	0,23136	0,76445
11	0	0,60457	0,82984	0,37967
-2	3	0,67366	0,70220	0,64521
0	3	0,67523	0,68340	0,66714
5	3	0,69293	0,63086	0,75493
5	4	0,70953	0,62165	0,79729
6	4	0,71518	0,63667	0,79354
7	4	0,70655	0,61337	0,79954
4	5	0,67822	0,62650	0,72987
5	5	0,69842	0,61533	0,78137
7	5	0,70975	0,61899	0,80040
8	5	0,71420	0,62348	0,80453
9	5	0,72817	0,63397	0,80212
10	5	0,71139	0,62324	0,79925
8	6	0,70178	0,60175	0,80161
9	6	0,70766	0,60703	0,80819
11	6	0,710226	0,61909	0,80114
11	7	0,70340	0,59440	0,81223

Παρατηρούμε ικανοποιητικές τιμές ακρίβειας και καλές τιμές ειδικότητας όμως χαμηλή επίδοση στην ευαισθησία. Αντίστοιχα, παρουσιάζουμε ενδεικτικά κάποια αποτελέσματα της Μηχανής Διανυσμάτων Υποστήριξης κατόπιν εκπαίδευσης με σύνολο εκπαίδευσης που περιλαμβάνει ίσο πλήθος καρκινικών και καλοηθών μαζών. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 4.

Πίνακας 4: Επιδόσεις SVM με εκπαίδευση σε όλη βάση

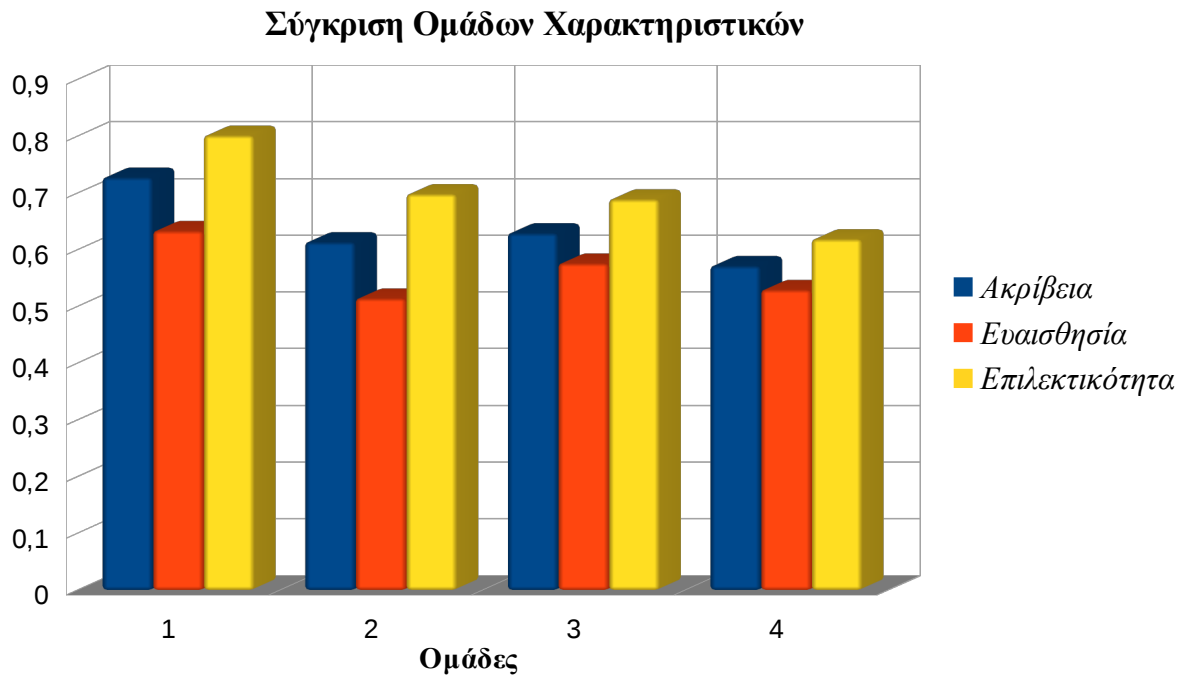
Παράμετρος C (2 [^])	Παράμετρος g (2 [^])	Ακρίβεια	Ευαισθησία	Ειδικότητα
-2	-4	0.52883	1	0
4	-2	0.52883	1	0
11	0	0.56094	0.86886	0.21535
11	7	0.60227	0.70949	0.48186
10	0	0.56298	0.86909	0.21931
3	0	0.56176	0.87290	0.21253
3	1	0.60017	0.71341	0.47305
4	0	0.55892	0.86790	0.21224
5	4	0.56136	0.87628	0.20781
6	-1	0.52895	1	0
7	0	0.55800	0.86438	0.21389
8	1	0.59999	0.70970	0.47683
9	0	0.56082	0.87268	0.21100
10	5	0.56298	0.86909	0.21931
10	2	0.59603	0.63257	0.55505
2	4	0.65794	0.50191	0.83311
2	5	0.62632	0.43131	0.84539
3	4	0.666837	0.53867	0.810645
3	5	0.633568	0.44681	0.843168

Έχοντας τώρα εξασφαλίσει τις κατάλληλες παραμέτρους για τον SVM προχωράμε στη σύγκριση των τριών ομάδων χαρακτηριστικών. Οι τρεις αυτές ομάδες είναι:

- Τα 10 μορφολογικά χαρακτηριστικά.
- Τα 75 χαρακτηριστικά Ενέργειας Υφής Laws

- Τα 10 χαρακτηριστικά Grey Level Run-length

Τα αποτελέσματα αυτής της σύγκρισης φαίνονται στα παρακάτω διαγράμματα. Στην Εικόνα 15 παρουσιάζονται οι επιδόσεις του SVM στην ακρίβεια, την ευαισθησία και την Ειδικότητα για το σύνολο των 1214 μαστογραφικών μαζών.



Εικόνα 16: Σύγκριση Επιδόσεων του SVM για χρήση όλων των χαρακτηριστικών, των μορφολογικών, των Laws και Run-Length (Ομάδες 1,2,3 και 4 αντίστοιχα)

Ένα επιπλέον σκέλος της διερεύνησής μας είναι η αποτελεσματικότητα των μεθοδολογιών για συγκεκριμένες κατηγορίες μαστογραφικών απεικονίσεων. Στη συγκεκριμένη βάση υπάρχουν 4 κατηγορίες διαφορετικής πυκνότητας μαζών και 5 κατηγορίες διακριτικότητας. Υπενθυμίζουμε ότι η κλίμακα έχει εύρος από το 1 για τις πιο δύσκολες περιπτώσεις ως το 5 για τις πιο προφανείς. Από τη σύγκριση των παραπάνω κατηγοριών για την υλοποίηση για SVM προκύπτουν τα αποτελέσματα των παρακάτω πινάκων. Ο Πίνακας 5 αποτυπώνει τις μεταβολές στην ακρίβεια για τις διάφορες τιμές της μεταβλητής πυκνότητα και διακριτικότητα, ο Πίνακας 6 την ευαισθησία και ο Πίνακας 7 την διακριτικότητα. Αρχικά παρουσιάζονται τα αποτελέσματα για την υλοποίηση με όλα τα διαθέσιμα χαρακτηριστικά.

Πίνακας 5: Ακρίβεια SVM με χρήση όλων των διαθέσιμων χαρακτηριστικών

		Διακριτικότητα (Subtlety)				
		1	2	3	4	5
Πυκνότητα (Density)	1	1	0.53302	0.52929	0.39832	0.75802
	2	0.68960	0.69548	0.53220	0.61604	0.64753
	3	0.75615	0.52311	0.62767	0.57366	0.54297
	4	0.48036	0.64638	0.62047	0.47678	0.63838

Παρατηρούμε μια αύξηση της ακρίβεια του SVM καθώς η διακριτικότητα αυξάνεται, γεγονός που συνάδει με τον ορισμό της που υπονοεί χαμηλότερη ευκρίνεια της μαστογραφικής απεικόνισης.

Πίνακας 6: Ευαισθησία SVM με χρήση όλων των διαθέσιμων χαρακτηριστικών

		Διακριτικότητα (Subtlety)				
		1	2	3	4	5
Πυκνότητα (Density)	1	1	0.03435	0.45963	0.37915	0.78289
	2	0.87233	0.92841	0.46200	0.60645	0.72648
	3	0.97993	0.61672	0.47853	0.52196	0.71216
	4	0.75683	0.72055	0.04200	0.46874	0.50833

Πίνακας 7: Ειδικότητα SVM με χρήση όλων των διαθέσιμων χαρακτηριστικών

		Διακριτικότητα (Subtlety)				
		1	2	3	4	5
Πυκνότητα (Density)	1	0	0.84537	0.60270	0.41909	0.56185
	2	0.27776	0.10419	0.58686	0.62518	0.52329
	3	0	0.39611	0.73759	0.61955	0.29819
	4	0.10247	0.56869	0.86289	0.48976	0.73450

Στη συνέχεια παρουσιάζουμε τα αποτελέσματα όταν χρησιμοποιούνται μόνο τα Μορφολογικά χαρακτηριστικά:

Πίνακας 8: Ακρίβεια SVM με χρήση μόνο Μορφολογικών χαρακτηριστικών

		Διακριτικότητα (Subtlety)				
		1	2	3	4	5
Πυκνότητα (Density)	1	0	0.44961	0.49930	0.58845	0.69591
	2	0.44210	0.70252	0.47458	0.55586	0.63743
	3	0.65815	0.52015	0.69951	0.56396	0.50005
	4	0.39987	0.55232	0.69149	0.41281	0.54482

Παρατηρούμε ότι όντως η απόδοση της ταξινόμησης καθώς μεγαλώνει η τιμή της διακριτικότητας, αυξάνεται.

Πίνακας 9: Ευαισθησία SVM με χρήση μόνο Μορφολογικών χαρακτηριστικών

		Διακριτικότητα (Subtlety)				
Πυκνότητα (Density)		1	2	3	4	5
	1	0	0.16269	0,43633	0.37232	0.69653
	2	0,7	0.87571	0.45552	0,50803	0.68322
	3	0.79262	0,60400	0.55172	0.50803	0,54584
	4	0.57750	0,53852	0.376	0,47948	0.43533

Παρατηρούμε αρκετά χαμηλές τιμές κυρίως στην ομάδα με πυκνότητα 1 και διακριτικότητα 1, που όμως δικαιολογείται από τον πολύ χαμηλό αριθμό περιπτώσεων μαζών που υπάρχουν σε αυτή την κατηγορία.

Πίνακας 10: Ειδικότητα SVM με χρήση μόνο Μορφολογικών χαρακτηριστικών

		Διακριτικότητα (Subtlety)				
Πυκνότητα (Density)		1	2	3	4	5
	1	1	0.55790	0.46889	0.59259	0.72657
	2	0.03857	0.38551	0.51440	0.56204	0.68644
	3	0.09281	0.51485	0.64901	0.61977	0.54438
	4	0.16738	0.49315	0.80704	0.49995	0.90883

Στη συνέχεια παρουσιάζονται οι ίδιες ενδείξεις της ακρίβειας, ευαισθησίας και Ειδικότητας του SVM όταν σαν είσοδο στην ταξινομητή επιλέξουμε μόνο τα χαρακτηριστικά υφής. Αρχικά, στους Πίνακες 11, 12 και 13 φαίνονται οι επιδόσεις για τα χαρακτηριστικά Ενέργειας Υφής Laws και στους Πίνακες 14, 15 και 16 για τα χαρακτηριστικά με βάση τους πίνακες Run-length.

Πίνακας 11: Ακρίβεια SVM με χρήση μόνο Laws χαρακτηριστικών

		Διακριτικότητα (Subtlety)				
Πυκνότητα (Density)		1	2	3	4	5
	1	0	0.45745	0.37633	0.46652	0.68564
	2	0.76214	0.63468	0.54152	0.65089	0.65086
	3	0.69463	0.47789	0.55478	0.54439	0.55127
	4	0.56045	0.68320	0.57455	0.52881	0.65254

Πίνακας 12: Ευαισθησία SVM με χρήση μόνο Laws χαρακτηριστικών

		Διακριτικότητα (Subtlety)				
		1	2	3	4	5
Πυκνότητα (Density)	1	0	0.08354	0.19350	0.40672	0.73439
	2	0,94511	0.84035	0.42959	0.58328	0.72967
	3	0.81786	0.56866	0.40791	0.46381	0.65925
	4	0.76083	0.74154	0.12933	0.52251	0.55928

Πίνακας 13: Ειδικότητα SVM με χρήση μόνο Laws χαρακτηριστικών

		Διακριτικότητα (Subtlety)				
		1	2	3	4	5
Πυκνότητα (Density)	1	0	0.68405	0.53672	0.51351	0.60708
	2	0.37676	0.14241	0.62820	0.71594	0.52709
	3	0	0.35974	0.66212	0.61531	0.39414
	4	0,28595	0.63276	0.76212	0.53788	0.73066

Πίνακας 14: Ακρίβεια SVM με χρήση μόνο Run-Length χαρακτηριστικών

		Διακριτικότητα (Subtlety)				
		1	2	3	4	5
Πυκνότητα (Density)	1	1	0.49346	0.69909	0.58243	0.70710
	2	0.49402	0.46314	0.54591	0.61671	0.56570
	3	0.76062	0.51046	0.57275	0.52238	0.51237
	4	0.70138	0.58308	0.58199	0.43536	0.36564

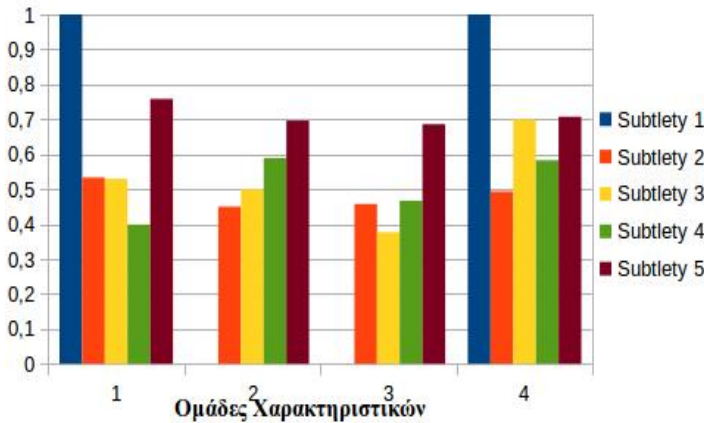
Πίνακας 15: Ευαισθησία SVM με χρήση μόνο Run-Length χαρακτηριστικών

		Διακριτικότητα (Subtlety)				
		1	2	3	4	5
Πυκνότητα (Density)	1	1	0.36030	0.54779	0.53439	0.71016
	2	0.47055	0.45436	0.53517	0.51699	0.48258
	3	0.75265	0.42227	0.59182	0.58233	0.47550
	4	0.61666	0.44121	0.38811	0.40801	0.52807

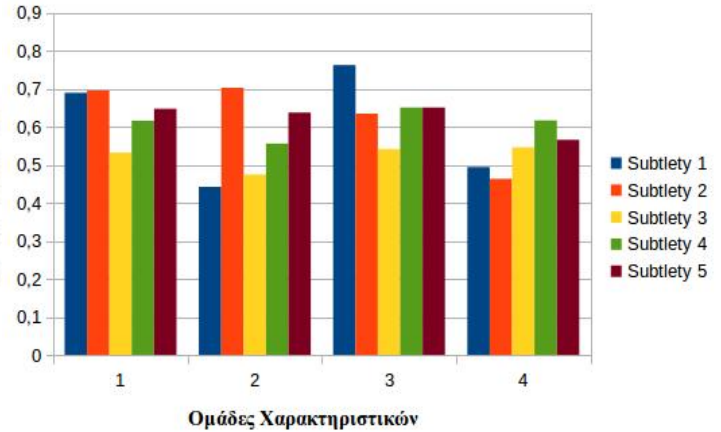
Πίνακας 16: Ειδικότητα SVM με χρήση μόνο Run-Length χαρακτηριστικών

		Διακριτικότητα (Subtlety)				
		1	2	3	4	5
Πυκνότητα (Density)	1	0	0.59490	0.82975	0.62066	0.86105
	2	0.86105	0.49062	0.55436	0.71216	0.69653
	3	0	0.62599	0.55925	0.46968	0.56542
	4	0.83923	0.75048	0.66196	0.45760	0.23216

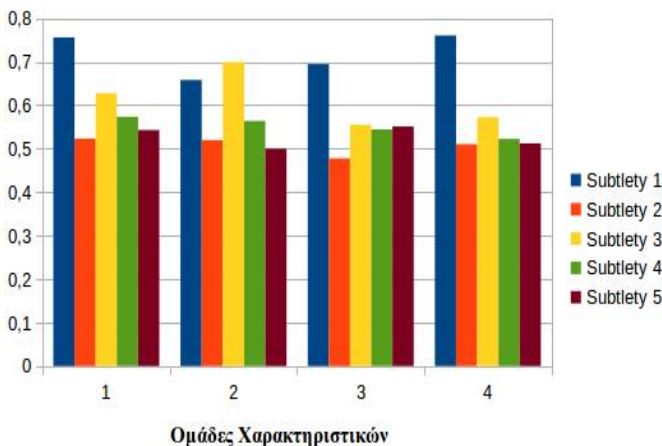
Συνοπτικά τα αποτελέσματα που καταγράφηκαν για τις τιμές της ακρίβειας φαίνονται στις παρακάτω εικόνες:



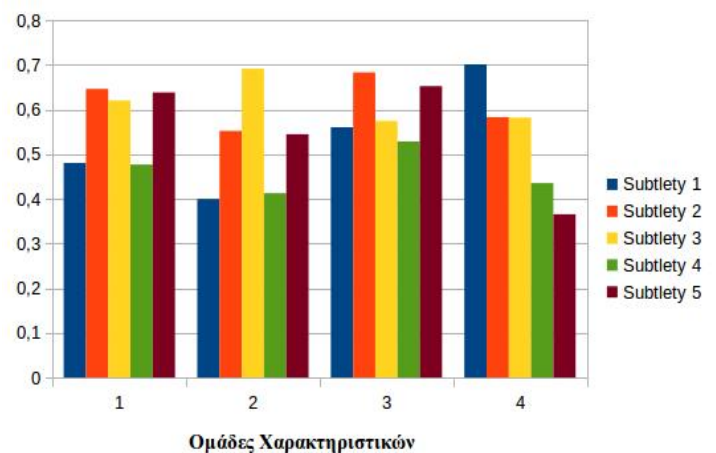
Εικόνα 17: Σύγκριση Ακρίβειας του SVM για πυκνότητα 1 και χρήση όλων των χαρακτηριστικών, των μορφολογικών, των Laws και Run-Length (Ομάδες 1,2,3 και 4 αντίστοιχα)



Εικόνα 18: Σύγκριση Ακρίβειας του SVM για πυκνότητα 2 και χρήση όλων των χαρακτηριστικών, των μορφολογικών, των Laws και Run-Length (Ομάδες 1,2,3 και 4 αντίστοιχα)



Εικόνα 19: Σύγκριση Ακρίβειας του SVM για πυκνότητα 3 και χρήση όλων των χαρακτηριστικών, των μορφολογικών, των Laws και Run-Length (Ομάδες 1,2,3 και 4 αντίστοιχα)



Εικόνα 20: Σύγκριση Ακρίβειας του SVM για πυκνότητα 4 και χρήση όλων των χαρακτηριστικών, των μορφολογικών, των Laws και Run-Length (Ομάδες 1,2,3 και 4 αντίστοιχα)

4.3.2 Υλοποίηση σε Νευρωνικά Δίκτυα

Στην ενότητα αυτή παρουσιάζονται τα αποτελέσματα της κατηγοριοποίησης των μαστογραφικών μαζών όταν ως ταξινομητής χρησιμοποιείται Νευρωνικό Δίκτυο. Πριν τη διαδικασία της ταξινόμησης απαραίτητη είναι η αρχικοποίηση του Δικτύου, δηλαδή αρχικοποίηση των παραμέτρων του. Ένα από τα προβλήματα που αντιμετωπίζουμε κατά την αρχικοποίηση ενός

Νευρωνικού Δικτύου είναι ο ορισμός του πλήθους των κρυμμένων επιπέδων του καθώς και το πλήθος των νευρώνων που αυτά περιέχουν. Αν οι νευρώνες είναι πολύ λίγοι σε σχέση με την πολυπλοκότητα του προβλήματος είναι πιθανόν να έχουμε υποπροσαρμογή (underfitting) του νευρωνικού. Θέτοντάς το διαφορετικά το σύστημα αδυνατεί να ανιχνεύσει τις εξαρτήσεις σε ένα περίπλοκο σύνολο δεδομένων. Αυτό το γεγονός οδηγεί προφανώς σε πολύ χαμηλά ποσοστά επίδοσης. Αν οι νευρώνες από την άλλη πλευρά είναι υπερβολικά πολλοί, θα έχουμε υπερπροσαρμογή. Αυτό με τη σειρά του σημαίνει ότι ενώ το λάθος στο σύνολο εκπαίδευσης θα είναι πολύ μικρό, ενώ το λάθος γενίκευσης σε νέα δεδομένα μεγάλο. Επιπρόσθετα, επιλέγοντας πολλά κρυφά επίπεδα αυξάνεται η πολυπλοκότητα και άρα ο χρόνος εκπαίδευσης του Νευρωνικού καθώς επίσης αυξάνονται οι πιθανότητες να πέσει σε λανθασμένο τοπικό ελάχιστο [9]. Είναι προφανές, λοιπόν, το πόσο κρίσιμο είναι το σημείο της σωστής παραμετροποίησης. Ακολουθώντας Μετά από έλεγχο των παραμέτρων του νευρωνικού καταλήξαμε στο να χρησιμοποιήσουμε 2 κρυφά επίπεδα νευρώνων με 30 νευρώνες το κάθε επίπεδο. Τα αποτελέσματα της ταξινόμησης δείχνουν αρκετά μικρότερη επιτυχία στην ακρίβεια του ταξινομητή σε σύγκριση με αυτή του SVM, καθώς στο νευρωνικό μας για όλες τις μαστογραφικές μάζες είχε ποσοστό ακρίβειας 63,01%.

Στα υποσύνολα της βάσης με τις πιο ευδιάκριτες μάζες το ποσοστό ακρίβειας έφτασε το 75,61% ενώ το ποσοστό ευαισθησίας το 78.3%. Στους Πίνακες 17, 18 και 19 παρουσιάζονται οι τιμές της ακρίβειας, της ευαισθησίας και ειδικότητας για τις διάφορες κατηγορίες πυκνότητας και διακριτικότητας των μαζών.

Πίνακας 17: Ακρίβεια Νευρωνικού Δικτύου με χρήση όλων των χαρακτηριστικών

		Διακριτικότητα (Subtlety)				
Πυκνότητα (Density)		1	2	3	4	5
	1	1	0.53302	0.52929	0.69832	0.69802
	2	0.68960	0.69548	0.53220	0.61604	0.64753
	3	0.75615	0.52311	0.62767	0.57366	0.54297
	4	0.48036	0.64638	0.62047	0.47678	0.63838

Πίνακας 18: Ευαισθησία Νευρωνικού Δικτύου με χρήση όλων των Μορφολογικών χαρακτηριστικών

		Διακριτικότητα (Subtlety)				
		1	2	3	4	5
Πυκνότητα (Density)	1	1	0.03435	0.45963	0.37915	0.78289
	2	0.87233	0.92841	0.46200	0.60645	0.72648
	3	0.97993	0.61672	0.47853	0.52196	0.71216
	4	0.75683	0.72055	0.04200	0.46874	0.50833

Πίνακας 19: Ειδικότητα Νευρωνικού Δικτύου με χρήση όλων των χαρακτηριστικών

		Διακριτικότητα (Subtlety)				
		1	2	3	4	5
Πυκνότητα (Density)	1	0	0.84537	0.60270	0.41909	0.56185
	2	0.27776	0.10419	0.58686	0.62518	0.52329
	3	0	0.39611	0.73759	0.61955	0.29819
	4	0.10247	0.56869	0.86289	0.48976	0.73450

Στους Πίνακες 20,21 και 22 φαίνονται οι επιδόσεις του Νευρωνικού όταν για την εκπαίδευσή του χρησιμοποιούνται μόνον μορφολογικά χαρακτηριστικά.

Πίνακας 20: Ακρίβεια Νευρωνικού Δικτύου με χρήση μόνο Μορφολογικών χαρακτηριστικών

		Διακριτικότητα (Subtlety)				
		1	2	3	4	5
Πυκνότητα (Density)	1	0	0.40992	0.45284	0.49673	0.70772
	2	0.49009	0.73354	0.48838	0.53557	0.68450
	3	0.62771	0.56531	0.60792	0.56746	0.54538
	4	0.41265	0.51531	0.67996	0.49133	0.69639

Πίνακας 21: Ευαισθησία Νευρωνικού Δικτύου με χρήση μόνο Μορφολογικών χαρακτηριστικών

		Διακριτικότητα (Subtlety)				
		1	2	3	4	5
Πυκνότητα (Density)	1	0	0.16269	0.43633	0.37232	0.69653
	2	0.70000	0.87571	0.45552	0.50803	0.68322
	3	0.79262	0.60400	0.55172	0.50803	0.54584
	4	0.57750	0.53852	0.37600	0.47948	0.43533

Πίνακας 22: Ειδικότητα Νευρωνικού Δικτύου με χρήση Μόνο Μορφολογικών χαρακτηριστικών

		Διακριτικότητα (Subtlety)				
		1	2	3	4	5
Πυκνότητα (Density)	1	0	0.55790	0.468891	0.59259	0.72657
	2	0.03857	0.38551	0.51440	0.56204	0.68644
	3	0	0.51485	0.64901	0.61977	0.54438
	4	0.16738	0.49315	0.80704	0.49995	0.90883

Στους Πίνακες 23,24 και 25 φαίνονται οι επιδόσεις του Νευρωνικού όταν για την εκπαίδευσή του χρησιμοποιούνται μόνον τα χαρακτηριστικά ενέργειας υφής Laws.

Πίνακας 23: Ακρίβεια Νευρωνικού Δικτύου με χρήση μόνο χαρακτηριστικών Laws

		Διακριτικότητα (Subtlety)				
		1	2	3	4	5
Πυκνότητα (Density)	1	0	0.45745	0.37633	0.46652	0.68564
	2	0.76214	0.63468	0.54152	0.65089	0.65086
	3	0.69463	0.47789	0.55478	0.54439	0.55127
	4	0.56045	0.68320	0.57455	0.52881	0.65254

Πίνακας 24: Ευαισθησία Νευρωνικού Δικτύου με χρήση μόνο χαρακτηριστικών Laws

		Διακριτικότητα (Subtlety)				
		1	2	3	4	5
Πυκνότητα (Density)	1	1	0.02750	0.22280	0.27461	0.82044
	2	0.91811	0.92757	0.33869	0.61435	0.78958
	3	1	0.71644	0.31706	0.39833	0.82250
	4	0.77711	0.78277	0.10822	0.36561	0.2993

Πίνακας 25: Ειδικότητα Νευρωνικού Δικτύου με χρήση μόνο χαρακτηριστικών Laws

		Διακριτικότητα (Subtlety)				
		1	2	3	4	5
Πυκνότητα (Density)	1	0	0.68405	0.53672	0.51351	0.60708
	2	0.37676	0.14241	0.62820	0.71594	0.52709
	3	0	0.35974	0.66212	0.61531	0.39414
	4	0.28595	0.63276	0.76212	0.53788	0.73066

Στους Πίνακες 23,24 και 25 φαίνονται οι επιδόσεις του Νευρωνικού όταν για την εκπαίδευσή του χρησιμοποιούνται μόνον τα χαρακτηριστικά Run-length επιπέδου γκρι.

Πίνακας 26: Ακρίβεια Νευρωνικού Δικτύου με χρήση μόνο Run-length χαρακτηριστικών

		Διακριτικότητα (Subtlety)				
		1	2	3	4	5
Πυκνότητα (Density)	1	1	0.49346	0.69909	0.58243	0.70710
	2	0.49402	0.46314	0.54591	0.61671	0.56570
	3	0.76062	0.51046	0.57275	0.52238	0.51237
	4	0.70138	0.58308	0.58199	0.43536	0.36564

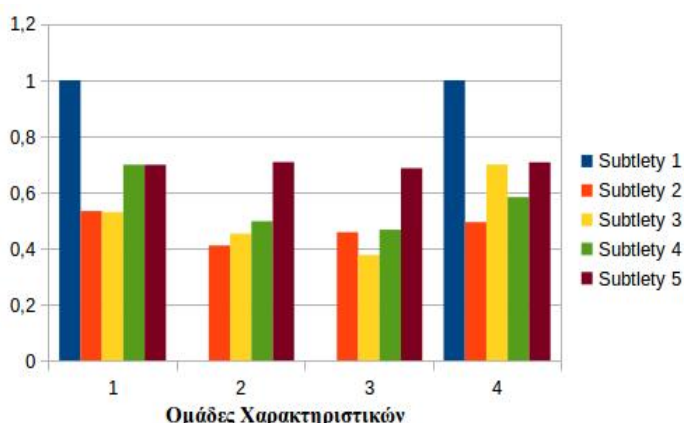
Πίνακας 27: Ευαισθησία Νευρωνικού Δικτύου με χρήση μόνο Run-length χαρακτηριστικών

		Διακριτικότητα (Subtlety)				
		1	2	3	4	5
Πυκνότητα (Density)	1	1	0.36030	0.54779	0.53439	0.61012
	2	0.47055	0.45436	0.53517	0.51699	0.48258
	3	0.75265	0.42227	0.59182	0.58233	0.47550
	4	0.61666	0.44121	0.38811	0.40801	0.52807

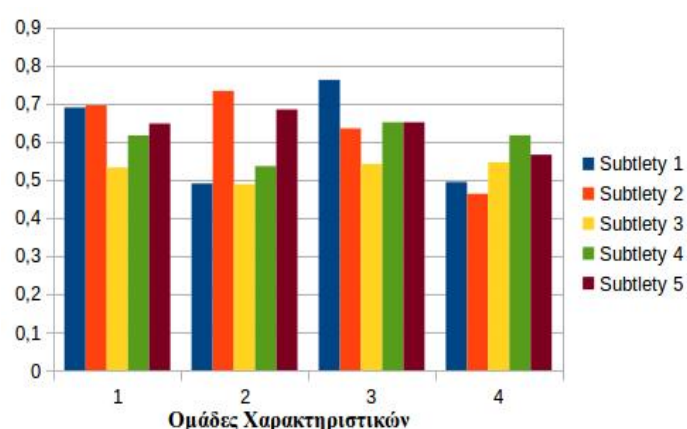
Πίνακας 28: Ειδικότητα Νευρωνικού Δικτύου με χρήση μόνο Run-length χαρακτηριστικών

		Διακριτικότητα (Subtlety)				
		1	2	3	4	5
Πυκνότητα (Density)	1	1	0.59490	0.82975	0.62066	0.86105
	2	0.53904	0.49062	0.55436	0.71216	0.69653
	3	1	0.62599	0.55925	0.46968	0.56542
	4	0.83923	0.75048	0.66196	0.45760	0.23216

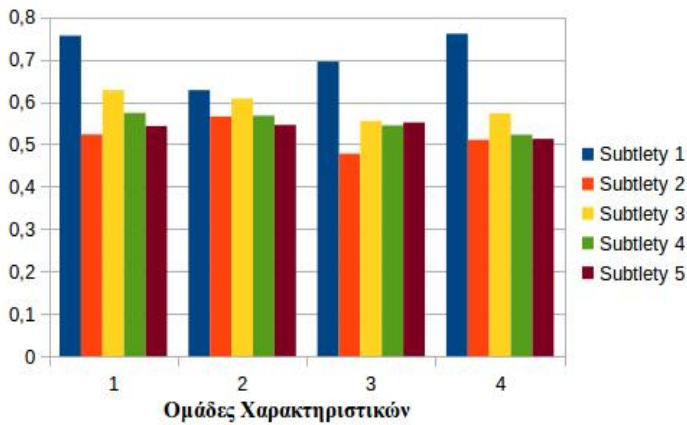
Συνοπτικά τα αποτελέσματα που καταγράφηκαν για τις τιμές της ακρίβειας φαίνονται στις παρακάτω εικόνες:



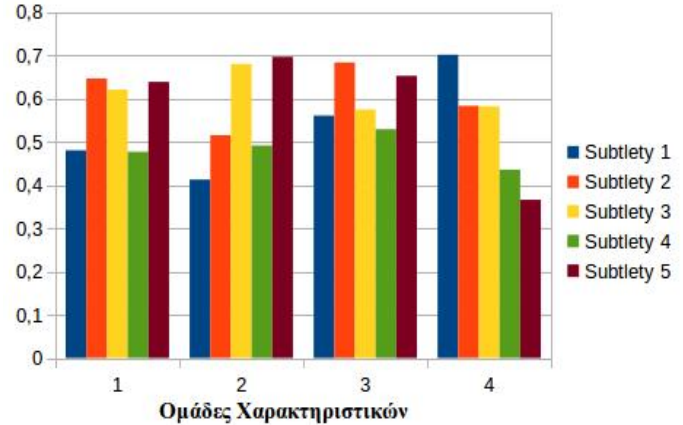
Εικόνα 21: Σύγκριση Ακρίβειας του Νευρωνικού Δικτύου για πυκνότητα 1 και χρήση όλων των χαρακτηριστικών, των μορφολογικών, των Laws και Run-Length (Ομάδες 1,2,3 και 4 αντίστοιχα)



Εικόνα 22: Σύγκριση Ακρίβειας του Νευρωνικού Δικτύου για πυκνότητα 2 και χρήση όλων των χαρακτηριστικών, των μορφολογικών, των Laws και Run-Length (Ομάδες 1,2,3 και 4 αντίστοιχα)



Εικόνα 24: Σύγκριση Ακρίβειας του Νευρωνικού Δικτύου για πυκνότητα 3 και χρήση όλων των χαρακτηριστικών, των μορφολογικών, των Laws και Run-Length (Ομάδες 1,2,3 και 4 αντίστοιχα)



Εικόνα 23: Σύγκριση Ακρίβειας του Νευρωνικού Δικτύου για πυκνότητα 4 και χρήση όλων των χαρακτηριστικών, των μορφολογικών, των Laws και Run-Length (Ομάδες 1,2,3 και 4 αντίστοιχα)

4.3.3 Υλοποίηση σε Ταξινομητή K-κοντινότερων Γειτόνων

Ο ταξινομητής k-κοντινότερων σημείων δίνει στην υλοποίησή μας τα χαμηλότερα ποσοστά ακρίβειας, για όλο το σύνολο των 1214 μαστογραφικών μαζών, 62,29%. Όλα τα αποτελέσματα για εκπαίδευση του ταξινομητή με χρήση και των 95 διαθέσιμων χαρακτηριστικών παρουσιάζονται στους Πίνακες 29, 30 και 31.

Πίνακας 29: Ακρίβεια kNN με χρήση όλων των διαθέσιμων χαρακτηριστικών

		Διακριτικότητα (Subtlety)				
Πυκνότητα (Density)		1	2	3	4	5
	1	1	0.54966	0.61389	0.59793	0.62084
	2	0.95860	0.68989	0.73645	0.59848	0.59179
	3	0.51197	0.42522	0.74219	0.60681	0.60676
	4	1	0.69166	0.66565	0.38562	0.68105

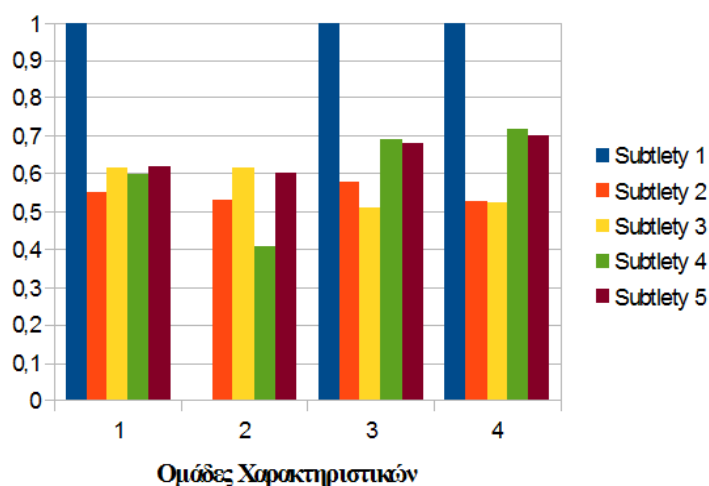
Πίνακας 30: Ευαισθησία kNN με χρήση όλων των διαθέσιμων χαρακτηριστικών

		Διακριτικότητα (Subtlety)				
		1	2	3	4	5
Πυκνότητα (Density)	1	1	0.06700	0.50447	0,49222	0.65518
	2	0.99288	0.73524	0.48474	0,46328	0.67109
	3	0.51212	0.42928	0.74257	0.51998	0.67822
	4	0	0.60016	0.23573	0.03202	0.81169

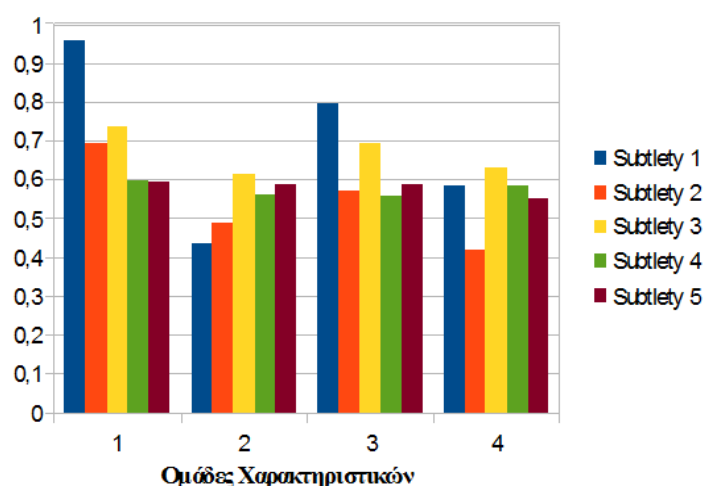
Πίνακας 31: Ειδικότητα kNN με χρήση όλων των διαθέσιμων χαρακτηριστικών

		Διακριτικότητα (Subtlety)				
		1	2	3	4	5
Πυκνότητα (Density)	1	0	0.75690	0.68266	0.65983	0.56228
	2	0.87666	0.60066	0.85742	0.71331	0.48174
	3	0	0.42040	0.74205	0.66780	0.52987
	4	0	0.78316	0.81293	0.59861	0.53200

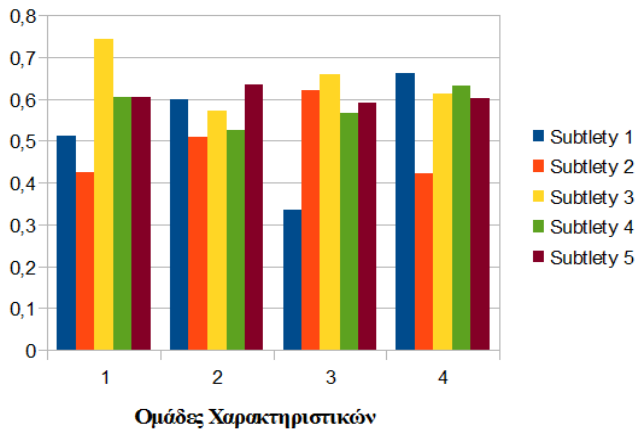
Συνοπτικά τα αποτελέσματα που καταγράφηκαν για τις τιμές της ακρίβειας φαίνονται στις παρακάτω εικόνες:



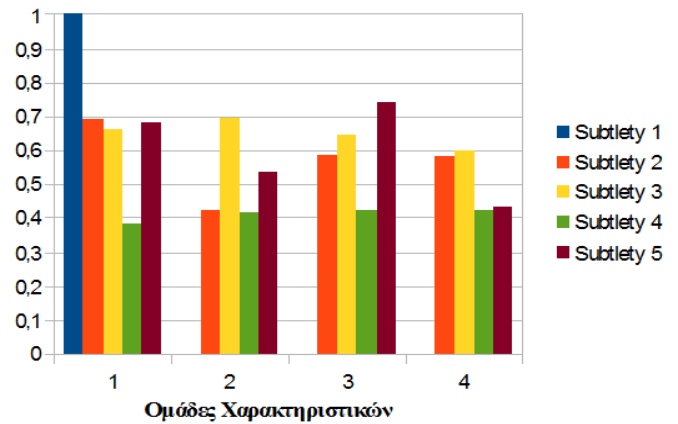
Εικόνα 25: Σύγκριση Ακρίβειας του kNN για πυκνότητα 1 και χρήση όλων των χαρακτηριστικών, των μορφολογικών, των Laws και Run-Length (Ομάδες 1,2,3 και 4 αντίστοιχα)



Εικόνα 26: Σύγκριση Ακρίβειας του kNN για πυκνότητα 2 και χρήση όλων των χαρακτηριστικών, των μορφολογικών, των Laws και Run-Length (Ομάδες 1,2,3 και 4 αντίστοιχα)

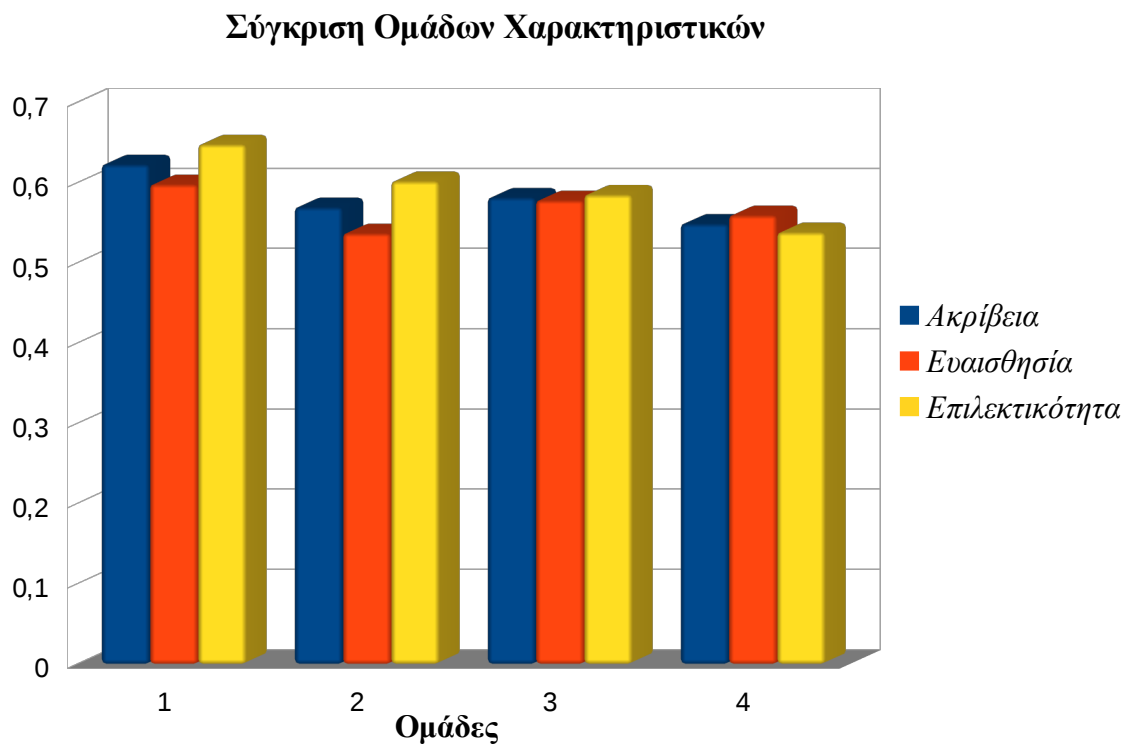


Εικόνα 27: Σύγκριση Ακρίβειας του kNN για πυκνότητα 3 και χρήση όλων των χαρακτηριστικών, των μορφολογικών, των Laws και Run-Length (Ομάδες 1,2,3 και 4 αντίστοιχα)



Εικόνα 28: Σύγκριση Ακρίβειας του kNN για πυκνότητα 4 και χρήση όλων των χαρακτηριστικών, των μορφολογικών, των Laws και Run-Length (Ομάδες 1,2,3 και 4 αντίστοιχα)

Ακολουθεί ένα γράφημα με τις ενδείξεις ακρίβειας, ευαισθησίας και ειδικότητας του αλγορίθμου για τις τρεις ομάδες χαρακτηριστικών. (Εικόνα 29)



Εικόνα 29: Σύγκριση Επιδόσεων του kNN για χρήση όλων των χαρακτηριστικών, των μορφολογικών, των Laws και Run-Length (Ομάδες 1,2,3 και 4 αντίστοιχα)

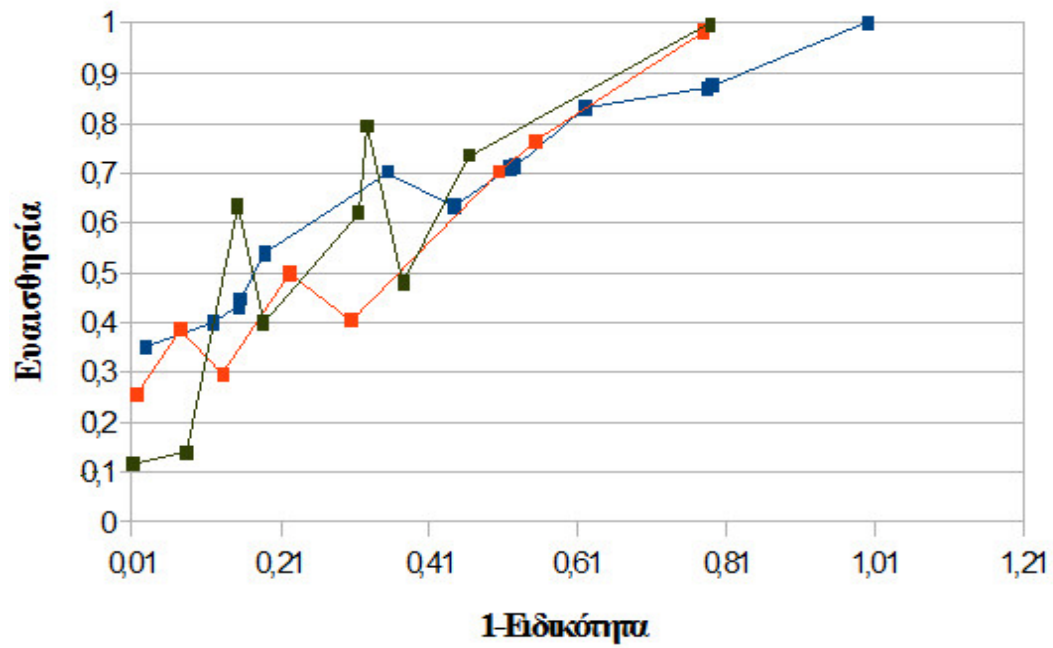
4.4 Συμπεράσματα

Η υλοποίηση και ο έλεγχος του συστήματος ταξινόμησης είχε αρκετά ενδιαφέροντα αποτελέσματα. Καταρχήν, κατά την σύγκριση των ταξινομητών SVM, Νευρωνικό Δίκτυο και kNN σε όλη τη βάση μαστογραφιών, φαίνεται ξεκάθαρα ότι το SVM έχει καλύτερα αποτελέσματα από τους υπόλοιπους δύο ταξινομητές. Συγκεκριμένα, το SVM είχε τιμή ακριβείας 72.82%, με ευαισθησία 63.4% και ειδικότητα 80.21%, ενώ σε κάποια υποσύνολα της βάσης μας έφτασε το 87.57%. Όσον αφορά τη σύγκριση μεταξύ των 3 ομάδων χαρακτηριστικών που διαχωρίσαμε, παρατηρούμε ότι τις καλύτερες επιδόσεις έχει το SVM που εκπαιδεύεται με χαρακτηριστικά ενέργειας υφής Laws, έπειτα αυτό που εκπαιδεύεται με μορφολογικά χαρακτηριστικά και τέλος με χαρακτηριστικά γκρι επιπέδου Run-length. Λαμβάνοντας υπόψιν ότι τα μορφολογικά χαρακτηριστικά ήταν ισόποσα με τα Run-length, κι άρα συγκρίσιμα, βλέπουμε ένα προβάδισμα στη διακριτική ικανότητα των μορφολογικών χαρακτηριστικών.

Οι επιδόσεις του Νευρωνικού Δικτύου ήταν λίγο χαμηλότερες. Η ακρίβειά του για έλεγχο σε όλη τη βάση ήταν 63.01% ενώ σε κάποια υποσύνολα έφτασε το 75.61% με ευαισθησία 78.3% και ειδικότητα 86.29%. Για εκπαίδευση μόνο με μορφολογικά χαρακτηριστικά έχουμε σε κάποια υποσύνολα μέγιστη ακρίβεια, ευαισθησία και ειδικότητα 0.73, 0.87 και 0.9 αντίστοιχα. Μόνο με χαρακτηριστικά Laws το Νευρωνικό Δίκτυο σε κάποια υποσύνολα άγγιζε το 0.76, 0.93, 0.76 για ακρίβεια, ευαισθησία και ειδικότητα αντίστοιχα. Τέλος, για εκπαίδευση μόνο με χαρακτηριστικά Run-length έχουμε σε κάποια υποσύνολα μέγιστη ακρίβεια, ευαισθησία και ειδικότητα 0.76, 0.75 και 0.83 αντίστοιχα.

Οι επιδόσεις του ταξινομητή kNN ήταν σαφώς χαμηλότερες με τιμή ακριβείας για όλο το σύνολο το 62.29%. Και σε αυτόν τον ταξινομητή παρατηρούμε βελτιωμένες αποδόσεις όταν χρησιμοποιήθηκαν χαρακτηριστικά Laws. Στη συνέχεια ακολουθούν τα μορφολογικά και τέλος τα χαρακτηριστικά Run-length.

Το γεγονός της μεγαλύτερης απόδοσης των κατά πολύ περισσότερων χαρακτηριστικών Laws έρχεται να στηρίξει η παρατήρησή μας πως κατά τη χρησιμοποίηση των μορφολογικών και Run-length χαρακτηριστικών, η απόδοση βελτιωνόταν χωρίς τη χρήση της συνάρτησης ttest, η οποία μείωνε κι άλλο το πλήθος τους.



Εικόνα 30: ROC καμπύλη για τους τρεις ταξινομητές (με μπλε χρώμα το SVM, πράσινο το NN και κόκκινο ο kNN)

Κεφάλαιο 5

Συμπεράσματα και Μελλοντικές Προοπτικές

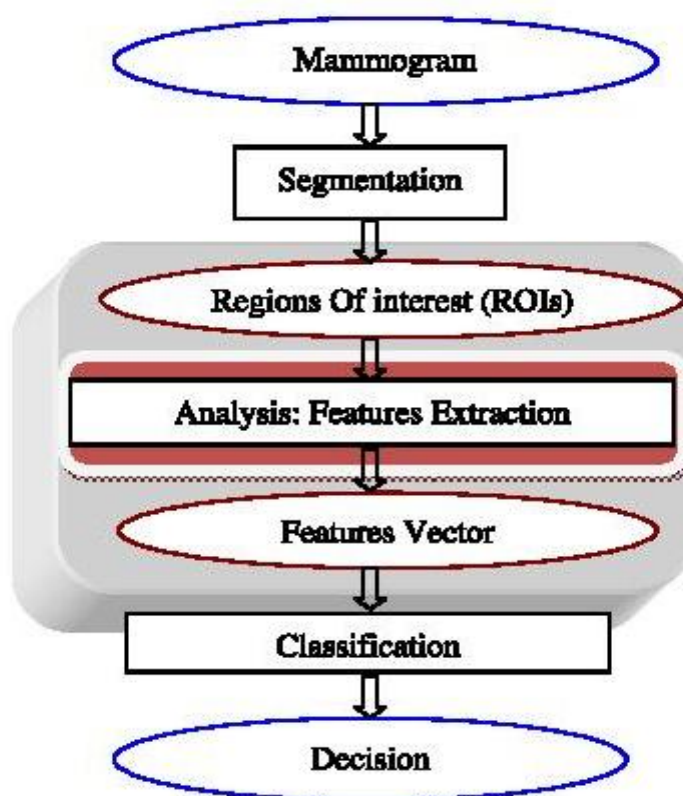
5.1 Γενικά Συμπεράσματα

Τα συστήματα υποβοηθούμενης διάγνωσης αποτελούν και θα συνεχίσουν να αποτελούν ένα αντικείμενο συνεχούς προόδου και καινοτομίας στις ιατρικές πρακτικές. Στον τομέα της υποβοηθούμενης ταξινόμησης μαστογραφικών μαζών, όπως έχουμε ήδη αναφέρει, υπάρχει έλλειψη επαρκών μελετών για τη σύγκριση των μεθοδολογιών ταξινόμησης με χρήση Μηχανικής Μάθησης σε ένα ενιαίο, κοινό σύνολο. Το κενό αυτό προσπαθεί να καλύψει σε ένα βαθμό η διπλωματική αυτή εργασία.

Ο στόχος της είναι διττός. Αρχικά να συγκρίνει την επίδραση στην απόδοση των διαφορετικών χαρακτηριστικών που έχουν προταθεί και, δεύτερον, να αναδείξει τον πιο αποτελεσματικό ταξινομητή για την κατηγοριοποίηση των μαζών ως καρκινικές ή καλοήθειες. Όσον αφορά την πρώτη διερεύνηση, τα αποτελέσματα της υλοποίησής μας έδειξαν ότι τη μεγαλύτερη διακριτική ικανότητα έχουν τα χαρακτηριστικά ενέργειας υψής Laws, έπειτα τα μορφολογικά χαρακτηριστικά και τέλος τα χαρακτηριστικά γκρι επιπέδου Run-length. Η δεύτερη απάντηση με βάση τα αποτελέσματα της εργασίας είναι η πιο ξεκάθαρη. Σε όλα τα υποσύνολα ελέγχου το SVM φαίνεται να ξεπερνάει σε απόδοση το Νευρωνικό Δίκτυο, το οποίο με τη σειρά του προηγείται του ταξινομητή k-κοντινότερων γειτόνων. Τα αποτελέσματα της υλοποίησης είναι ενθαρρυντικά αλλά δεν καθορίζουν με απόλυτη βεβαιότητα την πιο αποτελεσματική μέθοδο, κυρίως λόγω του ανόμοιου πλήθους χαρακτηριστικών σε όλες τις επιλεγθείσες κατηγορίες. Περαιτέρω τεχνικές που μπορούν να χρησιμοποιηθούν με σκοπό να βελτιωθεί η απόδοση του συστήματος ταξινόμησης αλλά και να ολοκληρωθεί το σύστημα CAD περιγράφονται στις παρακάτω ενότητες.

5.2 Κατάτμηση με Χρήση του Αλγορίθμου K-means

Στην προσπάθειά μας προς την επίτευξη υψηλότερης απόδοσης στην ταξινόμηση των μαστογραφικών μαζών, αλλά και για να ολοκληρώσουμε την υλοποίηση του συστήματος CADx (Εικόνα 18) επιχειρήσαμε να ασχοληθούμε με το υποσύστημα της κατάτμησης των μαζών σε μια μαστογραφία. Το υποσύστημα αυτό έπεται του εντοπισμού της Περιοχής Ενδιαφέροντος στη μαστογραφία και στοχεύει στον ακριβή καθορισμό των ορίων της μάζας.



Εικόνα 31: Γραφική αναπαράσταση ολοκληρωμένου CADx

Για την υλοποίηση του υποσυστήματος αυτόματης κατάτμησης έχουν προταθεί διάφορες τεχνικές όπως ο αλγόριθμος watershed και η μέθοδος επιπέδου συνόλου από τους Li et al [10], η Region Growing από τον Robb [11], η μέθοδος κατωφλίωσης από τους Suzuki and Toriwaki [12] καθώς και οι πιο πρόσφατες τεχνικές βασισμένες στο πεδίο της συχνότητας [13]. Ο αλγόριθμος που τελικά επιλέχθηκε είναι ο αλγόριθμος K-means με βάση τη μελέτη των Dalmiya et al [14].

Ο αλγόριθμος K-means είναι μια απλή αλλά κομψή μέθοδος κατάτμησης. Επιπρόσθετα, η

ταχύτητα εκτέλεσης είναι πολύ υψηλή. Μειονέκτημά του είναι η ανάγκη επιλογής σωστών αρχικών κέντρων (centroids) των συμπλεγμάτων (clusters) καθώς διαφορετικά υπάρχει πιθανότητα ο αλγόριθμος να μην καταφέρει να συγκλίνει. Γεγονός που συμβαίνει σε περιπτώσεις εικόνων με μεγάλο θόρυβο, όπως η μαστογραφία.

Ο αλγόριθμος K-means είναι μη επιβλεπόμενη μέθοδος ομαδοποίησης η οποία ταξινομεί τα σημεία εισόδου σε διαφορετικές κλάσεις με βάση την απόσταση μεταξύ τους. Ο αλγόριθμος λαμβάνει τα σημεία αυτά από έναν διανυσματικό χώρο και προσπαθεί να βρει μια φυσική ομαδοποίηση σε αυτά. Τα σημεία ομαδοποιούνται γύρω από κέντρα ζ_i , $i \in (1,2,...,k)$ τα οποία είναι αυτά που ελαχιστοποιούν την παρακάτω απόσταση:

$$V = \sum_{i=1}^k \sum_{x_j \in S_i} (x_j - \zeta_i)^2$$

όπου k ο αριθμός των κλάσεων S_i , $i = 1, 2, \dots, k$ και ζ_i το κέντρο ή ο μέσος όρος όλων των σημείων $x \in S_i$.

Τα βήματα που ακολουθήσαμε για την υλοποίηση του αλγορίθμου είναι τα εξής:

1. Υπολογισμός ιστογράμματος όλων των εντάσεων στην περιοχή της μάζας.
2. Αρχικοποίηση των κέντρων των κλάσεων με k εντάσεις. Στην υλοποίησή μας θεωρήσαμε ως αρχικά κέντρα τα κέντρα των ιστών.
3. Ομαδοποίηση σημείων με βάση την απόσταση των εντάσεών τους από την ένταση του κέντρου της κάθε κλάσης.

$$C^i = \arg \min_j \|x^i - \zeta_j\|^2$$

4. Υπολογισμός νέου κέντρου της κλάσης :

$$\zeta_i = \frac{\sum_{i=1}^m 1(c_i = j) x^i}{\sum_{i=1}^m 1(c_{i-1} = j)}$$

όπου k το πλήθος των κλάσεων που χρησιμοποιούμε, το i διατρέχει όλες τις εντάσεις της εικόνας, το j διατρέχει όλα τα κέντρα των κλάσεων και ζ_i είναι τα κέντρα των εντάσεων.

Η παραπάνω διαδικασία επαναλαμβάνεται ως ότου να μην υπάρχει καμία αλλαγή στα κέντρα των κλάσεων που χρησιμοποιούνται.

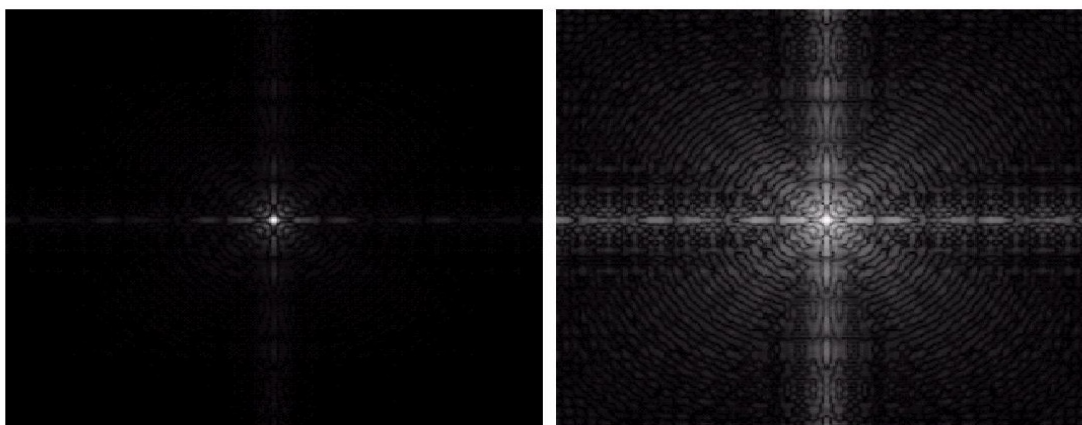
Στον αλγόριθμο εισήχθησαν Περιοχές Ενδιαφέροντος με μαστογραφικές μάζες και τα καταγράφηκαν τα αποτελέσματα της κατάτμησης. Για την καλύτερη δυνατή απόδοση του αλγορίθμου προηγήθηκε προεπεξεργασία των μαστογραφικών εικόνων ώστε να αυξηθεί η αντίθεση στις εντάσεις των pixels. Οι μετασχηματισμοί που χρησιμοποιήθηκαν για την αύξηση της αντίθεσης της εικόνας είναι οι:

1. Λογαριθμικός Μετασχηματισμός

Ο Λογαριθμικός Μετασχηματισμός έχει γενική μορφή:

$$s=c*\log(1+r)$$

Πρακτικά ο μετασχηματισμός αυτός διευρύνει τις τιμές των σκοτεινών περιοχών μιας εικόνας και συμπιέζει τις πολύ φωτεινές. Ο μετασχηματισμός Fourier μας επιτρέπει να διακρίνουμε καθαρά την επίδραση του Λογαριθμικού Μετασχηματισμού στο πεδίο της συχνότητας όπως φαίνεται στην Εικόνα 17.



Εικόνα 32: Μετασχηματισμός Fourier Λογαριθμικού Μετασχηματισμού

2. Συνάρτηση Εξίσωσης Ιστογράμματος

Η Συνάρτηση Εξίσωσης Ιστογράμματος μετασχηματίζει την εικόνα έτσι ώστε το ιστόγραμμα της μετασχηματισμένης εικόνας να περιέχει συγκεκριμένο αριθμό επιπέδων γκρι. Στην υλοποίησή μας χρησιμοποιήσαμε την προεπιλογή του Matlab για αριθμό επιπέδων 64. Η εικόνα μετασχηματίζεται με τέτοιον τρόπο ώστε προσεγγιστικά ίσος αριθμός pixel να αντιστοιχίζονται σε κάθε ένα επίπεδο γκρι, έτσι το ιστόγραμμα είναι σχεδόν επίπεδο. Στο πεδίο της συχνότητας ο αλγόριθμος

αντιστοιχίζει ένα στενό εύρος χαμηλών επιπέδων γκρι σε ένα μεγαλύτερο.

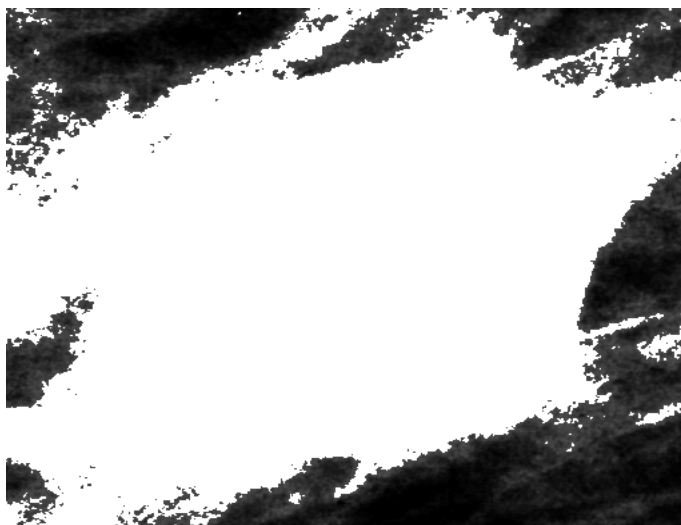
Όλα τα στάδια από τα οποία περιήλθε η Περιοχή Ενδιαφέροντος φαίνονται αναλυτικά στις Εικόνες 18-20. Στην Εικόνα 21 φαίνεται το αποτέλεσμα της κατάτμησης με χρήση του αλγορίθμου K-means και πλήθος κλάσεων ίσο με 5.



Εικόνα 33: Αρχική Εικόνα (1594_LMLO_ROI)



Εικόνα 34: Εικόνα μετά από τον Λογαριθμικό Μετασχηματισμό



Εικόνα 35: Εικόνα μετά από την Εξίσωση Ιστογράμματος

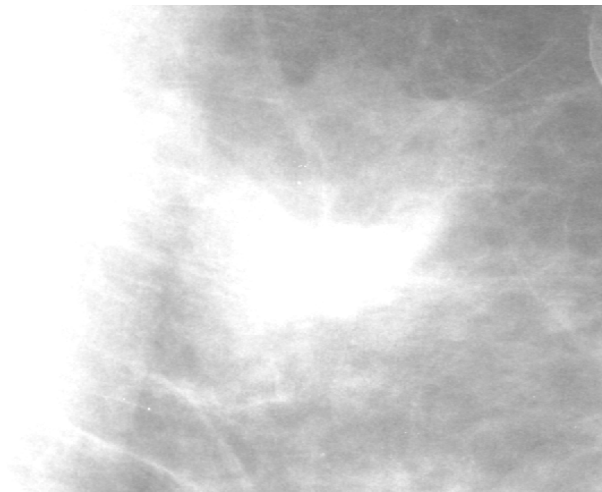


Εικόνα 36: Περίγραμμα Τελικής Εικόνας

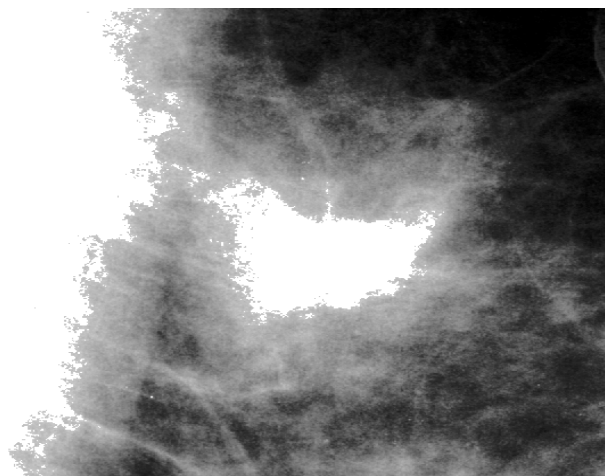
Ενδεικτικά παρουσιάζεται ένα ακόμα παράδειγμα του αυτόματης κατάτμησης του αλγορίθμου K-means στις Εικόνες 22 και 23.



Εικόνα 37: Μαστογραφική Μάζα (4016_LMLO_ROI)



Εικόνα 38: Εικόνα μετά από το Λογαριθμικό Αλγόριθμο



Εικόνα 39: Εικόνα μετά από την Εξίσωση Ιστογράμματος



Εικόνα 40: Μαστογραφική Μάζα μετά από κατάτμηση με τον Αλγόριθμο K-means

5.3 Μελλοντικές Προεκτάσεις

Η παρούσα μελέτη αποτέλεσε μια πρώτη προσπάθεια για σύγκριση των μεθοδολογιών που έχουν προταθεί τα τελευταία χρόνια στον τομέα της ταξινόμησης μαστογραφικών μαζών σε μεγάλη κλίμακα. Τα αποτελέσματα ήταν ικανοποιητικά και σε έναν βαθμό προσέφεραν απάντηση στο θέμα της επιλογής του κατάλληλου ταξινομητή καθώς και στην αποτελεσματικότητα της κάθε ομάδας των γνωστών χαρακτηριστικών.

Μελλοντικοί στόχοι περιλαμβάνουν την αύξηση των χαρακτηριστικών που χρησιμοποιούνται σαν είσοδος στον ταξινομητή για ακόμα καλύτερη ευελιξία, απόδοση και αντικειμενικότητα. Επίσης η μελέτη πάνω σε ένα ακόμα μεγαλύτερο σύνολο μαστογραφιών είναι πολύ σημαντική αφού πολλές υποκατηγορίες μαστογραφικών μαζών δεν περιείχαν τον απαραίτητο αριθμό περιπτώσεων για τη σωστή εκπαίδευση του συστήματος.

Μια ακόμα προοπτική αναφέρεται στην προηγούμενη ενότητα όπου διερευνάται η δυνατότητα αυτόματης κατάτμησης της μάζας από μια ευρύτερη Περιοχή Ενδιαφέροντος, με σκοπό την υποβοήθηση του ακτινολόγου στην χρονοβόρα και απαιτητική διεργασία της κατάτμησης.

Επιπρόσθετα, στη διαδικασία της διάγνωσης μπορούν να συνεκτιμηθούν χαρακτηριστικά που δεν λήφθηκαν υπόψιν στην παρούσα εργασία όπως η ηλικία ή το ιστορικό του ασθενή και διαφορετικοί τρόποι απεικόνισης του εσωτερικού του μαστού, όπως η Μαγνητική Τομογραφία και το Υπερηχογράφημα.

Βιβλιογραφία

[1] American Cancer Society, 2010

[2] Shantanu Gauri, Vandana Dialani, Priscilla J. Slanetz and Ronald L. Eisenberg, “Architectural Distortion of the Breast”, © American Roentgen Ray Society, AJR:201, November 2013

[3] Radiology Assistant

[4] Lori Mann Bruce and Reza R. Adhami, “Classifying Mammographic Mass Shapes Using the Wavelet Transform Modulus-Maxima Method”, IEEE Transactions on Medical Imaging, Vol. 18, No. 12, December 1999

[5] Γιάννης Ανδρεάδης, “Ανάπτυξη μεθοδολογιών για την υποβοηθούμενη διάγνωση συμπλεγμάτων μικροασβεστώσεων του μαστού”, Ιανουάριος 2014

[6] Gautam S. Muralidhar, Tamara Miner Haygood and Mia K. Markey, “Computer-Aided Detection of Breast Cancer – Have All Bases Been Covered?”, Libertas Academica, Breast Cancer: Basic and Clinical Research, 2008

[7] Wenmin Li, Jiawei Han, Jian Pei, “CMAR: accurate and efficient classification based on multiple class-association rules”, 2001. ICDM 2001, Proceedings IEEE International Conference on data Mining

[8] Fred Winsberg, Milton Elkin, Josiah Macy Jr, Victoria Bordaz and William Weymouth, “Detection of Radiographic Abnormalities in Mammograms by Means of Optical Scanning and Computer Analysis”, Radiology 89: 211-215, August 1967

- [9] Approximating Number of Hidden layer neurons in Multiple Hidden Layer BPNN Architecture
Saurabh Karsoliya Student, Department of Computer Science & Engineering, Maulana Azad National Institute of Technology, Bhopal, India
- [10] Li, N., M. Liu and Y. Li, Image segmentation algorithm using watershed transform and level set method. In Proc. ICASSP 2007. IEEE International Conference on Acoustics, Speech and Signal Processing (2007), 1: I-613-I-616
- [11] Robb, R.A., 2000. Biomedical Imaging, Visualization and Analysis. Wiley-Liss, USA.
- [12] Suzuki, H. and J. Toriwaki, 1991. Automatic Segmentation of Head MRI Images by Knowledge Guided Thresholding. Computer Med. Imaging Graph: The official J. Computerized Imaging Society, 15 (4): 233-240.
- [13] Jung, C., R., “Combining wavelets and watersheds for robust multi scale image segmentation”, Image and Vision Computing 25(2007), pp. 24–33
- [14] Shruti Dalmiya, Avijit Dasgupta and Soumya Kanti Datta, “ Application of Wavelet based K-means Algorithm in Mammogram Segmentation”. International Journal of Computer Applications (0975 – 8887) Volume 52– No.15, August 2012
- [15] Jinshan Tang and Xiaoming Liu, “Classification of Breast Mass in Mammography with an Improved Level Set Segmentation by Combining Morphological Features and Texture Features”, 2011
- [16] Man To Wong, Xiangjian He, Hung Nguyen, and Wei-Chang Yeh, “Mass Classification in Digitized Mammograms Using Texture Features and Artificial Neural Network”, T. Huang et al. (Eds.): ICONIP 2012, Part V, LNCS 7667, pp. 151–158. © Springer-Verlag Berlin Heidelberg 2012

- [17] Jae Young Choi and Yong Man Ro, “Multiresolution local binary pattern texture analysis combined with variable selection for application to false-positive reduction in computer-aided detection of breast masses on mammograms”, *Phys. Med. Biol.* **57** 7029 2012
- [18] Matthias Elter and Alexander Horsch, “CADx of mammographic masses and clustered microcalcifications: A review”, © American Association of Physicists in Medicine 2009
- [19] D. Cascio, F. Fauci, R. Magro, G. Raso, R. Bellotti, F. De Carlo, S. Tangaro, G. De Nunzio, M. Quarta, G. Forni, A. Lauria, M. E. Fantacci, A. Retico, G. L. Masala, P. Oliva, S. Bagnasco, S. C. Cheran, and E. Lopez Torres, “Mammogram Segmentation by Contour Searching and Mass Lesions Classification With Neural Network”, *IEEE Transactions on Nuclear Science*, Vol. 53, No. 5, October 2006
- [20] Mingqiang Yang, Kidiyo Kpalma, Joseph Ronsin. A Survey of Shape Feature Extraction Techniques. Peng-Yeng Yin. *Pattern Recognition*, IN-TECH, pp.43-90, 2008. <hal-00446037>
- [21] Harris Georgiou, Michael Mavroforakis, Nikos Dimitropoulos, Dionisis Cavouras, Sergios Theodoridis, ”Multi-scaled morphological features for the characterization of mammographic masses using statistical classification schemes”, *Artificial Intelligence in Medicine* (2007) 41, 39—55
- [22] Charles T. Zahn and Ralph Z. Roskies, “Fourier Descriptors for Plane Closed Curves”, *IEEE Transactions on Computers*, Vol. c-21, No. 3, March 1972
- [23] Denise Guliato, Juliano D. de Carvalho, Rangaraj M. Rangayyan and Sergio A. Santiago, “Feature Extraction from a Signature Based on the Turning Angle Function for the Classification of Breast Tumors”, *Journal of Digital Imaging*, Vol 21, No 2 (June), 2008: pp 129Y144
- [24] Pasquale Delogua, Maria Evelina Fantaccia, Parnian Kasaeb and Alessandra Reticoa, “Characterization of mammographic masses using a gradient-based segmentation algorithm and a

neural classifier”, arXiv:0803.1417v1 [physics.med-ph], 10 March 2008

[25] A. Karahaliou, S. Skiadopoulos, I. Boniatis, P. Sakellatopoulos, E. Likaki, G. Panayiotakis and L. Costaridou, “Texture analysis of tissue surrounding microcalcifications on mammograms for breast cancer diagnosis”, *The British Journal of Radiology*, 80 (2007), 648–656

[26] Haralick, R.M., Shanmugam, K. Dinstein, Its'Hak, “Textural Features for Image Classification”, *IEEE Transactions on Systems, Man and Cybernetics*, 1973

[27] Atef Boujelben, Ali Cherif Chahaabani, Hedi Tmar, Mohamed Abid, Feature Extraction from contours shape for tumor analyzing in Mammographic images, *Digital Image Computing: Techniques and Applications*, © 2009 IEEE DOI 10.1109/DICTA.2009.71

[28] Ping Zhang , Brijesh Verma , Kuldeep Kumar, “Neural vs. statistical classifier in conjunction with genetic algorithm based feature selection”, 2004 Elsevier B.V. All rights reserved. doi:10.1016/j.patrec.2004.09.053

[28] Shruti Dalmiya, Avijit Dasgupta, Soumya Kanti Datta, “Application of Wavelet based K-means Algorithm in Mammogram Segmentation”, *International Journal of Computer Applications* (0975 – 8887) Volume 52– No.15, August 2012

[29] Saurabh Karsoliya, Approximating Number of Hidden layer neurons in Multiple Hidden Layer BPNN Architecture, *International Journal of Engineering Trends and Technology- Volume3Issue6-2012*

[30] Huo Z, Giger ML, Vyborny CJ, Wolverton DE, Schmidt RA, Doi K., “Automated computerized classification of malignant and benign masses on digitized mammograms”, *Acad Radiol*.1998 Mar;5(3):155-68

[31] Jun Wei, Lubomir M. Hadjiiski, Berkman Sahiner, Heang-Ping Chan, Jun Ge, Marilyn A.

Roubidoux, Mark A. Helvie, Chuan Zhou, Yi-Ta Wu, Chintana Paramagul, Yiheng Zhang, “Computer-Aided Detection Systems for Breast Masses: Comparison of Performances on Full-Field Digital Mammograms and Digitized Screen-Film Mammograms”, *Acad Radiol* 2007; 14:659–669 doi:10.1016/j.acra.2007.02.017

[32] Naga R. Mudigonda, Rangaraj M. Rangayyan, and J. E. Leo Desautels, “Gradient and Texture Analysis for the Classification of Mammographic Masses ”, *IEEE TRANSACTIONS ON MEDICAL IMAGING*, VOL. 19, NO. 10, OCTOBER 2000

[33] Ning Li, Miaomiao Liu and Youfu Li, “Image Segmentation Algorithm using Watershed Transform and Level Set Method”, *ICASSP 2007, C2007 IEEE*

[34] Man To Wong, Xiangjian He, Hung Nguyen, and Wei-Chang Yeh, “Mass Classification in Digitized Mammograms Using Texture Features and Artificial Neural Network”, T. Huang et al. (Eds.): *ICONIP 2012, Part V, LNCS 7667*, pp. 151–158, 2012. © Springer-Verlag Berlin Heidelberg 2012

[35] Hidetomo Suzuki and Jun-ichiro Toriwaki, “Automatic segmentation of head MRI images by knowledge guided thresholding”, *Computerized Medical Imaging and Graphics* (Impact Factor: 1.5). 07/1991; 15(4):233-40. DOI: 10.1016/0895-6111(91)90081-6

[36] De-Shuang Huang, Yong Gan, Phalguni Gupta, M. Michael Gromiha, “Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence”, 7th International Conference, *ICIC 2011, Zhengzhou, China, August 11-14, 2011, Revised Selected Papers*

[37] Manavalan Radhakrishnan and Thangavel Kuttiannan, “Comparative Analysis of Feature Extraction Methods for the Classification of Prostate Cancer from TRUS Medical Images”, *IJCSI International Journal of Computer Science Issues*, Vol. 9, Issue 1, No 2, January 2012

[38] Leonardo de Oliveira Martins, Geraldo Braz Junior, Aristofanes Correa Silva, Anselmo Cardoso de Paiva, and Marcelo Gattass, “Detection of Masses in Digital Mammograms using K-means and Support Vector Machine”, *Electronic Letters on Computer Vision and Image Analysis*

8(2):39-50, 2009

[39] Jun Wei , Heang-Ping Chan , Berkman Sahiner , Lubomir M. Hadjiiski , Mark A. Helvie , Marilyn A. Roubidoux, Chuan Zhou , and Jun Ge, “Dual system approach to computer-aided detection of breast masses on mammograms”,*Medical Physics* 33 , 4157 (2006); doi: 10.1118/1.2357838

[40] Mehul P. Sampat, Mia K. Markey, and Alan C. Bovik, “Computer-Aided Detection and Diagnosis in Mammography ”, 2005 by Elsevier Academic Press

[41] Rafael C Gonzalez, Richard E. Woods, “Digital Image Processing”, © 2002 by Prentice-Hall, Inc

[42] Andreadis I, Spyrou GM, Antaraki A, Zografos GC, Koulocheri D, Giannakopoulou G, Nikita KS, Ligomenides PA. ”Combining SVM and Rule Based Classifiers for optimal classification in breast cancer diagnosis.” In: 8th Hellenic European Conference on Computer Mathematics and its Applications - HERCMA 2007. Athens, Greece; 2007

[43] Andreadis II, Nikita KS, Giannakopoulou G, Koulocheri D, Zografos GC, Antaraki A, Ligomenides PA, Spyrou GM. “Computer aided insights on obscure cases of breast cancer diagnosis.” In: IST 2008 - IEEE Workshop on Imaging Systems and Techniques Proceedings. Chania, Crete; 2008

[44] Andreadis II, Spyrou GM, Nikita KS. “A comparative study of image features for classification of breast microcalcifications”. *Measurement Science and Technology*. 2011

[45] Andreadis II, Spyrou GM, Ligomenides PA, Nikita KS. “Generation of clustered microcalcifications’ atlases for benign and malignant cases”. In: 13th IEEE International Conference on BioInformatics and BioEngineering (BIBE 2013). Chania, Greece; 2013

[46] Andreadis II, Spyrou GM, Ligomenides PA, Nikita KS. “Variations on breast density and subtlety of the findings require different computational intelligence pipelines for the diagnosis of

clustered microcalcifications”. In: 13th IEEE International Conference on BioInformatics and BioEngineering (BIBE 2013). Chania, Greece; 2013

[47] Andreadis I, Spyrou GM, Nikita KS. “Investigating the performance of a CADx scheme for mammography in specific BIRADS categories”. In: IEEE International Conference on Imaging Systems and Techniques. ; 2014

[48] Andreadis II, Spyrou GM, Nikita KS, “A CADx Scheme for Mammography Empowered with Topological Information from Clustered Microcalcifications’ Atlases”. IEEE Journal of Biomedical and Health Informatics, Volume:19 Issue:1, 2013