

Ατομική Διπλωματική Εργασία

**ΕΞΟΥΥΞΗ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΤΗΝ
ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΤΟΥ ΚΑΡΚΙΝΟΥ
ΤΟΥ ΕΝΔΟΜΗΤΡΙΟΥ ΚΑΤΑ ΤΗ ΔΙΑΡΚΕΙΑ
ΥΣΤΕΡΟΣΚΟΠΙΚΗΣ ΕΞΕΤΑΣΗΣ**

Δέσπω Θεοφάνους

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ



ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μάιος 2011

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΕΞΟΥΣΗ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΤΗΝ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ
ΤΟΥ ΚΑΡΚΙΝΟΥ ΤΟΥ ΕΝΔΟΜΗΤΡΙΟΥ ΚΑΤΑ ΤΗ
ΔΙΑΡΚΕΙΑ ΥΣΤΕΡΟΣΚΟΠΙΚΗΣ ΕΞΕΤΑΣΗΣ

Δέσπω Θεοφάνους

Επιβλέπων Καθηγητής
Κωνσταντίνος Παττίχης

Η Ατομική Διπλωματική Εργασία υποβλήθηκε προς μερική εκπλήρωση των
απαιτήσεων απόκτησης του πτυχίου Πληροφορικής του Τμήματος Πληροφορικής του
Πανεπιστημίου Κύπρου

Μάιος 2010

Ευχαριστίες

Με την παρούσα εργασία ολοκληρώνεται το στάδιο των προπτυχιακών μου σπουδών στο Τμήμα Πληροφορικής του Πανεπιστημίου Κύπρου.

Κρίνω λοιπόν σκόπιμο να ευχαριστήσω θερμά τον επιβλέποντα Καθηγητή Κωνσταντίνο Παττίχη για τη διαρκή υποστήριξη και καθοδήγηση για την ολοκλήρωση της εργασίας.

Θα ήθελα επίσης να ευχαριστήσω ιδιαίτερα τον Καθηγητή Μάριο Νεοφύτου ο οποίος ήταν πάντοτε πρόθυμος να με βοηθήσει, με την παροχή πληροφοριών, χάρη στην εμπειρία και την γνώση του στο τομέα αυτό, και ο οποίος διαδραμάτισε εξαιρετικά σημαντικό ρόλο στην εκπόνηση της εργασίας μου.

Επίσης θα ήθελα να ευχαριστήσω θερμά την οικογένεια μου για τη συμπαράσταση, την κατανόηση και την υπομονή τους σε όλη τη διάρκεια αυτής της εργασίας αλλά και των σπουδών μου. Με την δική τους πολύτιμη βοήθεια, κατάφερα να ξεπεράσω διάφορα προβλήματα τα οποία συναντούσα κατά καιρούς.

Εν κατακλείδι, θα ήθελα να εκφέρω ένα μεγάλο ευχαριστώ στον άνθρωπό μου, ο οποίος είναι πάντα δίπλα μου και που η στήριξή του, η αγάπη του και η κατανόηση του αποτέλεσαν τα αναγκαία εφόδια, τα οποία είχα ανάγκη κατά τη διάρκεια των σπουδών μου.

Περίληψη

Αναμφισβήτητα, ο καρκίνος του ενδομητρίου αποτελεί τον συχνότερο γυναικολογικό καρκίνο, μετά τον καρκίνο του μαστού, στις αναπτυγμένες χώρες. Ο καρκίνος ξεκινά από το ενδομήτριο, το οποίο αποτελεί τον εσωτερικό χιτώνα της μήτρας. Ο καρκίνος του ενδομητρίου είναι από τις περιπτώσεις καρκίνου που μπορεί να διαγνωσθεί σε πρώιμο στάδιο και να προληφθεί σχεδόν πλήρως με τακτική εξέταση. Η κακοήθεια του καρκίνου του ενδομητρίου εμφανίζεται συνήθως σε γυναίκες που βρίσκονται στην ηλικία των 50 με 60 ετών και είναι η συχνότερη κακοήθεια των γυναικείων γεννητικών οργάνων. Ο καρκίνος αυτός όμως αρχίζει από την επιφάνειά του ενδομητρίου και αυτό έχει ως αποτέλεσμα, ο καρκίνος αλλά και ο αριθμός θανάτων λόγω της κακοήθειας αυτής να μειωθεί στο ελάχιστο. Αυτό κάνει την έρευνά μας πολύ σημαντική στην πρόληψη και ανίχνευση του καρκίνου του ενδομητρίου.

Στόχος της συγκεκριμένης μελέτης, είναι η υποστήριξη διάγνωσης των γιατρών μέσω της επεξεργασίας και εξόρυξης δεδομένων σε βάσεις δεδομένων με στοιχεία τα οποία προκύπτουν μετά από εφαρμογή αλγόριθμων χαρακτηριστικών υφής στις εικόνες που δημιουργούνται μετά από επεξεργασία και μετασχηματισμό του χρώματος των υστεροσκοπικών εικόνων ενδομητρίου. Είναι αναμφίβολα δεδομένο ότι ο γιατρός αποτελεί το βασικότερο στοιχείο στη διάγνωση και αντιμετώπιση περιστατικών γυναικολογικού καρκίνου. Συνεπώς, η βασική ιδέα της μελέτης αυτής είναι να βοηθήσει τον κλινικό γιατρό να αναγνωρίζει και να διαχωρίζει φυσιολογικές και παθολογικές περιοχές ενδιαφέροντος.

Η εργασία αυτή, βασίζεται στην υπάρχουσα μελέτη της Μεταπτυχιακής φοιτήτριας Αντωνίας Λοΐζου, του Τμήματος Πληροφορικής του Πανεπιστημίου Κύπρου. Στην υπάρχουσα εργασία υπήρχαν δεδομένα από περισσότερα χρωματικά κανάλια, από τα οποία επιλέχτηκε το χρωματικό κανάλι Υ για να μελετηθεί στην συγκεκριμένη εργασία, και να εξαχθούν αποτελέσματα.

Στην μελέτη αυτή χρησιμοποιήθηκαν οι βάσεις δεδομένων που δημιουργήθηκαν

μετά από επεξεργασία 202 εικόνων υστεροσκόπησης ενδομητρίου 40 γυναικών, με 101 εικόνες ενδιαφέροντος που παρουσιάζουν φυσιολογικές περιοχές και 101 εικόνες ενδιαφέροντος που παρουσιάζουν ύποπτες περιοχές καρκίνου και την εφαρμογή τριών αλγόριθμων χαρακτηριστικών υφής σε διάφορα συστήματα απεικόνισης των 202 αυτών εικόνων. Οι αλγόριθμοι αυτοί μπορούν να χαρακτηρίσουν μια εικόνα και να εξαγάγουν χρήσιμες πληροφορίες για το περιεχόμενό της.

Στην διαδικασία εξόρυξης δεδομένων/εξαγωγής αποτελεσμάτων κατηγοριοποίησης χρησιμοποιήθηκε ο αλγόριθμος C4.5.

Η έρευνα δείχνει ότι η επεξεργασία ιατρικής εικόνας και η κατηγοριοποίηση της μπορεί να παίξει σημαντικό ρόλο στην αναγνώριση παθολογικών και φυσιολογικών εικόνων ενδομητρίου και άλλων οργάνων. Ουσιαστικά με τις πληροφορίες που δύναται να δώσουμε στον γιατρό μπορούμε να τον βοηθήσουμε να αυξήσει τη διαγνωστική του ακρίβεια σε δύσκολες περιπτώσεις γυναικολογικού καρκίνου. Η μελέτη μας επίσης μπορεί να βοηθήσει στη μελλοντική δημιουργία κοινής ιατρικής βάσης δεδομένων (με τα αποτελέσματα της εξόρυξης δεδομένων) από ανεξάρτητα κέντρα αντιμετώπισης γυναικολογικού καρκίνου για την πληρέστερη κατανόηση του προβλήματος καθώς και στη δημιουργία αυτόματων συστημάτων υπολογιστών τα οποία θα βοηθήσουν τον γιατρό σε πραγματικό χρόνο να δώσει μια πρώτη εκτίμηση του περιστατικού ή θα τον καθοδηγήσουν στον έλεγχο με ιστοπαθολογική εξέταση υπόπτων περιοχών καρκίνου.

Περιεχόμενα

Κεφάλαιο 1	Εισαγωγή.....	1
	1.1 Σύντομη Περιγραφή Προβλήματος.....	1
	1.2 Προέλευση Βάσης Δεδομένων.....	2
	1.3 Στόχος Εργασίας.....	3
	1.4 Εργαλεία που χρησιμοποιήθηκαν.....	3
	1.5 Αλγόριθμος που χρησιμοποιήθηκαν.....	3
	1.6 Προσδοκώμενα Αποτελέσματα.....	4
	1.7 Αρχιτεκτονική Συστήματος.....	4
	1.8 Δομή Εργασίας.....	5
Κεφάλαιο 2	Γυναικολογικός Καρκίνος.....	7
	2.1 Επισκόπηση.....	7
	2.2 Καρκίνος Ενδομητρίου	7
	2.3 Ενδοσκοπική Μέθοδος.....	9
	2.3.1 Γενικά.....	9
	2.3.2 Υστεροσκόπηση	11
	2.4 Σύνοψη.....	13
Κεφάλαιο 3	Χαρακτηριστικά Υφής	15
	3.1 Επισκόπηση.....	15
	3.2 Γενικές πληροφορίες για τα χαρακτηριστικά υφής.....	16
	3.3 Αλγόριθμος Statistical Features (SF)	19
	3.4 Αλγόριθμος Spatial Gray Level Dependence Matrices (SGLDM)	20
	3.4.1 Συμπεράσματα από τον αλγόριθμο SGLDM	24
	3.4.2 Μέτρα υπολογισμού χαρακτηριστικών υφής	25
	3.5 Αλγόριθμος Gray Level Difference Statistics (GLDS)	28
	3.5.1 Συμπεράσματα από τον αλγόριθμο GLDS	28

Κεφάλαιο 4	Αλγόριθμοι	30
	4.1 Επισκόπηση	30
	4.2 Γενικές Πληροφορίες Αλγορίθμου C4.5	30
	4.3 Κατασκευή δέντρων απόφασης	32
	4.4 Επιλογή χαρακτηριστικού ελέγχου	33
Κεφάλαιο 5	Μεθοδολογία	39
	5.1 Επισκόπηση	39
	5.2 Καταγραφή του ενδοσκοπικού βίντεο/ εικόνων	40
	5.3 Χειροκίνητη κατάτμηση περιοχών ενδιαφέροντος	41
	5.4 Προέλευση δείγματος εικόνων	43
	5.5 Εφαρμογή αλγορίθμων χαρακτηριστικών υφής	44
	5.6 Αλγόριθμοι κατηγοριοποίησης / εξόρυξης δεδομένων	45
	5.7 Μετρικές αποτελεσματικότητας της διάγνωσης από το σύστημα κατηγοριοποίησης	47
	5.8 Μετρικές αποτελεσμάτων από το σύστημα Weka	49
Κεφάλαιο 4	Αποτελέσματα	52
	6.1 Επισκόπηση	52
	6.2 Στατιστική Ανάλυση Gray Scale εικόνων ενδομητρίου σε γυναίκες	53
	6.3 Αποτελέσματα αλγορίθμου C4.5 για τον αλγόριθμο SGLDM	56
	6.4 Αποτελέσματα αλγορίθμου C4.5 για τον αλγόριθμο SF	59
	6.5 Αποτελέσματα αλγορίθμου C4.5 για τον αλγόριθμο GLDS	61
	6.6 Αποτελέσματα αλγορίθμου C4.5 για τον αλγόριθμο SGLDM και SF	63
	6.7 Αποτελέσματα αλγορίθμου C4.5 για τον αλγόριθμο	

SGLDM και GLDS	66
6.8 Αποτελέσματα αλγόριθμου C4.5 για τον αλγόριθμο SF και GLDS	69
6.9 Αποτελέσματα αλγόριθμου C4.5 για τον αλγόριθμο SGLDM, SF και GLDS	72
6.10 Συνοπτικά Αποτελέσματα	75
ΚΕΦΑΛΑΙΟ 7 Συμπεράσματα	76
7.1 Επισκόπηση	76
7.2 Συμπεράσματα	77
7.2.1 Πραγματικές εικόνες ενδομητρίου	77
7.2.2 Αλγόριθμοι κατηγοριοποίησης	77
7.2.3 Σύγκριση με προηγούμενη εργασία	79
7.3 Μελλοντική Εργασία	79
7.3.1 Επιλογή κατάλληλων χαρακτηριστικών υφής	79
7.3.2 Αυτόματη ανάλυση υστεροσκοπικού βίντεο και εξαγωγή αποτελεσμάτων	80
Βιβλιογραφία	83

Κεφάλαιο 1

Εισαγωγή

1.1 Σύντομη Περιγραφή Προβλήματος	1
1.2 Προέλευση Βάσης Δεδομένων	2
1.3 Στόχος Εργασίας	3
1.4 Εργαλεία που χρησιμοποιήθηκαν	3
1.5 Αλγόριθμοι που χρησιμοποιήθηκαν	3
1.6 Προσδοκώμενα Αποτελέσματα	4
1.7 Αρχιτεκτονική Συστήματος	4
1.8 Δομή Εργασίας	5

1.1 Σύντομη Περιγραφή Προβλήματος

Ο καρκίνος του γυναικείου γεννητικού συστήματος, είναι αναμφισβήτητο ότι είναι μία από τις κυριότερες αιτίες νοσηρότητας και θνησιμότητας από καρκίνο στο γυναικείο πληθυσμό. Η συγκεκριμένη μορφή καρκίνου, αποτελεί μία από τις συχνότερες μορφές γυναικολογικού καρκίνου στις βιομηχανικά ανεπτυγμένες χώρες. Η κακοήθεια του καρκίνου του ενδομητρίου εμφανίζεται συνήθως σε γυναίκες που βρίσκονται στην ηλικία των 50 με 60 ετών.

Ο καρκίνος του ενδομητρίου αρχίζει από την επιφάνειά του και αυτό έχει ως αποτέλεσμα, ο καρκίνος να μπορεί να διαγνωσθεί σε πρώιμο στάδιο και ο αριθμός θανάτων λόγω της κακοήθειας αυτής να μειωθεί στο ελάχιστο. Συνεπώς, η παρούσα έρευνα καθίσταται πάρα πολύ σημαντική, αφού αφορά την πρόληψη και ανίχνευση του καρκίνου του ενδομητρίου.

Αν και υπάρχουν αρκετές αναφορές για τον καρκίνο του γυναικείου

αναπαραγωγικού συστήματος καθώς και τρόποι θεραπείας του από ιατρική σκοπιά, ωστόσο δεν έχουν διεξαχθεί έρευνες που να ασχολούνται με την εύρεση και αναγνώριση του γυναικολογικού καρκίνου εισάγοντας τη χρήση των ηλεκτρονικών υπολογιστών.

Η συγκεκριμένη μελέτη θεωρείται η βάση για την εισαγωγή της χρήσης των υπολογιστών στην διαδικασία διάγνωσης και κατ' επέκταση αντιμετώπισης του γυναικολογικού καρκίνου του ενδομητρίου. Η βασική ιδέα της μελέτης αυτής είναι να βοηθήσει τον κλινικό γιατρό να αναγνωρίζει και να διαχωρίζει φυσιολογικές και παθολογικές περιοχές ενδιαφέροντος, αυξάνοντας την διαγνωστική ακρίβεια και εγκυρότητα της κάθε περίπτωσης.

1.2 Προέλευση βάσης δεδομένων

Η συγκεκριμένη μελέτη, βασίζεται σε βάσεις δεδομένων που προέρχονται από πληροφορίες 202 ιατρικών εικόνων.

Οι 202 εικόνες προέρχονται από υστεροσκόπηση του ενδομητρίου 40 γυναικών. Οι 101 εικόνες ενδιαφέροντος παρουσιάζουν φυσιολογικές περιοχές και 101 εικόνες ενδιαφέροντος παρουσιάζουν ύποπτες περιοχές καρκίνου. Όλες οι περιοχές ενδιαφέροντος παρουσιάζονται στη μορφή έγχρωμων εικόνων (σύστημα απεικόνισης RGB). Οι έγχρωμες εικόνες έχουν μετασχηματιστεί σε εικόνες γκρι τόνου στο χρωματικό κανάλι Y.

Στις εικόνες αυτές και συγκεκριμένα στις δεκατρείς μορφές απεικόνισης των 202 αυτών εικόνων εφαρμόστηκε ο αλγόριθμος χαρακτηριστικών υφής C4.5. Ο αλγόριθμος, ουσιαστικά χαρακτηρίζει μια εικόνα και εξάγει χρήσιμες πληροφορίες (χαρακτηριστικά υφής) για το περιεχόμενό της. Συνολικά για την έρευνα που διεξήχθηκε, χρησιμοποιήθηκαν δεκατρείς βάσεις δεδομένων που προέκυψαν από τα χαρακτηριστικά υφής των εικόνων στα διάφορα συστήματα απεικόνισης και των χρωματικών τους καναλιών καθώς και σε γκρι τόνο. Για την κάθε βάση δημιουργήθηκαν δέκα (τυχαία) μμοντέλα τα οποία χρησιμοποιήθηκαν για την εξόρυξη δεδομένων (εξαγωγή αποτελεσμάτων κατηγοριοποίησης) στις ιατρικές αυτές εικόνες

υστεροσκοπήσης ενδομητρίου.

1.3 Στόχος εργασίας

Στόχος της συγκεκριμένης έρευνας είναι η δημιουργία ενός διαγνωστικού συστήματος το οποίο θα βοηθήσει τον γιατρό στη λήψη της καλύτερης δυνατής απόφασης και να αυξήσει τη διαγνωστική του ακρίβεια σε δύσκολες περιπτώσεις γυναικολογικού καρκίνου. Το σύστημα αυτό βασίζεται στα χαρακτηριστικά υφής, των κατάλληλα μετασχηματισμένων εικόνων (που όπως δείχνει η έρευνα μας θα ήταν καλύτερα ο μετασχηματισμός αυτός να είναι στο χρωματικό κανάλι Υ) και στην εξόρυξη δεδομένων με την χρήση των δέντρων αποφάσεων (decisions trees).

Ουσιαστικά στόχος είναι η μελέτη αυτή να μπορεί να αυξήσει σημαντικά τη διαγνωστική ακρίβεια αναγνώρισης και διαχωρισμού περιπτώσεων όχι μόνο γυναικολογικού καρκίνου αλλά και άλλου είδους καρκίνων στο ανθρώπινο σώμα.

1.4 Εργαλεία που χρησιμοποιήθηκαν

Για υποστήριξη της βάσης δεδομένων, επεξεργασία/ εξόρυξη των δεδομένων και εξαγωγή αποτελεσμάτων κατηγοριοποίησης τους χρησιμοποιήθηκε πρόγραμμα Weka 3.6 Edition.

1.5 Αλγόριθμος που χρησιμοποιήθηκε

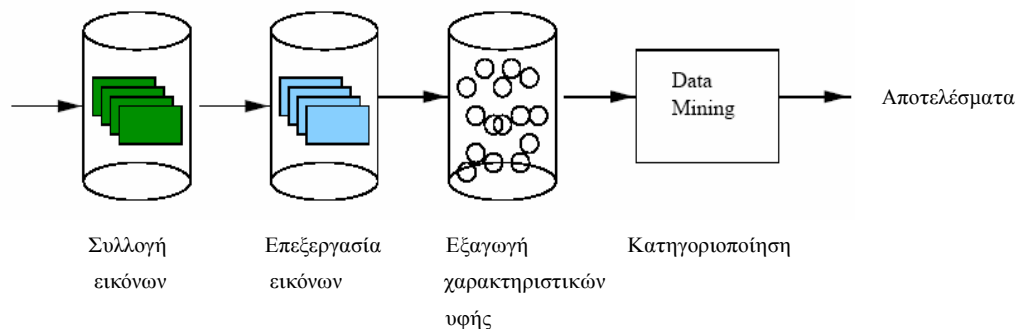
Ο αλγόριθμος που χρησιμοποιήθηκε είναι ο αλγόριθμος δέντρου αποφάσεων C4.5. Ο αλγόριθμος C4.5 χειρίζεται χαρακτηριστικά συνεχών τιμών καθώς επίσης επιλέγει την κατάλληλη μετρική για την επιλογή των χαρακτηριστικών διάσπασης. Τέλος, ο συγκεκριμένος αλγόριθμος έχει χρησιμοποιηθεί πάρα πολύ και έχει δώσει αποτελέσματα στα οποία φαίνεται ότι έχει βελτιώσει την υπολογιστική απόδοση.

1.6 Προσδοκόμενα αποτελέσματα

Με την ολοκλήρωση της παρούσας μελέτης, αναμένεται να έχουμε αποτελέσματα που θα βοηθήσουν τον γιατρό να αυξήσει τη διαγνωστική του ακρίβεια σε δύσκολες περιπτώσεις γυναικολογικού καρκίνου αλλά και για την πληρέστερη κατανόηση του προβλήματος του καρκίνου αυτού.

1.7 Αρχιτεκτονική Συστήματος

Τα μέρη της εν λόγω μελέτης αποτελούνται από τέσσερα βασικά στάδια:



Σχήμα 1.1 Αρχιτεκτονική Συστήματος

- Το πρώτο στάδιο περιέχει τη λήψη και αποθήκευση υστεροσκοπικών εικόνων ενδομητρίου: Κατά τη διάρκεια μιας υστεροσκοπικής επέμβασης ο γιατρός σε δύσκολες περιπτώσεις γυναικολογικού καρκίνου δημιουργεί άμεσες βιοψίες με στόχο την επαλήθευση της διάγνωσης του. Τα βίντεο / εικόνες αυτά/αυτές που συλλέγονται αποθηκεύονται στην ιατρική βάση δεδομένων και παράλληλα ο γιατρός επιλέγει ύποπτες περιοχές καρκίνου καθώς και φυσιολογικές περιοχές του ενδομητρίου. Στην εν λόγω έρευνα έχουμε συλλέξει 404 εικόνες από πραγματικό ενδομήτριο, 202 εικόνες ενδιαφέροντος που παρουσιάζουν φυσιολογικές περιοχές και 202 εικόνες ενδιαφέροντος που παρουσιάζουν ύποπτες περιοχές καρκίνου και έχουν εξαχθεί από ένα σύνολο 40 γυναικών.
- Το δεύτερο στάδιο περιέχει την προεπεξεργασία των υστεροσκοπικών αυτών εικόνων και την ευθυγράμμισή τους με τον αλγόριθμο γάμμα.
- Το τρίτο στάδιο περιλαμβάνει εξαγωγή χρήσιμων πληροφοριών από τις εικόνες καθώς και τη στατιστική τους επεξεργασία: Μέτα τον μετασχηματισμό των 404 υστεροσκοπικών εικόνων ενδομητρίου στα

διάφορα συστήματα απεικόνισης έχουν εφαρμοστεί τρεις αλγόριθμοι χαρακτηριστικών υφής στις εικόνες αυτές. Οι αλγόριθμοι αυτοί είναι ο Statistical Features (SF), ο Spatial Gray Level Dependence Matrices (SGLDM) και ο Gray Level Difference Statistics (GLDS). Τα χαρακτηριστικά υφής μπορούν να περιγράψουν με μεγάλη ακρίβεια την υφή της περιοχής ενδιαφέροντος και να δώσουν μοναδική ταυτότητα σε κάθε υπό εξέταση περιοχή. Με αυτά τα χαρακτηριστικά, η χρήση των υπολογιστών στην διαγνωστική χειρουργική, γίνεται υλοποιήσιμη και έξυπνα συστήματα επεξεργασίας εικόνων έχουν προταθεί με στόχο την αύξηση της διαγνωστικής ακρίβειας του περιστατικού.

- Τέλος το τέταρτο στάδιο περιέχει το διαχωρισμό των εικόνων, σε φυσιολογικές και παθολογικές, με στόχο την αύξηση της διακριτικής ικανότητας του γιατρού σε δύσκολες περιπτώσεις γυναικολογικού καρκίνου.

1.8 Δομή Εργασίας

Στην παρούσα διατριβή δίδεται αρχικά μια σύντομη περιγραφή του γυναικολογικού καρκίνου στο γυναικολογικό όργανο ενδομήτριο καθώς και χρήσιμες λεπτομέρειες για το ποια είναι η φύση του καρκίνου σε αυτό. Ακολούθως περιγράφεται η υστεροσκόπηση, μία μέθοδος παρατήρησης εσωτερικών οργάνων του ανθρώπινου σώματος, η οποία προσφέρει την ευκαιρία στον ιατρό για ολική και άμεση παρακολούθηση των οργάνων της γυναίκας. Με την μέθοδο αυτή συλλέχθηκαν οι υστεροσκοπικές εικόνες ενδομητρίου οι οποίες αποτέλεσαν το αντικείμενο της έρευνας μας.

Στη συνέχεια περιγράφεται και αναλύεται ο αλγόριθμος χαρακτηριστικών υφής C4.5 τον οποίο εφαρμόσαμε στις υστεροσκοπικές εικόνες ενδομητρίου που αναφέραμε πιο πάνω.

Στα χαρακτηριστικά υφής τα οποία έχουν εξαχθεί από τους τρεις αλγόριθμους χαρακτηριστικών υφής έχουν εφαρμοστεί δύο αλγόριθμοι δέντρων απόφασης με σκοπό την κατηγοριοποίηση των υστεροσκοπικών εικόνων που αποτέλεσαν το δείγμα

μας.

Ακολούθως αναλύεται η μεθοδολογία που έχουμε ακολουθήσει στην μελέτη. Πρώτα παρουσιάζεται η καταγραφή των ενδοσκοπικών βίντεο/εικόνων, στη συνέχεια προτείνεται το πρωτόκολλο ενδοσκοπικών εικόνων πριν από κάθε ενδοσκοπική εξέταση. Περαιτέρω καλύπτονται οι λειτουργίες της χειροκίνητης κατάτμησης περιοχών ενδιαφέροντος, ο μετασχηματισμός χρώματος των εικόνων που αποτέλεσαν το δείγμα μας και η εφαρμογή των αλγορίθμων χαρακτηριστικών υφής σε αυτά. Επειπλέον αναλύεται η μεθοδολογία που ακολουθήσαμε για την εξόρυξη δεδομένων/εξαγωγής αποτελεσμάτων κατηγοριοποίησης από τις βάσεις δεδομένων με τα χαρακτηριστικά υφής των υστεροσκοπικών εικόνων μας. Αναφορά επίσης γίνεται στις μετρικές αξιολόγησης των αποτελεσμάτων της εφαρμογής του αλγορίθμου C4.5 στα συγκεκριμένα χαρακτηριστικά υφής (βλ. Κεφάλαιο 5).

Τέλος παρουσιάζονται η στατιστική ανάλυση των χαρακτηριστικών υφής των εικόνων μας και τα αποτελέσματα που δημιουργήθηκαν τον αλγόριθμο που χρησιμοποιήθηκε.

Κεφάλαιο 2

Γυναικολογικός Καρκίνος

2.1	Επισκόπηση	7
2.2	Καρκίνος Ενδομητρίου	7
2.3	Ενδοσκοπική μέθοδος	9
2.3.1	Γενικά	9
2.3.2	Υστεροσκόπηση	11
2.4	Σύνοψη	13

2.1 Επισκόπηση

Στην ενότητα αυτή, γίνεται ανάλυση και επεξήγηση του γυναικολογικού καρκίνου που παρουσιάζεται στο γυναικολογικό όργανο ενδομήτριο, καθώς επίσης και της μεθόδου παρατήρησης και εντόπισης του καρκίνου στο ανθρώπινο σώμα, την υστεροσκόπηση.

2.2 Καρκίνος Ενδομητρίου

Ο καρκίνος του ενδομητρίου αποτελεί μία από τις συχνότερες μορφές καρκίνου, η οποία εμφανίζεται συνήθως σε γυναίκες που βρίσκονται στην ηλικία των 50 με 60 ετών.

Ο καρκίνος του ενδομητρίου αποτελεί τη συχνότερη κακοήθεια των γυναικείων γεννητικών οργάνων. Ο καρκίνος ξεκινά από το ενδομήτριο, δηλαδή τον εσωτερικό χιτώνα της μήτρας.

Ο καρκίνος του ενδομητρίου εμφανίζεται σε πολλές περιπτώσεις μετά την εμμηνόπαυση, με μέση ηλικία διάγνωσης τα 60 χρόνια. Επίσης, στους παράγοντες κινδύνου της νόσου αυτής, αν και επηρεάζει σε πολύ μικρότερο βαθμό απ' ότι στον καρκίνο της ωοθήκης και του μαστού, είναι η κληρονομική επιβάρυνση. Παράλληλα, στους παράγοντες κινδύνου καταλογίζονται η παχυσαρκία, η ατοκία, το σύνδρομο των πολυκυστικών ωοθηκών, η πρόωγη εμμηναρχή και η καθυστερημένη εμμηνόπαυση. Επιπλέον, οι γυναίκες που έχουν υπέρταση και διαβήτη θα πρέπει να ελέγχονται πιο συχνά για το λόγο ότι ανήκουν στην ομάδα ατόμων τα οποία διατρέχουν μεγαλύτερο κίνδυνο. Ιδιαίτερα οι γυναίκες άνω των 40 ετών, εάν έχουν διαταραχές περιόδου, θα πρέπει να επικοινωνούν όσον το δυνατόν πιο σύντομα με το γυναικολόγο τους, ο οποίος θα κρίνει αν θα πρέπει να γίνουν περαιτέρω εξετάσεις.

Η ταξινόμηση του καρκίνου του ενδομητρίου της μήτρας σε στάδια διευκολύνει την πρόγνωση και την επιλογή της θεραπείας. Τα στάδια αυτά κατηγοριοποιούνται ως εξής:

ΣΤΑΔΙΟ 0	Ο καρκίνος περιορίζεται στα επιθηλιακά κύτταρα.
ΣΤΑΔΙΟ 1	Ο καρκίνος περιορίζεται στο σώμα της μήτρας.
ΣΤΑΔΙΟ 2	Ο καρκίνος έχει επεκταθεί προς τον ενδοτράχηλο.
ΣΤΑΔΙΟ 3	Ο καρκίνος έχει προχωρήσει εκτός της μήτρας.
ΣΤΑΔΙΟ 4	Ο καρκίνος έχει καταλάβει τα όργανα και τα οστά της πυέλου.

Πίνακας 2.1 Στάδια καρκίνου του ενδομητρίου

Ένα μεγάλο ποσοστό των περιπτώσεων διγινώσκεται στο στάδιο 1, ενώ η συνολική 5ετής επιβίωσης το 2001 ήταν 76,5%. Ανιχνεύεται συχνά σε αρχικό στάδιο εξαιτίας του γεγονότος ότι ο καρκίνος αυτός δίνει συμπτώματα ,τα οποία οδηγούν τη γυναίκα στον γυναικολόγο της.

Συμπτώματα του καρκίνου του ενδομητρίου αποτελούν η ασταθής περίοδος μετά την ηλικία των 40 ετών, η σταγονοειδής αιμόρροια σε μετεμμηνοπαυσιακές γυναίκες και τα φουσκώματα κοιλιάς. Επίσης, σύμπτωμα του καρκίνου αποτελεί η αυξημένη ροή αίματος κατά τη διάρκεια της περιόδου, καθώς και η μεσοκυκλική αιμορραγία, η αιμορραγία δηλαδή ανάμεσα στις περιόδους.

Μια γυναίκα μπορεί να προλάβει την εμφάνιση του καρκίνου εάν κάνει την ετήσια γυναικολογική της εξέταση και της γίνεται μέτρηση του πάχους του ενδομητρίου της. Οι γυναίκες θα πρέπει μία φορά τον χρόνο να επισκεπτονται στο γυναικολόγο τους και να κάνουν το Τεστ Παπανικολάου για τον αποκλεισμό κακοήθειας από τον τράχηλο της μήτρας. Παράλληλα, η υστεροσκόπηση αποτελεί ένα από τα πιο σύγχρονα εργαλεία καθώς επιτρέπει την άμεση επισκόπηση της κοιλότητας της μήτρας και την δυνατότητα λήψης κατευθυνόμενων βιοψιών από ύποπτες περιοχές.

2.3 Ενδοσκοπική μέθοδος

2.3.1 Γενικά

Προσπάθειες για διεύρυνση του «εσωτερικού σκότους» του ορθού αναφέρονται από το 1700 π.χ., στην Αίγυπτο. Χρησιμοποίηση άκαμπτου σωλήνα για εξέταση του ορθού ασθενών με αιμορροΐδες δακτυλίου αναφέρεται και από τους αρχαίους Έλληνες και Ρωμαίους.

Η σχολή της νήσου Κω, την οποία διηύθυνε ο Ιπποκράτης (460-375 π. Χ.), έδωσε κάποιες περιγραφές ενδοσκόπησης, με τη χρήση πρωτογενών εργαλείων που είχαν επινοηθεί από τον ίδιο. Ενδοσκοπικά εργαλεία σε αρχέγονη μορφή καθώς και γραφές που αναφέρουν τη χρήση ενός σωλήνα δια μέσου του οποίου γινόταν επισκόπηση του κόλπου και του τραχήλου της μήτρας έχουν ανακαλυφθεί και σε άλλες περιοχές. Σήμερα είναι παγκόσμια αποδεκτό ότι ο Βοτζίνι (1806) είναι ο πατέρας της σύγχρονης ενδοσκόπησης, μετά την κατασκευή ενός σύνθετου οργάνου, που το ονόμασε «Lichtleiter», για την επισκόπηση του τραχήλου της μήτρας.

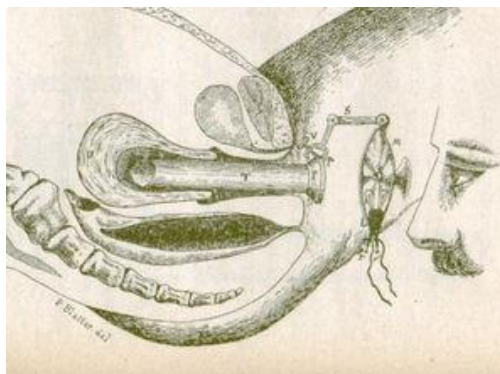
Ο Pantaleoni (1869) θεωρείται ο πρώτος που εφάρμοσε την επισκόπηση της ενδομητριακής κοιλότητας (υστεροσκόπηση) σε γυναίκες, χρησιμοποιώντας ένα ουρηθροσκόπιο με φως κεριού και καθρέφτη. Έκτοτε η υστεροσκόπηση προχώρησε με τη βελτίωση των ενδοσκοπίων, προσθήκη φακών και φωτιστικών πηγών, ώστε μετά το 1900 να κυκλοφορήσει και σύγγραμμα στο Παρίσι (Σχήμα 2.1).

Πατέρας της ενδοσκόπησης θεωρείται ο Mikulitz, μαθητής του χειρουργού Billroth, που επινόησε ένα άκαμπτο όργανο με φωτισμό και κατάφερε να πραγματοποιήσει ενδοσκοπική εξέταση στον οισοφάγο και στο ανώτερο τμήμα του στομάχου. Το 1932 οι Wolf και Schindler επινοούν ένα διοπτρικό σύστημα μεταφοράς της οπτικής δέσμης φωτός και επιτυγχάνουν να ενδοσκοπήσουν ολόκληρο τον στόμαχο.

Η επισκόπηση των οργάνων της κοιλιάς (λαπαροσκόπηση) άρχισε από τον Kelling (1902) στη Γερμανία σε πειραματόζωα, και από τον Jacoaeus (1912) σε ασθενείς με ασκίτη, ενώ ήδη ο Von Ott (1901) ανέφερε την πρώτη κοιλιοσκόπηση, δηλαδή την αντίστοιχη της σημερινής λαπαροσκόπησης. Σήμερα η διεθνής βιβλιογραφία αναγνωρίζει τους τρεις παρά πάνω ερευνητές ως πατέρες της λαπαροσκόπησης.

Η ανάπτυξη της υστεροσκόπησης, μεταξύ 1910 και 1945 ήταν ραγδαία με βελτίωση των ενδοσκοπίων, ανάπτυξη φωτιστικών πηγών, εργαλείων μικροχειρουργικής μέσω διάταξης της ενδομητριακής κοιλότητας και της κοιλιάς, καταγραφή των ενδοσκοπικών εικόνων και τέλος με μικροεπεμβάσεις.

Το 1957 ήταν έτος σταθμός. Ο Hirschowitz χρησιμοποίησε οπτικές ίνες. Από τότε η εξέλιξη ήταν ταχεία και εντυπωσιακή. Σήμερα υπάρχουν πολύ λεπτά ενδοσκόπια ($\delta = 9 \text{ mm}$) με ευρύ αυλό εργασίας και με ηλεκτρονική μεταφορά της ενδοσκοπικής εικόνας σε οθόνη [1].



Σχήμα 2.1 Η εξέλιξη της υστεροσκόπησης (S. Duplay, S. Clado, 1898).

Η εξέλιξη της ενδοσκόπησης το 1945 είναι πλέον αλματώδης (ανακάλυψη του ψυχρού φωτισμού, αυτόματα μηχανήματα παροχής CO₂ και υγρών μέσω διάταξης, χρήση των laser, ηλεκτροχειρουργικής, κάμερες, οθόνες, εκτυπωτές) (Σχήμα 2.2) [2].



Σχήμα 2.2 Μηχανήματα διαγνωστικής και εγχειρητικής λαπαροσκόπησης

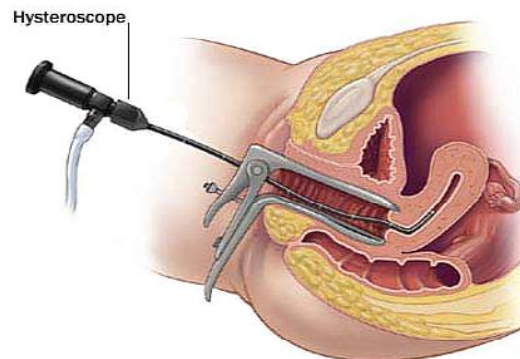
Παράλληλα στην υστεροσκόπηση αναπτύχθηκε η χειρουργική ενδοσκόπηση, με σκοπό να καταργηθούν οι πολλές κοιλιακές εγχειρήσεις, προς όφελος των ασθενών με ελάχιστη παραμονή τους στο νοσοκομείο, μείωση του κόστους νοσηλείας, άριστη μετεγχειρητική πορεία και επάνοδο στην εργασία τους σε 5-7 μέρες.

2.3.2 Υστεροσκόπηση

Η υστεροσκόπηση είναι η εξέταση του εσωτερικού του τραχήλου και της μήτρας, του ενδοτραχήλου και του ενδομητρίου, με τη χρήση μίας λεπτής κάμερας που εισάγεται διαμέσου του κόλπου και του τραχήλου. Με την υστεροσκόπηση η εξέταση γίνεται υπό άμεση όραση και μεγέθυνση με ακρίβεια και αξιοπιστία μεγαλύτερη από οποιαδήποτε άλλη μέθοδο (π.χ. υπερηχογράφημα, αξονική ή μαγνητική τομογραφία, απόξεση).

Το υστεροσκόπιο, που μοιάζει με ένα τηλεσκόπιο, εισάγεται από το άνοιγμα του τραχήλου μέσα στη κοιλότητα της μήτρας. Προκειμένου να απομακρυνθούν τα τοιχώματα της μήτρας

μεταξύ τους κι έτσι να επιτρέψουν να παρατηρήσει ο γυναικολόγος το εσωτερικό της κοιλότητας, εισάγεται στη μήτρα διοξείδιο του άνθρακα ή φυσιολογικός ορός ως διατατικό μέσο. [2]



Σχήμα 2.3 Υστεροσκοπική εξέταση

Με την υστεροσκόπηση μπορεί να παρατηρηθεί η μορφολογία (μέγεθος και σχήμα) της κοιλότητας της μήτρας, τα μητριαία στόμια των σαλπίγγων, η υφή και η ανάπτυξη του ενδομητρίου καθώς επίσης και ο τραχηλικός σωλήνας (ενδοτράχηλο).

Η κύρια ένδειξη της υστεροσκόπησης είναι η ανώμαλη αιμορραγία της μήτρας, όπου απαιτείται εκτίμηση της παθολογίας της ενδομήτριας κοιλότητας και αποκλεισμός κακοήθειας. Μη φυσιολογική αιμόρροια μπορεί να είναι δυσλειτουργικής αιτιολογίας ή να οφείλεται σε ύπαρξη πολυπόδων, ινομυωμάτων, ξένου σώματος(κυρίως ενδομητρίων σπειραμάτων), σε χορήγηση ορμονικής θεραπείας υποκατάστασης σε γυναίκες μετεμηνοπαυσιακής ηλικίας, σε χορήγηση ταμοξιφαίνης ή σε παρουσία καρκίνου του ενδομητρίου. Με την υστεροσκόπηση η διάγνωση όλων αυτών των καταστάσεων γίνεται ευχερώς και με ακρίβεια υπό άμεση όραση, ενώ είναι δυνατή η εστιακή βιοψία, εάν υπάρχει ένδειξη. Η υψηλή ευαισθησία της μεθόδου της καθιέρωσαν ως τον "χρυσό κανόνα" στην διάγνωση των ενδομητρίων παθήσεων και τείνει να αντικαταστήσει την κλασική απόξεση του ενδομητρίου.

Σε περιπτώσεις υπογονιμότητας και στο πλαίσιο του ελέγχου, υπερτερεί σαφώς της υστεροσαλπιγγογραφίας με άμεση διάγνωση των ενδομήτριων συμφύσεων, διαφραγμάτων, ινομυωμάτων, πολυπόδων αλλά και εκτίμηση της κατάστασης του ενδομητρίου.

2.4 Σύνοψη

Στο κεφάλαιο αυτό έχει γίνει αναφορά στον καρκίνο του ενδομητρίου, ο οποίος αποτελεί ένα από τα συχνότερα είδη γυναικολογικού καρκίνου. Απώτερος στόχος της παρούσας μελέτης είναι η υποστήριξη διάγνωσης των γιατρών, μέσω της επεξεργασίας και εξόρυξης δεδομένων σε διάφορες βάσεις δεδομένων με στοιχεία τα οποία προκύπτουν μετά από εφαρμογή αλγόριθμων χαρακτηριστικών υφής στις υστεροσκοπικές εικόνες αλλά και στις εικόνες που δημιουργούνται μετά από επεξεργασία και μετασχηματισμό του χρώματος των εικόνων αυτών. Αυτό θα είναι πολύ βοηθητικό για τον γιατρό ώστε να αυξήσει τη διαγνωστική του ακρίβεια σε δύσκολες περιπτώσεις γυναικολογικού καρκίνου. Επίσης θα επιτρέψει τη δημιουργία κοινής ιατρικής βάσης δεδομένων με ανεξάρτητα κέντρα αντιμετώπισης γυναικολογικού καρκίνου για την πληρέστερη κατανόηση του προβλήματος.

Στη βιβλιογραφία έχουν συμπεριληφθεί αρκετές αναφορές για τον καρκίνο του γυναικείου αναπαραγωγικού συστήματος καθώς και τρόποι αντιμετώπισης του από ιατρική σκοπιά. Η συγκεκριμένη μελέτη, η οποία όπως έχει αναφερθεί βασίζεται στη μελέτη της Αντωνίας Λοΐζου, μπορεί να γίνει η βάση για την αρχή της εφαρμογής νέων τεχνολογιών στην επεξεργασία εικόνας γυναικολογικού καρκίνου που θα αποτελέσει αφετηρία για άλλους ερευνητές για τη δημιουργία νέων εφαρμογών στον συγκεκριμένο τομέα.

Στόχος της έρευνας είναι να βοηθήσει το γιατρό στην αντιμετώπιση δύσκολων περιπτώσεων γυναικολογικού καρκίνου και να αυξήσει σημαντικά τη διαγνωστική ακρίβεια αναγνώρισης και διαχωρισμού περιπτώσεων γυναικολογικού καρκίνου αλλά και άλλου είδους καρκίνων στο ανθρώπινο σώμα.

Σε αυτό το κεφάλαιο πραγματοποιήθηκε παρουσίαση των βασικών αρχών της υστεροσκόπησης. Μέχρι και σήμερα δεν υπάρχουν ιατρικά πρωτόκολλα διεξαγωγής της υστεροσκοπικής εξέτασης. Δηλαδή, δεν υπάρχουν δημοσιευμένα η σειρά εφαρμογής του εκάστοτε γιατρού από την έναρξη της εξέτασης ή επέμβασης μέχρι την αναγνώριση ή τη θεραπεία του περιστατικού. Ιατρικά πρωτόκολλα έχουν δημοσιευθεί από την American Society for Gastrointestinal Endoscopy [3] και περιγράφουν βήμα προς

βήμα τη μέθοδο για γαστροενδοσκόπηση βοηθώντας έτσι τους γιατρούς στην καλύτερη παρατήρηση και θεραπεία των περιπτώσεων αυτών.

Κεφάλαιο 3

Χαρακτηριστικά Υφής

3.1	Επισκόπηση	15
3.2	Γενικές πληροφορίες για τα χαρακτηριστικά υφής	16
3.3	Αλγόριθμος Statistical Features (SF)	19
3.4	Αλγόριθμος Spatial Gray Level Dependence Matrices (SGLDM)	20
3.4.1	Συμπεράσματα από τον αλγόριθμο SGLDM	24
3.4.2	Μέτρα υπολογισμού χαρακτηριστικών υφής	25
3.5	Αλγόριθμος Gray Level Difference Statistics (GLDS)	28
3.5.1	Συμπεράσματα από τον αλγόριθμο GLDS	28

3.1 Επισκόπηση

Στο κεφάλαιο αυτό περιγράφονται οι αλγόριθμοι χαρακτηριστικών υφής, καθώς επίσης και η εφαρμογή τους στις υστεροσκοπικές εικόνες. Συγκεκριμένα περιγράφονται οι αλγόριθμοι Statistical Features (SF), Spatial Gray Level Dependence Matrices (SGLDM) και Gray Level Difference Statistics (GLDS).

Σε κάθε εικόνα περιέχονται πολλές πληροφορίες, οι οποίες δεν μπορούν να αναγνωριστούν από τους εμπειρογνώμονες με απλή παρατήρηση. Συνεπώς, οι έμπειροι επιστήμονες, με τη χρήση των χαρακτηριστικών υφής, χαρακτηρίζουν την εικόνα και εξαγάγουν χρήσιμες πληροφορίες για το περιεχόμενο της. Τις πληροφορίες αυτές τις χρησιμοποιούν έπειτα για έχουν πιο ακριβή αποτελέσματα στη διάγνυσή τους.

Οι αλγόριθμοι χαρακτηριστικών υφής μπορούν να οριστούν σαν μαθηματικές εξισώσεις οι

οποίες δέχονται σαν είσοδο τις τιμές του γκριζου που αποτελείται η εικόνα, καθώς επίσης και άλλες συνιστώσες, όπως για παράδειγμα τον αριθμό των στιγμάτων και τη μέση τιμή του γκρι.

Τα χαρακτηριστικά υφής τα εντοπίζουμε συχνά σε εικόνες γκρι τόνου όπως για παράδειγμα εικόνες υπερήχων, MRI, CT κ.α. Μερικά χαρακτηριστικά τους είναι η μέση τιμή, η διασπορά, η ομογένεια, η διακύμανση, ο μέσος όρος, η εντροπία κ.α.

Αν και τα χαρακτηριστικά υφής χρησιμοποιούνται ευρέως από πολλούς επιστήμονες, ωστόσο δεν υπάρχει μοναδικός και παγκοσμίως αποδεκτός ορισμός της υφής της εικόνας. Η υφή είναι ένα μέτρο της τραχύτητας, της ομαλότητας και της κανονικότητας της εικόνας.

Οι τεχνικές περιγραφής της υφής μπορούν να ομαδοποιηθούν σε τρεις μεγάλες κατηγορίες:

1. Στατιστικές
2. Φασματικές
3. Δομικές

Οι στατιστικές τεχνικές περιγραφής βασίζονται στα ιστογράμματα των περιοχών, τις επεκτάσεις τους και τις ροπές τους. Επίσης, μετρούν αντίθεση ύπαρξη κόκκων και τραχύτητα. Οι φασματικές τεχνικές βασίζονται στη συνάρτηση αυτοσυσχέτισης μιας περιοχής ή της κατανομής ισχύος στην περιοχή του μετασχηματισμού Fourier, με σκοπό να ανιχνεύσουν περιοδικότητα υφής. Τέλος, οι δομικές τεχνικές περιγράφουν την υφή με χρήση προτύπων συνοδευόμενων από συγκεκριμένους κανόνες τοποθέτησης [2].

Στη συνέχεια αναλύονται οι τρεις αλγόριθμοι υφής που εφαρμόσαμε για την έρευνά μας.

3.2 Γενικές πληροφορίες για τα χαρακτηριστικά υφής

Τα τελευταία χρόνια στο χώρο της ιατρικής απεικόνισης έγιναν μεγάλα βήματα στην αναπαράσταση τρισδιάστατων και δισδιάστατων δεδομένων. Αυτές οι απεικονιστικές μέθοδοι είναι για παράδειγμα η απεικόνιση με MRI, CT, US και X- rays. Οι εξέλιξη αυτή οδήγησε τους επιστήμονες στην υλοποίηση αλγορίθμων εξαγωγής

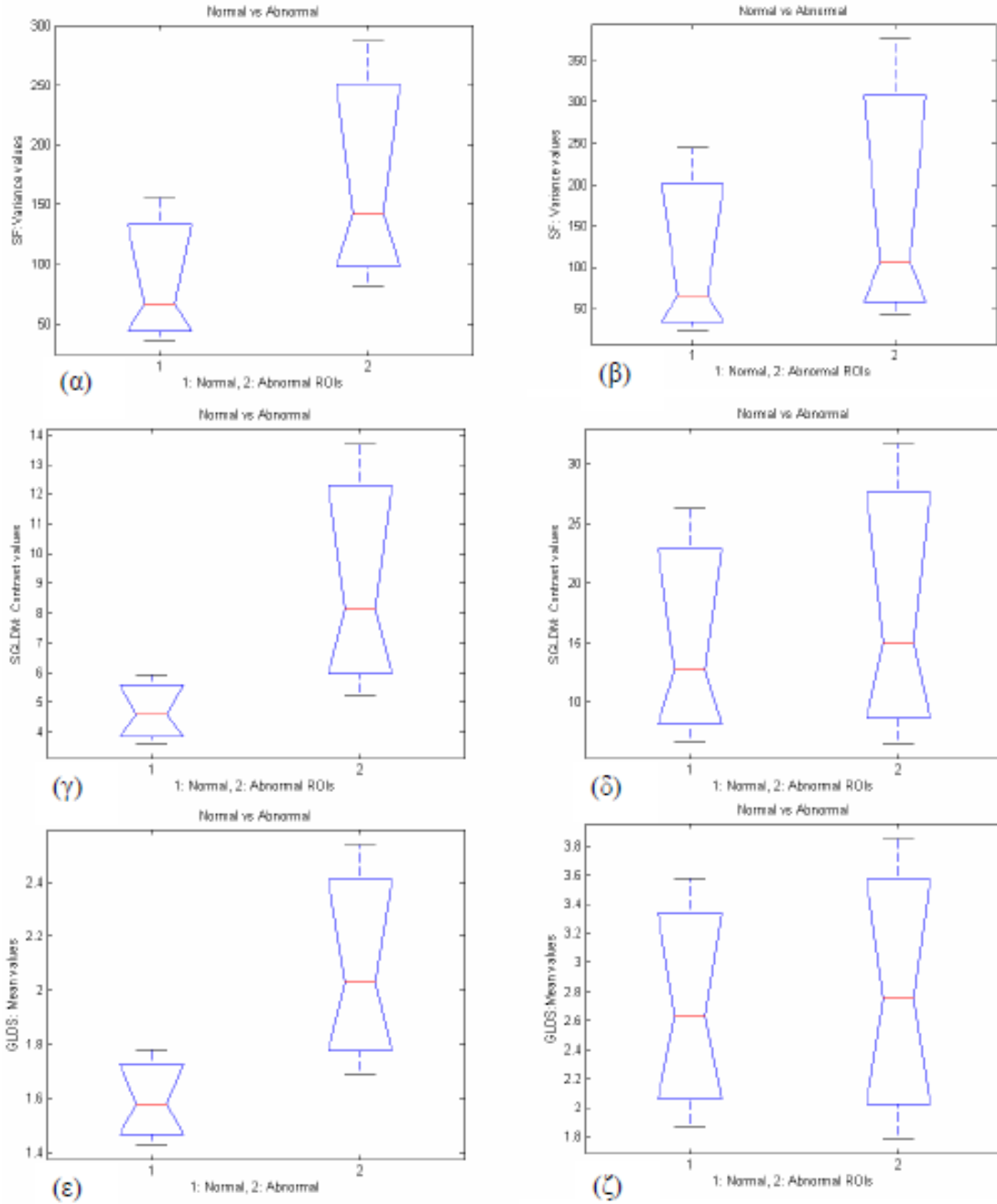
χαρακτηριστικών υφής για ιατρικές εφαρμογές [4]-[7].

Οι αλγόριθμοι εξαγωγής χαρακτηριστικών υφής έχουν χρησιμοποιηθεί ευρέως στον χαρακτηρισμό της ιατρικής εικόνας με πολύ καλά αποτελέσματα είτε σε τρισδιάστατο είτε σε δισδιάστατο μοντέλο απεικόνισης.

Στην διαγνωστική κολονοσκόπηση ο ειδικός αναγνωρίζει ύποπτες περιοχές ενδιαφέροντος βάσει της υφής του ιστού. Η τραχεία και η ομαλή επιφάνεια, το σχήμα, και η συμμετρικότητα είναι μερικά χαρακτηριστικά για αναγνώριση ύποπτων περιοχών ενδιαφέροντος. Γειτονικές επιφάνειες που παρουσιάζουν διαφορετικά χαρακτηριστικά μπορούν να διαχωρίσουν φυσιολογικές και παθολογικές περιοχές βάσει των χαρακτηριστικών υφής. Η εμπειρία του ειδικού είναι σημαντική στην αναγνώριση αυτών των περιοχών. Η χρήση των υπολογιστών με την εφαρμογή αλγορίθμων χαρακτηριστικών υφής βοηθά τον ειδικό στην αύξηση της αναγνώρισης ύποπτων περιοχών καρκίνου. Φυσικά πάντα η διάγνωση γίνεται με την ιστοπαθολογική εξέταση της εν λόγω περιοχής. Οι περιοχές ενδιαφέροντος είναι ξεχωριστές για κάθε περίπτωση όπως επίσης και τα χαρακτηριστικά υφής είναι μοναδικά για κάθε ιστό. Το γεγονός αυτό μας επιτρέπει να ξεχωρίσουμε τους ιστούς βάσει των χαρακτηριστικών υφής [8], [9].

Άλλες έρευνες που έχουν δημοσιευθεί και αφορούν ανάλυση εικόνων με αλγόριθμους χαρακτηριστικών υφής αφορούν την αναγνώριση ύποπτων περιοχών ενδιαφέροντος από διάφορα όργανα όπως καρκίνος στο συκώτι, καρκίνος του προστάτη [10], καρκίνος του μαστού [11], καρκίνου του εγκεφάλου [12], καρκίνος του τραχήλου της μήτρας [13] και καρδιακά νοσήματα [14]. Έχουν χρησιμοποιηθεί αλγόριθμοι εξαγωγής χαρακτηριστικών υφής με αλγόριθμους μμετασχηματισμού 2-D wavelet με πολύ καλά αποτελέσματα στον χαρακτηρισμό των εικόνων.

Εν κατακλείδι, τα χαρακτηριστικά υφής μπορούν να περιγράψουν με μεγάλη ακρίβεια την υφή της περιοχής ενδιαφέροντος και να δώσουν μια μοναδική ταυτότητα σε κάθε υπό εξέταση περιοχή. Με αυτά τα χαρακτηριστικά, η χρήση των υπολογιστών στην διαγνωστική χειρουργική, γίνεται υλοποιήσιμη και έξυπνα συστήματα επεξεργασίας εικόνων έχουν προταθεί για την αύξηση της διαγνωστικής ακρίβειας του περιστατικού.



Σχήμα 3.1 Γραφικές παραστάσεις των χαρακτηριστικών υφής εικόνων ενδομητρίου, πριν και μετά την εφαρμογή του αλγόριθμου γάμμα. Τα σχήματα (α), (γ) και (ε) παρουσιάζουν τα χαρακτηριστικά υφής, SF: variance, SGLDM: contrast και GLDS: mean πριν την εφαρμογή του αλγόριθμου γάμμα. Τα σχήματα (β), (δ) και (ζ) παρουσιάζουν τα ίδια χαρακτηριστικά υφής μετά την εφαρμογή του αλγόριθμου γάμμα. (Σημείωση: Τα σχήματα αντιπροσωπεύουν το median, lower και upper quartiles. Η κόκκινη γραμμή ενώνει τις πιο κοντινές “παρατηρήσεις” στο διάστημα 1.5 του inter-quartile range (IQR) των lower και upper quartiles.)

3.3 Αλγόριθμος Statistical Features (SF)

Ο αλγόριθμος Statistical Features (SF) στηρίζεται στην εξαγωγή παραμέτρων με βάση των γκρι αποχρώσεων του ιστογράμματος [11], [15]. Τα χαρακτηριστικά περιγράφουν την γκρι κατανομή χωρίς να εξαρτώνται από τη χωρική κατανομή τους. Αυτό έχει σαν αποτέλεσμα να μπορούν να περιγράφουν μόνο την echogensity των χαρακτηριστικών και την ολική μεταβολή των χαρακτηριστικών των περιοχών ενδιαφέροντος.

Τα συγκεκριμένα χαρακτηριστικά περιγράφονται πιο κάτω με μαθηματικούς τύπους:

- The Mean gray level (MEA)

$$MEA = \frac{1}{N} \sum_i \sum_j g(i,j)$$

Η πιο πάνω παράμετρος περιγράφει το μέσο όρο των γκρι επιπέδων στα σημεία (pixels) της περιοχής ενδιαφέροντος.

- Variance (VAR)

$$VAR = \frac{1}{N} \sum_i \sum_j (g(i,j) - MEA)^2$$

Η παράμετρος της απόκλισης περιγράφει το πλάτος του ιστογράμματος.

- Kurtosis (KUR)

$$KUR = \frac{1}{N} \left[\frac{\sum_i \sum_j (g(i,j) - \mu)^4}{\sigma^4} \right]$$

- Το Median (MED) είναι ο μέσος όρος των γκρι επιπέδων της περιοχής ενδιαφέροντος.

- Skewness (SKE)

$$SKE = \frac{[\sum_i \sum_j (g(i,j) - \mu)^2]}{N}$$

- Energy (ENER)

$$ENER = \sum_I p(i)^2$$

- Entropy (ENT)

$$ENT = -\sum_i p(i) \log(p(i))$$

Όπου $g(i,j)$ είναι η τιμή του pixel (i,j) , N είναι ο ολικός αριθμός των pixels στην περιοχή ενδιαφέροντος, μ και σ είναι η μέση τιμή και η standard deviation όλων των pixels, και p είναι η πιθανότητα κατανομής.

3.4 Αλγόριθμος Spatial Gray Level Dependence Matrices (SGLDM)

Ο αλγόριθμος Spatial Gray Level Dependence Matrices (SGLDM) [16] βασίζεται στον υπολογισμό της συνάρτησης δευτέρας τάξης δεσμευμένης πυκνότητας πιθανότητας $f(i,j;d,\theta)$. Κάθε στοιχείο $f(i,j;d,\theta)$ είναι η πιθανότητα, δύο σημείων που βρίσκονται σε διεύθυνση που καθορίζεται από τη γωνία θ και έχουν απόσταση d μεταξύ τους, να έχουν τιμές γκρι επιπέδων i και j .

Με την υπόθεση ότι έχουμε μια εικόνα προς ανάλυση με διαστάσεις $N_x \times N_y$, και ότι κάθε σημείο έχει μια ακέραια (διακριτή) τιμή γκρι επιπέδου που βρίσκεται μέσα στο διάστημα 1 μέχρι N_g (1,2,3,4,... N_g).

Εάν $L_x = \{1,2,..., N_x\}$ είναι η κλίμακα (πεδίο) του X άξονα,
 $L_y = \{1,2,..., N_y\}$ είναι η κλίμακα (πεδίο) του Y άξονα, G
 $= \{1,2,..., N_g\}$ τα επίπεδα γκρι της εικόνας.

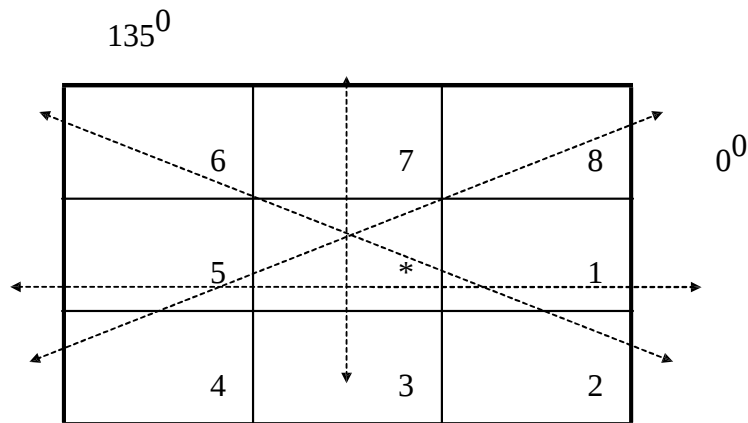
Τότε το σύνολο $L_x \times L_y$ είναι το σύνολο των σημείων (pixel) της εικόνας.

Γενικότερα, μια συνάρτηση μπορεί να παρουσιαστεί σαν μια συνάρτηση $I(x,y)$ η οποία παίρνει και μια διακριτή τιμή από το διάστημα $1..Ng$ για κάθε σημείο,

$$I: L_X \times L_Y \rightarrow G.$$

Οι υπολογισμοί των μέτρων $f(i,j;d,\theta)$ για μια εικόνα γίνονται για τέσσερις διαφορετικές γωνίες. Για κάθε d γίνεται ο υπολογισμός για γωνίες $\theta=0^0, 45^0, 90^0, 135^0$. Οι υπολογισμοί αυτοί έχουν σαν αποτέλεσμα τέσσερις πίνακες (Nearest-neighbour gray-tone spatial-dependence matrices). Οι πίνακες αυτοί βασίζονται στην αρχή του πλησιέστερου γείτονα σε ένα σημείο. Στο Σχήμα 3.1 φαίνονται ο πλησιέστεροι γείτονες (κοντινότερα σημεία) για ένα τυχαίο σημείο της εικόνας.

Φαίνονται οι πλησιέστεροι γείτονες για $d=1$ και $\theta=0^0, 45^0, 90^0, 135^0$. Το συγκεκριμένο δεν ισχύει για τα σημεία που βρίσκονται στα άκρα μιας εικόνας [2].



Σχήμα 3.1 (Κατευθύνσεις επαλήθευσης των σημείων για τις τέσσερις διαφορετικές γωνίες. Όπως φαίνεται για $\theta=0^0$ κοντινότεροι γείτονες είναι τα σημεία 1 και 5, για $\theta=45^0$ τα σημεία 4 και 8, για $\theta=90^0$ τα σημεία 7 και 3, για $\theta=135^0$ τα σημεία 2 και 6)

Αρχικά υπολογίζεται η συχνότητα με την οποία παρατηρείται αλλαγή από το επίπεδο i στο επίπεδο j και αποθηκεύεται σε ένα πίνακα $P(i,j;d,\theta)$.

Για τις τέσσερις διαφορετικές γωνίες η συχνότητα υπολογίζεται από πιο κάτω σχέσεις:

$$\begin{aligned}
 P(i,j,d,0^0) = & \# \{ ((k,l),(m,n)) \in (L_y \times L_x) \\
 & \times (L_y \times L_x) : k-m=0, |l-n|=d, \\
 & I(k,l)=i, I(m,n)=j \}
 \end{aligned} \tag{3.1}$$

$$\begin{aligned}
 P(i,j,d,45^0) = & \# \{ ((k,l),(m,n)) \in (L_y \times L_x) \\
 & \times (L_y \times L_x) : (k-m=d, l-n=-d) \\
 & \text{or } (k-m=-d, l-n=d), \\
 & I(k,l)=i, I(m,n)=j \}
 \end{aligned} \tag{3.2}$$

$$\begin{aligned}
 P(i,j,d,90^0) = & \# \{ ((k,l),(m,n)) \in (L_y \times L_x) \\
 & \times (L_y \times L_x) : |k-m|=d, l-n=0, \\
 & I(k,l)=i, I(m,n)=j \}
 \end{aligned} \tag{3.3}$$

$$\begin{aligned}
 P(i,j,d,135^0) = & \# \{ ((k,l),(m,n)) \in (L_y \times L_x) \\
 & \times (L_y \times L_x) : (k-m=d, l-n=d), \\
 & \text{or } (k-m=-d, l-n=-d), \\
 & I(k,l)=i, I(m,n)=j \}
 \end{aligned} \tag{3.4}$$

Το σύμβολο # υποδηλώνει τον αριθμό των στοιχείων στη θέση (i,j). Επομένως, για ένα σταθερό d έχουμε τέσσερις διαφορετικούς πίνακες. Στο Σχήμα 3.2 φαίνεται ο υπολογισμός του πίνακα P(i,j;1,0) για μια εικόνα 4x4.

(1,1)	(1,2)	(1,3)	(1,4)
(2,1)	(2,2)	(2,3)	(2,4)
(3,1)	(3,2)	(3,3)	(3,4)
(4,1)	(4,2)	(4,3)	(4,4)

$$L_y = \{1, 2, 3, 4\}$$

$$L_x = \{1, 2, 3, 4\}$$

$$d=1 \quad \theta=0^0$$

$$R_H = \{[(k,l),(m,n)] \in (L_x, L_y) \times (L_x, L_y) : k-m=0, |l-n|=1\}$$

$$= \{[(1,1),(1,2)], [(1,2),(1,1)], [(1,2),(1,3)], [(1,3),(1,2)],$$

$$[(1,3),(1,4)], [(1,4),(1,3)], [(2,1),(2,2)], [(2,2),(2,1)],$$

$$[(2,2),(2,3)], [(2,3),(2,2)], [(2,3),(2,4)], [(2,4),(2,3)],$$

$$[(3,1),(3,2)], [(3,2),(3,1)], [(3,2),(3,3)], [(3,3),(3,2)],$$

$$[(3,3),(3,4)], [(3,4),(3,3)], [(4,1),(4,2)], [(4,2),(4,1)],$$

$$[(4,2),(4,3)], [(4,3),(4,2)], [(4,3),(4,4)], [(4,4),(4,3)]\}$$

Σχήμα 3.2 Σύνολο όλων των αποστάσεων σε εικόνα μεγέθους 4x4 για $d=1$ και $\theta=0$

0^0	0	0	1	1	90^0
	0	0	1	1	
	0	2	2	2	
135^0	2	2	3	3	45^0

$$P_H = \begin{bmatrix} 4 & 2 & 1 & 0 \\ 2 & 4 & 0 & 0 \\ 1 & 0 & 6 & 1 \\ 0 & 0 & 1 & 2 \end{bmatrix} \quad P_V = \begin{bmatrix} 6 & 0 & 2 & 0 \\ 0 & 4 & 2 & 0 \\ 2 & 2 & 2 & 2 \\ 0 & 0 & 2 & 0 \end{bmatrix}$$

$$P_{LD} = \begin{bmatrix} 2 & 1 & 3 & 0 \\ 1 & 2 & 1 & 0 \\ 3 & 1 & 0 & 2 \\ 0 & 0 & 2 & 0 \end{bmatrix} \quad P_{RD} = \begin{bmatrix} 4 & 1 & 0 & 0 \\ 1 & 2 & 2 & 0 \\ 0 & 2 & 4 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Σχήμα 3.3 Πίνακες αλγορίθμου SGLDM για εικόνα 4x4 και απόσταση d=1

Στο Σχήμα 3.3 παρουσιάζονται οι τέσσερις πίνακες για d=1 σε μια εικόνα 4x4. Όπως διακρίνεται στο πιο πάνω σχήμα οι πίνακες είναι συμμετρικοί, ενώ ισχύει $P(i,j;d,a)=P(j,i;d,a)$. Οι πίνακες είναι μεγέθους $N_g \times N_g$ (όπου N_g είναι ο μέγιστος αριθμός γκρι επιπέδων).

Όταν υπολογιστεί ο πίνακας συχνοτήτων για μια συγκεκριμένη γωνία θ , γίνεται βαθμονόμηση του πίνακα, διαιρώντας κάθε στοιχείο με μια σταθερά R, όπου R είναι το άθροισμα των τιμών όλων των στοιχείων του πίνακα [2].

3.4.1 Συμπεράσματα από τον αλγόριθμο SGLDM

Μελετώντας τον αλγόριθμο GLDM μπορούμε να παρατηρήσουμε ότι σε περίπτωση που έχουμε κοκκώδη υφή εικόνας (μεγάλα Speckle Pattern) και η απόσταση d είναι μικρή σε σύγκριση με τους κόκκους της υφής, τότε τα σημεία με απόσταση d έχουν πολύ κοντινά επίπεδα γκρι. Από αυτό, μπορούμε να βγάλουμε συμπέρασμα ότι ο πίνακας $P(i,j;d,\theta)$ θα έχει μεγάλες τιμές όπου $i \cong j$, δηλαδή οι μεγάλες τιμές του πίνακα θα είναι συγκεντρωμένες γύρω από την κύρια διαγώνιο του πίνακα.

Αντίστοιχα για ομαλή υφή (μικρά Speckle Pattern) στα σημεία με απόσταση d θα έχουμε διαφορές στα γκρι επίπεδα, έτσι και οι τιμές του πίνακα θα είναι σχεδόν ίδιες για όλα τα (i,j) . Ανάλογα ισχύουν και για τις τιμές του κανονικοποιημένου πίνακα $p(i,j;d,\theta)$. Ένας καλός τρόπος ανάλυσης της υφής είναι ο υπολογισμός για διαφορετικές τιμές του d , κάποιων χαρακτηριστικών μέτρων γύρω από την κύρια διαγώνιο του πίνακα.

3.4.2 Μέτρα υπολογισμού χαρακτηριστικών υφής

Πιο κάτω παρουσιάζεται μια σειρά από μέτρα χαρακτηρισμού της υφής που μπορεί να μας δώσουν οι πίνακες του αλγόριθμου. [2]

Ορισμοί:

$p(i,j)$ (i,j) τιμή του πίνακα είναι η κανονικοποιημένη τιμή από τον αρχικό πίνακα συχνοτήτων $= P(i,j)/R$.

$p_x(i)$ i τιμή είναι το άθροισμα πιθανοτήτων της i -οστής γραμμής του πίνακα πιθανοτήτων $p(i,j) = \sum_{j=1}^{N_g} p(i,j)$.

N_g Αριθμός των διακριτών γκρι επιπέδων σε μια εικόνα.

$\sum_i$ και $\sum_j$ αντιστοιχούν στα $\sum_{i=1}^{N_g}$ και $\sum_{j=1}^{N_g}$

$$p_y(i) = \sum_{j=1}^{N_g} p(i,j)$$

$$p_{x+y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j), \quad k=2,3,\dots,2N_g$$

$$i+j=k$$

$$p_{x-y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j), \quad k=0,1,\dots,N_g-1$$

$$|i-j|=k$$

Μέτρα χαρακτηρισμού της υφής:

- *Angular Second Moment:*

$$f_1 = \sum_i \sum_j \{p(i, j)\}^2 \quad (3.5)$$

- *Contrast (Αντίθεση):*

$$f_2 = \sum_{i=1}^{N_g-1} n^2 \left\{ \sum_{i=1}^{N_g} \sum_{\substack{j=1 \\ |i-j|=k}}^{N_g} p(i, j) \right\} \quad (3.6)$$

- *Correlation (Συσχέτιση):*

$$f_3 = \frac{\sum_i \sum_j (ij) p(i, j) - \mu_x \mu_y}{\sigma_x \sigma_y} \quad (3.7)$$

Όπου μ_x, μ_y, σ_x και σ_y είναι οι μέσοι όροι και οι τυπικές αποκλίσεις των p_x, p_y αντίστοιχα.

- *Sum of squares: Variance (Άθροισμα τετραγώνων):*

$$f_4 = \sum_i \sum_j (i - \bar{i})^2 p(i, j) \quad (3.8)$$

- *Inverse Difference Moment:*

$$f_5 = \sum_i \sum_j \frac{1}{1 + (i - j)^2} p(i, j) \quad (3.9)$$

- *Sum Average:*

$$f_6 = \sum_{i=2}^{2N_g} i p_{x+y}(i) \quad (3.10)$$

- *Sum Variance:*

$$f_7 = \sum_{i=2}^{2N_g} (i - f_8)^2 p_{x+y}(i) \quad (3.11)$$

- *Sum Entropy* (άθροισμα εντροπίας):

$$f_8 = -\sum_{i=2}^{2N_g} p_{x+y}(i) \log\{p_{x+y}(i)\} \quad (3.12)$$

- *Entropy* (Εντροπία):

$$f_9 = -\sum_i \sum_j p(i, j) \log(p(i, j)) \quad (3.13)$$

- *Difference Variance*:

$$f_{10} = \text{variance of } p_{x-y} \quad (3.14)$$

- *Difference Entropy*:

$$f_{11} = -\sum_{i=0}^{N_g-1} p_{x-y}(i) \log\{p_{x-y}(i)\} \quad (3.15)$$

- *Information Measures of Correlation*:

$$f_{12} = \frac{HXY - HXY1}{\max\{HX, HY\}} \quad (3.16)$$

- $f_{13} = (1 - \exp[-2.0(HXY2 - HXY)])^{1/2} \quad (3.17)$

- $HXY = -\sum_i \sum_j p(i, j) \log(p(i, j)) \quad (3.18)$

Όπου HX και HY είναι οι εντροπίες των p_x, p_y και

- $HXY1 = -\sum_i \sum_j p(i, j) \log\{p_x(i) p_y(j)\} \quad (3.19)$

- $HXY2 = -\sum_i \sum_j p_x(i) p_y(j) \log\{p_x(i) p_y(j)\} \quad (3.20)$

- *Maximal Correlation Coefficient*:

$$f_{14} = (\text{δεύτερη μεγαλύτερη ζυγή τιμή του } Q)^{1/2} \quad (3.21)$$

όπου

$$Q = \sum_k \frac{p(i, k) p(j, k)}{p_x(i) p_y(k)} \quad (3.22)$$

3.5 Αλγόριθμος Gray Level Difference Statistics (GLDS)

Περιγραφή του GLDS [17] μπορεί να γίνει, υποθέτοντας ότι $I(x,y)$ είναι η συνάρτηση έντασης της εικόνας (όπου $I(x,y)$ το γκρι επίπεδο στο σημείο (x,y)). Για μια δεδομένη μετατόπιση $\delta \equiv (\Delta x, \Delta y)$ (όπου $\Delta x, \Delta y$ ακέραιοι αριθμοί) έχουμε την διαφορά:

$$I_\delta \equiv |I(x,y) - I(x+\Delta x, y+\Delta y)| \quad (3.23)$$

p_δ είναι η πυκνότητα πιθανότητας της $I_\delta(x,y)$. Εάν έχουμε m διακριτά επίπεδα του γκρι τότε η p_δ έχει τη μορφή διανύσματος m -διαστάσεων. όπου το σημείο i του διανύσματος έχει την τιμή της πιθανότητας ότι η τιμή της $I_\delta(x,y)$ θα είναι ίση με i .

Αρχικά υπολογίζεται η συχνότητα που έχουμε μια τιμή i στην I_δ , στη συνέχεια υπολογίζεται η πιθανότητα να συμβεί κάθε τιμή i , αυτό γίνεται διαιρώντας κάθε στοιχείο του αρχικού πίνακα με το άθροισμα των τιμών όλων των στοιχείων (βαθμονόμηση)[2].

3.5.1 Συμπεράσματα από τον αλγόριθμο GLDS

Όταν έχουμε κοκκώδη υφή και μικρό δ σε σχέση με το μέγεθος των κόκκων της υφής, τότε το ζεύγος των συγκρινόμενων σημείων στην I_δ πρέπει να έχει πολύ κοντινά επίπεδα του γκρι, έτσι το I_δ θα έχει πολύ μικρές τιμές. Οι τιμές του πίνακα (διάνυσμα) πιθανότητας P_δ θα είναι συγκεντρωμένες γύρω από το μηδέν ($i=0$). Ανάλογα για ομαλή υφή οι τιμές στην P_δ πρέπει να είναι διασκορπισμένες σε όλο τον πίνακα.

Μια καλή τεχνική για την ανάλυση της υφής μιας εικόνας, είναι να υπολογίσουμε για διαφορετικές τιμές του δ , κάποιο μέτρο που να δείχνει κατά πόσο οι τιμές στην P_δ είναι διασκορπισμένες σε όλο τον πίνακα ή συγκεντρωμένες γύρω από το μηδέν.

Τέσσερα τέτοια μέτρα χαρακτηρισμού είναι τα ακόλουθα [2]:

- *Contrast* (Αντίθεση):

$$CON \equiv \sum i^2 p_{\delta}(i) \quad (3.24)$$

- *Angular Second Moment*:

$$ASM \equiv \sum p_{\delta}(i)^2 \quad (3.25)$$

- *Entropy* (Εντροπία):

$$ENT \equiv -\sum p_{\delta}(i) \log(p_{\delta}(i)) \quad (3.26)$$

- *Mean* (Μέσος):

$$MEAN \equiv (1/m) \sum i p_{\delta}(i) \quad (3.27)$$

Κεφάλαιο 4

Αλγόριθμος C4.5

4.1 Επισκόπηση	30
4.2 Γενικές Πληροφορίες Αλγορίθμου C4.5	30
4.3 Κατασκευή δέντρων απόφασης	32
4.4 Επιλογή χαρακτηριστικού ελέγχου	33

4.1 Επισκόπηση

Στο κεφάλαιο αυτό περιγράφονται ο αλγόριθμος C4.5. Συγκεκριμένα, δίνονται γενικές πληροφορίες για τους παραπάνω αλγόριθμο αλλά και πληροφορίες σχετικά με τον τρόπο που λειτουργεί.

4.2 Γενικές Πληροφορίες Αλγορίθμου C4.5

Το πρόγραμμα του αλγορίθμου C4.5 [18] παράγει έναν κατηγοριοποιητή (classifier) σε μορφή δέντρου απόφασης. Δέντρο απόφασης ορίζεται ένα δέντρο δομής δεδομένων με τις ακόλουθες ιδιότητες (properties):

- Ένα κόμβο φύλλου ο οποίος σηματοδοτείται (labeled) με το όνομα μιας κατηγορίας / κλάσης.
- Ένα κόμβο ελέγχου που έχει δύο ή περισσότερα αποτελέσματα/εξόδου (outcomes) που συνδέονται με ένα υποδέντρο (subtree).
 - ο Η ρίζα και κάθε εσωτερικός κόμβος απόφασης σηματοδοτούνται με το όνομα ενός χαρακτηριστικού (attribute).
 - ο Κάθε εσωτερικός κόμβος έχει ένα σύνολο τουλάχιστον δύο υποκόμβων,

όπου τα κλαδιά στους κόμβους αυτούς χωρίζουν τις τιμές ή τα σύνολα τιμών των χαρακτηριστικών (attributes) εκείνου του κόμβου έτσι ώστε η ένωση των τιμών αυτών να αποτελεί το σύνολο όλων των πιθανών τιμών του κάθε χαρακτηριστικού (attribute).

Ένα δέντρο απόφασης μπορεί να χρησιμοποιηθεί για να κατηγοριοποιήσει μια περίπτωση (case) ξεκινώντας από τη ρίζα του δέντρου και προχωρώντας προς τα κάτω (στους κόμβους) έως ότου φθάσει σε ένα φύλλο. Σε κάθε κόμβο απόφασης, καθορίζεται το αποτέλεσμα/ έξοδος της περίπτωσης (case) και όταν αυτή η διαδικασία τελικά καταλήξει σε ένα φύλλο, η κατηγορία της περίπτωσης (case) καθορίζεται αυτή με την οποία το φύλλο είναι σηματοδοτημένο.

Ο C4.5 αλγόριθμος δέντρων απόφασης:

1. Επιλέγει ως ρίζα του δέντρου το χαρακτηριστικό (attribute) του οποίου οι τιμές χωρίζουν καλύτερα το σύνολο εκπαίδευσης (training set) σε υποσύνολα (ο καλύτερος διαχωρισμός είναι αυτός όπου όλα τα στοιχεία σε κάθε υποσύνολο ανήκουν στην ίδια κατηγορία/ κλάση).
2. Επαναληψη του σημείου 1 για κάθε υποκόμβο έως ότου ικανοποιείται το κριτήριο για τερματισμό (stopping criterio) της διαδικασίας.

Το κριτήριο αυτό για τον C4.5 είναι ο ελάχιστος αριθμός στοιχείων που πρέπει να βρίσκονται σε ένα σύνολο τα οποία δεν μπορούν να διαχωριστούν περαιτέρω.

Ο C4.5 αλγόριθμος μπορεί να εκφράσει το πρότυπο κατηγοριοποίησης ως κανόνες κατηγοριοποίησης που προκαλούνται από το ήδη χτισμένο δέντρο απόφασης. Το C4.5 πρόγραμμα χρησιμοποιεί μια απλουστευμένη μορφή κανόνα $L - R$, όπου το L είναι ένας ικανοποιητικός όρος (που διαμορφώνεται από την ένωση των όρων ενός attribute-value) για την κατηγοριοποίηση των αντικειμένων που ανήκουν στην κατηγορία R . Επιπλέον μία από τις κατηγορίες ορίζεται ως η προεπιλογή (default). Για την κατηγοριοποίηση μιας περίπτωσης (case) χρησιμοποιώντας ένα πρότυπο παραγωγής κανόνα, εξετάζεται η διατεταγμένη λίστα κανόνων για να βρεθεί ο πρώτος του οποίου η αριστερή πλευρά ικανοποιείται από την περίπτωση (case). Εάν

δεν υπάρχει κάποιος κανόνας του οποίου ικανοποιείται η αριστερή πλευρά, η περίπτωση (case) προβλέπεται να ανήκει στην κατηγορία προεπιλογής. Το πρόγραμμα προβλέπει επίσης την ορθότητα της κατηγοριοποίησης των άγνωστων περιπτώσεων (unseen cases) που ικανοποιούν την αριστερή πλευρά του κανόνα.

Ουσιαστικά ο C4.5 αλγόριθμος παρέχει αποδοτικές διαδικασίες για την μετατροπή ενός δέντρου απόφασης σε ένα σύνολο κανόνων κατηγοριοποίησης. Απλά heuristics, τα οποία γενικεύουν και συνδυάζουν τους κανόνες που παράγονται από τα δέντρα απόφασης, ενεργούν ως υποκατάστατο του κλαδέματος για τα μεταβλητά (univariate) δέντρα του C4.5.

4.3 Κατασκευή δέντρων απόφασης [2]

Για την κατασκευή δέντρων απόφασης, χρησιμοποιείται πολύ συχνά η μέθοδος του “διαίρει και βασίλευε”. Η μέθοδος αυτή επιτυγχάνει την κατασκευή ενός δέντρου απόφασης από ένα σύνολο περιπτώσεων εκπαίδευσης (training cases).

Έστω ότι το S είναι ένα σύνολο περιπτώσεων εκπαίδευσης (training cases) και $\{C_1, C_2, \dots, C_k\}$ οι διάφορες κατηγορίες. Υπάρχουν λοιπόν τρεις δυνατότητες:

- Το S να περιέχει μια ή περισσότερες περιπτώσεις (cases), όπου όλες ανήκουν σε μια ενιαία κατηγορία/ κλάση C_j . Τότε το δέντρο απόφασης για το S είναι ένα φύλλο που σηματοδοτείται από την κατηγορία/ κλάση C_j .
- Το S να μην περιέχει καμία περίπτωση (case):

Το δέντρο απόφασης είναι πάλι ένα φύλλο, αλλά η κατηγορία/ κλάση που συνδέεται με το φύλλο πρέπει να καθοριστεί από άλλες πληροφορίες εκτός από το S . Ο C4.5 αλγόριθμος χρησιμοποιεί τη συχνότερη κατηγορία/ κλάση στο γονέα αυτού του κόμβου για την σηματοδότηση του φύλλου.

- Το S περιέχει περιπτώσεις (cases) που ανήκουν σε ένα μείγμα κατηγοριών/κλάσεων:

Εδώ η ιδέα είναι να χωριστεί το S σε υποσύνολα από περιπτώσεις που να αποτελούν συλλογή μιας ενιαίας-κατηγορίας περιπτώσεων. Επιλέγεται ένα χαρακτηριστικό ελέγχου το $testT10$, που βασίζεται σε ένα ενιαίο χαρακτηριστικό (attribute), το οποίο έχει ένα ή περισσότερα αποτελέσματα/ εξόδους $\{O1, O2..., On\}$. Το S χωρίζεται στα υποσύνολα $S1, S2..., Sn$, όπου το Si περιέχει όλες τις περιπτώσεις στο S που έχουν το αποτέλεσμα Oi του επιλεγμένου χαρακτηριστικού ελέγχου. Το δέντρο απόφασης για το S αποτελείται από έναν κόμβο απόφασης που προσδιορίζει το χαρακτηριστικό ελέγχου, και ένα κλαδί για κάθε πιθανό αποτέλεσμα. Η διαδικασία αυτή της δημιουργίας του δέντρου εφαρμόζεται κατ' επανάληψη σε κάθε υποσύνολο των περιπτώσεων εκπαίδευσης, έτσι ώστε το i -οστό (i th) κλαδί να οδηγεί στο δέντρο απόφασης που κατασκευάζεται από το Si υποσύνολο του συνόλου S των περιπτώσεων εκπαίδευσης.

Η επιτυχής διαίρεση του συνόλου S των περιπτώσεων εκπαίδευσης προχωρά έως ότου όλα τα υποσύνολα να αποτελούνται από περιπτώσεις (cases) οι οποίες ανήκουν σε μια μόνο κατηγορία/ κλάση. Αυτή η μέθοδος κατασκευής δέντρων απόφασης είναι ένας άπληστος αλγόριθμος. Μόλις επιλεγεί το χαρακτηριστικό ελέγχου με βάση του οποίου θα χωριστεί το τρέχον σύνολο περιπτώσεων εκπαίδευσης, η μέθοδος διαίρεσης σταματά και οι εναλλακτικές επιλογές χαρακτηριστικών ελέγχου δεν εξερευνούνται. Έτσι οι μόνες πληροφορίες διαθέσιμες είναι η διανομή των κατηγοριών/ κλάσεων στο S και τα υποσύνολά τους. [2]

4.4 Επιλογή χαρακτηριστικού ελέγχου [2]

Η επιλογή ενός χαρακτηριστικού (attribute) ως χαρακτηριστικό ελέγχου αποτελείται από 3 βήματα τα οποία παρουσιάζονται πιο κάτω.

1. Το πρώτο βήμα αφορά την επιλογή των πιθανών χαρακτηριστικών (attribute) ως χαρακτηριστικών ελέγχου. Συγκεκριμένα, ο αλγόριθμος C4.5 περιέχει μηχανισμούς για την πρόταση των εξής τριών τύπων (χαρακτηριστικών) ελέγχου:
 - Τον "τυποποιημένο" έλεγχο βασισμένο σε ένα χαρακτηριστικό (attribute)

με διακριτές (discrete) τιμές, με ένα αποτέλεσμα/ έξοδο και ένα κλαδί για κάθε πιθανή τιμή του συγκεκριμένου χαρακτηριστικού (attribute).

- Ένα πιο σύνθετο έλεγχο, βασισμένο σε ένα χαρακτηριστικό (attribute) με διακριτές (discrete) τιμές, στον οποίο οι πιθανές τιμές διατίθενται σε έναν μεταβλητό αριθμό ομάδων με ένα αποτέλεσμα/ έξοδο για κάθε ομάδα παρά για κάθε τιμή.
- Στην περίπτωση που ένα χαρακτηριστικό A έχει συνεχείς τιμές, ο C4.5 προτείνει ένα binary χαρακτηριστικό ελέγχου, με αποτελέσματα/ εξόδους $A \leq Z$ και $A > Z$, βασισμένο στην σύγκριση της τιμής του χαρακτηριστικού A με την τιμή ενός κατώτατου ορίου Z . Οι πιθανές τιμές του Z μπορούν να βρεθούν με την ταξινόμηση των μοναδικών τιμών του χαρακτηριστικού A που βρίσκονται στο σύνολο S και ακολούθως αναγνωρίζοντας ένα κατώτερο όριο ανάμεσα σε κάθε ζεύγος παρακείμενων (adjacent) τιμών. Έτσι αν οι περιπτώσεις (cases) στο S έχουν d μοναδικές τιμές για το A , τότε θα έχουμε υπόψη μας $d-1$ κατώτατα όρια.

Έλεγχος βασισμένος σε συνεχείς χαρακτηριστικά (attributes): Οι περιπτώσεις εκπαίδευσης (training cases) T αρχικά ταξινομούνται στις τιμές του χαρακτηριστικού A που λαμβάνουμε υπόψη. Υπάρχει ένας πεπερασμένος αριθμός από τέτοιες τιμές έστω $\{v_1, v_2, \dots, v_n\}$. Έστω ότι εξετάζεται ένα συνεχές χαρακτηριστικό (attribute) A_1 που παίρνει τις μοναδικές N τιμές (π.χ. για $A_1=3, A_1=5, A_1=7, N=3$). Από την στιγμή που οι τιμές του χαρακτηριστικού ταξινομούνται υπάρχουν $N-1$ πιθανά κατώτατα όρια στο $t=(v_i + v_{i+1})/2$, όπου v_i είναι η τιμή του χαρακτηριστικού A και $v_i < v_{j+1} < v_j$ (επειδή οι τιμές είναι ταξινομημένες). Εάν κάποιο χαρακτηριστικό επιτρέπει μόνο δυαδικές διασπάσεις τότε κάθε κατώτατο όριο παρέχει μοναδικά υποσύνολα K_1 και K_2 στον κόμβο K . Ο C4.5 αλγόριθμος επιλέγει ως κατώτατο όριο τη μεγαλύτερη τιμή του A σε ολόκληρο το σύνολο εκπαίδευσης S η οποία δεν υπερβαίνει τον μέσο όρο των τιμών, παρά τον μέσο όρο τον ίδιο εξασφαλίζοντας έτσι όλες οι τιμές κατώτατων ορίων που εμφανίζονται στα δέντρα ή/και στους κανόνες κατηγοριοποίησης είναι τιμές που πραγματικά υπάρχουν στα δεδομένα.

2. Ανάθεση αποτελεσμάτων στα χαρακτηριστικά (attributes) που επιλέγονται σύμφωνα με μια μετρική ποιότητας:

Όλα τα χαρακτηριστικά ελέγχου αξιολογούνται με βάση την μετρική αυτή η οποία προκύπτει από το χωρισμό των περιπτώσεων εκπαίδευσης που παράγουν τα χαρακτηριστικά αυτά. Αλλά για οποιοδήποτε χωρισμό του συνόλου S τουλάχιστον δύο από τα υποσύνολα πρέπει να περιέχουν τον ελάχιστο αριθμό των δύο περιπτώσεων (training cases) αποφεύγοντας έτσι τις near-trivial διασπάσεις (splits).

- *Information Gain Metric:* Οι πληροφορίες που μεταβιβάζονται από ένα μήνυμα εξαρτώνται από την πιθανότητά τους και μπορούν να μετρηθούν σε bits σαν μείον το λογάριθμο στη βάση 2 εκείνης της πιθανότητας. Έτσι, παραδείγματος χάριν, εάν υπάρχουν οκτώ εξίσου πιθανά μηνύματα τότε οι πληροφορίες που μεταβιβάζονται από οποιοδήποτε από αυτά είναι $-\log_2(8)$ ή 3 bits. Έστω ένα τυχαίο παράδειγμα από το σύνολο S των περιπτώσεων εκπαίδευσης (training cases) που ανήκει στην κατηγορία C_i όπου $i = 1, \dots, k$. Το μήνυμα αυτό θα έχει πιθανότητα

$$\frac{|C_i|}{|S|} \quad (4.1)$$

Όπου S είναι ο αριθμός των περιπτώσεων που ανήκουν στην κατηγορία C_i . Οι πληροφορίες που περιλαμβάνονται σε αυτό το μήνυμα είναι [19]:

$$-\log_2 \frac{|C_i|}{|S|} \text{ bits} \quad (4.2)$$

Γενικά, εάν μας δίνεται μια κατανομή πιθανότητας (probability distribution) [19]:

$$P = (p_1 = \frac{|C_1|}{|S|}, p_2 = \frac{|C_2|}{|S|}, \dots, p_k = \frac{|C_k|}{|S|}) \quad (4.3)$$

το Information που μεταβιβάζεται από αυτήν την κατανομή (γνωστή ως εντροπία (Entropy) του P), είναι [19]:

$$I(P) = -\sum_{i=1}^k p_i * \log(p_i) \quad (4.4)$$

Όταν εφαρμόζεται στο σύνολο S των περιπτώσεων εκπαίδευσης (training cases), το I(P) μετρά το μέσο ποσό πληροφοριών (bits) που απαιτείται για να προσδιορίσει την κατηγορία/ κλάση ενός παραδείγματος του S (είναι μια σταθμισμένη (weighted) πρόσθεση k διαφορετικών εξισώσεων).

Έστω ότι λαμβάνουμε υπόψη μια παρόμοια μέτρηση μετά που το S έχει χωριστεί σύμφωνα με τα n αποτελέσματα/ εξόδους ενός χαρακτηριστικού ελέγχου X. Η αναμενόμενη πληροφορία μπορεί να βρεθεί ως μια σταθμισμένη (weighted) πρόσθεση των υποσυνόλων [19]:

$$Info(X, S) = \sum_i^n \frac{|S_i|}{|S|} Info(S_i) \quad (4.5)$$

όπου S1, S2..., Sm είναι ο διαχωρισμός του S που προκαλείται από την τιμή του X.

Η ποσότητα Gain(X,S) υπολογίζει την πληροφορία που κερδίζεται διαχωρίζοντας το S με βάση το χαρακτηριστικό ελέγχου X και ορίζεται ως [19]:

$$Gain(X, S) = Info(S) - Info(X, S) \quad (4.6)$$

Η έννοια Gain (εξίσωση 4.6) ευνοεί τα χαρακτηριστικά (features) που έχουν έναν μεγάλο αριθμό αποτελεσμάτων/ εξόδων (όταν το n είναι μεγάλο - στην εξίσωση 4.5). Μπορούμε να το δούμε αυτό αν ασχοληθούμε με το θέμα μίας ιατρικής διάγνωσης (κάτι με το οποίο θα ασχοληθούμε στην μελέτη μας) στην οποία ένα από τα χαρακτηριστικά των δεδομένων, που χρησιμοποιούμε για την διάγνωση αυτή, περιλαμβάνει την αναγνώριση ενός ασθενή. Από την στιγμή που μια τέτοια αναγνώριση είναι μοναδική (μία για κάθε ασθενή) ο διαχωρισμός οποιουδήποτε S συνόλου περιπτώσεων εκπαίδευσης (training cases) με βάση τις τιμές αυτού του χαρακτηριστικού θα μας οδηγήσει σε ένα μεγάλο αριθμό υποσυνόλων που όλα θα περιέχουν μία μόνο περίπτωση εκπαίδευσης X (αυτήν του κάθε ασθενή ξεχωριστά). Από την στιγμή που αυτά τα υποσύνολα μίας περίπτωσης θα έχουν $\text{info}(X,S)=0$, το information gain $\text{Gain}(X,S)$ θα είναι μέγιστο. Από την σκοπιά της πρόβλεψης ένα τέτοιος διαχωρισμός θα ήταν αχρείαστος.

Γι' αυτό χρησιμοποιούμε αντί το Gain την ακόλουθη αναλογία [19]:

$$\text{GainRatio}(X, S) = \frac{\text{Gain}(X, S)}{\text{SplitInfo}(S, T)} \quad (4.7)$$

Εξετάζουμε λοιπόν το περιεχόμενο των πληροφοριών ενός μηνύματος που δείχνει όχι την κατηγορία στην οποία ανήκει η περίπτωση αλλά το αποτέλεσμα/έξοδο του χαρακτηριστικού ελέγχου αυτής στο χαρακτηριστικό (feature) X . όπως θα παρατηρήσουμε, το $\text{GainRatio}(X,S)$ λοιπόν είναι το ποσοστό των πληροφοριών που παράγεται από μία διάσπαση η οποία είναι χρήσιμη για την κατηγοριοποίηση.

Σε αναλογία με την εξίσωση (4.4), έχουμε

[27]:

$$\text{SplitInfo}(X, S) = - \sum_i^n \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|} \quad (4.8)$$

3. Επιλογής του χαρακτηριστικού (attribute) με το υψηλότερο αποτέλεσμα:

Μπορούμε να χρησιμοποιήσουμε την έννοια GainRatio για να ταξινομήσουμε τα χαρακτηριστικά (attributes) και να χτίσουμε τα δέντρα απόφασης όπου σε κάθε κόμβο θα βρίσκονται τα χαρακτηριστικά (attributes) με μέγιστο GainRatio μεταξύ των χαρακτηριστικών (attributes) τα οποία δεν έχουν ληφθεί ακόμα υπόψη στην πορεία από την ρίζα προς τον συγκεκριμένο κόμβο.

Επίσης μπορούμε να εξετάσουμε την περίπτωση των χαρακτηριστικών (features) με συνεχείς τιμές. Έστω ότι το χαρακτηριστικό (feature) X έχει μια συνεχή σειρά. Εξετάζουμε τις τιμές για αυτό το χαρακτηριστικό (feature) στο σύνολο εκπαίδευσης (training set). Έστω ότι είναι, στην ταξινομημένη κατά αύξουσα σειρά $A_1, A_2, \dots, A_m$. Κατόπιν για κάθε τιμή A_j όπου $j = 1 \dots m$, χωρίζουμε τα δεδομένα σε 2 σύνολα: το πρώτο σύνολο έχει τις τιμές του X μέχρι και το A_j και το δεύτερο σύνολο έχει τις τιμές του X που είναι μεγαλύτερες από το A_j . Για κάθε ένα από αυτά τα m υποσύνολα υπολογίζουμε το $\text{GainRatio}(X(j), S)$ $j = 1, 2, \dots, m$, και επιλέγουμε το υποσύνολο το οποίο μεγιστοποιεί το gain. Εάν όλα τα χαρακτηριστικά (features) είναι συνεχή, θα λάβουμε ένα δυαδικό δέντρο.

Το κριτήριο gain ratio, που χρησιμοποιείται στον αλγόριθμο C4.5, είναι robust και μας δίνει μια καλύτερη επιλογή ιδιότητας ελέγχου από το κριτήριο gain. [2]

Κεφάλαιο 5

Μεθοδολογία

5.1 Επισκόπηση	39
5.2 Καταγραφή του ενδοσκοπικού βίντεο/ εικόνων	40
5.3 Χειροκίνητη κατάτμηση περιοχών ενδιαφέροντος	41
5.4 Προέλευση δείγματος εικόνων	43
5.5 Εφαρμογή αλγορίθμων χαρακτηριστικών υφής	44
5.6 Αλγόριθμοι κατηγοριοποίησης / εξόρυξης δεδομένων	45
5.7 Μετρικές αποτελεσματικότητας της διάγνωσης από το σύστημα κατηγοριοποίησης	47
5.8 Μετρικές αποτελεσμάτων από το σύστημα Weka	49

5.1 Επισκόπηση

Στο κεφάλαιο αυτό περιγράφονται η μεθοδολογία που χρησιμοποιήθηκε για την επεξεργασία των υστεροσκοπικών εικόνων.

Αρχικά, γίνεται καταγραφή του ενδοσκοπικού βίντεο/ εικόνων, ενώ στη συνέχεια της ενότητας αυτής καλύπτονται οι λειτουργίες της χειροκίνητης κατάτμησης περιοχών ενδιαφέροντος. Επίσης, γίνεται αναφορά για την προέλευση των δειγμάτων εικόνων καθώς και η εφαρμογή των αλγορίθμων χαρακτηριστικών υφής

Τέλος θα αναλύσουμε την μεθοδολογία που ακολουθήσαμε για την εξόρυξη δεδομένων/ εξαγωγής αποτελεσμάτων κατηγοριοποίησης από τις βάσεις δεδομένων (στις οποίες αναφερθήκαμε πιο πάνω) που χρησιμοποιήσαμε για την μελέτη αυτή.

Επιπλέον θα αναφέρουμε τις μετρικές αξιολόγησης των αποτελεσμάτων της εφαρμογής του αλγορίθμου C4.5 στις δεκατρείς βάσεις δεδομένων με τα χαρακτηριστικά υφής των εικόνων μας και αντίστοιχα στα δέκα (τυχαία) μοντέλα ανά βάση.

5.2 Καταγραφή του ενδοσκοπικού βίντεο/ εικόνων

Για την καταγραφή του βίντεο χρησιμοποιήσαμε το ιατρικό τηλεσκόπιο από την εταιρεία Wolf [20]. Τα χαρακτηριστικά του τηλεσκοπίου είναι: 2,8 mm διάμετρος και 30^0 γωνία παρακολούθησης. Το ενδοσκοπικό βίντεο καταγράφηκε με τη χρήση της ιατρικής κάμερας Circon IP4.1. Όλα τα βίντεο είχαν τις ίδιες συνθήκες λήψης, δηλαδή, η εστίαση και η φωτεινότητα ήταν στο βέλτιστο σημείο βάσει της γνώσης και εμπειρίας του γιατρού. Ο κατασκευαστής της ιατρικής κάμερας προτείνει την αρχικοποίηση της κάμερας με την εφαρμογή της μεθόδου white balanced χρησιμοποιώντας μια λευκή επιφάνεια για την βαθμονόμηση της κάμερας πριν από κάθε επέμβαση. Εμείς έχουμε χρησιμοποιήσει το λευκό χρώμα της παλέτας χρωμάτων της Edmund Industrial Optics Company. Η πηγή φωτός είναι της εταιρείας ACMI Corporation και είναι 300 Watt Xenon Light Source.

Το αναλογικό σήμα εξόδου της κάμερας με ανάλυση 475 οριζόντιες γραμμές μετατρέπεται σε ψηφιακό σήμα με ανάλυση τα 720x576 pixels χρησιμοποιώντας 24 bits για το χρώμα και 25 frames το δευτερόλεπτο με ανάλυση εικόνας για την πανοραμική λήψη τα 15 pixels/mm περίπου και για την κοντινή λήψη ανάλυσης εικόνας έχουμε 21 pixels/mm περίπου. Το βίντεο αυτό αποθηκεύεται με μορφή AVI στον υπολογιστή μας. Η ψηφιοποίηση του βίντεο γίνεται με τον κωδικοποιητή Digital Video Creator 120 frame grabber ο οποίος είναι συνδεδεμένος με την πύλη εισόδου IEEE 1394 του υπολογιστή. Οι συνθήκες σκόπευσης και λήψης των εικόνων ελέγχονται από τον γιατρό και προσομοιώνουν τις πραγματικές συνθήκες ενδοσκοπικής επέμβασης [1].

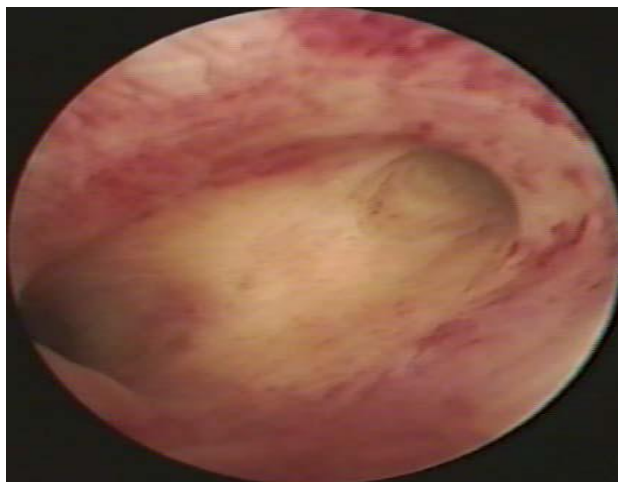


Σχήμα 5.1 Ολοκληρωμένος σταθμός ενδοσκοπικού συστήματος. Περιλαμβάνει: ιατρική κάμερα, οθόνη, κεφάλι της κάμερας, μονάδα παραγωγής φωτός, μονάδα ελέγχου φυσιολογικού ορού.

5.3 Χειροκίνητη κατάτμηση περιοχών ενδιαφέροντος

Κατά τη διάρκεια μιας υστεροσκοπικής επέμβασης, ο γιατρός παρακολουθώντας το βίντεο σε πραγματικό χρόνο από την οθόνη του, χρειάζεται να παγώσει την αντίστοιχη εικόνα για να παρατηρήσει με μεγαλύτερη λεπτομέρεια όλες τις ύποπτες περιοχές ενδιαφέροντος. Η ιατρική κάμερα είναι συνδεδεμένη με ένα υπολογιστή όπου μέσω μιας κάρτας μετατροπής αναλογικού σε ψηφιακού σήματος (frame grabber), το βίντεο ψηφιοποιείται και αποθηκεύεται.

Οι επιλεγμένες εικόνες (ανάλυση εικόνας 720X576X24bits) αποθηκεύονται και αυτές στον υπολογιστή οι οποίες μετά μπορούν να επεξεργαστούν. Ο γιατρός επιλέγει τις περιοχές ενδιαφέροντος ανάλογα με το μέγεθος της ύποπτης περιοχής και τις απομονώνει (Σχήμα 5.2) [1].

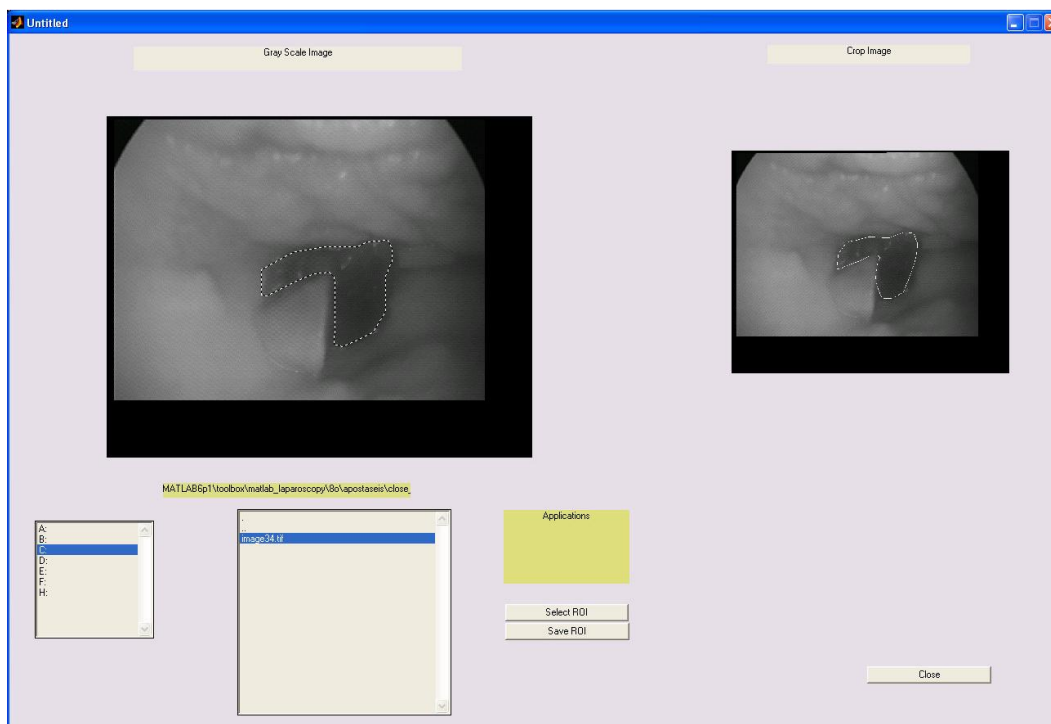


Σχήμα 5.2 Παγωμένη εικόνα ενδομητρίου και μια ύποπτη περιοχή παθολογίας (64X64 pixels).

Στην αποθηκευμένη εικόνα μπορούμε να τμηματοποιήσουμε ύποπτες περιοχές γυναικολογικού καρκίνου με ένα απλό σύστημα κατάτμησης. Πιο κάτω αναλύεται ένα σύστημα κατάτμησης με χειροκίνητη τμηματοποίηση των περιοχών ενδιαφέροντος που παρουσιάζει την αρχική εικόνα καθώς και την τμηματοποιημένη περιοχή ενδιαφέροντος, Σχήμα 5.3 [1].

Θα ήταν ιδιαίτερα σημαντικό να αναφερθεί η δυσκολία της αυτόματης τμηματοποίησης υστεροσκοπικών εικόνων λόγω της ιδιομορφίας της απεικόνισης όπως επίσης και τις πολυχρωμίες που παρουσιάζουν οι ιστοί.

Στην εικόνα που ακολουθεί, παρουσιάζεται μέρος του συστήματος χειροκίνητης κατάτμησης της περιοχής ενδιαφέροντος από τον ιατρό.



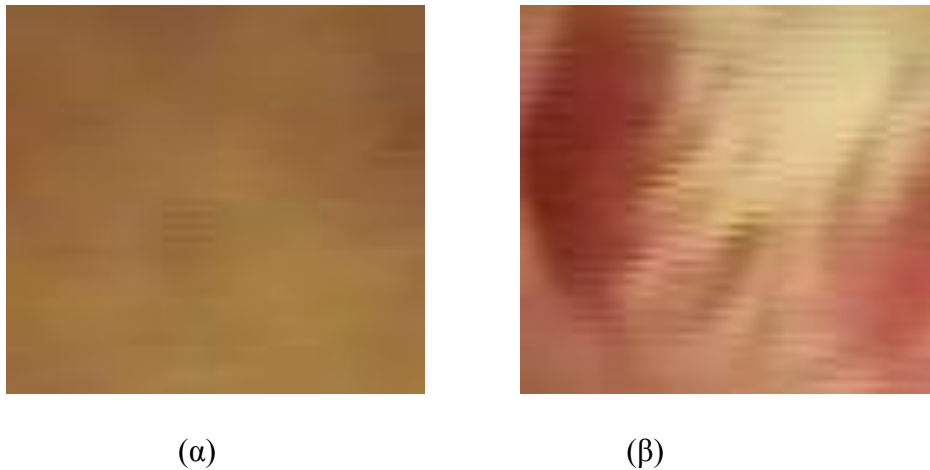
Σχήμα 5.3 Μέρος του συστήματος χειροκίνητης κατάτμησης της περιοχής ενδιαφέροντος από τον ιατρό.

5.4 Προέλευση δείγματος εικόνων

Κατά τη διάρκεια μιας υστεροσκοπικής επέμβασης ο γιατρός σε δύσκολες περιπτώσεις γυναικολογικού καρκίνου δημιουργεί άμεσες βιοψίες με στόχο την επαλήθευση της διάγνωσης του. Με την χρήση ηλεκτρονικών υπολογιστών και την ανάλυση των περιοχών ενδιαφέροντος θα αυξήσει την διαγνωστική ακρίβεια του εν λόγω περιστατικού.

Τα βίντεο / εικόνες αυτά/αυτές αποθηκεύονται στην ιατρική βάση δεδομένων και παράλληλα ο γιατρός επιλέγει ύποπτες περιοχές καρκίνου καθώς και φυσιολογικές περιοχές του ενδομητρίου. Στην εν λόγω έρευνα έχουμε συλλέξει 404 εικόνες από πραγματικό ενδομήτριο, 202 εικόνες ενδιαφέροντος που παρουσιάζουν φυσιολογικές περιοχές και 202 εικόνες ενδιαφέροντος που παρουσιάζουν ύποπτες περιοχές καρκίνου. Αυτές οι περιοχές ενδιαφέροντος αποτελούν το δείγμα μας και έχουν εξαχθεί από ένα σύνολο 40 γυναικών. Σε 30 περιπτώσεις τα περιστατικά αυτά

παρουσιάζουν την ίδια παθολογία, καρκίνο του ενδομητρίου, και χρησιμοποιώντας την ίδια μεθοδολογία όπως της πραγματικής υστεροσκοπικής εξέτασης εξαγάγαμε τις περιοχές ενδιαφέροντος. Οι άλλες 10 γυναίκες παρουσιάζουν φυσιολογικό ενδομήτριο, έπειτα και από τα αποτελέσματα τις ιστοπαθολογικής εξέτασης. Η ηλικία των γυναικών αυτών είναι από 22 μέχρι 50 χρόνων.



Σχήμα 5.4 Οι εικόνες (α) και (β) παρουσιάζουν φυσιολογική και παθολογική περιοχή του ενδομητρίου

Στο Σχήμα 5.4 (α) και (β) παρουσιάζουμε 2 περιοχές ενδιαφέροντος από μία υστεροσκοπική εξέταση του ενδομητρίου με παθολογική και φυσιολογική περιοχή πριν την εφαρμογή του αλγορίθμου γάμμα. [1]

5.5 Εφαρμογή αλγορίθμων χαρακτηριστικών υφής

Η εικόνα περιέχει πληροφορίες που είναι δύσκολο να αναγνωριστούν με μια απλή παρατήρηση. Έτσι χρησιμοποιούνται αλγόριθμοι χαρακτηριστικών υφής για να χαρακτηρίσουν την εικόνα και να εξαγάγουν χρήσιμες πληροφορίες για το περιεχόμενό της. Οι αλγόριθμοι χαρακτηριστικών υφής είναι μαθηματικές εξισώσεις που δέχονται σαν είσοδο τις τιμές του γκριζού που αποτελείται η εικόνα καθώς και άλλες συνιστώσες, όπως τον αριθμό των στιγμάτων, τη μέση τιμή του γκρι κτλ.

Τα χαρακτηριστικά υφής μπορούν να περιγράψουν με μεγάλη ακρίβεια την υφή της περιοχής ενδιαφέροντος και να δώσουν μοναδική ταυτότητα σε κάθε υπό εξέταση περιοχή. Με αυτά τα χαρακτηριστικά, η χρήση των υπολογιστών στην διαγνωστική χειρουργική, γίνεται υλοποιήσιμη και έξυπνα συστήματα επεξεργασίας εικόνων έχουν προταθεί με στόχο την αύξηση της διαγνωστικής ακρίβειας του περιστατικού. Τα χαρακτηριστικά υφής τα συναντούμε σε εικόνες γκρι τόνου όπως εικόνες υπερήχων, MRI, CT κ.α. Μερικά χαρακτηριστικά είναι η μέση τιμή, η διασπορά, η ομογένεια, η διακύμανση, ο μέσος όρος, η εντροπία κ.α [2].

Μέτα τον μετασχηματισμό των 202 υστεροσκοπικών εικόνων ενδομητρίου στα διάφορα συστήματα απεικόνισης εφαρμόσαμε τρεις αλγόριθμους χαρακτηριστικών υφής στις εικόνες αυτές. Οι αλγόριθμοι αυτοί είναι ο *Statistical Features (SF)*, ο *Spatial Gray Level Dependence Matrices (SGLDM)* και ο *Gray Level Difference Statistics (GLDS)*. Με τον αλγόριθμο *Statistical Features (SF)* έχουμε εξαγάγει οκτώ χαρακτηριστικά υφής (Mean, Variance, Median, Mode, Skewness, Kurtosis, Energy και Entropy), με τον αλγόριθμο *Spatial Gray Level Dependence Matrices (SGLDM)* έχουμε εξαγάγει δεκατρία χαρακτηριστικά υφής (ASM, Contrast, Correlation, Variance, Homogeneity, Sum Average, Sun Variance, Entropy, Sum Entropy, Dif. Variance, Dif. Entropy, Inf. Correlation1, και Inf. Correlation2) και τέλος με τον αλγόριθμο *Gray Level Difference Statistics (GLDS)* έχουμε εξαγάγει πέντε χαρακτηριστικά υφής (Homogeneity, Contrast, Energy, Entropy και Mean). Συμπερασματικά κάθε μοντέλο μιας βάσης δεδομένων μας για το χρωματικό κανάλι Y και το σύστημα απεικόνισης Gray Scale περιέχει αυτά τα εικοσιέξι χαρακτηριστικά υφής.

5.6 Αλγόριθμοι κατηγοριοποίησης / εξόρυξης δεδομένων

Η κατηγοριοποίηση δεδομένων είναι μία διαδικασία που αποτελείται από δύο φάσεις. Στην πρώτη φάση δημιουργήσαμε ένα μοντέλο περιγράφοντας ένα σύνολο από καθορισμένες κατηγορίες. Στην μελέτη αυτή οι καθορισμένες κατηγορίες είναι οι εξής:

- Abnormal: Παθολογική περιοχή του ενδομητρίου.
- Normal: Φυσιολογική περιοχή του ενδομητρίου.

Το μοντέλο κτίστηκε αναλύοντας μία ομάδα από εγγραφές – δείγματα όπου το κάθε δείγμα ανήκει σε μία κατηγορία. Τα δείγματα αυτά είναι τα δείγματα εκπαίδευσης. Επειδή σε αυτή τη φάση δίδεται η κατηγορία στην οποία ανήκει το κάθε δείγμα η διαδικασία αυτής της μάθησης ονομάζεται επιβλεπόμενη μάθηση (supervised learning).

Το αποτέλεσμα της μάθησης αυτής στη δική μας μελέτη δίδεται στην μορφή ενός δέντρου απόφασης (decision tree) και στη μορφή κανόνων κατηγοριοποίησης οι οποίοι πηγάζουν από το δέντρο αυτό. Ο αλγόριθμος που χρησιμοποιήθηκε για την δημιουργία του μοντέλου (των δέντρων απόφασης και των κανόνων κατηγοριοποίησης) είναι ο C4.5 με τη χρήση των εργαλείων που αναφέραμε πιο πάνω.

Στην δεύτερη φάση το μοντέλο χρησιμοποιήθηκε για κατηγοριοποίηση, με την εφαρμογή του μοντέλου αυτού σε μια ομάδα από δείγματα για τα οποία δεν γνωρίζουμε σε ποια κατηγορία ανήκουν. Τα δείγματα αυτά είναι τα δείγματα αξιολόγησης. Ακολούθως συγκρίναμε την κατηγοριοποίηση που δίνει το μοντέλο με την κατηγορία που ήδη υπάρχει στα δείγματα. Τα δείγματα (αξιολόγησης) στα οποία εφαρμόστηκε το μοντέλο για έλεγχο είναι διαφορετικά από τα δείγματα (εκπαίδευσης) που χρησιμοποιήθηκαν για να δημιουργηθεί το μοντέλο. Στόχος της μελέτης, είναι να μπορεί το μοντέλο να χρησιμοποιηθεί για την κατηγοριοποίηση άγνωστων δειγμάτων για τα οποία δεν έχουμε ένδειξη σε ποια κατηγορία ανήκουν.

Συνολικά είχαμε 202 εικόνες υστεροσκόπησης ενδομητρίου. Οι 101 εικόνες (συγκεκριμένα τα είκοσιέξι χαρακτηριστικά υφής τους), χρησιμοποιήθηκαν για δημιουργία/ εκπαίδευση του μοντέλου (δέντρων απόφασης και κανόνων κατηγοριοποίησης) και οι υπόλοιπες 101 για την αξιολόγηση του.

5.7 Μετρικές αποτελεσματικότητας της διάγνωσης από το σύστημα κατηγοριοποίησης

Αφού πήραμε τα αποτελέσματα από τους αλγόριθμους εξόρυξης δεδομένων χρησιμοποιήσαμε διάφορες μετρικές για την αξιολόγηση των αποτελεσμάτων των δύο αλγορίθμων. Οι μετρήσεις ή αναλύσεις που χρησιμοποιούνται για την αξιολόγηση των κατηγοριοποιήσεων των αλγορίθμων είναι:

Αρχικά, δίνεται ο ορισμός των εξής παραμέτρων : (α) True Positive (TP) - όταν η θετική διάγνωση του συστήματος συμπίπτει με την θετική διάγνωση του ιατρού (β) False Positive (FP) - όταν το σύστημα δίνει θετική διάγνωση αλλά η διάγνωση του ιατρού είναι αρνητική (γ) False Negative (FN) - όταν το σύστημα δίνει αρνητική διάγνωση αλλά η διάγνωση του ιατρού είναι θετική (δ) True Negative (TN) - όταν η αρνητική διάγνωση του συστήματος συμφωνεί με την αρνητική διάγνωση του ιατρού. Ακολουθώντας θα δώσουμε τον ορισμό και την σχέση μεταξύ των διάφορων μετρήσεων [23].

(1)Accuracy

(2)Precision

(3)Sensitivity

(4)Specificity

Τα δύο βασικά πρότυπα στα οποία μια νέα διαγνωστική δοκιμή του συστήματος πρέπει να ανταποκριθεί είναι η ισχύς (validity) και η αξιοπιστία (reliability), οι οποίες αναφέρονται στην ιατρική βιβλιογραφία ως accuracy και precision. Η ισχύς απεικονίζει εάν η δοκιμή μετρά πραγματικά αυτό που πρέπει να μετρηθεί, και η ακρίβεια μετρά το βαθμό στον οποίο η δοκιμή είναι σωστή. Η αξιοπιστία (reliability) και η ακρίβεια (precision) απεικονίζουν εάν τα αποτελέσματα της διαγνωστικής δοκιμής μπορούν να επαναληφθούν από τους ίδιους ή διαφορετικούς ειδικούς. Η μέτρηση accuracy είναι η σχέση μεταξύ των αποτελεσμάτων του συστήματος και των διαγνώσεων (θετικών και αρνητικών) των ιατρών. Από τη σύγκριση και το γεγονός αν συμπίπτουν αυτά τα δύο, πηγάζουν οι μετρήσεις κλειδιά

για μια διαγνωστική δοκιμή - sensitivity, specificity, positive predictive value, και negative predictive value. Ο πίνακας απόφασης (Confusion Matrix) διευκολύνει την ανάλυση μιας απόφασης του συστήματος ανάλογα εάν η ασθένεια είναι παρούσα ή απύουσα σύμφωνα με την διάγνωση του ιατρού.

CM	Διάγνωση ιατρού	
Διάγνωση συστήματος	D- (Normal)	D+ (Abnormal)
T- (Normal)	TN	FN
T+ (Abnormal)	FP	TP

Πίνακας 5.1 Πίνακας απόφασης (confusion matrix)

Το ενδιαφέρον του ιατρού εστιάζεται στην πιθανότητα παρουσίας (D+) ή απουσίας (D-) της ασθένειας όταν η διαγνωστική δοκιμή είναι θετική (T+) ή αρνητική (T-). Ο πίνακας απόφασης (Πίνακας 5.1) επεξηγεί τη σχέση μεταξύ των τεσσάρων πιθανών αποτελεσμάτων: True Positive (TP), False Positive (FP), True Negative (TN), και False Negative (FN). Από τον προσδιορισμό του αριθμού αποτελεσμάτων TP, FP, TN, και FN μετά από μια σειρά διαγνωστικών δοκιμών, μπορούν να παραχθούν αρκετές αναλογίες. Οι δύο πρώτες - sensitivity και specificity είναι ανεξάρτητες από την επικράτηση (prevalance) της ασθένειας στις γυναίκες των οποίων μελετούνται οι εικόνες ενδοσκοπησης. (1) *Sensitivity*, ή true-positive ratio, προσδιορίζει το ποσοστό των ασθενών οι οποίες όντως πάσχουν από καρκίνο του ενδομήτριου στις οποίες ορθά το σύστημα κάνει θετική διάγνωση. (2) *Specificity*, ή true-negative ratio, προσδιορίζει το ποσοστό των ασθενών οι οποίες δεν πάσχουν από καρκίνο του ενδομήτριου στις οποίες ορθά το σύστημα δίνει αρνητική διάγνωση. Οι μετρήσεις Sensitivity και Specificity είναι ανεξάρτητες από την επικράτηση της ασθένειας επειδή καθορίζονται μόνο από τους ασθενείς που πάσχουν από καρκίνο του ενδομήτριου (sensitivity) ή μόνο από τους ασθενείς που δεν πάσχουν (specificity). Έτσι στον Πίνακα 5.1 υπολογίζονται κάθετα στην στήλη παρουσίας της ασθένειας D+ η μετρική Sensitivity ή στην στήλη απουσίας της ασθένειας D- η μετρική Specificity.

Οι τέσσερις μετρήσεις που έχουμε περιγράψει πιο πάνω χαρακτηρίζουν την αποτελεσματικότητα μιας διαγνωστικής δοκιμής. Εντούτοις, όσον αφορά στον κάθε ασθενή, ενδιαφερόμαστε για την πιθανότητα αν ο ασθενής έχει την ασθένεια, λαμβάνοντας υπόψη ένα συγκεκριμένο αποτέλεσμα της διαγνωστικής δοκιμής. Εδώ, δεν ξέρουμε την αλήθεια εκ των προτέρων (a priori). Υποθέτουμε μόνο, με την βοήθεια της πιθανότητας της ύπαρξης ορθότητας (TP ή TN) βασισμένοι στην ακρίβεια (accuracy) της δοκιμής. Κατά συνέπεια, εξετάζουμε τον πίνακα απόφασης στην οριζόντια του διάστασης και υπολογίζουμε δύο τύπους τιμών πρόβλεψης.

Οι διαφορές αυτές μετρήσεις (με ποσοστό επί τοις 100, εκτός από το Likelihood Ratio) ορίζονται πιο κάτω:

- Accuracy: $\%CC = 100 \times (TP + TN) / (TP + TN + FP + FN)$
Είναι το ποσοστό των ασθενών για το οποίο η διαγνωστική δοκιμή είναι σωστή
- Precision: $PR = 100 \times TP / (TP + FP)$
- Sensitivity: $SE = 100 \times TP / (TP + FN)$
Είναι το ποσοστό που δείχνει κατά πόσο η ασθένεια θα διαγνωσθεί δεδομένου ότι όντως υπάρχει.
- Specificity: $SP = 100 \times TN / (TN + FP)$
Είναι το ποσοστό που δείχνει κατά πόσο θα διαγνωσθεί ότι η ασθένεια δεν υπάρχει δεδομένου ότι όντως δεν υπάρχει.

5.8 Μετρικές αποτελεσμάτων από το σύστημα Weka

Αφού χρησιμοποιήσαμε το σύστημα Weka για να εφαρμόσουμε τον αλγόριθμο C4.5 στα δεδομένα μας, πήραμε κάποια αποτελέσματα σε μορφή πινάκων, τα οποία παρουσιάζονται στο κεφάλαιο 6.

Στο συγκεκριμένο υποκεφάλαιο γίνεται επεξήγηση των μετρικών που εμφανίζονται στους πίνακες.

Πιο κάτω παρουσιάζεται παράδειγμα πίνακα αποτελεσμάτων όπως εμφανίζει το σύστημα Weka μετά την εφαρμογή του αλγορίθμου C4.5.

```

=== Stratified cross-validation ===

Correctly Classified Instances      9      64.2857 %
Incorrectly Classified Instances    5      35.7143 %
Kappa statistic                     0.186
Mean absolute error                 0.2857
Root mean squared error            0.4818
Relative absolute error             60      %
Root relative squared error        97.6586 %
Total Number of Instances          14

```

Σχήμα 5.5 Πίνακας αποτελεσμάτων από το σύστημα Weka μετά την εφαρμογή του αλγορίθμου C4.5 [21]

Το Correctly Classified Instances, δίνει το ποσοστό των περιπτώσεων που έχουν κατηγοριοποιηθεί ορθά, ενώ το Incorrectly Classified Instances δίνει το ποσοστό των περιπτώσεων που δεν έχουν κατηγοριοποιηθεί ορθά. Το Kappa statistic μετρά κατά πόσο συμφωνεί η πρόβλεψη με το πραγματικό αποτέλεσμα (1.0 σημαίνει πλήρης συμφωνία). Επίσης, το Mean absolute error δίνει την απόλυτη τιμή λάθους ενώ το Root mean squared error δίνει τη ρίζα του μέσου τετραγωνικού σφάλματος. Οι μετρικές Relative absolute error, Root relative squared error, Coverage of cases (0.95 level), Mean rel. region size (0.95 level) αφορούν μαθηματικές εξισώσεις οι οποίες δεν δίνουν κάποιο χρήσιμο αποτέλεσμα σχετικά με τη συγκεκριμένη μελέτη. Τέλος, το πεδίο Total Number of Instances δίνει το συνολικό αριθμό περιπτώσεων που έχουν εξαχθεί από τον αλγόριθμο.

Επιπρόσθετα, το σύστημα εμφανίζει ακόμη ένα πίνακα του που έχει την εξής μορφή:

```

=== Detailed Accuracy By Class ===

TP Rate  FP Rate  Precision  Recall  F-Measure  Class
0.778    0.6     0.7        0.778   0.737      yes
0.4      0.222    0.5        0.4     0.444      no

```

Σχήμα 5.6 Πίνακας αποτελεσμάτων από το σύστημα Weka [21]

Το TP rate (True Positive) δίνει το ποσοστό των παραδειγμάτων τα οποία ταξινομήθηκαν ως x (Class), ανάμεσα σε όλα τα παραδείγματα που ανήκουν πραγματικά στην τάξη x . Ουσιαστικά δίνει το πόσο μέρος της τάξης συνελήφθη, και είναι ισοδύναμο με το Recall.

Το FP rate (False Positive) δίνει το ποσοστό των παραδειγμάτων τα οποία ταξινομήθηκαν ως κατηγορίας x , αλλά ανήκουν σε διαφορετική κατηγορία, ανάμεσα σε όλα τα παραδείγματα που δεν είναι της κλάσης x .

Το Precision δίνει το ποσοστό από τα παραδείγματα που ανήκουν πραγματικά στην τάξη x , μεταξύ όλων αυτών που είχαν ταξινομηθεί ως κλάσης x .

Το F-Measure είναι η τιμή που προκύπτει από το $2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})$, είναι δηλαδή συνδυασμός του Precision και του Recall. [21]

Κεφάλαιο 6

Αποτελέσματα

6.1	Επισκόπηση	49
6.2	Στατιστική Ανάλυση Gray Scale εικόνων ενδομητρίου σε γυναίκες	50
6.3	Αποτελέσματα αλγόριθμου C4.5 για τον αλγόριθμο SGLDM	56
6.4	Αποτελέσματα αλγόριθμου C4.5 για τον αλγόριθμο SF	62
6.5	Αποτελέσματα αλγόριθμου C4.5 για τον αλγόριθμο GLDS	67
6.6	Αποτελέσματα αλγόριθμου C4.5 για τον αλγόριθμο SGLDM και SF	72
6.7	Αποτελέσματα αλγόριθμου C4.5 για τον αλγόριθμο SGLDM και GLDS	78
6.8	Αποτελέσματα αλγόριθμου C4.5 για τον αλγόριθμο SF και GLDS	84
6.9	Αποτελέσματα αλγόριθμου C4.5 για τον αλγόριθμο SGLDM, SF, GLDS	90
6.10	Συνοπτικά Αποτελέσματα	96

6.1 Επισκόπηση

Στο κεφάλαιο αυτό παρουσιάζονται τα αποτελέσματα που δημιουργήθηκαν με την με την εφαρμογή του C4.5 αλγόριθμου.

Η εκπαίδευση των δέντρων απόφασης και κατ' επέκταση η δημιουργία των κανόνων κατηγοριοποίησης γίνεται με βάση 101 εικόνες ενδοσκοπικής ενδομητρίου και με βάση τις υπόλοιπες 101 γίνεται η αξιολόγηση του δέντρου και των κανόνων. Σε κάθε φύλλο των δέντρων απόφασης που παράγονται από τον αλγόριθμο C4.5 δίδεται ο αριθμός των περιπτώσεων που λήφθηκαν υπόψη και ο αριθμός των περιπτώσεων που έχουν κατηγοριοποιηθεί λανθασμένα από την κατηγορία/κλάση με την οποία έχει σηματοδοθεί το συγκεκριμένο φύλλο.

Για τα δέντρα απόφασης και τους κανόνες κατηγοριοποίησης τους οι δύο αλγόριθμοι δημιουργούν αντίστοιχους πίνακες απόφασης. Με βάση τους πίνακες αυτούς

βλέπουμε την αποτελεσματικότητα κατηγοριοποίησης των δεδομένων αξιολόγησης, από το συγκεκριμένο δέντρο ή κανόνες κατηγοριοποίησης αντίστοιχα, σύμφωνα με τις μετρικές τις οποίες αναφέραμε πιο πάνω. Τέλος δίδονται τα ποσοστά λάθους της πρόβλεψης του καρκίνου του ενδομήτριου από τους δύο αλγόριθμους για κάθε ένα μοντέλο της βάσης δεδομένων μας κάθε φορά.

Τα αποτελέσματα κατηγοριοποίησης των χαρακτηριστικών υφής των εικόνων μας στα διάφορα μοντέλα απεικόνισης παρουσιάζονται σε μορφή πινάκων. Πιο αναλυτικά παρουσιάζουμε τα αποτελέσματα κατηγοριοποίησης των χαρακτηριστικών υφής των εικόνων γκρι τόνου και των εικόνων σε σύστημα απεικόνισης χρωματικού καναλιού Υ. Αφού γίνει η παρουσίαση των αποτελεσμάτων αυτών θα σχολιαστούν τα σημαντικότερα ευρήματα της μελέτης και θα αξιολογηθούν οι δύο αλγόριθμοι που χρησιμοποιήθηκαν για την εξαγωγή των αποτελεσμάτων σύμφωνα με τις μετρικές αξιολόγησης (Accuracy, Precision, Sensitivity, Specificity, Negative Predictive Value, Likelihood ratio, %FP, %FN). Για κάθε δέντρο απόφασης και κανόνα κατηγοριοποίησης που εξάγουμε χρησιμοποιηθήκαν και οι δυο αλγόριθμοι έτσι ώστε να μπορεί να γίνει σύγκριση τους στο τέλος.

Στη συνέχεια παρουσιάζονται τα αποτελέσματα του αλγορίθμου C4.5 για κάθε ένα από τους αλγόριθμους SGLDM, SF και GLDS καθώς και συνδιασμοί αλγορίθμων.

6.2 Στατιστική Ανάλυση Gray Scale εικόνων ενδομητρίου σε γυναίκες

Τα χαρακτηριστικά υφής έχουν εξαχθεί μετά από μετατροπή των έγχρωμων περιοχών ενδιαφέροντος σε αποχρώσεις του γκρι.

Ο Πίνακας 6.1 παρουσιάζει τη στατιστική ανάλυση μεταξύ του φυσιολογικού και παθολογικού ιστού μετά την εφαρμογή του αλγορίθμου γάμμα. Παρατηρούμε ότι υπάρχει στατιστική διαφορά στα περισσότερα χαρακτηριστικά υφής.

Η πιο πάνω στατιστική ανάλυση παρουσιάζει ιδιαίτερο ενδιαφέρον και εξάγονται ενδιαφέροντα συμπεράσματα. Στον αλγόριθμο SF το median της διακύμανση (variance) είναι μικρότερο από την τιμή του στην παθολογική περιοχή, 66 με 106 αντίστοιχα. Στον αλγόριθμο SGLDM παρατηρούμε την ίδια απόκλιση της διακύμανση (variance, 64/104) ενώ στον αλγόριθμο GLDS η median τιμή του contrast είναι μικρότερη στις φυσιολογικές εικόνες σε σχέση με τις παθολογικές, με τιμές 12 και 15 αντίστοιχα. Επίσης παρατηρήσαμε ότι και στους τρεις αλγόριθμους η εντροπία των δύο κατηγοριών εικόνων παραμένει σταθερή.

	Πραγματικές εικόνες				
	Φυσιολογικό Ενδομήτριο		Παθολογικό Ενδομήτριο		
	Mean	STD	Mean	STD	H ¹
SF					
Mean	55,81	44,59	52,78	46,69	0
Variance	264,36	514,09	335,67	585,83	1
Median	54,69	44,31	51,41	47,40	0
Mode	56,29	55,05	49,95	54,30	1
Skewness	0,20	0,64	0,35	0,77	1
Kurtosis	3,23	4,02	3,35	1,96	0
Energy	0,05	0,05	0,05	0,05	0
Entropy	3,44	0,77	3,58	0,76	1
SGLDM					
ASM	0,01	0,018	0,012	0,03	0
Contrast	29,07	56,52	28,63	43,06	0
Correlation	0,86	0,12	0,92	0,07	1
Variance	260,33	507,47	330,17	577,98	1
Homogeneity	0,36	0,13	0,38	0,15	0
Sum Average	113,62	89,19	107,56	93,40	0
Sum Variance	1012	1992	1292	2285	1
Sum Entropy	4,07	0,79	4,22	0,78	1
Dif. Variance	13,74	29,62	14,09	23,12	0
Dif. Entropy	1,99	0,50	2,00	0,54	0
InF. Correlation 1	-0,31	0,09	-0,36	0,07	1
InF. Correlation 2	0,89	0,10	0,93	0,06	1
Entropy	5,77	1,15	5,86	1,21	0
GLDS					
Homogeneity	0,37	0,13	0,38	0,15	0
Contrast	29,00	56,38	28,57	42,97	0
Energy	0,17	0,08	0,18	0,10	0
Entropy	2,08	0,50	2,07	0,55	0
Mean	3,09	2,07	3,08	1,94	0

Πίνακας 6.1 Χαρακτηριστικά υφής και στατιστική ανάλυση φυσιολογικού και παθολογικού ενδομητρίου από 202 εικόνες gray scale περιοχών ενδιαφέροντος

6.3 Αποτελέσματα αλγόριθμου C4.5 για τον αλγόριθμο SGLDM

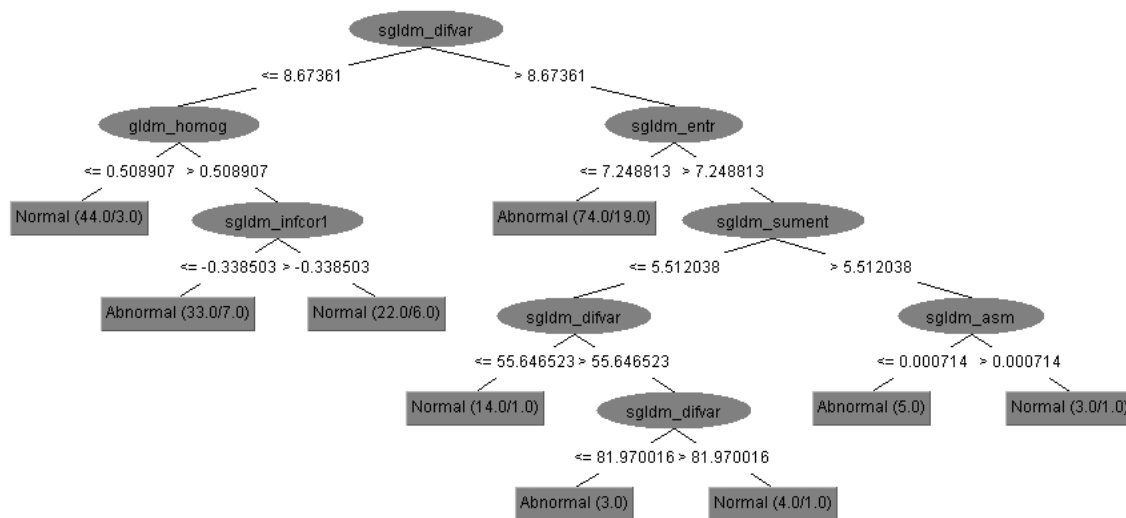
Πιο κάτω παρουσιάζονται τα αποτελέσματα του αλγόριθμου C4.5 για τον αλγόριθμο SGLDM. Συγκεκριμένα, παρουσιάζεται το δέντρο που έχει δημιουργηθεί από την εφαρμογή του αλγόριθμου C4.5 στα δεδομένα.

Στη συνέχεια παρουσιάζονται τα αποτελέσματα από την εφαρμογή του αλγόριθμου C4.5 για διάφορες τιμές του ποσοστού των εικόνων που δίνεται ως δείγμα εκπαίδευσης του συστήματος.

Δέντρο

```
sgldm_difvar <= 8.67361
|   sgldm_homog <= 0.508907: Normal (44.0/3.0)
|   sgldm_homog > 0.508907
|       |   sgldm_infcor1 <= -0.338503: Abnormal (33.0/7.0)
|       |   sgldm_infcor1 > -0.338503: Normal (22.0/6.0)
sgldm_difvar > 8.67361
|   sgldm_entr <= 7.248813: Abnormal (74.0/19.0)
|   sgldm_entr > 7.248813
|       |   sgldm_sument <= 5.512038
|       |       |   sgldm_difvar <= 55.646523: Normal (14.0/1.0)
|       |       |   sgldm_difvar > 55.646523
|       |       |       |   sgldm_difvar <= 81.970016: Abnormal (3.0)
|       |       |       |   sgldm_difvar > 81.970016: Normal (4.0/1.0)
|       |       |   sgldm_sument > 5.512038
|       |       |       |   sgldm_asm <= 0.000714: Abnormal (5.0)
|       |       |       |   sgldm_asm > 0.000714: Normal (3.0/1.0)
```

Αριθμός Φύλλων: 9
Μέγεθος Δέντρου : 17



Σχήμα 6.1 Δέντρο το οποίο δημιουργήθηκε από την εφαρμογή του αλγορίθμου C4.5 στα χαρακτηριστικά υφής που πάρθηκαν από τον αλγόριθμο SGLDM. Ο αριθμός που βρίσκεται σε παρένθεση υποδηλώνει των αριθμό των περιπτώσεων που βρέθηκαν να είναι Normal ή Abnormal αντίστοιχα.

	Δείγμ.Εκπ. 52% Δείγμ.Επαλ. 48%	Δείγμ.Εκπ. 60% Δείγμ.Επαλ. 40%	Δείγμ.Εκπ. 66% Δείγμ.Επαλ. 34%	Δείγμ.Εκπ. 72% Δείγμ.Επαλ. 28%
Correctly Classified Instances	60.8%	67%	61%	72%
Incorrectly Classified Instances	40%	40%	39%	28%
Kappa statistic	0.2	0.3	0.2	0.4
Mean absolute error	4	0.4	0.4	0.4
Root mean squared error	0.6	0.5	0.5	0.5
Relative absolute error	80%	74%	81%	71%
Root relative squared error	120%	107%	107%	91%
Coverage of cases (0.95 level)	71%	88%	93%	96%
Mean rel. region size (0.95 level)	62%	83%	91%	93%
Total Number of Instances	97	81	69	57

Πίνακας 6.2 Αποτελέσματα από την εφαρμογή του αλγορίθμου C4.5 στα χαρακτηριστικά υφής που πάρθηκαν από τον αλγόριθμο SGLDM

		TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
Δείγμα Εκπαίδευσης 52% Δείγμα Επαλήθευσης 48%		0.6	0.4	0.6	0.6	0.6	0.6	Normal
		0.6	0.4	0.6	0.6	0.6	0.6	Abnormal
	Weighted Avg	0.6	0.4	0.6	0.6	0.6	0.6	
Δείγμα Εκπαίδευσης 60% Δείγμα Επαλήθευσης 40%		0.7	0.4	0.7	0.7	0.7	0.6	Normal
		0.6	0.3	0.6	0.6	0.6	0.6	Abnormal
	Weighted Avg	0.6	0.3	0.6	0.6	0.6	0.6	
Δείγμα Εκπαίδευσης 66% Δείγμα Επαλήθευσης 34%		0.5	0.3	0.7	0.5	0.6	0.6	Normal
		0.7	0.4	0.5	0.6	0.6	0.6	Abnormal
	Weighted Avg	0.6	0.4	0.6	0.6	0.6	0.6	
Δείγμα Εκπαίδευσης 72% Δείγμα Επαλήθευσης 28%		0.7	0.3	0.7	0.7	0.7	0.7	Normal
		0.6	0.2	0.6	0.6	0.6	0.7	Abnormal
	Weighted Avg	0.7	0.2	0.7	0.7	0.7	0.7	

Πίνακας 6.3 Αποτελέσματα από την εφαρμογή του αλγορίθμου C4.5 στα χαρακτηριστικά υφής που πάρθηκαν από τον αλγόριθμο SGLDM

6.4 Αποτελέσματα αλγόριθμου C4.5 για τον αλγόριθμο SF

Πιο κάτω παρουσιάζονται τα αποτελέσματα του αλγορίθμου C4.5 για τον αλγόριθμο SF. Συγκεκριμένα, παρουσιάζεται το δέντρο που έχει δημιουργηθεί από την εφαρμογή του αλγορίθμου C4.5 στα δεδομένα.

Στη συνέχεια παρουσιάζονται τα αποτελέσματα από την εφαρμογή του αλγορίθμου C4.5 για διάφορες τιμές του ποσοστού των εικόνων που δίνεται ως δείγμα εκπαίδευσης του συστήματος.

Δέντρο

fos_var $\leq$ 140.291494

| fos_mode $\leq$ 7: Abnormal (26.0/8.0)

| fos_mode $>$ 7: Normal (84.0/24.0)

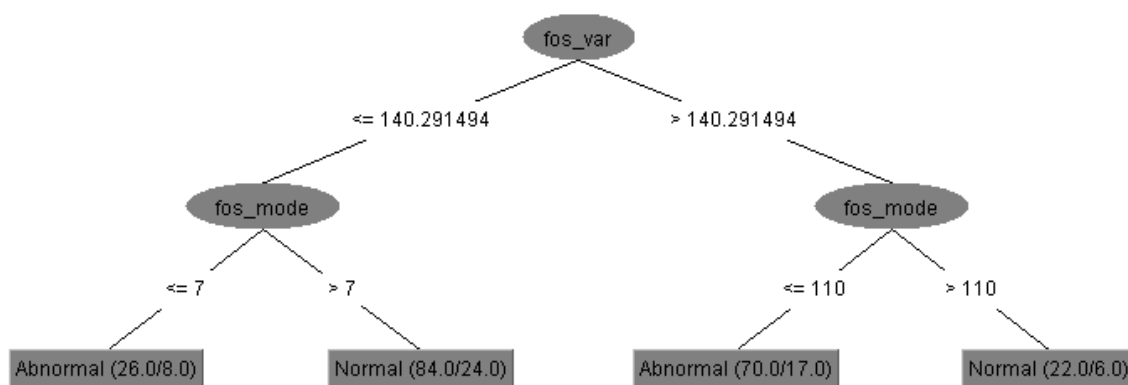
fos_var $>$ 140.291494

| fos_mode $\leq$ 110: Abnormal (70.0/17.0)

| fos_mode $>$ 110: Normal (22.0/6.0)

Αριθμός Φύλλων: 4

Μέγεθος Δέντρου : 7



Σχήμα 6.2 Δέντρο το οποίο δημιουργήθηκε από την εφαρμογή του αλγορίθμου C4.5 στα χαρακτηριστικά υφής που πάρθηκαν από τον αλγόριθμο SF. Ο αριθμός που βρίσκεται σε παρένθεση υποδηλώνει των αριθμό των περιπτώσεων που βρέθηκαν να είναι Normal ή Abnormal αντίστοιχα.

	Δείγμ.Εκπ. 52% Δείγμ.Επαλ. 48%	Δείγμ.Εκπ. 60% Δείγμ.Επαλ. 40%	Δείγμ.Εκπ. 66% Δείγμ.Επαλ. 34%	Δείγμ.Εκπ. 72% Δείγμ.Επαλ. 28%
Correctly Classified Instances	48%	47%	62%	42%
Incorrectly Classified Instances	51%	53%	37%	37%
Kappa statistic	0	0	0.2	0
Mean absolute error	0.5	0.5	0.4	0.5
Root mean squared error	0.5	0.5	0.5	0.5
Relative absolute error	100%	100%	89%	100%
Root relative squared error	100%	100%	100%	100%
Coverage of cases (0.95 level)	100%	100%	98%	100%
Mean rel. region size (0.95 level)	100%	100%	98%	100%
Total Number of Instances	97	81	69	57

Πίνακας 6.4 Αποτελέσματα από την εφαρμογή του αλγορίθμου C4.5 στα χαρακτηριστικά υφής που πάρθηκαν από τον αλγόριθμο SF

		TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
Δείγμα Εκπαίδευσης 52% Δείγμα Επαλήθευσης 48%		0	0	0	0	0	0.5	Normal
		1	1	0.4	1	0.6	0.5	Abnormal
	Weighted Avg	0.4	0.4	0.2	0.4	0.3	0.5	
Δείγμα Εκπαίδευσης 60% Δείγμα Επαλήθευσης 40%		0	0	0	0	0	0.5	Normal
		1	1	0.4	1	0.6	0.5	Abnormal
	Weighted Avg	0.4	0.4	0.2	0.4	0.3	0.5	
Δείγμα Εκπαίδευσης 66% Δείγμα Επαλήθευσης 34%		0.6	0.3	0.6	0.6	0.6	0.6	Normal
		0.6	0.3	0.5	0.6	0.6	0.6	Abnormal
	Weighted Avg	0.6	0.3	0.6	0.6	0.6	0.6	
Δείγμα Εκπαίδευσης 72% Δείγμα Επαλήθευσης 28%		0	0	0	0	0	0.5	Normal
		1	1	0.4	1	0.5	0.5	Abnormal
	Weighted Avg	0.4	0.4	0.1	0.4	0.2	0.5	

Πίνακας 6.5 Αποτελέσματα από την εφαρμογή του αλγορίθμου C4.5 στα χαρακτηριστικά υφής που πάρθηκαν από τον αλγόριθμο SF

6.5 Αποτελέσματα αλγόριθμου C4.5 για τον αλγόριθμο GLDS

Πιο κάτω παρουσιάζονται τα αποτελέσματα του αλγορίθμου C4.5 για τον αλγόριθμο GLDS. Συγκεκριμένα, παρουσιάζεται το δέντρο που έχει δημιουργηθεί από την εφαρμογή του αλγορίθμου C4.5 στα δεδομένα. Στη συνέχεια παρουσιάζονται τα αποτελέσματα από την εφαρμογή του αλγορίθμου C4.5 για διάφορες τιμές του ποσοστού των εικόνων που δίνεται ως δείγμα εκπαίδευσης του συστήματος.

Δέντρο

glds_con $\leq$ 23.196725

| glds_hom $\leq$ 0.519061

| | glds_con $\leq$ 9.735096: Normal (36.0/2.0)

| | glds_con $>$ 9.735096

| | | glds_hom $\leq$ 0.27884: Normal (7.0)

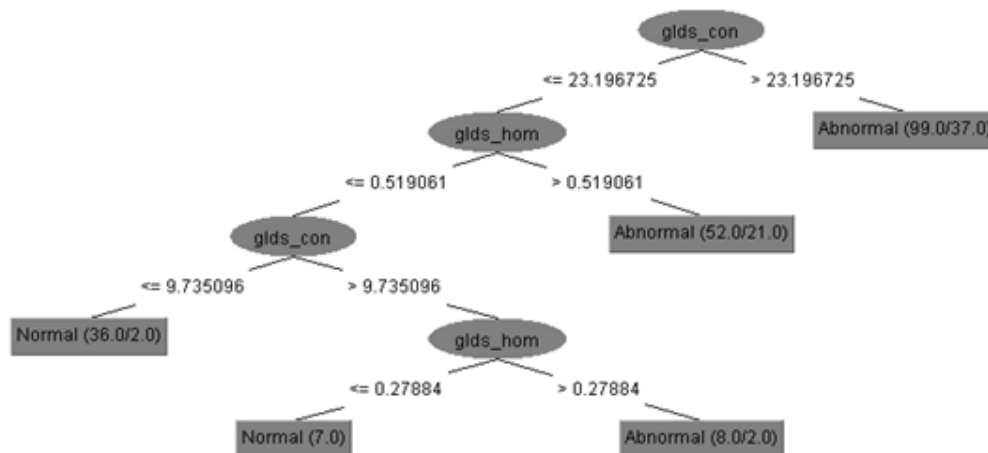
| | | glds_hom $>$ 0.27884: Abnormal (8.0/2.0)

| glds_hom $>$ 0.519061: Abnormal (52.0/21.0)

glds_con $>$ 23.196725: Abnormal (99.0/37.0)

Αριθμός Φύλλων: 5

Μέγεθος Δέντρου : 9



Σχήμα 6.3 Δέντρο το οποίο δημιουργήθηκε από την εφαρμογή του αλγορίθμου C4.5 στα χαρακτηριστικά υφής που πάρθηκαν από τον αλγόριθμο GLDS. Ο αριθμός που βρίσκεται σε παρένθεση υποδηλώνει των αριθμό των περιπτώσεων που βρέθηκαν να είναι Normal ή Abnormal αντίστοιχα.

	Δείγμ.Εκπ. 52% Δείγμ.Επαλ. 48%	Δείγμ.Εκπ. 60% Δείγμ.Επαλ. 40%	Δείγμ.Εκπ. 66% Δείγμ.Επαλ. 34%	Δείγμ.Εκπ. 72% Δείγμ.Επαλ. 28%
Correctly Classified Instances	60%	47%	45%	42%
Incorrectly Classified Instances	40%	53%	55%	58%
Kappa statistic	0,2	0	0	0
Mean absolute error	0.4	0.5	0.5	0.5
Root mean squared error	0.5	0.5	0.5	0.5
Relative absolute error	85%	100%	100%	100%
Root relative squared error	109%	100%	100%	100%
Coverage of cases (0.95 level)	91%	100%	100%	100%
Mean rel. region size (0.95 level)	91%	100%	100%	100%
Total Number of Instances	97	81	69	57

Πίνακας 6.6 Αποτελέσματα από την εφαρμογή του αλγορίθμου C4.5 στα χαρακτηριστικά υφής που πάρθηκαν από τον αλγόριθμο GLDS

		TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
Δείγμα Εκπαίδευσης 52% Δείγμα Επαλήθευσης 48%		0.5	0.3	0.6	0.5	0.5	0.5	Normal
		0.6	0.4	0.5	0.6	0.6	0.5	Abnormal
	Weighted Avg	0.5	0.3	0.6	0.5	0.5	0.5	
Δείγμα Εκπαίδευσης 60% Δείγμα Επαλήθευσης 40%		0	0	0	0	0	0.5	Normal
		1	1	0.4	1	0.6	0.5	Abnormal
	Weighted Avg	0.4	0.4	0.2	0.4	0.3	0.5	
Δείγμα Εκπαίδευσης 66% Δείγμα Επαλήθευσης 34%		0	0	0	0	0	0.5	Normal
		1	1	0.4	1	0.6	0.5	Abnormal
	Weighted Avg	0.4	0.4	0.2	0.4	0.2	0.5	
Δείγμα Εκπαίδευσης 72% Δείγμα Επαλήθευσης 28%		0	0	0	0	0	0.5	Normal
		1	1	0.4	1	0.5	0.5	Abnormal
	Weighted Avg	0.4	0.4	0.1	0.4	0.2	0.5	

Πίνακας 6.7 Αποτελέσματα από την εφαρμογή του αλγορίθμου C4.5 στα χαρακτηριστικά υφής που πάρθηκαν από τον αλγόριθμο GLDS

6.6 Αποτελέσματα αλγόριθμου C4.5 για τον αλγόριθμο SGLDM και SF

Πιο κάτω παρουσιάζονται τα αποτελέσματα του αλγορίθμου C4.5 για το συνδιασμό αλγορίθμων SGLDM και SF. Συγκεκριμένα, παρουσιάζεται το δέντρο που έχει δημιουργηθεί από την εφαρμογή του αλγορίθμου C4.5 στα δεδομένα.

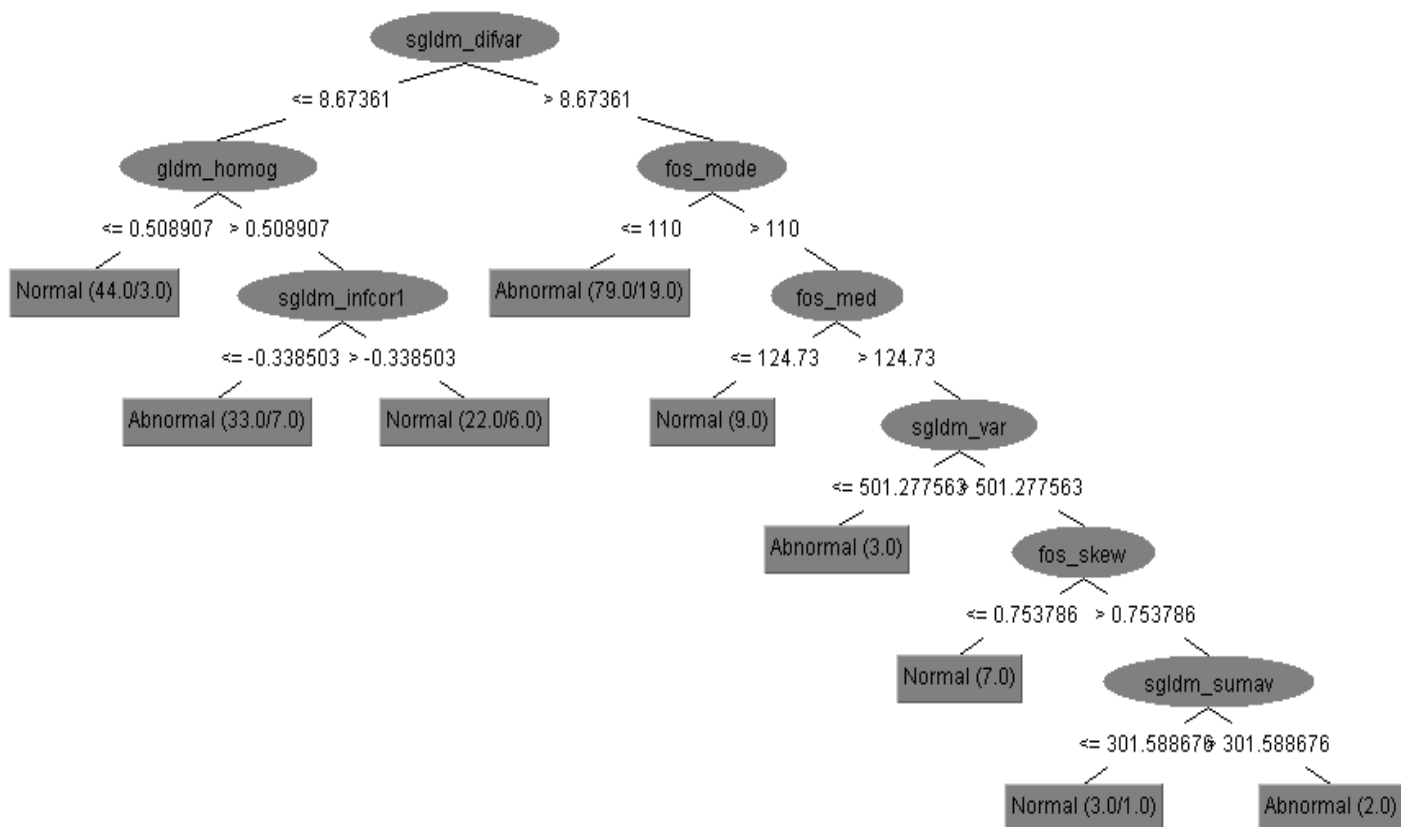
Στη συνέχεια παρουσιάζονται τα αποτελέσματα από την εφαρμογή του αλγορίθμου C4.5 για διάφορες τιμές του ποσοστού των εικόνων που δίνεται ως δείγμα εκπαίδευσης του συστήματος.

Δέντρο

```
sgldm_difvar <= 8.67361
|   sgldm_homog <= 0.508907: Normal (44.0/3.0)
|   sgldm_homog > 0.508907
|       |   sgldm_infcor1 <= -0.338503: Abnormal (33.0/7.0)
|       |   sgldm_infcor1 > -0.338503: Normal (22.0/6.0)
sgldm_difvar > 8.67361
|   fos_mode <= 110: Abnormal (79.0/19.0)
|   fos_mode > 110
|       |   fos_med <= 124.73: Normal (9.0)
|       |   fos_med > 124.73
|       |       |   sgldm_var <= 501.277563: Abnormal (3.0)
|       |       |   sgldm_var > 501.277563
|       |       |       |   fos_skew <= 0.753786: Normal (7.0)
|       |       |       |   fos_skew > 0.753786
|       |       |       |       |   sgldm_sumav <= 301.588676: Normal(3.0/1.0)
|       |       |       |       |   sgldm_sumav > 301.588676: Abnormal (2.0)
```

Αριθμός Φύλλων: 9

Μέγεθος Δέντρου : 17



Σχήμα 6.4 Δέντρο το οποίο δημιουργήθηκε από την εφαρμογή του αλγορίθμου C4.5 στα χαρακτηριστικά υφής που πάρθηκαν από τους αλγορίθμους SGLDM και SF. Ο αριθμός που βρίσκεται σε παρένθεση υποδηλώνει τον αριθμό των περιπτώσεων που βρέθηκαν να είναι Normal ή Abnormal αντίστοιχα.

	Δείγμ.Εκπ. 52% Δείγμ.Επαλ. 48%	Δείγμ.Εκπ. 60% Δείγμ.Επαλ. 40%	Δείγμ.Εκπ. 66% Δείγμ.Επαλ. 34%	Δείγμ.Εκπ. 72% Δείγμ.Επαλ. 28%
Correctly Classified Instances	67%	66%	62%	63%
Incorrectly Classified Instances	33%	33%	37%	36%
Kappa statistic	0,3	0,3	0,2	0,2
Mean absolute error	0.3	0.3	0.4	0.4
Root mean squared error	0.5	0.5	0.5	0.5
Relative absolute error	68%	73%	79%	84%
Root relative squared error	105%	109%	106,%	106%
Coverage of cases (0.95 level)	85%	82%	88%	91%
Mean rel. region size (0.95 level)	75%	78%	80%	93%
Total Number of Instances	97	81	69	57

Πίνακας 6.8 Αποτελέσματα από την εφαρμογή του αλγορίθμου C4.5 στα χαρακτηριστικά υφής που πάρθηκαν από τους αλγορίθμους SGLDM και SF

		TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
Δείγμα Εκπαίδευσης 52% Δείγμα Επαλήθευσης 48%		0.6	0.2	0.7	0.6	0.6	0.7	Normal
		0.7	0.3	0.6	0.7	0.6	0.7	Abnormal
	Weighted Avg	0.6	0.3	0.6	0.6	0.6	0.7	
Δείγμα Εκπαίδευσης 60% Δείγμα Επαλήθευσης 40%		0.6	0.3	0.6	0.6	0.6	0.6	Normal
		0.6	0.3	0.6	0.6	0.6	0.6	Abnormal
	Weighted Avg	0.6	0.3	0.6	0.6	0.6	0.6	
Δείγμα Εκπαίδευσης 66% Δείγμα Επαλήθευσης 34%		0.5	0.2	0.7	0.5	0.5	0.6	Normal
		0.7	0.5	0.5	0.7	0.6	0.6	Abnormal
	Weighted Avg	0.6	0.3	0.6	0.6	0.6	0.6	
Δείγμα Εκπαίδευσης 72% Δείγμα Επαλήθευσης 28%		0.5	0.2	0.7	0.5	0.6	0.6	Normal
		0.7	0.4	0.5	0.7	0.6	0.6	Abnormal
	Weighted Avg	0.6	0.3	0.6		0.6	0.6	

Πίνακας 6.9 Αποτελέσματα από την εφαρμογή του αλγορίθμου C4.5 στα χαρακτηριστικά υφής που πάρθηκαν από τους αλγορίθμους SGLDM και SF

6.7 Αποτελέσματα αλγόριθμου C4.5 για τον αλγόριθμο SGLDM και GLDS

Πιο κάτω παρουσιάζονται τα αποτελέσματα του αλγορίθμου C4.5 για το συνδυασμό αλγορίθμων SGLDM και GLDS. Συγκεκριμένα, παρουσιάζεται το δέντρο που έχει δημιουργηθεί από την εφαρμογή του αλγορίθμου C4.5 στα δεδομένα.

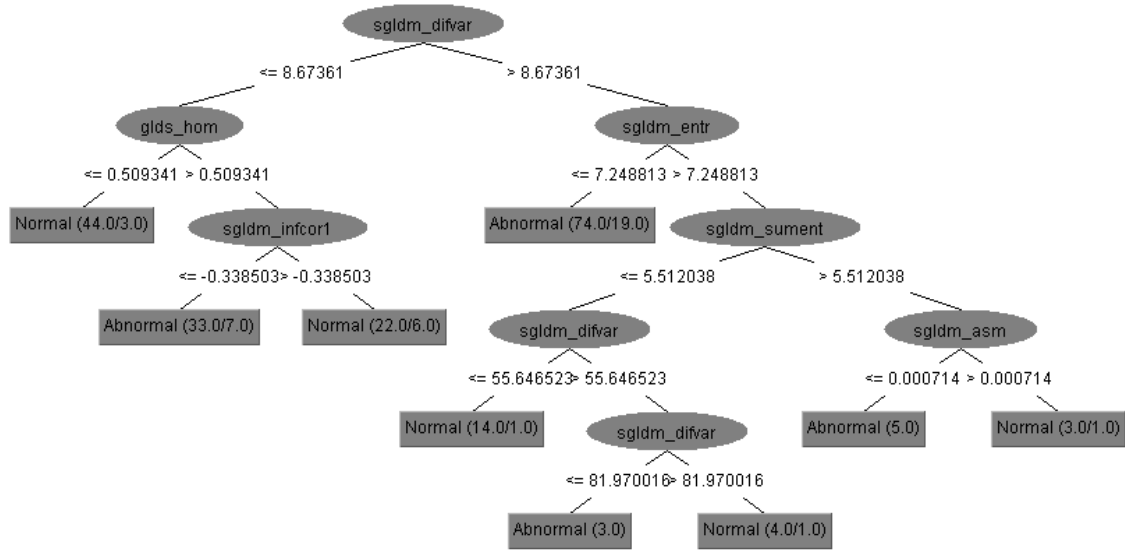
Στη συνέχεια παρουσιάζονται τα αποτελέσματα από την εφαρμογή του αλγορίθμου C4.5 για διάφορες τιμές του ποσοστού των εικόνων που δίνεται ως δείγμα εκπαίδευσης του συστήματος.

Δέντρο

```
sgldm_difvar <= 8.67361
|   glds_hom <= 0.509341: Normal (44.0/3.0)
|   glds_hom > 0.509341
|   |   sgldm_infcor1 <= -0.338503: Abnormal (33.0/7.0)
|   |   sgldm_infcor1 > -0.338503: Normal (22.0/6.0)
sgldm_difvar > 8.67361
|   sgldm_entr <= 7.248813: Abnormal (74.0/19.0)
|   sgldm_entr > 7.248813
|   |   sgldm_sument <= 5.512038
|   |   |   sgldm_difvar <= 55.646523: Normal (14.0/1.0)
|   |   |   sgldm_difvar > 55.646523
|   |   |   |   sgldm_difvar <= 81.970016: Abnormal (3.0)
|   |   |   |   sgldm_difvar > 81.970016: Normal (4.0/1.0)
|   |   |   sgldm_sument > 5.512038
|   |   |   sgldm_asm <= 0.000714: Abnormal (5.0)
|   |   |   sgldm_asm > 0.000714: Normal (3.0/1.0)
```

Αριθμός Φύλλων: 9

Μέγεθος Δέντρου : 17



Σχήμα 6.5 Δέντρο το οποίο δημιουργήθηκε από την εφαρμογή του αλγορίθμου C4.5 στα χαρακτηριστικά υψής που πάρθηκαν από τους αλγορίθμους SGLDM και GLDS. Ο αριθμός που βρίσκεται σε παρένθεση υποδηλώνει των αριθμό των περιπτώσεων που βρέθηκαν να είναι Normal ή Abnormal αντίστοιχα.

	Δείγμ.Εκπ. 52% Δείγμ.Επαλ. 48%	Δείγμ.Εκπ. 60% Δείγμ.Επαλ. 40%	Δείγμ.Εκπ. 66% Δείγμ.Επαλ. 34%	Δείγμ.Εκπ. 72% Δείγμ.Επαλ. 28%
Correctly Classified Instances	65%	66%	61%	73%
Incorrectly Classified Instances	35%	33%	39%	26%
Kappa statistic	0,3	0,3	0,2	0,4
Mean absolute error	0.3	0.3	0.4	0.3
Root mean squared error	0.5	0.5	0.5	0.4
Relative absolute error	72%	73%	81%	67%
Root relative squared error	111%	107%	106%	93%
Coverage of cases (0.95 level)	80%	87%	92%	94%
Mean rel. region size (0.95 level)	71%	82%	91%	92%
Total Number of Instances	97	81	69	57

Πίνακας 6.10 Αποτελέσματα από την εφαρμογή του αλγορίθμου C4.5 στα
χαρακτηριστικά υφής που πάρθηκαν από τους αλγορίθμους SGLDM και GLDS

		TP Rate	FP Rate	Precision	Recall	F- Measure	ROC Area	Class
Δείγμα Εκπαίδευσης 52% Δείγμα Επαλήθευσης 48%		0.6	0.3	0.6	0.6	0.6	0.6	Normal
		0.6	0.3	0.6	0.6	0.6	0.6	Abnormal
	Weighted Avg	0.6	0.3	0.6	0.6	0.6	0.6	
Δείγμα Εκπαίδευσης 60% Δείγμα Επαλήθευσης 40%		0.6	0.3	0.6	0.6	0.6	0.6	Normal
		0.6	0.3	0.6	0.6	0.6	0.6	Abnormal
	Weighted Avg	0.6	0.3	0.6	0.6	0.6	0.6	
Δείγμα Εκπαίδευσης 66% Δείγμα Επαλήθευσης 34%		0.5	0.3	0.6	0.5	0.6	0.6	Normal
		0.6	0.4	0.5	0.6	0.6	0.6	Abnormal
	Weighted Avg	0.6	0.3	0.6	0.6	0.6	0.6	
Δείγμα Εκπαίδευσης 72% Δείγμα Επαλήθευσης 28%		0.7	0.3	0.7	0.7	0.7	0.6	Normal
		0.6	0.2	0.6	0.6	0.6	0.6	Abnormal
	Weighted Avg	0.7	0.2	0.7	0.7	0.7	0.6	

Πίνακας 6.11 Αποτελέσματα από την εφαρμογή του αλγορίθμου C4.5 στα
χαρακτηριστικά υφής που πάρθηκαν από τους αλγορίθμους SGLDM και GLDS

6.8 Αποτελέσματα αλγόριθμου C4.5 για τον αλγόριθμο SF και GLDS

Πιο κάτω παρουσιάζονται τα αποτελέσματα του αλγορίθμου C4.5 για το συνδιασμό αλγορίθμων SF και GLDS. Συγκεκριμένα, παρουσιάζεται το δέντρο που έχει δημιουργηθεί από την εφαρμογή του αλγορίθμου C4.5 στα δεδομένα.

Στη συνέχεια παρουσιάζονται τα αποτελέσματα από την εφαρμογή του αλγορίθμου C4.5 για διάφορες τιμές του ποσοστού των εικόνων που δίνεται ως δείγμα εκπαίδευσης του συστήματος.

Δέντρο

fos_var <= 140.291494

| glds_hom <= 0.519061

| | glds_con <= 21.638733: Normal (40.0/3.0)

| | glds_con > 21.638733

| | | fos_mode <= 45

| | | | fos_ener <= 0.025974: Normal (2.0)

| | | | fos_ener > 0.025974: Abnormal (9.0/1.0)

| | | fos_mode > 45: Normal (7.0)

| glds_hom > 0.519061

| | fos_var <= 11.375965

| | | fos_med <= 7.098058

| | | | fos_skew <= 1.06152

| | | | | fos_mode <= 4: Abnormal (6.0)

| | | | | fos_mode > 4

| | | | | | fos_var <= 3.17148: Normal (2.0)

| | | | | | fos_var > 3.17148: Abnormal (7.0/1.0)

| | | | | fos_skew > 1.06152: Normal (2.0)

| | | | fos_med > 7.098058

| | | | | glds_con <= 2.253406: Normal (11.0)

| | | | | glds_con > 2.253406

| | | | | | fos_mean <= 14.903076: Abnormal (2.0)

| | | | | | fos_mean > 14.903076: Normal (3.0/1.0)

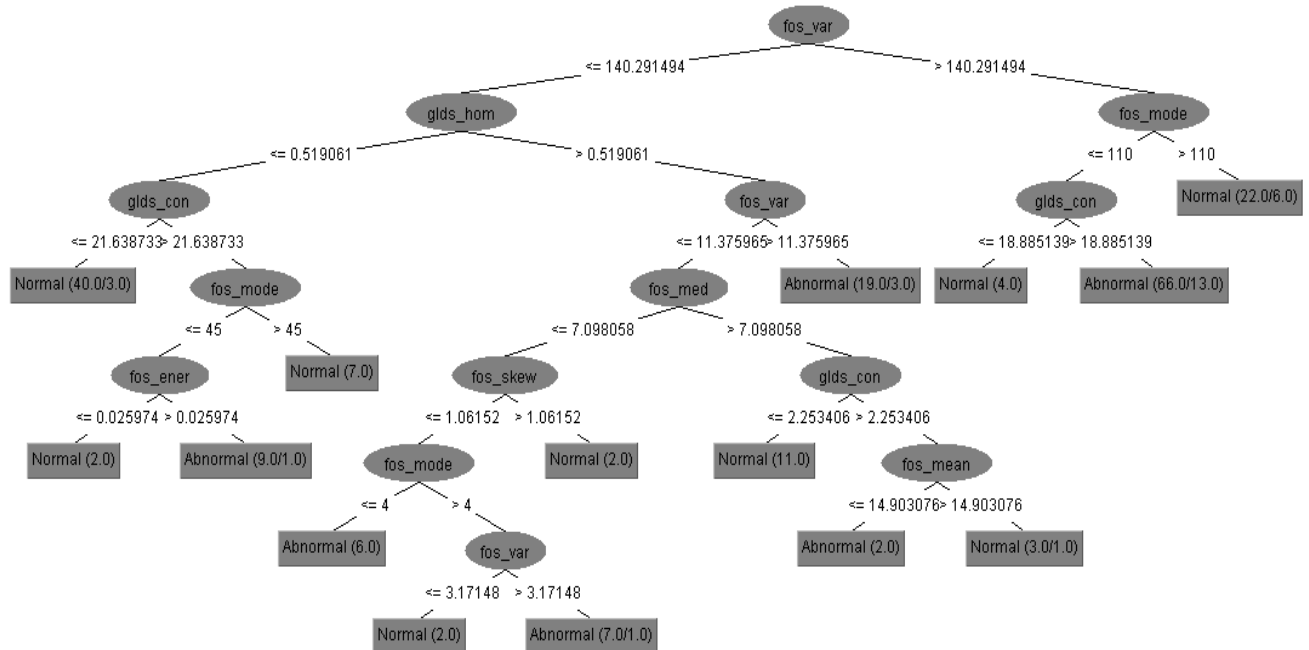
```

|      |      fos_var > 11.375965: Abnormal (19.0/3.0)
fos_var > 140.291494
|      fos_mode <= 110
|      |      glds_con <= 18.885139: Normal (4.0)
|      |      glds_con > 18.885139: Abnormal (66.0/13.0)
|      fos_mode > 110: Normal (22.0/6.0)

```

Αριθμός Φύλλων: 15

Μέγεθος Δέντρου : 29



Σχήμα 6.6 Δέντρο το οποίο δημιουργήθηκε από την εφαρμογή του αλγορίθμου C4.5 στα χαρακτηριστικά υφής που πάρθηκαν από τους αλγορίθμους SF και GLDS. Ο αριθμός που βρίσκεται σε παρένθεση υποδηλώνει των αριθμό των περιπτώσεων που βρέθηκαν να είναι Normal ή Abnormal αντίστοιχα.

	Δείγμ.Εκπ. 52% Δείγμ.Επαλ. 48%	Δείγμ.Εκπ. 60% Δείγμ.Επαλ. 40%	Δείγμ.Εκπ. 66% Δείγμ.Επαλ. 34%	Δείγμ.Εκπ. 72% Δείγμ.Επαλ. 28%
Correctly Classified Instances	59,7938%	46,9136%	65,2174%	42,1053%
Incorrectly Classified Instances	40,2062%	53,0864%	34,7826%	57,8947%
Kappa statistic	0,2037	0	0,3094	0
Mean absolute error	0.397	0.5013	0.3807	0.5049
Root mean squared error	0.5355	0.5017	0.5126	0.5058
Relative absolute error	79,3344%	100,0041%	75,7473%	100,0132%
Root relative squared error	106,975%	100,0069%	101,8488%	100,0181%
Coverage of cases (0.95 level)	87,6289%	100%	91,3043%	100%
Mean rel. region size (0.95 level)	77,3196%	100%	85,5072%	100%
Total Number of Instances	97	81	69	57

Πίνακας 6.12 Αποτελέσματα από την εφαρμογή του αλγορίθμου C4.5 στα χαρακτηριστικά υφής που πάρθηκαν από τους αλγορίθμους SF και GLDS

		TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
Δείγμα Εκπαίδευσης 52% Δείγμα Επαλήθευσης 48%		0.44	0.234	0.667	0.44	0.53	0.662	Normal
		0.766	0.56	0.563	0.766	0.649	0.662	Abnormal
	Weighted Avg	0.598	0.392	0.616	0.598	0.588	0.662	
Δείγμα Εκπαίδευσης 60% Δείγμα Επαλήθευσης 40%		0	0	0	0	0	0.5	Normal
		1	1	0.469	1	0.639	0.5	Abnormal
	Weighted Avg	0.469	0.469	0.22	0.469	0.3	0.5	
Δείγμα Εκπαίδευσης 66% Δείγμα Επαλήθευσης 34%		0.605	0.29	0.719	0.605	0.657	0.689	Normal
		0.71	0.395	0.595	0.71	0.647	0.689	Abnormal
	Weighted Avg	0.652	0.337	0.663	0.652	0.653	0.689	
Δείγμα Εκπαίδευσης 72% Δείγμα Επαλήθευσης 28%		0	0	0	0	0	0.5	Normal
		1	1	0.421	1	0.593	0.5	Abnormal
	Weighted Avg	0.421	0.421	0.177	0.421	0.25	0.5	

Πίνακας 6.13 Αποτελέσματα από την εφαρμογή του αλγορίθμου C4.5 στα χαρακτηριστικά υφής που πάρθηκαν από τους αλγορίθμους SF και GLDS

6.9 Αποτελέσματα αλγόριθμου C4.5 για τον αλγόριθμο SGLDM, SF και GLDS

Πιο κάτω παρουσιάζονται τα αποτελέσματα του αλγορίθμου C4.5 για το συνδυασμό αλγορίθμων SGLDM, SF και GLDS. Συγκεκριμένα, παρουσιάζεται το δέντρο που έχει δημιουργηθεί από την εφαρμογή του αλγορίθμου C4.5 στα δεδομένα.

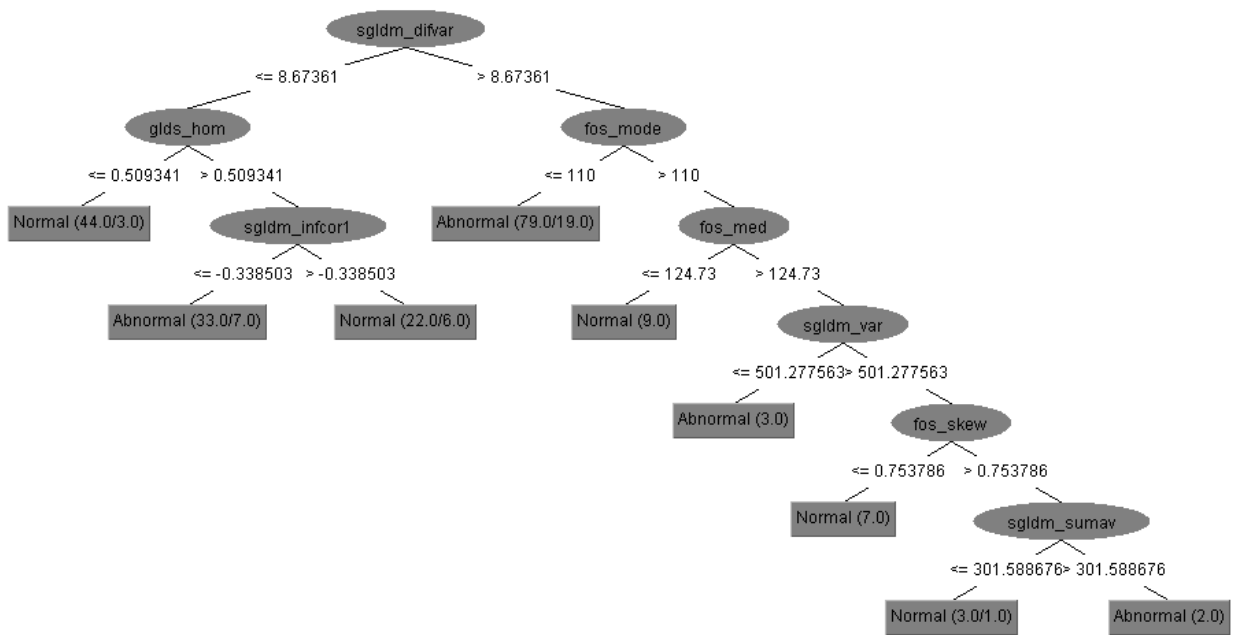
Στη συνέχεια παρουσιάζονται τα αποτελέσματα από την εφαρμογή του αλγορίθμου C4.5 για διάφορες τιμές του ποσοστού των εικόνων που δίνεται ως δείγμα εκπαίδευσης του συστήματος.

Δέντρο

```
sgldm_difvar <= 8.67361
|   glds_hom <= 0.509341: Normal (44.0/3.0)
|   glds_hom > 0.509341
|   |   sgldm_infcor1 <= -0.338503: Abnormal (33.0/7.0)
|   |   sgldm_infcor1 > -0.338503: Normal (22.0/6.0)
sgldm_difvar > 8.67361
|   fos_mode <= 110: Abnormal (79.0/19.0)
|   fos_mode > 110
|   |   fos_med <= 124.73: Normal (9.0)
|   |   fos_med > 124.73
|   |   |   sgldm_var <= 501.277563: Abnormal (3.0)
|   |   |   sgldm_var > 501.277563
|   |   |   |   fos_skew <= 0.753786: Normal (7.0)
|   |   |   |   fos_skew > 0.753786
|   |   |   |   |   sgldm_sumav <= 301.588676: Normal (3.0/1.0)
|   |   |   |   |   sgldm_sumav > 301.588676: Abnormal (2.0)
```

Αριθμός Φύλλων: 9

Μέγεθος Δέντρου : 17



Σχήμα 6.7 Δέντρο το οποίο δημιουργήθηκε από την εφαρμογή του αλγορίθμου C4.5 στα χαρακτηριστικά υφής που πάρθηκαν από τους αλγορίθμους SGLDM, SF και GLDS. Ο αριθμός που βρίσκεται σε παρένθεση υποδηλώνει των αριθμό των περιπτώσεων που βρέθηκαν να είναι Normal ή Abnormal αντίστοιχα.

	Δείγμ.Εκπ. 52% Δείγμ.Επαλ. 48%	Δείγμ.Εκπ. 60% Δείγμ.Επαλ. 40%	Δείγμ.Εκπ. 66% Δείγμ.Επαλ. 34%	Δείγμ.Εκπ. 72% Δείγμ.Επαλ. 28%
Correctly Classified Instances	67%	66%	62%	63%
Incorrectly Classified Instances	32%	33%	37%	36%
Kappa statistic	0,3	0,3	0,2	0,2
Mean absolute error	0.3	0.3	0.4	0.4
Root mean squared error	0.5	0.5	0.5	0.5
Relative absolute error	68%	73%	79%	84%
Root relative squared error	105%	109%	106%	106%
Coverage of cases (0.95 level)	85%	82%	88%	91%
Mean rel. region size (0.95 level)	75%	78%	80%	93%
Total Number of Instances	97	81	69	57

Πίνακας 6.14 Αποτελέσματα από την εφαρμογή του αλγορίθμου C4.5 στα
χαρακτηριστικά υφής που πάρθηκαν από τους αλγορίθμους SGLDM, SF και GLDS

		TP Rate	FP Rate	Precision	Recall	F- Measure	ROC Area	Class
Δείγμα Εκπαίδευσης 52% Δείγμα Επαλήθευσης 48%		0.6	0.2	0.7	0.6	0.6	0.7	Normal
		0.7	0.3	0.6	0.7	0.6	0.7	Abnormal
	Weighted Avg	0.6	0.3	0.6	0.6	0.6	0.7	
Δείγμα Εκπαίδευσης 60% Δείγμα Επαλήθευσης 40%		0.6	0.3	0.6	0.6	0.6	0.6	Normal
		0.6	0.3	0.6	0.6	0.6	0.6	Abnormal
	Weighted Avg	0.6	0.3	0.6	0.6	0.6	0.6	
Δείγμα Εκπαίδευσης 66% Δείγμα Επαλήθευσης 34%		0.5	0.2	0.7	0.5	0.5	0.6	Normal
		0.7	0.5	0.5	0.7	0.6	0.6	Abnormal
	Weighted Avg	0.6	0.3	0.6	0.6	0.6	0.6	
Δείγμα Εκπαίδευσης 72% Δείγμα Επαλήθευσης 28%		0.5	0.2	0.7	0.5	0.6	0.6	Normal
		0.7	0.4	0.5	0.7	0.6	0.6	Abnormal
	Weighted Avg	0.6	0.3	0.6	0.6	0.6	0.6	

Πίνακας 6.15 Αποτελέσματα από την εφαρμογή του αλγορίθμου C4.5 στα
χαρακτηριστικά υφής που πάρθηκαν από τους αλγορίθμους SGLDM, SF και GLDS

6.10 Συνοπτικά Αποτελέσματα

Στην συγκεκριμένη ενότητα καταγράφονται συνοπτικά τα αποτελέσματα από την εφαρμογή του αλγορίθμου C4.5 στα διάφορα δεδομένα.

Συγκεκριμένα, έχει γίνει έρευνα σχετικά με τη μεγαλύτερη τιμή του πεδίου Correctly Classified Instances (CC), για ποσοστό δείγματος εικόνων εκπαίδευσης από 50-72%, σε κάθε ένα από τους πιο κάτω συνδυασμούς αλγορίθμων:

1. SGLDM
2. SF
3. GLDS
4. SGLDM και SF
5. SGLDM και GLDS
6. SF και GLDS
7. SGLDM, SF και GLDS

Για κάθε ένα από τους παραπάνω αλγορίθμους, έχουν μελετηθεί όλες οι τιμές από το 50-72%, ως δείγμα εικόνων εκπαίδευσης του συστήματος, ώστε να βρεθεί η μέγιστη τιμή. Πιο κάτω παραθέτονται τα αποτελέσματα της έρευνας.

Αλγόριθμος				
	Μέγιστη Τιμή %CC	% Δείγμα εκπαίδευσης	Ελάχιστη Τιμή %CC	% Δείγμα εκπαίδευσης
SGLDM	72	72	48	58
SF	62	66	41,5	68
GLDS	60,5	55	41,5	68
SGLDM, SF	70	61	48	58
SGLDM, GLDS	74	72	48	58
SF, GLDS	67	65	41,5	68
SGLDM, SF, GLDS	70	61	48	58

Πίνακας 6.59 Μέγιστες και ελάχιστες τιμές του CC για ποσοστό δείγματος του συστήματος από 50% μέχρι 72% δείγμα εκπαίδευσης

Κεφάλαιο 7

Συμπεράσματα / Μελλοντική Εργασία

7.1 Επισκόπηση	76
7.2 Συμπεράσματα	77
7.2.1 Πραγματικές εικόνες ενδομητρίου	77
7.2.2 Αλγόριθμοι κατηγοριοποίησης	77
7.2.3 Σύγκριση με προηγούμενη εργασία	79
7.3 Μελλοντική Εργασία	79
7.3.1 Επιλογή κατάλληλων χαρακτηριστικών υφής	79
7.3.2 Αυτόματη ανάλυση υστεροσκοπικού βίντεο και εξαγωγή αποτελεσμάτων	80

7.1 Επισκόπηση

Στο κεφάλαιο αυτό καταγράφονται τα κύρια συμπεράσματα που έχουν εξαχθεί από την εφαρμογή του αλγορίθμου C4.5 στα χαρακτηριστικά υφής των 3 αλγορίθμων SGLDM,SF και GLDS.

Επίσης, παρουσιάζονται πληροφορίες για τη μελλοντική έρευνα και τις προοπτικές που υπάρχουν στον τομέα της ιατρικής.

7.2 Συμπεράσματα

7.2.1 Πραγματικές εικόνες ενδομητρίου

Έχοντας συλλέξει σε πραγματικό χρόνο εικόνες ενδομητρίου, με φυσιολογικές και παθολογικές περιοχές ενδιαφέροντος, η ανάλυσή τους έγινε με την ίδια σειρά, δηλαδή η εξαγωγή των χαρακτηριστικών υφής των δυο κατηγοριών και η στατιστική ανάλυσή τους. Μετά οι εικόνες έτυχαν διόρθωσης με τον αλγόριθμο γάμμα και ακολούθησε η ίδια μεθοδολογία. Η στατιστική ανάλυση έδειξε ότι υπάρχει στατιστική διαφορά σε αρκετά χαρακτηριστικά υφής πριν και μετά τη διόρθωση με τον αλγόριθμο γάμμα.

Ο αλγόριθμος γάμμα εφαρμόστηκε με τη συλλογή πραγματικών εικόνων από διαφορετικά ιατρικά μηχανήματα. Για να γίνει κατορθωτή η σύγκριση εικόνων από διαφορετικά μηχανήματα οι εικόνες διορθώθηκαν με τον αλγόριθμο γάμμα και έτσι έγινε κατορθωτό η δημιουργία ενός συστήματος ανάλυσης ενδοσκοπικών εικόνων ανεξάρτητα από τα ιατρικά μηχανήματα [3].

Με την ανάλυση των εικόνων και την επεξεργασία των χαρακτηριστικών υφής παρατηρήθηκε ότι οι δύο περιοχές ενδιαφέροντος, φυσιολογικός και παθολογικός ιστός μπορούν να διαχωριστούν αυτόματα εισάγοντας τα χαρακτηριστικά υφής στον C4.5 αλγόριθμο με ένα ενδιαφέρον ποσοστό επιτυχίας (74%) σχετικά με τα χαρακτηριστικά υφής των αλγορίθμων SGLDM και GLDS.

7.2.2 Αλγόριθμος κατηγοριοποίησης

Έχοντας αναλύσει τις εικόνες από πραγματικό ενδομήτριο, έχουν εξαχθεί τα χαρακτηριστικά υφής που περιγράφουν τις περιοχές ενδιαφέροντος. Έχοντας περιγράψει τις εικόνες βάσει αυτών των χαρακτηριστικών υφής εφαρμόσαμε τον αλγόριθμο κατηγοριοποίησης C4.5 με σκοπό τον αυτόματο διαχωρισμό των περιοχών ενδιαφέροντος. Η χρήση του αλγορίθμου C4.5 έδειξε τη μεγάλη διαχωριστικότητα με ποσοστό 73,6842% των περιοχών ενδιαφέροντος, φυσιολογικού και παθολογικού ιστού. Το ποσοστό 73,6842% έχει επιτευχθεί με τη χρήση του μοντέλου βάσης δεδομένων Y1 το οποίο περιλαμβάνει τα 26 χαρακτηριστικά υφής των εικόνων (μετά από διόρθωση τους από τον αλγόριθμο γάμμα) των αλγορίθμων SGLDM και GLDS, οι οποίες είναι

μετασχηματισμένες στο χρωματικό κανάλι Υ. Το δεύτερο καλύτερο αποτέλεσμα έχει επιτευχθεί με τη χρήση των χαρακτηριστικών υφής από τον αλγόριθμο SGLDM με ποσοστό επιτυχίας 72% δεδομένου ότι έχουμε δώσει στο σύστημα 72% δείγμα από τις εικόνες, ώστε να εκπαιδευτεί.

Τα αποτελέσματα αυτά έδειξαν τη μελλοντική χρήση ανάλυσης εικόνας σε περιπτώσεις γυναικολογικού καρκίνου αλλά και σε άλλες εφαρμογές της ιατρικής όπως την ενδοσκόπηση και ανάλυση εικόνων ήπατος, μορφές καρκίνου με παρακολούθηση με χρήση ιατρικών απεικονιστικών μεθόδων.

Η έρευνα μας έδειξε ότι η επεξεργασία ιατρικής εικόνας μπορεί να παίζει σημαντικό ρόλο στην αναγνώριση παθολογικών και φυσιολογικών εικόνων ενδομητρίου και άλλων οργάνων. Αυτό θα δημιουργήσει αυτόματα συστήματα υπολογιστών τα οποία θα βοηθήσουν τον γιατρό σε πραγματικό χρόνο να δώσει μια πρώτη εκτίμηση του περιστατικού ή θα τον καθοδηγήσει στον έλεγχο με ιστοπαθολογική εξέταση υπόπτων περιοχών καρκίνου.

Ουσιαστικά έχοντας αφενός υπόψη τα ενδιαφέροντα αποτελέσματα της έρευνας αυτής αφετέρου πρέπει να σημειωθεί πως παρόλο που η εξόρυξη δεδομένων μπορεί να παρέχει σημαντικές πληροφορίες και υποστήριξη στο ιατρικό προσωπικό αναγνωρίζοντας ύποπτες περιοχές που δεν γίνονται άμεσα αντιληπτές υπάρχει όμως περιορισμός στο τι μπορεί να καταφέρει η εξόρυξη δεδομένων. Δεν είναι όλα τα αποτελέσματα μιας εξόρυξης δεδομένων ενδιαφέρον. Για να είναι ένα αποτέλεσμα ενδιαφέρον πρέπει να είναι λογικό και να μπορεί να τύχει επεξεργασίας (actionable). Γι' αυτό και η εξόρυξη δεδομένων χρειάζεται την ανθρώπινη επέμβαση για την εκμετάλλευση της εξορυχθείσας γνώσης. Για παράδειγμα η εξόρυξη δεδομένων μπορεί να παρέχει βοήθεια στην ολοκλήρωση μίας διάγνωσης ή την περιγραφή μίας θεραπείας αλλά δεν μπορεί να αντικαταστήσει την διαίσθηση και τις ερμηνευτικές δεξιότητες (interpretive skills) του γιατρού [23].

7.2.3 Σύγκριση με προηγούμενη εργασία

Όπως έχει προαναφερθεί, η παρούσα εργασία βασίζεται στη μελέτη που έκανε η μεταπτυχιακή φοιτήτρια Αντωνία Λοΐζου τον Ιούνιο του 2007. Σύμφωνα με τη μελέτη και τα αποτελέσματα της Αντωνίας, η χρήση του αλγόριθμου C4.5 έδειξε τη μεγάλη διαχωρισιμότητα με ποσοστό 78,7% των περιοχών ενδιαφέροντος, φυσιολογικού και παθολογικού ιστού ενώ ο SQL Server 2005 Decision Tree είχε μικρή διαφορά αφού έφθασε στο 77.2% στις μετασχηματισμένες στο χρωματικό κανάλι Υ εικόνες ενδομήτριου. Το ποσοστό 78,7% είχε επιτευχθεί με τη χρήση του μοντέλου βάσης δεδομένων Y1 το οποίο περιλάμβανε τα 26 χαρακτηριστικά υφής των εικόνων (μετά από διόρθωση τους από τον αλγόριθμο γάμμα) οι οποίες είναι μετασχηματισμένες στο χρωματικό κανάλι Υ.

Σύμφωνα με τα αποτελέσματα της παρούσας εργασίας, ο αλγόριθμος C4.5 έδειξε ποσοστό διαχωρισιμότητα 74% των περιοχών ενδιαφέροντος, ποσοστό το οποίο δικαιολογείται αφού έχουν χρησιμοποιηθεί συνολικά 202 εικόνες ενδομητρίου, σε αντίθεση με την παραπάνω εργασία που είχαν χρησιμοποιηθεί 404 εικόνες ενδομητρίου.. Είναι σαφές, ότι όσο μεγαλύτερος είναι ο αριθμός εικόνων, τόσο πιο καλά εκπαιδεύεται το σύστημα, και κατά συνέπεια δίνει καλύτερα αποτελέσματα (μεγαλύτερο ποσοστό διαχωρισιμότητας).

7.3 Μελλοντική εργασία

7.3.1 Επιλογή κατάλληλων χαρακτηριστικών υφής

Η επιλογή του σωστού συνόλου χαρακτηριστικών (features) για την κατηγοριοποίηση/ ταξινόμηση είναι ένα από τα σημαντικότερα προβλήματα στο σχεδιασμό ενός καλού κατηγοριοποιητή (classifier). Οι αλγόριθμοι κατασκευής δέντρων απόφασης (όπως για παράδειγμα ο C4.5) έχουν ενσωματώσει στη φάση εκπαίδευσης τους μια αυτόματη στρατηγική επιλογής χαρακτηριστικών ενώ κάποιοι άλλοι στατιστικοί αλγόριθμοι κατηγοριοποίησης απαιτούν το υποσύνολο χαρακτηριστικών να επιλεχτεί σε μια φάση προεπεξεργασίας δεδομένων. Είναι γνωστό ότι άσχετα χαρακτηριστικά μπορούν να υποβιβάσουν την απόδοση κάποιου

αλγόριθμου. Θα ήταν καλό στο μέλλον να γίνει μία αξιολόγηση της επιρροής της προεπιλογής χαρακτηριστικών στην ακρίβεια πρόβλεψης των δύο αλγορίθμων μας. Ο αριθμός χαρακτηριστικών υφής που χρησιμοποιούνται για την κατηγοριοποίηση των εικόνων μας θα μπορούσε ίσως να μειωθεί κάτι που είναι σημαντικό για την επεξεργασία εικόνας αφού ο υπολογισμός χαρακτηριστικών είναι μια χρονοβόρα διαδικασία. Οι χρήστες δέντρων απόφασης και οι ερευνητές έχουν αναγνωρίσει τον αντίκτυπο της εφαρμογής ενός πλήρους συνόλου χαρακτηριστικών σε μια διαδικασία κατασκευής δέντρων απόφασης εναντίον της εφαρμογής μόνο ενός συνετά επιλεγμένου υποσυνόλου. Είναι συχνή η περίπτωση το τελευταίο να παράγει δέντρα απόφασης με χαμηλότερα λάθη κατηγοριοποίησης, ιδιαίτερα όταν επιλεχτεί το υποσύνολο από έναν domain εμπειρογνώμονα. [25] Στην μελέτη αυτή έχει χρησιμοποιηθεί ένας συνδυασμός χαρακτηριστικών υφής τα οποία έχουν εξαχθεί από τρεις αλγόριθμους χαρακτηριστικών υφής (βλ. Κεφάλαιο 3). Ο συνδυασμός αυτός είναι καλό να παραμείνει και ως μελλοντική εργασία προτείνεται η επιλογή ενός υποσυνόλου χαρακτηριστικών υφής από αυτό τον συνδυασμό 26 χαρακτηριστικών υφής που έχουμε χρησιμοποιήσει. Αυτό θα οδηγήσει και στην μείωση των διαστάσεων των βάσεων δεδομένων ιατρικών εικόνων. Τα μελλοντικά πειράματά μας προορίζεται να αξιολογήσουν την επίδραση της χρησιμοποίησης του C4.5 ως μοντέλο επιλογής ορισμένων χαρακτηριστικών [26] από τα 26 χαρακτηριστικά μας ως βήμα προεπεξεργασίας δεδομένων πριν την κατασκευή των δέντρων απόφασης για βελτίωση της ακρίβειας της πρόβλεψης.

7.3.2 Αυτόματη ανάλυση υστεροσκοπικού βίντεο και εξαγωγή αποτελεσμάτων

Με την παρουσίαση της έρευνάς μας και με την ανάπτυξη της τεχνολογίας, νέες εφαρμογές μπορούν να υλοποιηθούν στο πεδίο της αυτόματης αναγνώρισης ύποπτων περιοχών σε ενδοσκοπικές εικόνες. Θα μπορούσε να υλοποιηθεί ένα σύστημα αυτόματης ανάλυσης υστεροσκοπικού βίντεο και εξαγωγής αποτελεσμάτων (εξόρυξης δεδομένων). Με το σύστημα αυτό να καταγράφεται σε πραγματικό χρόνο η επέμβαση και να αναλύονται σε κλάσματα δευτερολέπτου όλα τα frames για τυχόν ανωμαλίες που μπορεί να παρουσιάσουν τα όργανα.

Θα μπορούσε να παρουσιάζεται στην οθόνη το ενδοσκοπικό βίντεο και να αναλύει παράθυρο προς παράθυρο όλη την επιφάνειά του ιστού που θα καταγράφεται.

Ακολούθως εισάγοντας τα χαρακτηριστικά υφής, των κατάλληλα μετασχηματισμένων εικόνων (που όπως δείχνει η έρευνα μας θα ήταν καλύτερα ο μετασχηματισμός αυτός να είναι στο χρωματικό κανάλι Υ ή στο σύστημα απεικόνισης YCrCb), σε ένα ειδικά υλοποιημένο περιβάλλον με τους αλγόριθμους τους οποίους χρησιμοποιήσαμε να διαχωρίζονται ύποπτες περιοχές γυναικολογικού καρκίνου καθώς και να υπολογίζεται με μεγάλη ακρίβεια το ποσοστό παθολογίας που παρουσιάζει το κάθε σημείο που θα εξετάζεται ή ακόμα και ένα ενδεικτικό όριο επικινδυνότητας του περιστατικού πριν αυτό παρουσιάσει παθολογία.

Με αυτή τη διαδικασία όλες οι περιοχές του εκάστοτε οργάνου θα καταγράφονται με ακρίβεια και παράλληλα με την ιστολογική εξέταση θα υπάρχει αύξηση της διαγνωστικής ακρίβειας για όλα τα περιστατικά γυναικολογικού καρκίνου.

Αυτό θα βοηθήσει το γιατρό να επιλέγει και μη φαινομενικά ύποπτες περιοχές έτσι ώστε να οδηγείται σε καλύτερη διάγνωση και θεραπεία του περιστατικού. Ακόμη οι αναλύσεις αυτές και το βίντεο θα ήταν καλό να μπορούν να μεταφέρονται ασύρματα και ενσύρματα σε ένα δεύτερο ειδικό γιατρό για μια δεύτερη γνώμη με στόχο την πιο αξιόπιστη ανάλυση του περιστατικού και της ιδιομορφίας του ιστού.

Επίσης έξυπνες βάσεις δεδομένων με έξυπνα συστήματα εξόρυξης δεδομένων μπορούν σε πραγματικό χρόνο να αναλύσουν και να συγκρίνουν παρόμοια περιστατικά για να εξάγουν συμπεράσματα για το γυναικολογικό καρκίνο δημιουργώντας έτσι ένα δίκτυο μεταφοράς και ανάλυσης παρόμοιων περιστατικών σε όλο τον κόσμο.

Μια άλλη επιλογή για την βελτίωση της ακρίβεια πρόβλεψης θα ήταν ένα υβριδικό ευφυές σύστημα όπου τα αποτελέσματα πρόβλεψης, των μεθόδων εξόρυξης δεδομένων, να βελτιώνονται από τις απόψεις των ειδικών (γιατρών), οι οποίες θα ενσωματώνονται στο σύστημα αυτό.

Τέλος το σύστημα θα μπορούσε να είναι διαθέσιμο στο ευρύ κοινό μέσω ενός διαδικτυακού χώρου που θα χρησιμοποιείται ως ένα διαδικτυακό (web-based) σύστημα υποστήριξης απόφασης (decision support system - DSS). Σε αυτό το

διαδικτυακό DSS, θα μπορούσαν να ενσωματωθούν περισσότερα από ένα μοντέλα πρόβλεψης καθώς και τα υβρίδιά τους όχι μόνο για την πρόβλεψη του καρκίνου του ενδομητρίου αλλά για όλους τους τύπους καρκίνου.

Επαγωγικά λοιπόν σκεπτόμενοι καταλήγουμε στο συμπέρασμα ότι παρόλο που οι μέθοδοι εξόρυξης δεδομένων είναι σε θέση να εξάγουν πρότυπα και σχέσεις που κρύβονται βαθιά μέσα σε μεγάλα ιατρικά σύνολα δεδομένων, χωρίς τη συνεργασία και την ανατροφοδότηση από τον ιατρό, τα αποτελέσματά τους ίσως και να μην μπορούν να χρησιμοποιηθούν. Τα πρότυπα που βρίσκονται μέσω των μεθόδων εξόρυξης δεδομένων θα πρέπει να αξιολογούνται από ιατρούς που έχουν πολλά χρόνια εμπειρίας στον τομέα προβλήματος του καρκίνου για να αποφασίσουν εάν είναι λογικά, μπορούν να χρησιμοποιηθούν και είναι νέα (δεν έχουν εξαχθεί προηγουμένως) και έτσι θα μπορούν να τροφοδοτήσουν νέες βιολογικές και κλινικές ερευνητικές κατευθύνσεις. Άρα ένα σύστημα ενσωμάτωσης των δύο γνώσεων (εξόρυξης δεδομένων και ιατρού) είναι απαραίτητο. Εν ολίγοις, η εξόρυξη δεδομένων δεν στοχεύει να αντικαταστήσει τους ιατρούς και τους ερευνητές, αλλά να συμπληρώσει τις ανεκτίμητες προσπάθειές τους να σώσουν περισσότερες ανθρώπινες ζωές [23].

Βιβλιογραφία

- [1] Λοΐζου Α. Εξόρυξη Δεδομένων για την καταγοριοποίηση του Ενδομητρίου σε υστεροσκοπική εξέταση, Διατριβή Μεταπτυχιακού, Τμήμα Πληροφορικής, Πανεπιστήμιο Κύπρου, Λευκωσία, 2007.
- [2] Νεοφύτου Μ. Σ., Διδακτορική Διατριβή - Ανάπτυξη και Αξιολόγηση Ολοκληρωμένου Συστήματος Ψηφιακής Ανάλυσης Ενδοσκοπικής Εικόνας Μακροβιοψίας για την Υποστήριξη διάγνωσης γυναικολογικού Καρκίνου, Εθνικό Μετσόβιο Πουτεχνείο, Αθήνα 2007.
- [3] **American Society for Gastrointestinal Endoscopy** <http://www.asge.org>
- [4] Tjoa P.M., Krishnan M.S.: **Feature extraction for the analysis of colon status from the endoscopic images**. BioMedical Engineering OnLine, Apr. 2003. <http://www.biomedical-engineering-online.com/content/2/1/9>.
- [5] Plataniotis K.N., Venetsanopoulos A.N., **Color Image Processing and Applications**, Springer Verlag. Berlin, ISBN 3-540-66953-1, August 2000.
- [6] Gonzalez, Woods: **Digital Image Processing**, 2nd Edition, Prentice Hall, 2002.
- [7] Bankman I.N., Spisz T.S., Pavlopoulos S., **Two-dimensional shape and texture quantification**, Handbook of medical imaging, Academic Press, Inc, October 2000.
- [8] Karkanis S.A., Galousi K., Maroulis D., **Classification of Endoscopic Images Based on Texture Spectrum**. ACAI99, Workshop on Machine Learning in Medical Applications, Chania, Greece pp. 63-69, 5-16 July 1999.
- [9] Karkanis S.A., Iakovidis D.K., Maroulis D.E., Karras A.D., Tzivras M., **Computer- aided tumor detection in endoscopic video using color**

- wavelet features.** IEEE Transactions on Information Technology in Biomedicine, 7(3), pp. 141-152, 2003.
- [10] Houston A.G., Premkumar S. B., **Statistical interpretation of texture for medical applications.** Presented at the Biomedical Image Processing and Three Dimensional Microscopy, San Jose, CA, 1991.
 - [11] Enderwick C., Micheli-Tzanakou E., **Classification of mammographic tissue using shape and texture features.** Proc. 19th Annu. Int. Conf. IEEE Engineering Medicine Biology Society, pp. 810–813, 1997.
 - [12] Lachmann F., Barillot C., **Brain tissue classification from MRI data by means of texture analysis.** In Proc. Medical Imaging VI: Image Processing, vol. 1652. Newport Beach, CA, 1992, pp. 72–83.
 - [13] Ji Q., Engel J., Craine E., **Texture analysis for classification of cervix lesions.** IEEE Trans. Med. Imag., vol. 19, pp. 1144–1149, Nov. 2000.
 - [14] Fortin C., Ohley W., **Automatic segmentation of cardiac images: Texture mapping.** Proc. IEEE 17th Annu. Northeast Bioeng. Conf., 1991.
 - [15] Ilgner J., Palm C., Schutz A., Spitzer K., Westhofen M., Lehmann T., **Colour Texture Analysis for Quantitative Laryngoscopy,** 123: pp. 730-734, Acta Otolaryngol 2003.
 - [16] Haralick R.M., Shanmugam K., Dinstein I., **Texture Features for Image Classification.** IEEE Trans. on Systems, Man., and Cybernetics, SMC-3, pp. 610-621, Nov. 1973.
 - [17] Wu C.M., Chen Y.C., Hsieh K.S., **Texture features for classification of ultrasonic liver images.** IEEE Trans. Med. Imaging, 11, pp.141-152, 1992.
 - [18] Quinlan J. R., **C4.5: Programs for Machine Learning.** San Mateo, CA:

Morgan, Kaufmann, 1993.

- [19] Vanden B. F., **Classification Trees: C4.5**. IRIDIA, Universit Libre de Bruxelles, 2003.
- [20] Web link: <http://www.acmicorp.com>
- [21] WEKA Documentation:
<http://ignum.dl.sourceforge.net/project/weka/documentation>
- [22] WEKA : <http://www.cs.waikato.ac.nz/ml/weka/>
- [23] Douglas CW, McNeil BJ, **Clinical decision analysis methods applied to diagnostic tests in dentistry**, J DentEduc41:10S-1\2, 1983.
- [24] Richards G, Rayward-Smith VJ, Sonksen PH, Carey S, Weng C. **Data mining for indicators of early mortality in a database of clinical records**. Artif Intell Med 2001;22:215-31.
- [25] P. Perner, **Improving the Accuracy of Decision Tree Induction by Feature Pre- Selection**, Applied Artificial Intelligence 2001, vol. 15, No. 8, p. 747-760.
- [26] Marcela X. Ribeiro, André G. R. Balan, Joaquim C. Felipe, Agma J. M. Traina, Caetano Traina Jr, **Mining Statistical Association Rules to Select the Most Relevant Medical Image Features**, 2009, Volume 165/2009, 113-131, DOI: 10.1007/978-3-540-88067-7_7