

Σχολή Επιστήμης της Πληροφορίας & Πληροφορικής



ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Τμήμα Πληροφορικής

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

**Εξόρυξη Δεδομένων και Συστήματα Υποστήριξης
Αποφάσεων για την διάγνωση του Καρκίνου του
Τραχήλου της Μητρας**



*Clinical Decision Support System
for Cervical Cancer*

Επιβλέπων Καθηγητής: Έξαρχος Θεμιστοκλής

Ονοματεπώνυμο: Μαγουνάκη Ουρανία

A.M : Π2015140

Κέρκυρα, Σεπτέμβριος 2019

Περιεχόμενα

Ευχαριστίες	6
Πρόλογος	7
Εισαγωγή	8
ΚΕΦΑΛΑΙΟ 1^ο: Βιβλιογραφική Ανασκόπηση	10
1.1 Τεχνική Μηχανικής Μάθησης για την ανίχνευση καρκίνου του τραχήλου της μήτρας χρησιμοποιώντας k-NN και Τεχνητό Νευρωνικό Δίκτυο	10
1.2 Πρόγνωση του καρκίνου του τραχήλου της μήτρας με την χρήση της Εξόρυξης δεδομένων	12
1.3 Σύστημα Υποστήριξης Κλινικών αποφάσεων με αυτοματοποιημένη επεξεργασία κειμένου για την πρόγνωση του καρκίνου του τραχήλου της μήτρας	16
1.4 Αυτοματοποιημένη σύσταση για την εξέταση και επιτήρηση του καρκίνου του τραχήλου της μήτρας	20
1.5 Ένα ευφυές σύστημα υποστήριξης κλινικών αποφάσεων για συγκεκριμένες προβλέψεις ασθενών για την βελτίωση ανίχνευσης νεοπλασιών του τραχήλου της μήτρας	22
1.6 Σύστημα υποστήριξης αποφάσεων με χρήση NLP για τον έλεγχο και την επιτήρηση του καρκίνου του τραχήλου της μήτρας	25
1.7 Εφαρμογή της βαθιάς μάθησης για την ταξινόμηση εικόνων από κολποσκόπηση	27
1.8 Προσδιορισμός της πιθανότητας καρκίνου του τραχήλου της μήτρας με χρήση μεθόδων μηχανικής μάθησης	28
1.9 Εύρεση συμπτωμάτων καρκίνου του τραχήλου της μήτρας στο σουηδικό κλινικό κείμενο χρησιμοποιώντας μια προσέγγιση μηχανικής μάθησης και NegEx	29
1.10 Ένα σύστημα υποστήριξης αποφάσεων βασισμένο σε ένα σύνολο Random Forests για τη βελτίωση της διαχείρισης των γυναικών με μη φυσιολογικά ευρήματα στην εξέταση του καρκίνου του τραχήλου της μήτρας	31
1.11 Αξιολόγηση της απόδοσης των τεχνικών μηχανικής μάθησης για τον έλεγχο του καρκίνου του τραχήλου της μήτρας	33

1.12 Παρακολούθηση ενσωματώσεων βαθιάς μάθησης για την πρόβλεψη του καρκίνου του τραχήλου της μήτρας	35
1.13 Πρόβλεψη του καρκίνου του τραχήλου της μήτρας με χρήση υβριδικής τεχνικής επαγωγής: Μια λύση για τα μοντέλα κληρονομικής νόσου του ανθρώπου	36
1.14 Έλεγχος του καρκίνου του τραχήλου της μήτρας με χρήση τεχνητής νοημοσύνης βασισμένη στην ανάλυση των ψηφιοποιημένων εικόνων κηλίδας Παπανικολάου ...	38
1.15 Ένα κβαντικό υβριδικό PSO σε συνδυασμό με την ασαφή προσέγγιση k-NN για την επιλογή χαρακτηριστικών και την ταξινόμηση κυττάρων στην ανίχνευση του καρκίνου του τραχήλου της μήτρας	40
1.16 Βελτίωση της ακρίβειας ενός συστήματος υποστήριξης κλινικών αποφάσεων για την εξέταση και επιτήρηση του καρκίνου του τραχήλου της μήτρας	41
ΚΕΦΑΛΑΙΟ 2^ο : Περιγραφή Συνόλου Δεδομένων, Μεθοδολογίας και Αλγορίθμων ...	44
2.1 Σύνολο Δεδομένων	44
2.2 Μεθοδολογία	44
2.3 Αποτελέσματα	45
2.4 Περιγραφή Αλγορίθμων	46
2.4.1 J48/C4.5	46
2.4.2 NaiveBayes	46
2.4.3 K-NN	47
2.4.4 Random Forest	48
ΚΕΦΑΛΑΙΟ 3^ο : Ολοκληρωμένο Σύστημα	57
Συζήτηση/Συμπεράσματα	60
Παράρτημα Κώδικας	61
Βιβλιογραφία	70

Κατάλογος Εικόνων και Πινάκων

Εικόνα 1.1.α : Κανονικό Κύτταρο	11
Εικόνα 1.1.β : Κατακερματισμένο Κύτταρο	11
Εικόνα 1.2.α : Πυραμίδα DIKW (Data Information Knowledge Wisdom)	13
Πίνακας 1.2.α : Πίνακας Χαρακτηριστικών	15
Εικόνα 1.3.α : Συνδεσιμότητα Συστήματος με Ηλεκτρονικό Αρχείο Υγείας	17
Εικόνα 1.3.β : Αναγνώριση Παραμέτρων	18
Εικόνα 1.4 : Εξαγωγή εννοιών από έξι διαφορετικές πηγές	21
Πίνακας 1.5 : Ταξινόμηση Ακρίβειας στα σετ εκπαίδευσης, επικύρωσης και στα σετ δοκιμής για τους 6 ταξινομητές που δημιουργήθηκαν (προσέγγιση μεμονωμένου ταξινομητή)	25
Εικόνα 1.10α : Αποτελέσματα Μελέτης (SN, SP, PPV, NPV, OA)	32
Πίνακας 1.11.α : Συνολικός αριθμός τύπων εικόνων Παπανικολάου για φυσιολογικά και μη κύτταρα	34
Πίνακας 1.11.β : Αποτελέσματα Αλγορίθμων	34
Πίνακας 1.13 : Αποτελέσματα πρόβλεψης αλγορίθμων	37
Πίνακας 1.14 : Αποτελέσματα για 10 folds cross επικύρωση για διαφορετικούς αλγορίθμους	39
Πίνακας 1.16 : Αποτελέσματα ακρίβειας συστήματος	42
Εικόνα 2.3.2 Τύποι Δεσμευμένης Πιθανότητας	46
Εικόνα 2.3.3 Διάγραμμα για $k=3$ πλησιέστερους γείτονες	47
Εικόνα 2.3.4 : Random Forest	48
Εικόνα 2.4.1 : Τα πιο συχνά χαρακτηριστικά δεδομένων με εφαρμογή SMOTE	48
Εικόνα 2.4.2 : Τα πιο συχνά χαρακτηριστικά των δεδομένων χωρίς SMOTE	49
Εικόνα 2.4.3 : Ακρίβεια αλγορίθμων για 250 φυσιολογικά και 250 παθολογικά	50
Εικόνα 2.4.4 : Ακρίβεια αλγορίθμων για 220 φυσιολογικά και 220 παθολογικά	50
Εικόνα 2.4.5 : Ακρίβεια αλγορίθμων για 165 φυσιολογικά και 165 παθολογικά	51
Εικόνα 2.4.6 : Ακρίβεια αλγορίθμων για 55 φυσιολογικά και 55 παθολογικά	51
Εικόνα 2.4.7 : Ακρίβεια αλγορίθμων για 165 φυσιολογικά και 55 παθολογικά	52
Εικόνα 2.4.8 : Ακρίβεια αλγορίθμων για 250 φυσιολογικά και 55 παθολογικά	52
Εικόνα 2.4.9 : Ακρίβεια αλγορίθμων για 220 φυσιολογικά και 55 παθολογικά	53
Εικόνα 2.4.10 : Αποτελέσματα ευαισθησίας και πιστότητας του αλγορίθμου Naive Bayes	53
Εικόνα 2.4.11 : Αποτελέσματα ευαισθησίας και πιστότητας του αλγορίθμου J48	53
Εικόνα 2.4.12 : Αποτελέσματα ευαισθησίας και πιστότητας του αλγορίθμου Random Forest	53
Εικόνα 2.4.13 : Αποτελέσματα ευαισθησίας και πιστότητας του αλγορίθμου K-NN	54
Εικόνα 2.4.14 : Αποτελέσματα ευαισθησίας και πιστότητας του αλγορίθμου Naive Bayes	54
Εικόνα 2.4.15 : Αποτελέσματα ευαισθησίας και πιστότητας του αλγορίθμου J48	54
Εικόνα 2.4.16 : Αποτελέσματα ευαισθησίας και πιστότητας του αλγορίθμου Random Forest	55
Εικόνα 2.4.17 : Αποτελέσματα ευαισθησίας και πιστότητας του αλγορίθμου K-NN	55
Εικόνα 2.4.18 : Αποτελέσματα εξειδίκευσης αλγορίθμων για τα δεδομένα χωρίς SMOTE	55
Εικόνα 2.4.19 : Αποτελέσματα εξειδίκευσης αλγορίθμων για τα δεδομένα χωρίς SMOTE	55
Εικόνα 2.4.20 : Αποτελέσματα εξειδίκευσης αλγορίθμων για τα δεδομένα με SMOTE	56
Εικόνα 2.4.21 : Συνοπτικά αποτελέσματα Αλγορίθμων	56

<i>Εικόνα 2.4.22 : Δέντρο J48</i>	<i>56</i>
<i>Εικόνα 3.1 : Αρχική σελίδα συστήματος</i>	<i>57</i>
<i>Εικόνα 3.2 : Μεγάλη πιθανότητα εμφάνισης καρκίνου</i>	<i>57</i>
<i>Εικόνα 3.3 : Μικρή πιθανότητα εμφάνισης καρκίνου</i>	<i>58</i>
<i>Εικόνα 3.4 : Αναδυόμενο παράθυρο συμπλήρωσης όλων των ερωτήσεων από τον χρήστη</i>	<i>58</i>
<i>Εικόνα 3.5 : Σελίδα με τα στατιστικά στοιχεία</i>	<i>59</i>

*“Μια αστραπή η ζωή μας...
...μα προλαβαίνουμε”
Νίκος Καζαντζάκης*

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Θεμιστοκλή Έξαρχο για την εμπιστοσύνη που μου έδειξε να συνεργαστούμε για την διεκπεραίωση της πτυχιακής μου εργασίας . Επίσης, για τις συμβουλές που μου παρείχε καθόλη την διάρκεια αλλά και για τον χρόνο που διέθεσε προκειμένου να με καθοδηγήσει και να με βοηθήσει σε οποιοδήποτε απορία ή πρόβλημα αντιμετώπισα .

Επιπλέον, θα ήθελα να ευχαριστήσω όλους τους καθηγητές για τις γνώσεις που μου προσέφεραν κατά την διάρκεια των φοιτητικών μου χρόνων.

Εν κατακλείδι, ένα τεράστιο ευχαριστώ στην οικογένεια μου για την στήριξη τους όλα αυτά τα χρόνια.

Μαγουνάκη Ουρανία
Κέρκυρα, 2019

Πρόλογος

Η παρούσα πτυχιακή εργασία αφορά την υλοποίηση ενός κλινικού συστήματος υποστήριξης αποφάσεων για την διάγνωση του καρκίνου του τραχήλου της μήτρας.

Αρχικά, στο πρώτο κεφάλαιο παρουσιάζεται σχετική βιβλιογραφική ανασκόπηση η οποία περιλαμβάνει έρευνες που έχουν γίνει χρησιμοποιώντας την μηχανική μάθηση και όχι μόνο.

Στο δεύτερο κεφάλαιο περιγράφεται το σύνολο δεδομένων καθώς και η διαδικασία που ακολουθήθηκε για την υλοποίηση του συστήματος αλλά και η περιγραφή των αλγορίθμων.

Στο τρίτο κεφάλαιο παρουσιάζεται το σύστημα το οποίο αναπτύχθηκε με τις γλώσσες HTML, CSS και Javascript.

Εισαγωγή

Αντικείμενο της παρούσας εργασίας είναι η μελέτη συστημάτων υποστήριξης αποφάσεων (DSS) καθώς και η χρήση τεχνικών εξόρυξης δεδομένων για την πρόληψη του καρκίνου του τραχήλου της μήτρας.

Ο καρκίνος του τραχήλου της μήτρας επηρεάζει τα κύτταρα που καλύπτουν το κάτω μέρος της μήτρας (τράχηλος). Υπάρχουν διάφοροι τύποι καρκίνου του τραχήλου της μήτρας όπως το καρκίνωμα πλακωδών κυττάρων και το αδenoκαρκίνωμα το οποίο αναπτύσσεται στα κύτταρα των αδένων.

Όσον αφορά τα συστήματα υποστήριξης αποφάσεων αποτελούν ένα είδος πληροφοριακών συστημάτων τα οποία υποστηρίζουν την λήψη ημιδομημένων και αδόμητων αποφάσεων με την χρήση αλγορίθμων. Επισημαίνεται ακόμη ότι για την λήψη αποφάσεων τα CDSS χρησιμοποιούν μεθόδους Τεχνητής Νοημοσύνης. Προκειμένου να γίνει πλήρως κατανοητός ο τρόπος που λειτουργούν τα DSS είναι αναγκαίο να διευκρινιστεί ο όρος της Τεχνητής Νοημοσύνης στην οποία υπόκειται και η Μηχανική Μάθηση. Αντικείμενο της είναι η σχεδίαση και η υλοποίηση συστημάτων τα οποία διαχειρίζονται μεγάλο όγκο δεδομένων με βάση την ανθρώπινη συμπεριφορά και κατηγοριοποιούνται/ομαδοποιούνται με τεχνικές εξόρυξης ώστε η πληροφορία που θα προκύψει να είναι κατανοητή από τον ανθρώπινο νου.

Επιπλέον, γίνεται μία βιβλιογραφική ανασκόπηση σχετικά με την ανάπτυξη διαφόρων συστημάτων, την χρήση Νευρωνικών Δικτύων, Μηχανικής Μάθησης και άλλων τεχνικών για την διάγνωση του καρκίνου του τραχήλου της μήτρας. Στην συνέχεια περιγράφεται το σύνολο δεδομένων και γίνεται εφαρμογή αλγορίθμων με την χρήση του WEKA¹.

Σκοπός της παρούσας πτυχιακής εργασίας είναι η διερεύνηση των παραγόντων που επηρεάζουν την εμφάνιση καρκίνου τραχήλου της μήτρας και η δημιουργία ενός συστήματος υποστήριξης αποφάσεων για τη διάγνωση του συγκεκριμένου είδους καρκίνου.

Περιγραφή Προβλήματος

Σύμφωνα με τον παγκόσμιο οργανισμό υγείας (WHO)² το 2018 ο καρκίνος του τραχήλου της μήτρας κατατάσσεται τέταρτος ενώ σημειώθηκαν 570.000 νέες περιπτώσεις εκ των οποίων περίπου το 90% αφορούσαν χώρες με χαμηλό και μεσαίο εισόδημα. Προκαλείται κυρίως από τον ιο του ανθρώπινου θηλώματος (HPV) ο οποίος πρόκειται για μια σεξουαλικά μεταδιδόμενη λοίμωξη η οποία μεταξύ άλλων μπορεί να προκαλέσει ακόμη και καρκίνο του στόματος. Ωστόσο, άλλοι παράγοντες που συμβάλλουν στην εμφάνιση του καρκίνου είναι το κάπνισμα, η ηλικία καθώς και η χρήση αντισυλληπτικών. Όσον αφορά τα συμπτώματα είναι η μη φυσιολογική κολπική αιμορραγία έπειτα από σεξουαλική επαφή, ο πόνος στα πόδια, η απώλεια βάρους και όρεξης κ.α.

¹ Weka, *Machine Learning Software in Java*

² World Health Organization, *Cervical Cancer, Early diagnosis and screening*

Συγκεκριμένα, όπως αναφέρεται στο παγκόσμιο παρατηρητήριο³ για τον καρκίνο το οποίο περιλαμβάνει παγκόσμιες στατιστικές, η χώρα με τα μεγαλύτερα ποσοστά καρκίνου του τραχήλου της μήτρας το 2018 είναι η Σουαζιλάνδη με ποσοστό 75.3% .

Ο έλεγχος για την εύρεση καρκίνου επιτυγχάνεται με διάφορους τρόπους όπως την διμηνιαία εξέταση πυέλου δηλαδή την εξέταση από τον γιατρό σχετικά με τις ανωμαλίες που μπορεί να εμφανίσει ο τράχηλος ή η μήτρα καθώς και άλλα όργανα, τη δοκιμή HPV δηλαδή τη λήψη δείγματος από τον τράχηλο και τη μετέπειτα εξέταση του, το τεστ Παπανικολάου που αποτελεί την πιο συνηθισμένη δοκιμασία δηλαδή την συλλογή δείγματος κυττάρων από τον τράχηλο και τέλος την οπτική εξέταση με οξύ (VIA) για τυχόν ανωμαλίες στον τράχηλο. Επιπλέον, άλλες μέθοδοι για την διάγνωση του καρκίνου είναι η κολποσκόπηση, η βιοψία LEEP και η αξονική τομογραφία.

Ο σκοπός ανάπτυξης των συστημάτων υποστήριξης κλινικών αποφάσεων καθώς και η χρήση της μηχανικής μάθησης οφείλεται κυρίως στην πολυπλοκότητα σχετικά με την λήψη αποφάσεων του γιατρού που αφορούν περίπλοκες περιπτώσεις ασθενών. Ακόμη, η δημιουργία ενός συστήματος υποστήριξης αποφάσεων μπορεί να αυξήσει την ακρίβεια στην διάγνωση και στον εντοπισμό γυναικών που διατρέχουν υψηλό κίνδυνο εμφάνισης καρκίνου. Ωστόσο, επισημαίνεται ότι τα συστήματα δεν αντικαθιστούν τον γιατρό αλλά ουσιαστικά αποτελούν ένα βοηθητικό εργαλείο σχετικά με την ιατρική διάγνωση. Τονίζεται ακόμη, ότι ο λόγος που χρησιμοποιείται η μηχανική μάθηση δεν οφείλεται μόνο στα σφάλματα που μπορεί να γίνουν λόγω μη σωστής επικοινωνίας μεταξύ των γιατρών αλλά και στην ανάλυση σύνθετων δεδομένων.

Καταληκτικά, μέσω της τεχνητής νοημοσύνης μπορεί να επιτευχθεί η ενημέρωση του γιατρού για το στάδιο του καρκίνου, κατά πόσο η χημειοθεραπεία αποδίδει αλλά και στατιστικά στοιχεία σχετικά με τις παραγώγους που παίζουν πιο σημαντικό ρόλο στην εμφάνιση του καρκίνου.

³ World Cancer Research Fund, *Cervical Cancer Statistics*

ΚΕΦΑΛΑΙΟ 1^ο : Βιβλιογραφική Ανασκόπηση

Στο παρόν κεφάλαιο παρουσιάζεται η βιβλιογραφική ανασκόπηση σχετικά με τις έρευνες που έχουν γίνει αξιοποιώντας την μηχανική μάθηση και όχι μόνο, για την ανίχνευση/διάγνωση του καρκίνου του τραχήλου της μήτρας. Από τις έρευνες έχουν προκύψει αξιοσημείωτα αποτελέσματα τα οποία αναλύονται παρακάτω. Ωστόσο, επισημαίνεται ότι η χρήση της πληροφορικής έχει συνεισφέρει αρκετά στην διάγνωση του καρκίνου του τραχήλου της μήτρας.

1.1 Τεχνική Μηχανικής Μάθησης για την ανίχνευση καρκίνου του τραχήλου της μήτρας χρησιμοποιώντας k-NN και Τεχνητό Νευρωνικό Δίκτυο⁴

Η παρούσα εργασία αφορά την χρήση τεχνικών μηχανικής μάθησης προκειμένου να γίνει διάγνωση του καρκίνου του τραχήλου της μήτρας. Η προτεινόμενη τεχνική εξάγει τα χαρακτηριστικά του χρώματος και της μορφολογίας του πυρήνα και του κυτταροπλάσματος. Ο καρκίνος του τραχήλου της μήτρας είναι ο δεύτερος συνηθέστερος τύπος καρκίνου που επηρεάζει τις γυναίκες ενώ μπορεί να θεραπευτεί μόνο αν ανιχνευθεί σε αρχικό στάδιο. Το προτεινόμενο σύστημα βασίζεται στο MATLAB 2016 και ασχολείται με την εμφύτευση ενός αυτοματοποιημένου συστήματος το οποίο συμβάλει στην ανίχνευση του καρκίνου του τραχήλου της μήτρας χρησιμοποιώντας ψηφιακή επεξεργασία εικόνας.

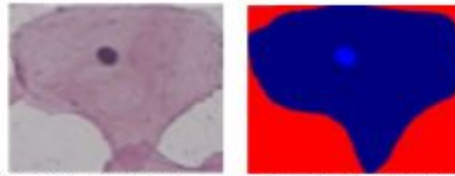
Αρχικά, μετατρέπεται το χρώμα (κλίμακα RGB) σε γκρι, έπειτα ενισχύεται, κλιμακώνεται τμηματοποιείται και ανιχνεύεται το κυτταρικό όριο καθώς ανιχνεύεται το επικαλυπτόμενο κυτταρόπλασμα και οι πυρήνες. Έπειτα γίνεται εξαγωγή και ταξινόμηση των κυττάρων βάσει των χαρακτηριστικών. Επισημαίνεται ότι τα περισσότερα διακριτά σύνολα χαρακτηριστικών χρησιμοποιούνται ως είσοδο σε ένα τεχνητό νευρωνικό δίκτυο (ANN) και στον αλγόριθμο K-NN και στην συνέχεια εξάγονται χαρακτηριστικά που αφορούν το σχήμα όπως είναι η στρογγυλότητα, η περιοχή, το πλάτος, το ύψος, τα χαρακτηριστικά υφής και τα χαρακτηριστικά χρώματος. Ειδικότερα, τα χαρακτηριστικά που συλλέγονται είναι τα χρωματικά, χαρακτηριστικά που αφορούν το σχήμα του πυρήνα αλλά και του κυτταροπλάσματος (μέγεθος πυρήνα, περιοχή, σχήμα κ.α). Ωστόσο, και οι δύο κατηγορίες χαρακτηριστικών εκπαιδεύονται από το ANN και τον K-NN ταξινομητή.

Τονίζεται, ότι η τμηματοποίηση βασίζεται στην τεχνική ***Fuzzy c-mean clustering (ασαφής ομαδοποίηση)***, ουσιαστικά κατά την τεχνική αυτή, στοιχεία των δεδομένων μπορούν να ανήκουν σε δύο ή περισσότερες ομάδες. Η κατάτμηση είναι σημαντική καθώς στο συγκεκριμένο βήμα η ανάλυση της εικόνας χωρίζει τα αντικείμενα από το φόντο ώστε να διακριθεί η περιοχή που παρουσιάζει το μεγαλύτερο ενδιαφέρον και να αποκτηθούν τα

⁴ Priyanka K Malli, Dr. Suvarna Nandyal, “Machine Learning Technique for detection of Cervical Cancer using k-NN and Artificial Neural Network”, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) (2017)

χαρακτηριστικά (στο παρόν άρθρο χρησιμοποιείται η μέθοδος *thresholding* παρόλου που υπάρχουν και άλλες μέθοδοι) .

Προκειμένου να γίνει πιο κατανοητός ο τρόπος που γίνεται η ομαδοποίηση παρουσιάζεται σχετική εικόνα



Εικόνα 1.1.α : Κανονικό κύτταρο Εικόνα 1.1.β : Κατακερματισμένο Κύτταρο

Φάση εκπαίδευσης

Στην παρούσα φάση υπολογίζεται η περιοχή που αφορά τον πυρήνα των φυσιολογικών κυττάρων και αντίστοιχα των μη φυσιολογικών. Έπειτα, οι μικρές περιοχές οι οποίες παραμένουν στην εικόνα και μοιάζουν με πυρήνα αφαιρούνται πρώτες. Το αρνητικό της εικόνας είναι καταλληλότερο καθώς απομακρύνονται τα ανεπιθύμητα συμπλέγματα με χρήση της μορφολογικής διάβρωσης. Η προκύπτουσα εικόνα εκτός από τον πυρήνα περιλαμβάνει αντικείμενα τα οποία αφαιρούνται με την λειτουργία της διάβρωσης. Στη συνέχεια η διαβρωμένη εικόνα αφαιρείται από την αρνητική ώστε να χρησιμοποιήσει τον μετρητή των τεμαχισμένων αντικειμένων. Η εικόνα που θα προκύψει αποτελείται από την περιοχή του πυρήνα, η οποία είναι η μεγαλύτερη συγκριτικά με τις άλλες περιοχές που διακρίνονται. Στη συνέχεια γίνεται η επισήμανση των αντικειμένων επιλέγοντας το μεγαλύτερο σύμπλεγμα εικονοστοιχείων μεταξύ των διαφόρων ομάδων. Μετά την εξαγωγή των χαρακτηριστικών από την εικόνα δείγματος, εφαρμόζεται η μέθοδος *thresholding* για τα εξαγόμενα χαρακτηριστικά. Το *threshold* υπολογίζεται χρησιμοποιώντας τα ακόλουθα χαρακτηριστικά χρωμάτων και χαρακτηριστικά σχήματος.

- Περιοχή πυρήνα και περιοχή κυτταροπλάσματος : υπολογίζεται μετρώντας τα αντίστοιχα εικονοστοιχεία της κατατμημένης εικόνας.
- Αναλογία N/C : αφορά το πόσο μικρή είναι η περιοχή του πυρήνα σε σχέση με την περιοχή του κυτταροπλάσματος.
- Φωτεινότητα πυρήνα και κυτταροπλάσματος : υπολογίζεται ως η μέση αντιληπτή φωτεινότητα του ανθρώπινου ματιού με βάση την ένταση των χρωμάτων.
- Μεγαλύτερη διάμετρος πυρήνα και κυτταροπλάσματος.
- Μικρότερη διάμετρος πυρήνα και κυτταροπλάσματος.
- Επιμήκυνση πυρήνα και κυτταροπλάσματος : υπολογίζεται ως ο λόγος μεταξύ της μικρότερης διαμέτρου και της μεγαλύτερης.
- Στρογγυλότητα πυρήνα και κυτταροπλάσματος.
- Περιφέρεια πυρήνα και κυτταροπλάσματος.
- Θέση πυρήνα.
- Πυρήνας και κυτταρόπλασμα Μέγιστα / Ελάχιστα.

Το *threshold* υπολογίζεται χρησιμοποιώντας 20 χαρακτηριστικά.

Η εκπαιδευμένη εικόνα ταξινομείται χρησιμοποιώντας τον K-NN και ANN για τους διάφορους τύπους καρκίνου. Οι εκπαιδευμένοι πίνακες αποτελούνται από 3 παραμέτρους το μέγεθος F το οποίο αποτελείται από 20 χαρακτηριστικά που έχουν ήδη εκπαιδευτεί, το C που περιέχει όλα τα ονόματα των κατηγοριών και το info που περιέχει όλα τα ονόματα των αρχείων. Στον αλγόριθμο K-NN δίνεται ως είσοδος ο πίνακας εκπαίδευσης ώστε να ταξινομήσει το αποτέλεσμα του χαρακτηριστικού της τρέχουσας εικόνας σε σχέση με το F και το C, και παράγει την κλάση του C. Σε περίπτωση που τα χαρακτηριστικά γνωρίσματα του σχήματος και του χρώματος που φορτώθηκαν ταιριάζουν με τον εκπαιδευμένο πίνακα τότε δείχνει τον τύπο καρκίνου. Αντίστοιχα, το τεχνητό νευρωνικό δίκτυο εκπαιδεύεται με τον πίνακα εκπαίδευσης και αναμένει τις τάξεις βασισμένες σε διανυσματική μορφή και διανύσματα χαρακτηριστικών. Μετά την εκπαίδευση του νευρωνικού δικτύου ακολουθεί η ταξινόμηση των χαρακτηριστικών 20 τύπων. Στη συνέχεια, τα χαρακτηριστικά δίνονται ως είσοδο και ο αριθμός των κλάσεων λαμβάνεται ως επίπεδα εξόδου ενώ το κρυφό στρώμα αποτελείται από τον αριθμό των εικόνων που έχουν εκπαιδευτεί και χαρτογραφείται με τον αριθμό των επιπέδων. Όσο εκπαιδεύονται οι εικόνες τα κρυμμένα στρώματα αυξάνονται. Εν κατακλείδι φορτώνεται το προκατασκευασμένο δίκτυο και υπολογίζονται το αποτέλεσμα με προσομοίωση της τρέχουσας δέσμης δικτύων.

Τα αποτελέσματα που προέκυψαν από τον αλγόριθμο k-NN είναι μέγιστη ακρίβεια 100% για καρκίνωμα (carcinoma in situ) και ελάχιστη 72,73% για δυσπλασία ενώ τα αποτελέσματα από τα νευρωνικά δίκτυα είναι 95% ακρίβεια για καρκίνωμα και 25% για φυσιολογικούς επιφανειακούς τύπους καρκίνου. Συνοπτικά, παρατηρήθηκε ότι η μεγαλύτερη ακρίβεια επιτυγχάνεται με τον αλγόριθμο πλησιέστερων γειτόνων (k-NN, 88,04%) ενώ τα νευρωνικά δίκτυα παρουσίασαν μικρότερη ακρίβεια σε ποσοστό 54%. Όμως όπως επισημαίνεται το σύστημα μπορεί να βελτιωθεί με την ταξινόμηση του σταδίου του καρκίνου.

1.2 Πρόγνωση του καρκίνου του τραχήλου της μήτρας με την χρήση της Εξόρυξης δεδομένων⁵

Στην παρούσα εργασία χρησιμοποιούνται τεχνικές εξόρυξης δεδομένων προκειμένου να γίνει πρόβλεψη σχετικά με τους ασθενείς που πάσχουν από καρκίνο του τραχήλου της μήτρας. Η εξόρυξη δεδομένων χρησιμοποιείται κυρίως γιατί δεν είναι εφικτή η παραδοσιακή γλώσσα ερωτήσεων σε συστήματα βάσης δεδομένων ώστε να εξαχθούν οι απαραίτητες τιμές.

Η διαδικασία που ακολουθείται στην εξόρυξη δεδομένων είναι η εξής .

Αρχικά, επιλέγονται τα πραγματικά δεδομένα στα οποία θα εφαρμοστεί η διαδικασία της εξόρυξης ενώ κάποια από αυτά τα οποία θεωρούνται καταλληλότερα θα χρησιμοποιηθούν για περαιτέρω επεξεργασία (δεδομένα στόχων). Στην παρούσα φάση οποιοδήποτε δεδομένο είναι λανθασμένο αφαιρείται προκειμένου οι τεχνικές εξόρυξης να παρέχουν ακριβή αποτελέσματα. Όσον αφορά τα δεδομένα στόχου (target data) κάποιο μέρος μπορεί να μετατραπεί σε άλλη μορφή ώστε να επιτευχθεί καλύτερη κατανόηση και συμβατότητα με τις

⁵ Dipti N. Punjani, Dr. Kishor H. Atkotiya, “Cervical Cancer Prediction using Data Mining”, International Journal for Research in Applied Science & Engineering Technology (IJRASET) (2017)

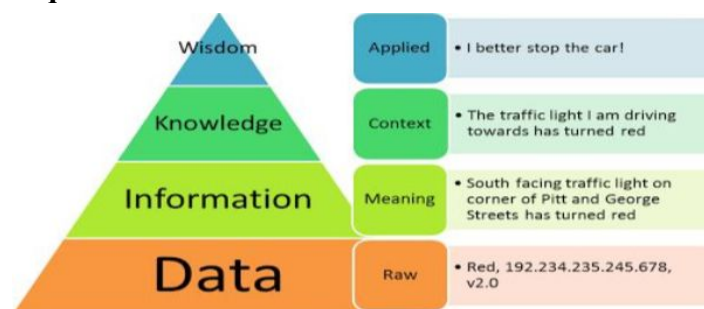
τεχνικές που χρησιμοποιούνται, ενώ σε περιπτώσεις στις οποίες τα δεδομένα δεν συλλέγονται από μία πηγή τότε είναι αναγκαίο να μετατραπούν στον ίδιο τύπο.

Έπειτα, αφού τα δεδομένα μετασχηματιστούν (κατά τον μετασχηματισμό δεν χάνονται τιμές) χρησιμοποιούνται τεχνικές εξόρυξης με σκοπό την ανάλυση τους για να εξαχθεί το σύνολο προτύπων. Η μορφή των συγκεκριμένων συνόλων εξαρτάται από την τεχνική εξόρυξης που χρησιμοποιείται, ενώ σε περιπτώσεις που δεν είναι απαραίτητη η χρήση των αποτελεσμάτων που προκύπτουν από τα αυτά, καθώς δεν χρησιμοποιούνται για περαιτέρω επεξεργασία, τότε δεν γίνεται χρήση τεχνικών εξόρυξης δεδομένων (data mining). Όσον αφορά την μορφή των συνόλων (πρότυπα) προκειμένου να είναι πλήρως κατανοητά αναπαριστώνται με την χρήση γραφημάτων, πινάκων και άλλων μορφών οπτικοποίησης.

Εν κατακλείδι, το τελευταίο βήμα είναι η ερμηνεία και η αξιολόγηση όπου αξιολογούνται οι τεχνικές που χρησιμοποιήθηκαν ως προς το πόσο ακριβείς, αποτελεσματικές, αξιόπιστες και επίκαιρες είναι. Στο παρόν βήμα όπως αναφέρεται χαρακτηριστικά βρίσκεται η καλύτερη τεχνική εξόρυξης δεδομένων η οποία έχει τις καλύτερες επιδόσεις και απαιτήσεις πόρων.

Στην επόμενη φάση προσδιορίζεται ο σκοπός της εφαρμογής τεχνικών εξόρυξης ο οποίος είναι η εξαγωγή μοτίβων από τα δεδομένα.

Παρατίθεται σχετική εικόνα της πυραμίδας **DIKW (Data Information Knowledge Wisdom)**.
Data Mining Techniques



Εικόνα 1.2.α : Πυραμίδα DIKW (Data Information Knowledge Wisdom)

Οι 4 βασικές τεχνικές εξόρυξης που αναλύονται είναι οι εξής.

- **Ταξινόμηση (Classification)**

Η διαδικασία της ταξινόμησης περιλαμβάνει τρία μέρη τα οποία είναι η δημιουργία του μοντέλου ταξινόμησης , η δομική του μοντέλου και η χρήση του. Η δημιουργία του μοντέλου γίνεται με διάφορους αλγορίθμους (*ID3 Decision Tree, Naive Bayes, K-NN*) οι οποίοι λαμβάνουν τα δεδομένα εκπαίδευσης. Στην συνέχεια, αναπτύσσεται ο ταξινομητής στον οποίο δίνονται δεδομένα δοκιμής ώστε να προσδιοριστεί εάν ταξινομεί με ακρίβεια. Τα αποτελέσματα που προκύπτουν συγκρίνονται με τα πραγματικά που είναι διαθέσιμα στα δεδομένα δοκιμής τα οποία (δεδομένα δοκιμής) μπορεί να είναι είτε υποσύνολο δεδομένων εκπαίδευσης είτε διαφορετικό σύνολο δεδομένων. Εφόσον ο ταξινομητής παρέχει ικανοποιητικό επίπεδο ακρίβειας τότε ταξινομούνται νέα δεδομένα οι ετικέτες των οποίων είναι άγνωστες.

Παράδειγμα : Δεδομένα ασθενών οι οποίοι πάσχουν από τον καρκίνο του τραχήλου της μήτρας μπορούν να δημιουργήσουν ένα μοντέλο ταξινόμητή. Το συγκεκριμένο μοντέλο μπορεί να ταξινομεί τους ασθενείς ανάλογα με τον κίνδυνο σε τρεις κατηγορίες Χαμηλή, Μέτρια, Υψηλή. Κατόπιν, οι λεπτομέρειες κάθε ασθενούς του οποίου η κατάσταση του

καρκίνου είναι άγνωστη μπορούν να εισαχθούν στον ταξινομητή ώστε να προβλεφθεί ο κίνδυνος να έχει καρκίνο. Επομένως, το επίπεδο κινδύνου παρέχει την δυνατότητα στις γυναίκες ασθενείς να γνωρίσουν την πιθανότητα καρκίνου του τραχήλου της μήτρας παρέχοντας τα δεδομένα τους.

- **Παλινδρόμηση (Regression)**

Η παλινδρόμηση είναι παρόμοια με την ταξινόμηση αλλά η διαφορά τους έγκειται στο γεγονός ότι προβλέπει βασιζόμενη σε αριθμητικές τιμές. Ορισμένοι από τους αλγόριθμους βασίζονται στην εύρεση σχέσεων μεταξύ αριθμητικών και μη τιμών.

Παράδειγμα : Το κάπνισμα αποτελεί έναν από τους λόγους αύξησης του καρκίνου του τραχήλου της μήτρας. Ο ρυθμός που καπνίζει η ασθενής σε σχέση με τον ρυθμό με τον οποίο αναπτύσσεται ο καρκίνος του τραχήλου της μήτρας μπορεί να αποτελέσει αντικείμενο σχεδιασμού από κάποιον ερευνητή. Η ανάλυση παλινδρόμησης βοηθάει στην πρόβλεψη της ποσότητας της λοίμωξης, των τιμών της αιμοσφαιρίνης, της ζάχαρης, του καρδιακού ρυθμού κ.α. Συμπερασματικά, η παλινδρόμηση παρέχει μεγαλύτερη ακρίβεια στην λήψη αποφάσεων.

- **Ομαδοποίηση (Clustering)**

Σε περίπτωση που δεν υπάρχουν συγκεκριμένα στοιχεία για ανάλυση τότε χρησιμοποιείται η τεχνική της ομαδοποίησης. Η ομαδοποίηση αποτελεί μία διαδικασία διερεύνησης νέων δεδομένων, εύρεσης ομοιότητας και διαφορών μεταξύ τους καθώς και ομαδοποίησης τους σε ένα σύνολο ζωνών που ονομάζονται συμπλέγματα. Τα συμπλέγματα περιλαμβάνουν ένα σύνολο δεδομένων με τις ιδιότητες τους να είναι συγκριτικά παρόμοιες με αυτές που βρίσκονται στα υπόλοιπα σύνολα. Οι αλγόριθμοι είναι αρκετά χρήσιμοι ιδίως όταν τα δεδομένα εκπαίδευσης δεν είναι διαθέσιμα.

Παράδειγμα : Όταν σε μία ομάδα ασθενών τα δεδομένα δοκιμών δεν είναι κρίσιμα αλλά εμπλέκονται περισσότερο στο κάπνισμα, στην κατανάλωση αλκοόλ και σε παρόμοιες συνήθειες τότε μπορούν να παρακινηθούν να αποφεύγουν τέτοιου είδους συνήθειες. Άλλη ομάδα ασθενών που χρησιμοποιεί αντισυλληπτικά εδώ και πολλά χρόνια μπορεί να παρακινηθεί να χρησιμοποιήσει άλλες εγκαταστάσεις ελέγχου των γεννήσεων.

- **Εξόρυξη Κανόνα σύνδεσης (Association Rule Mining)**

Η άντληση κανόνων συσχέτισης γίνεται βρίσκοντας τις συχνές σειρές στοιχείων (frequent item set). Ένα συνηθισμένο σύνολο αντικειμένων αναφέρεται σε ένα σύνολο τα στοιχεία του οποίου φαίνεται να είναι διαθέσιμα πάντα μαζί. Επίσης, η εξόρυξη κανόνων συσχέτισης χρησιμοποιείται για να διαπιστωθούν οι σχέσεις μεταξύ ενός συνόλου αντικειμένων (set of items) που ανήκουν σε ένα ίδιο σύνολο συχνών αντικειμένων.

Παράδειγμα : Στην περίπτωση του καρκίνου του τραχήλου της μήτρας, οι κανόνες συσχέτισης μπορούν να εξαχθούν όπως αν η ασθενής καπνίζει από τότε που χρησιμοποιούσε αντισυλληπτικά, διότι μπορεί από τότε να διατρέχει κίνδυνο. Γενικότερα, αν κάποια ασθενής επηρεάζεται από τους ιούς (A) και (B) και στο

συχνοτικό σύνολο περιλαμβάνεται και το (C) τότε αν υπάρχει αντίστοιχος κανόνας συσχέτισης (A, B), (C) τότε ο γιατρός μπορεί να προτείνει να πάει για δοκιμή του ιού C.

Θεωρείται σημαντικό να σημειωθεί ότι τα δεδομένα που συλλέγονται από ερωτηματολόγια, συνεντεύξεις και συμπλήρωση φόρμας δεν είναι αξιόπιστα για την ερευνητική εργασία καθώς είναι ελλιπή. Η συγκεκριμένη ερευνητική εργασία βασίζεται στην εξεύρεση της σχέσης μεταξύ ορισμένων συμπτωμάτων, συνηθειών και προηγούμενου ιατρικού ιστορικού για τον εντοπισμό ασθενούς με χαμηλό, μέτριο ή υψηλό κίνδυνο καρκίνου του τραχήλου της μήτρας. Τα χαρακτηριστικά γνωρίσματα που χρησιμοποιήθηκαν ήταν άμεσα γνωστά σε μία γυναίκα ή θα μπορούσαν να εξαχθούν εύκολα από το προηγούμενο ιατρικό ιστορικό της.

Παρατίθεται σχετικός πίνακας με τα χαρακτηριστικά.

Age
No_of_Pregnancies
No_of_Deliveries
Smoking
Smoke_Freq_Yearly
Contraceptives
Contraceptives_Freq_No_Years
AIDS
Hepatitis_B
Risk

Πίνακας 1.2.α : Πίνακας Χαρακτηριστικών

Η βάση δεδομένων παρέχεται από το νοσοκομείο Universitario de Caracas και χρησιμοποιείται για ρεαλιστική ταξινόμηση ενώ παράλληλα απλοποιείται για να γίνει πιο προσιτή. Η μέθοδος που χρησιμοποιήθηκε είναι η **Tree Decision** με δεδομένα **50** γυναικών σε **R** περιβάλλον εφαρμογής για σκοπούς εκπαίδευσης και δοκιμής. Τα αποτελέσματα που προέκυψαν είναι ότι υπήρξαν **20** εγγραφές στην κατηγορία υψηλού κινδύνου, **20** εγγραφές στην κατηγορία μετρίου κινδύνου και **10** εγγραφές στην κατηγορία χαμηλού κινδύνου. Όμως, **12** εγγραφές ταξινομήθηκαν σωστά από τις **20** που αφορούν υψηλό κίνδυνο και **8** ταξινομήθηκαν λανθασμένα στην κατηγορία μετρίου κινδύνου, όλες οι εγγραφές ταξινομήθηκαν ορθά στην κατηγορία μετρίου κινδύνου ενώ στην τελευταία κατηγορία όλες οι εγγραφές ήταν λάθος ταξινομημένες διότι μία κατηγορία άνηκε στον υψηλό κίνδυνο και οι υπόλοιπες στον μέτριο. Ωστόσο, ο λόγος που δεν έχουν ταξινομηθεί σωστά είναι γιατί δεν ελήφθη υπόψη το ιστορικό δοκιμής για τον καρκίνο του τραχήλου της μήτρας.

Συμπερασματικά, ο ταξινομητής μπορεί να προβλέψει με απόλυτη ακρίβεια με απλή βάση δεδομένων η οποία μπορεί να χρησιμοποιηθεί από οποιαδήποτε γυναίκα χωρίς να επισκεφθεί κάποιον γιατρό ενώ τα δεδομένα όπως έχει προαναφερθεί πρέπει να είναι ακριβή και αξιόπιστα.

1.3 Σύστημα Υποστήριξης Κλινικών αποφάσεων με αυτοματοποιημένη επεξεργασία κειμένου για την πρόγνωση του καρκίνου του τραχήλου της μήτρας⁶

Το παρόν άρθρο πραγματεύεται την ανάπτυξη ενός συστήματος υποστήριξης κλινικών αποφάσεων προκειμένου να εντοπιστεί ο καρκίνος του τραχήλου της μήτρας ερμηνεύοντας αναφορές Παπανικολάου με ελεύθερο κείμενο.

Η Αμερικανική Αντικαρκινική Εταιρεία, η Ειδική Ομάδα Προληπτικών Υπηρεσιών των Ηνωμένων Πολιτειών και το Αμερικανικό Κολέγιο Μαιευτήρων και Γυναικολόγων έχουν δημοσιεύσει κατευθυντήριες οδηγίες ελέγχου καθώς και η Αμερικανική Κοινωνία Κολποσκόπησης και Τραχηλικής Παθολογίας έχει εκδώσει κατευθυντήριες οδηγίες για την παρακολούθηση και τη διαχείριση μη φυσιολογικών εξετάσεων του τραχήλου της μήτρας. Ωστόσο είναι περίπλοκες και δεν μπορούν να ανακληθούν εύκολα από τους παρόχους υπηρεσιών υγείας με αποτέλεσμα τον λανθασμένο έλεγχο των μη φυσιολογικών αποτελεσμάτων των ασθενών. Συνεπώς, ο λόγος ανάπτυξης του CDSS είναι να παράγει συστάσεις για όλους τους ασθενείς.

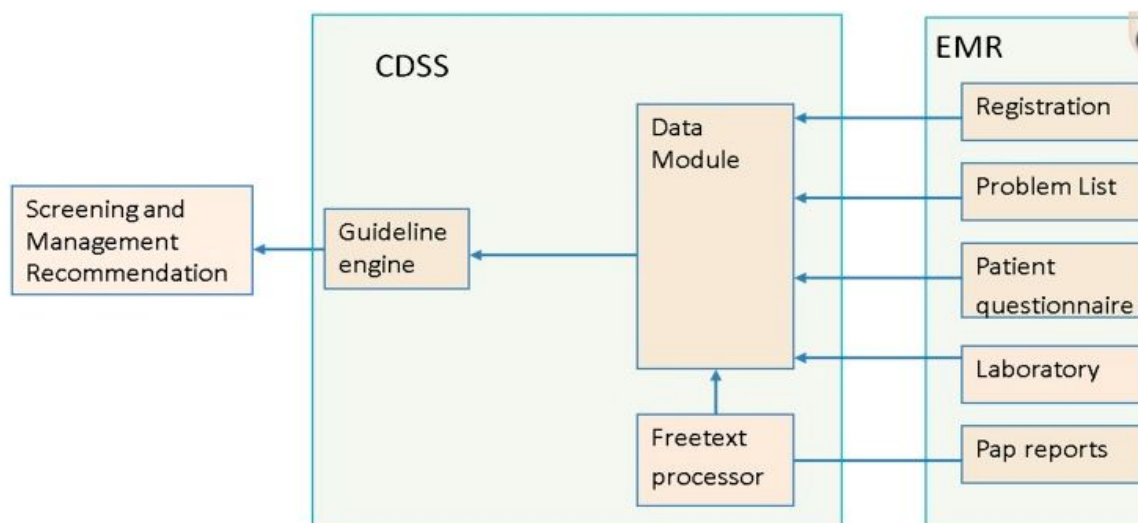
Το CDSS βοηθάει τον πάροχο υγειονομικής περίθαλψης στην λήψη αποφάσεων χρησιμοποιώντας δεδομένα ασθενών από το σύστημα ηλεκτρονικών ιατρικών αρχείων. Ωστόσο, τα προβλήματα που αναφέρονται είναι ότι το CDSS πρέπει να ερμηνεύσει αναφορές Pap που είναι σε μορφή ελεύθερου και υπάρχει έλλειψη σαφών κατευθυντηρίων οδηγιών. Παρόλα αυτά, χάρη στις κατευθυντήριες οδηγίες που αναπτύχθηκαν από εθνικές ενώσεις κατάφεραν να εφαρμοστούν στο συγκεκριμένο σύστημα αποφάσεων το οποίο βασίζεται σε κανόνες. Όσον αφορά την επεξεργασία φυσικής γλώσσας αυτή εξετάζεται παρακάτω.

Επισημαίνεται ότι τα δεδομένα ελεύθερου κειμένου (σημειώσεις γιατρών, φυσικές εξετάσεις) δεν χρησιμοποιήθηκαν σε μεγάλο βαθμό λόγω ανησυχίας σχετικά με την εγκυρότητα και την ακρίβεια. Αν και έχουν αναπτυχθεί διάφορα εργαλεία NLP (παρακολούθηση του συνδρόμου, διαγνωστική ταξινόμηση, εντοπισμός ανεπιθύμητων ενεργειών, προσδιορισμός μετεγχειρητικών επιπλοκών) για τέτοιες ανησυχίες, η εφαρμογή τους για λήψη αποφάσεων δεν υφίσταται λόγω απαίτησης για υψηλή ακρίβεια. Εντούτοις, δεδομένης της τυποποίησης της αναφοράς Παπ είναι δυνατό να κατασκευαστεί ένας επεξεργαστής ελεύθερου κειμένου που να είναι ακριβής για κλινική εφαρμογή.

Το σύστημα υποστήριξης κλινικών αποφάσεων που αναπτύχθηκε περιλαμβάνει τρεις ενότητες, μία βάση δεδομένων για την ερμηνεία των αναφορών Παπ και έναν βασικό κανόνα για την εκπροσώπηση των κατευθυντηρίων οδηγιών ελέγχου και διαχείρισης. Η μελέτη διεξήχθη στην κλινική Mayo και χρησιμοποιήθηκε ένα σύνολο αναφορών Παπ, συνεντεύξεις από γιατρούς και ένα πρότυπο για την δημιουργία αναφορών.

⁶ Kavishwar B. Waghlikar, Kathy L. MacLaughlin, Michael R. Henry, Robert A. Greenes, Ronald A. Hankey, Hongfang Liu et al., "Clinical decision support with automated text processing for cervical cancer screening", Journal of American Medical Informatics Association (2012)

Παρατίθεται σχετική εικόνα



Εικόνα 1.3.α : Συνδεσιμότητα Συστήματος με Ηλεκτρονικό Αρχείο Υγείας

Επίσης, αναφέρονται οι κανόνες ελεύθερου κειμένου. Ο επεξεργαστής ελεύθερου κειμένου περιλαμβάνει τον lexer και τον parser. Ο lexer μετατρέπει το κείμενο σε tokens και ο parser το επεξεργάζεται ώστε να παράγει την έξοδο. Συγκεκριμένα οι κανόνες του lexer αναπτύχθηκαν από μία ανάλυση τυχαία επιλεγμένων αναφορών Παπ από τα τελευταία 5 χρόνια στην κλινική Mayo, Rochester. Στη συνέχεια αφαιρέθηκαν οι μη αλφαριθμητικοί χαρακτήρες και μετατράπηκαν σε πεζά. Οι επαναλαμβανόμενες ομάδες λέξεων, εντοπίστηκαν σε τουλάχιστον 1% των αναφορών, χαρτογραφήθηκαν και χρησιμοποιήθηκαν από το τμήμα παθολογίας. Οι αναφορές παθολογίας δημιουργούνται μεταφράζοντας τους κωδικούς με πρότυπα κειμένου συμπεριλαμβανομένων τυχόν σχολίων που έγιναν από τον παθολόγο και αντιστοιχίζονται σε έννοιες που μεταφέρονται στην έκθεση Παπανικολάου. Αυτό το σύνολο κανόνων if-then αποτελεί τον lexer. Το σύνολο των κανόνων ανάλυσης κατασκευάστηκε με συνέντευξη από τον γιατρό και πρόκειται για ένα σύνολο κανόνων if-then που κωδικοποιούν την λογική που χρησιμοποιείται από τον γιατρό για την ερμηνεία των εννοιών που υπάρχουν στην αναφορά.

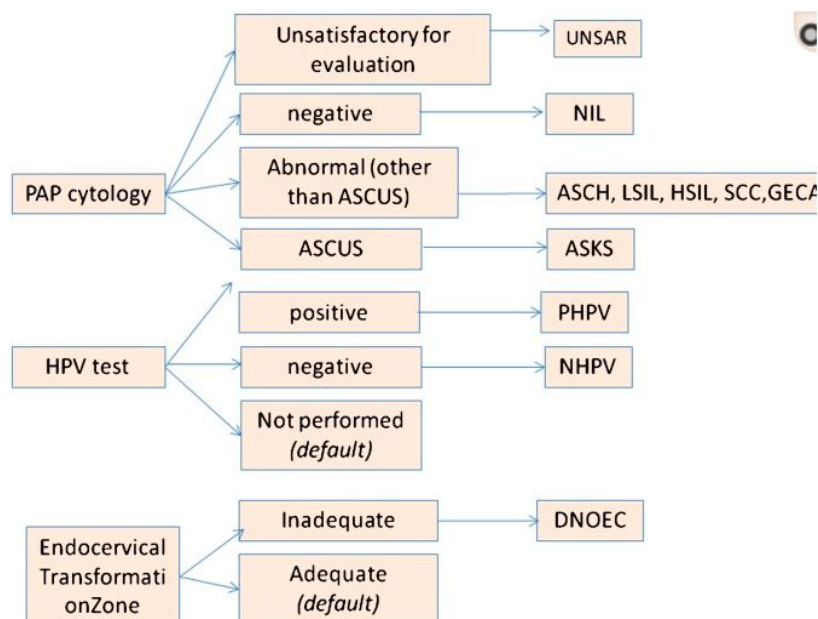
Συνεπώς, για να υπάρξει κάλυψη και ακρίβεια της βάσης δεδομένων ελεύθερου κειμένου οι αναφορές ομαδοποιήθηκαν σε κατηγορίες με βάσει τις τιμές παραμέτρων που εξάγονται από τον κανόνα. Από κάθε μία από τις κλάσεις επιλέχθηκε τυχαία ένας αριθμός εκθέσεων για χειροκίνητη επαλήθευση και ανάλυση σφαλμάτων.

Είναι αξιοσημείωτο το γεγονός ανάπτυξης διαγράμματος ροής για να αντιπροσωπεύει τις γνώσεις που περιέχονται στις κατευθυντήριες οδηγίες. Επιπλέον, για λόγους απλότητας δεν αναφέρονται παραμορφώσεις που οδήγησαν σε συστάσεις όπως λόγω χάρη για γυναίκες που έχουν υποβληθεί σε υστερεκτομή δεν συνίσταται τεστ Παπανικολάου.

Η αξιολόγηση του CDSS έγινε με ένα σετ δοκιμών 74 περιπτώσεων το οποίο κατασκευάστηκε με αναζήτηση αρχείων ασθενών στο EMR. Το CDSS δημιουργεί την βέλτιστη σύσταση για την αναφερόμενη ημερομηνία απόφασης καθώς κάθε σύσταση πρέπει να είναι συγκεκριμένη καθώς με την πάροδο του χρόνου αυξάνεται η ηλικία του ασθενούς και αλλάζουν τα δεδομένα στο EMR. Στην συνέχεια η σύσταση που δημιουργήθηκε,

συγκρίνεται με αυτή του γιατρού ενώ σε περίπτωση αναντιστοιχίας εξετάζεται από τον υπεύθυνο γιατρό και πραγματοποιείται ανάλυση σφάλματος.

Σε περίπτωση μη κυτταρολογικής ένδειξης νεοπλασίας, οι διαφάνειες χαρακτηρίζονται ως αρνητικές για ενδοεπιθηλιακή βλάβη ή κακοήθεια. Τα άτυπα πλακώδη κύτταρα χαρακτηρίζονται ως απροσδιόριστης σημασίας (ASC-US) ή δεν μπορούν να αποκλείσουν την πλακώδη ενδοεπιθηλιακή βλάβη υψηλής ποιότητας (ASC-H). Η κατηγοριοποίηση για τις κακοήθειες ενδοεπιθηλιακές βλάβες είναι είτε χαμηλού επιπέδου (LSIL) ή καρκίνωμα πλακωδών κυττάρων (SCC). Τα άτυπα αδενικά κύτταρα αναφέρονται ως ανωμαλία των αδενικών επιθηλιακών κυττάρων (GECA). Καταληκτικά, η ανωμαλία των αδενικών επιθηλιακών κυττάρων αναφέρεται ως ανεπαρκής συνιστώσα ζώνης μετασχηματισμού (DNOEC). Ωστόσο, το εύρος κανόνων της βάσης ελεύθερου κειμένου περιορίστηκε ώστε να απαιτούνται μόνο οι τιμές των παραμέτρων που είναι απαραίτητες. Οι συγκεκριμένες παράμετροι είναι 4 όπως περιγράφονται στο σχήμα που ακολουθεί και η αξιολόγηση τους περιλαμβάνει αναγνώριση 11 εννοιών στην έκθεση Pap.



Εικόνα 1.3.β : Αναγνώριση Παραμέτρων

Οι αναφορές που χρησιμοποιήθηκαν ως ανάλυση είναι 49.293 εκ των οποίων οι 41.910 (99,1%) ταξινομήθηκαν από το σύστημα σε 21 κατηγορίες που αντιστοιχούν στους συνδυασμούς των τιμών παραμέτρων που εξάγονται από τον κανόνα του ελεύθερου κειμένου. Το 91% των αναφορών έδειξαν φυσιολογική κυτταρολογία, το 2% δεν ήταν ικανοποιητικό για την αξιολόγηση ενώ το υπόλοιπο 7% έδειξε μη φυσιολογική κυτταρολογία. Επιπλέον, ελέγχθηκαν χειροκίνητα 10 τυχαίες αναφορές από κάθε κατηγορία δηλαδή έγινε επαλήθευση 210 αναφορών. Επίσης, το ποσοστό των σφαλμάτων ήταν 0,9% δηλαδή 395 αναφορές οι οποίες προήλθαν από ελλιπή στοιχεία ή λόγω σύντηξης πολλαπλών αναφορών. Τα αποτελέσματα υποδηλώνουν ότι η βάση κανόνων του ελεύθερου κειμένου καλύπτει όλα τα πρότυπα λέξεων ενώ τα αναγνωρισμένα πρότυπα λέξεων αποκλείουν μερικές ορθογραφικές παραλλαγές. Ακόμη, η επιλογή μιας προσέγγισης βασισμένης σε κανόνες για τον επεξεργαστή κειμένου οφείλεται στη διαθεσιμότητα του εσωτερικού

προτύπου που χρησιμοποιήθηκε για την δημιουργία αναφορών, στην ανάγκη να εκπροσωπείται η λογική του γιατρού, στην ανάγκη παροχής εξηγήσεων για τις αποφάσεις στους γιατρούς, στην απαίτηση υψηλής ακρίβειας για την αίτηση υποστήριξης απόφασης και στην παροχή συμπληρωματικών κανόνων προκειμένου να διασφαλιστεί ότι η έκθεση δεν είναι αλλοιωμένη.

Οι συστάσεις που έγιναν από το CDSS αντιστοιχούν σε εκείνες που έκανε ο γιατρός για 66 από τις 74 περιπτώσεις, έπειτα ο γιατρός επιβεβαίωσε ότι σε 7 από τις 8 περιπτώσεις η σύσταση του συστήματος είναι βέλτιστη και έχει σφάλει σε μια μόνο. Επιπρόσθετα, υπήρξε μία περίπτωση στην οποία ο γιατρός αγνόησε ότι η ηλικία της ασθενούς είναι μεγαλύτερη των 65 ετών και συνέστησε εξέταση Pap, σε άλλες δύο περιπτώσεις ο γιατρός δεν έλεγξε την προηγούμενη αναφορά και συνέστησε συνδοκιμή HPV/PAP σε 1 χρόνο, ενώ το CDSS ορθά βρήκε ότι η τρέχουσα και η προηγούμενη αναφορά ήταν θετικές για HPV παραπέμποντας την ασθενή σε κλινική γυναικολογίας. Ακόμη, υπήρξαν περιπτώσεις στις οποίες ο γιατρός δεν παρατήρησε στο ερωτηματολόγιο ότι η ασθενής είχε κάνει υστερεκτομή και συνέστησε λανθασμένα τεστ Παπ. Επίσης, σε ακόμη μία περίπτωση ο γιατρός εξέτασε τα αποτελέσματα σε διαφορετική ημερομηνία κάνοντας μη βέλτιστη σύσταση. Επιπλέον, συνέστησε παρακολούθηση Παπ σε 3 χρόνια σε αντίθεση με το CDSS το οποίο εντόπισε ότι η ασθενής είχε ιστορικό δυσπλασίας του τραχήλου της μήτρας και παρουσίαζε υψηλό κίνδυνο.

Σε περιπτώσεις ασθενών με παράγοντες κινδύνου περίθαλψης που επιστρέφουν για συστηματική εξέταση, το σύστημα ειδοποιεί τους παρόχους για την ύπαρξη παραγόντων κινδύνου παρέχοντας την δυνατότητα για επανεξέταση της απόφασης σχετικά με την επιστροφή του ασθενούς σε ρουτίνα διαλογής. Επομένως, η σύσταση του συστήματος θεωρήθηκε βέλτιστη .

Εντούτοις, το σύστημα λανθασμένα συνέστησε σε ασθενή να πάει σε γυναικολογική κλινική όταν η βέλτιστη απόφαση του γιατρού ήταν να επαναληφθεί η συνδοκιμή PAP/HPV σε ένα χρόνο, ο λόγος ήταν διότι ο έλεγχος HPV εκτελείται μερικές φορές ξεχωριστά και τα αποτελέσματα από το εργαστηριακό σύστημα δεν λήφθηκαν υπόψη. Ωστόσο, το CDSS βελτιώθηκε προκειμένου να αντιμετωπίζονται τετοιου ειδους καταστασεις.

Συμπερασματικά, ο γιατρός έχασε τις βέλτιστες συστάσεις για έξι από τις 74 περιπτώσεις. Το 76% των αναφορών του Pap έχουν φυσιολογική κυτταρολογία, επαρκή συνιστώσα ενδοκέρβιας ζώνης και δεν είναι θετικές για HPV. Το αναπτυγμένο CDSS εξαρτάται από την έκθεση Pap και μπορεί να μην είναι εύκολα φορητό σε άλλα ιδρύματα που έχουν διαφορετικά πρότυπα λέξεων στις εκθέσεις Παπ αλλά και από τη διαθεσιμότητα άλλων δεδομένων στο EMR, όπως μια σαφώς καθορισμένη λίστα προβλημάτων και πληροφορίες που παρέχονται από ασθενείς οι οποίες ενδέχεται να απουσιάζουν από άλλα ιδρύματα. Η αξιολόγηση αποκάλυψε ότι το σύστημα έκανε τις βέλτιστες συστάσεις διαλογής για 73 από τους 74 ασθενείς και εντόπισε δύο περιπτώσεις για παραπομπή γυναικολογίας που είχαν χαθεί από το γιατρό ενώ παράλληλα παρείχε βοήθεια στην τροποποίηση των συστάσεων από τον γιατρό σε έξι περιπτώσεις. Δεδομένης της υψηλής ακρίβειας του συστήματος, οι συγγραφείς θεωρούν ότι είναι κατάλληλος υποψήφιος για ανάπτυξη στην κλινική πρακτική. Σύμφωνα με τα αποτελέσματα της αξιολόγησης, το ανεπτυγμένο CDSS εκτελείται με ακρίβεια δεδομένης της τυποποιημένης αναφοράς των δοκιμών Pap και της διαθεσιμότητας

ρητών οδηγιών. Συνολικά, η μελέτη απέδειξε ότι το ελεύθερο κείμενο στο EMR μπορεί να χρησιμοποιηθεί αποτελεσματικά με μεθόδους NLP για την ανάπτυξη χρήσιμων εργαλείων υποστήριξης κλινικών αποφάσεων.

Όσον αφορά την μελλοντική προσέγγιση υπογραμμίζεται ότι μετά την επικύρωση με άλλους γιατρούς, το αναπτυγμένο CDSS θα αναπτυχθεί στα τμήματα εξωτερικών ασθενών της κλινικής Mayo, Rochester και θα συλλέγονται στοιχεία ανατροφοδότησης από τους χρήστες σχετικά με τις επιδόσεις του. Παράλληλα, το CDSS θα ενσωματωθεί με μια διασύνδεση στο EMR, η οποία απαριθμεί όλες τις υπενθυμίσεις προληπτικής φροντίδας για τους παρόχους υγειονομικής περίθαλψης και θα πραγματοποιηθεί ανάλυση αντίκτυπου συγκρίνοντας τα ποσοστά εξέτασης και παραπομπής πριν και μετά την εγκατάσταση του συστήματος. Εν κατακλείδι, στο μέλλον θα αντιγραφεί για άλλα ελεύθερα βασισμένα σε κείμενο λύματα, όπως ο έλεγχος του καρκίνου του παχέος εντέρου, η διαχείριση του, η διαταραχή του ύπνου και η διαχείριση του άσθματος.

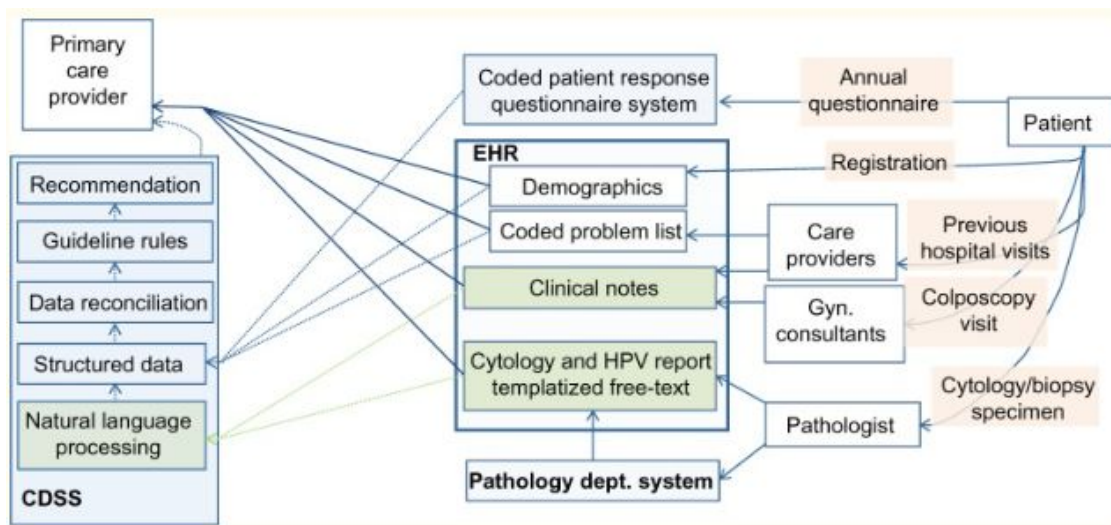
1.4 Αυτοματοποιημένη σύσταση για την εξέταση και επιτήρηση του καρκίνου του τραχήλου της μήτρας⁷

Το παρόν άρθρο πραγματεύεται την δημιουργία ενός συστήματος υποστήριξης κλινικών αποφάσεων βασισμένο σε δέντρο απόφασης το οποίο παρέχει ολοκληρωμένη καθοδήγηση για την περίθαλψη του ασθενούς που πάσχει από καρκίνο του τραχήλου της μήτρας.

Αρχικά, αναφέρεται ότι το 15% των γυναικών με καρκίνο του τραχήλου της μήτρας δεν είχαν την κατάλληλη παρακολούθηση των φυσιολογικών αποτελεσμάτων και 27% είχαν καθυστερήσει ή δεν παραπέμφθηκαν για κολποσκόπηση. Είναι αξιοσημείωτο το γεγονός ότι η έρευνα εγκρίθηκε από το θεσμικό όργανο αναθεώρησης της κλινικής Mayo και από το IRB ενώ έχει διεξαχθεί σύμφωνα με τις αρχές της Διακήρυξης του Ελσίνκι.

Το διάγραμμα ροής που παρουσιάζεται αποτελείται από κόμβους οι οποίοι αντιπροσωπεύουν έννοιες στις πληροφορίες του ασθενούς (π.χ ηλικία) ενώ οι ακμές αντιστοιχούν στις πιθανές τιμές των εννοιών (π.χ ηλικία άνω των 30 ετών). Όσον αφορά τις τιμές που ορίζονται στο διάγραμμα ροής εξάγονται από το EHR (Electronic Health Record). Συγκεκριμένα 10 έννοιες εξάγονται από έξι διαφορετικές πηγές (π.χ δημογραφικά στοιχεία, αναφορές παθολογίας του τραχήλου της μήτρας σε μορφή ελεύθερου κειμένου, αποτελέσματα για δοκιμές DNA υψηλού κινδύνου από ιό ανθρώπινου θηλώματος κ.α). Ωστόσο, για να διασφαλιστεί η ακρίβεια της εξαγωγής δεδομένων ο συμβιβασμός εκτελείται σε πολλαπλές πηγές όπως παρουσιάζεται στο παρακάτω διάγραμμα.

⁷ Kavishwar B. Waghoikar, Kathy L. MacLaughlin, Petra M. Casey, Thomas M. Kastner, Michael R. Henry, Ronald A. Hankey et al., “Automated Recommendation for Cervical Cancer Screening and Surveillance”, Cancer Informatics (2014)



Εικόνα 1.4 : Εξαγωγή εννοιών από έξι διαφορετικές πηγές

Η εμφάνιση της σύστασης πραγματοποιείται όταν ο πάροχος πλοηγείται στις αναφορές παθολογίας ενώ η εξήγηση της εμφανίζεται προκειμένου να ενημερωθεί. Επιπλέον, περιέχει ένα αποθετήριο θεσμικών κατευθυντήριων οδηγιών που περιλαμβάνει τοπικούς εμπειρογνώμονες για διαβουλεύσεις ή διευκρινίσεις και παρέχεται ενημέρωση στον πάροχο σχετικά με την σύσταση του CDSS. Η ανάπτυξη του συστήματος έγινε χρησιμοποιώντας τυχαίο σύνολο ασθενών κάθε φορά ενώ η επικύρωση με ένα ξεχωριστό τυχαίο σύνολο ασθενών (συνολικά 333 ασθενείς). Ακόμη, πραγματοποιήθηκε ανάλυση σφάλματος προκειμένου να εντοπιστούν ελλείψεις στην εξαγωγή δεδομένων ή στο διάγραμμα ροής για την βελτίωση του συστήματος. Στη συνέχεια διεξήχθη πιλοτική μελέτη με εννέα φορείς παροχής πρωτοβάθμιας περίθαλψης για μία περίοδο έξι μηνών και συγκεντρώθηκε η ανατροφοδότηση. Επισημαίνεται ότι οι επικυρώσεις έγιναν προτού ενημερωθεί το ενιαίο μοντέλο ώστε να συμπεριληφθεί η ενημέρωση ASCCP 2013.

Το διάγραμμα ροής που χρησιμοποιήθηκε καλύπτει 51 σενάρια αποφάσεων ενώ συμβάλλει στην δημιουργία συστήματος υποστήριξης κλινικών αποφάσεων. Συνεπώς, η χρήση πολλών μεταβλητών οδηγεί στην παραγωγή συστάσεων για σχεδόν όλα τα σενάρια αποφάσεων εκτός από ένα μικρό αριθμό ασθενών οι οποίοι αποτελούν εξαίρεση από το μοντέλο (έγκυες με κακοήθειες ενδοεπιθηλιακές αλλοιώσεις, γυναίκες με άτυπα αδενικά κύτταρα AGC ή άτυπα κυτταρικά ενδομήτρια και αδενοκαρκίνωμα AIS).

Εντούτοις, οι επαναληπτικές επικυρώσεις των συστάσεων του CDSS αποκάλυψαν ελλείψεις στο διάγραμμα ροής, συγκεκριμένα τα λάθη που προέκυψαν οφείλονταν σε σενάρια που δεν προβλέπονταν από τους εμπειρογνώμονες όταν αναθεώρησαν τις αφηγήσεις. Συνεπώς, το σύστημα είχε ακρίβεια 83% σε 333 περιπτώσεις. Όμως, αποκαταστάθηκαν τα σφάλματα και το σύστημα βελτιώθηκε. Οι μεταβλητές του διαγράμματος εξάγονται από πολλαπλές πηγές EHR. Συμπληρωματικά, διαπιστώθηκε ότι για τους ασθενείς με μη φυσιολογικά ευρήματα χρειάστηκε περίπου 1 λεπτό και 39 δευτερόλεπτα για την συλλογή των απαραίτητων πληροφοριών και την απόφαση σχετικά με την σύσταση. Επιπρόσθετα, οι μεταβλητές που απαιτούνται για την λήψη αποφάσεων καθίσταται εφικτό να συνδυαστούν από διαφορετικές πηγές, γεγονός το οποίο συμβάλλει στον εντοπισμό ασθενών με ιστορικό υστερεκτομής και

κινδύνου για καρκίνο του μαστού (4 από τις 10 έννοιες προέκυψαν από τις παρατηρήσεις του ελεύθερου κειμένου). Ένα πρόβλημα το οποίο κρίνεται αναγκαίο να αναφερθεί είναι ο συνεχής μεταβαλλόμενος χαρακτήρας των κατευθυντήριων οδηγιών ο οποίος απαιτεί εμπειρογνωμοσύνη και προσδιορισμό των συνεπειών που προκύπτουν στο ενιαίο μοντέλο, ως επακόλουθο είναι η χρονική καθυστέρηση για την ενημέρωση του συστήματος η οποία είναι μια συνεχής και επαναληπτική διαδικασία.

Καταληκτικά, υπογραμμίζεται ότι παρόλο που πολλά συστήματα υποστήριξης αποφάσεων προτείνουν κυτταρομετρία κάθε τρία χρόνια για τον καρκίνο του τραχήλου της μήτρας για όλους τους ασθενείς, πλέον θεωρείται λάθος καθώς υπάρχουν ομάδες ασθενών όπως γυναίκες ηλικίας 30 έως 65 ή ασθενείς με μη φυσιολογική κυτταρολογία οι οποίες απαιτούν συχνότητα δοκιμαστικής περιόδου πέντε ετών.

Τα αποτελέσματα που προέκυψαν είναι ότι οι εθνικές κατευθυντήριες οδηγίες για τον προσυμπτωματικό έλεγχο και έλεγχο του καρκίνου του τραχήλου της μήτρας είναι σαφείς, δωρεάν και πλήρεις για σχεδόν όλους τους ασθενείς που παρατηρούνται στην κλινική πρακτική. Η εφαρμογή τους για την πρόληψη του καρκίνου του τραχήλου της μήτρας απαιτεί τον προσδιορισμό πολλαπλών μεταβλητών. Όσον αφορά το σύστημα λήψης κλινικών αποφάσεων προέκυψε ότι είναι δυνατόν να εξοικονομήσει σημαντικό χρόνο και να βελτιώσει την ποιότητα της περίθαλψης εφόσον δεν απαιτείται η χρήση των αποτελεσμάτων που αφορούν τις δοκιμές από τους χρήστες. Επίσης, το διάγραμμα κατευθυντήριων οδηγιών μπορεί να χρησιμοποιηθεί για την ανάπτυξη παρόμοιων συστημάτων όπως για την πρόληψη του καρκίνου του παχέος εντέρου και του καρκίνου του πνεύμονα.

1.5 Ένα ευφύες σύστημα υποστήριξης κλινικών αποφάσεων για συγκεκριμένες προβλέψεις ασθενών για την βελτίωση ανίχνευσης νεοπλασιών του τραχήλου της μήτρας⁸

Το παρόν άρθρο πραγματεύεται την δημιουργία ενός ευφυούς συστήματος υποστήριξης κλινικών αποφάσεων βιοπληροφορικής σχετικά με τον καρκίνο του τραχήλου της μήτρας το οποίο συνδυάζει αποτελέσματα διάφορων διαγνωστικών τεχνικών με σκοπό την παροχή ειδικών προβλέψεων για ασθενείς σχετικά με διαγνωστικές ή προγνωστικές εκβάσεις.

Αρχικά, πριν αναλυθεί το προαναφερθέν σύστημα αναφέρονται τα δεδομένα που χρησιμοποιούν άλλα CDSS όπως δεδομένα σχετικά με τους παράγοντες κινδύνου, με τον ασθενή και την θεραπεία καθώς και τις μεθόδους στις οποίες βασίζονται όπως τα τεχνητά νευρωνικά δίκτυα, τεχνικές τεχνητής νοημοσύνης, τεχνικές αναγνώρισης προτύπων κτλ.

Τα δεδομένα που χρησιμοποιήθηκαν, συλλέχθηκαν τυχαία από γυναίκες οι οποίες συμμετείχαν στο ερευνητικό πρόγραμμα που διεξήγαγε το Τμήμα Κυτταρολογίας της Ιατρικής Σχολής του Πανεπιστημίου Αθηνών (Πανεπιστημιακό Νοσοκομείο ΑΤΤΙΚΩΝ) και το Τμήμα Μαιευτικής και Γυναικολογίας του Πανεπιστημιακού Νοσοκομείου Ιωαννίνων. Τα δεδομένα αυτά αφορούσαν αποτελέσματα μοριακών εξετάσεων, κυτταρολογικές διαγνώσεις,

⁸ Panagiotis Bountris, Maria Haritou, Abraham Pouliakis, Niki Margari, Maria Kyrgiou, Aris Spathis et al., “An Intelligent Clinical Decision Support System for Detection”, BioMed Research International (2014)

ηλικία ασθενούς κ.λπ, τα οποία αποθηκεύτηκαν σε μία βάση δεδομένων που δημιουργήθηκε καθαρά για ερευνητικούς σκοπούς με την εξαγωγή τους να είναι ανώνυμη.

Τα δείγματα λήφθηκαν από γυναίκες οι οποίες υποβλήθηκαν σε κολποσκόπηση για δύο λόγους, είτε διότι είχαν προηγούμενο μη φυσιολογικό τεστ Παπανικολάου είτε γιατί προσφέρθηκαν εθελοντικά και δέχτηκαν κολποσκοπική εξέταση ακόμη και αν το τεστ Παπανικολάου ήταν φυσιολογικό (π.χ γυναίκες με HPV). Υπογραμμίζεται ότι περιπτώσεις με αρνητικό τεστ Παπ και κολποσκόπηση θεωρήθηκαν κλινικά αρνητικές. Συνεπώς, όλες οι περιπτώσεις πλην των κλινικά αρνητικών διαγνώστηκαν από έναν ιστοπαθολόγο. Η βάση δεδομένων είχε πάνω από 5.000 ασθενείς με περισσότερες από 10.000 σειρές εξετάσεων εκ των οποίων κάθε σειρά περιελάμβανε διάφορες δοκιμές (π.χ δοκιμές NASBA, NuclisSENS για την αναγνώριση του E6/E7 mRNA των τύπων HPV 16,18,31,33,45 κτλ.). Στη συνέχεια τα ευρήματα ταξινομήθηκαν σύμφωνα με το σύστημα Bethesda ως εξής : εντός φυσιολογικών ορίων(WLN), άτυπα πλακώδη κύτταρα μη καθορισμένης σημασίας (ASCUS), χαμηλού βαθμού πλακώδη ενδοεπιθηλιακή βλάβη (LSIL), υψηλής ποιότητας πλακώδη ενδοεπιθηλιακή βλάβη (HSIL), καρκίνωμα πλακωδών κυττάρων (SCC) ή αδενοκαρκίνωμα (Adeno-Ca). Το σύστημα ταξινόμησης τριών κλινικών ενδοεπιθηλιακών νεοπλασιών (CIN) χρησιμοποιήθηκε για ιστολογική διάγνωση με την εξής ταξινόμηση: χωρίς ενδείξεις κακοήθειας (αρνητική), 1η βαθμίδα τραχηλικής ενδοεπιθηλιακής νεοπλασίας (CIN1) , βαθμός II ή III (CIN2/3), τραχηλικής ενδοεπιθηλιακής νεοπλασίας, καρκίνωμα πλακωδών κυττάρων (SCC) ή αδενοκαρκίνωμα (Adeno-Ca).

Ωστόσο, από τις 10.000 προαναφερθείσες σειρές εξετάσεων μόνο 740 περιπτώσεις ολοκληρώθηκαν και επιλέχθηκαν για την παρούσα μελέτη αποκλείοντας περιπτώσεις όπου τα αποτελέσματα ήταν ανεπαρκή ή μη έγκυρα. Κατά συνέπεια για κάθε μία από τις 740 περιπτώσεις δημιουργήθηκε ένα σύνολο 46 μεταβλητών που προέκυψαν από τις εφαρμοζόμενες δοκιμές (π.χ Pap test, HR-PPV DNA κ.α).

Σύμφωνα με τα αποτελέσματα καθεμία από τις 740 περιπτώσεις κατηγοριοποιήθηκε στις ακόλουθες κατηγορίες:

- **κατηγορία 1:** αρνητική ή κλινικά αρνητική,
- **κατηγορία 2:** CIN1,
- **κατηγορία 3 :** CIN2 ή CIN3 (CIN2/3)
- **κατηγορία 4:** καρκίνος (SCC ή Adeon-Ca).

Ο προσδιορισμός των 46 μεταβλητών έγινε με την χρήση διαφόρων αλγορίθμων επιλογής οι οποίοι ταξινομούνται σε δύο κατηγορίες την προσέγγιση φίλτρου και την προσέγγιση περιτυλίγματος. Οι μέθοδοι που χρησιμοποιήθηκαν αφορούσαν την περιοχή εμπειρικής καμπύλης λειτουργικών χαρακτηριστικών δέκτη (ROC), της τυχαίας κλίσης ταξινομητή καθώς και το πλαίσιο επιλογής ελάχιστης – μέγιστης συνάφειας (MRMR), μέθοδοι οι οποίες επιστρέφουν το καλύτερο υποσύνολο χαρακτηριστικών για μια επιλεγμένη τιμή. Δεδομένου ότι το σύνολο δεδομένων αποτελείται από 4 τάξεις, υιοθετήθηκε η στρατηγική ταξινόμησης πολλαπλών κατηγοριών “one against one” ώστε να εκτελεστεί η εργασία FS. Επομένως, το σύνολο δεδομένων διαιρέθηκε σε 6 χωρίσματα, εφαρμόστηκαν οι τεχνικές FS και

διερευνήθηκαν τα κοινά κορυφαία χαρακτηριστικά μεταξύ των 12 καλύτερων υποσυνόλων που επέστρεψαν οι δύο τεχνικές FS. Τα χαρακτηριστικά αυτά αποτέλεσαν το καλύτερο υποσύνολο που χρησιμοποιήθηκε για την ανάπτυξη του συστήματος.

Όσον αφορά τους αλγορίθμους ταξινόμησης που χρησιμοποιήθηκαν είναι οι εξής : k-πλησιέστερων γειτόνων (k-NN), ταξινομητή Bayesian (NB) και 3 τύποι ANNs συγκεκριμένα το δίκτυο πολλαπλών στρώσεων perceptron (MLP) , το δίκτυο λειτουργιών ακτινικής βάσης (RBF) και το πιθανοτικό νευρωνικό δίκτυο (PNN).

Το σύνολο δεδομένων χωρίστηκε σε 3 σύνολα, το σύνολο εκπαίδευσης (486 περιπτώσεις), το σύνολο επικύρωσης (126 περιπτώσεις) για βελτιστοποίηση των παραμέτρων κάθε ταξινομητή και 128 περιπτώσεις για αξιολόγηση της πρόβλεψης απόδοσης. Επισημαίνεται ότι, λόγω της ποικιλίας των αποτελεσμάτων των εξετάσεων, χρησιμοποιήθηκε ένα μεγάλο σετ κατάρτισης, συγκεκριμένα 66% του αρχικού συνόλου δεδομένων έτσι ώστε να παραχθούν αντιπροσωπευτικά δείγματα.

Το PNN επέδειξε καλές προγνωστικές επιδόσεις στις περισσότερες περιπτώσεις με κυταρολογική LSIL και ειδικά για γυναίκες με κυταρολογική LSIL που διαθέτουν CIN βαθμού 2 ή χειρότερου (CIN2 +) ενώ σε περιπτώσεις γυναικών με κυταρολογία ASCUS έδειξε ανεπαρκή απόδοση στον εντοπισμό της σωστής ιστολογίας. Το MLP είχε θετικά αποτελέσματα πρόβλεψης για περιπτώσεις με κυταρολογία ASCUS αλλά μειωμένη απόδοση σε σχέση με ολόκληρο το σύνολο δεδομένων.

Το CDSS αποτελείται από δύο υποσυστήματα το PNN που χρησιμοποιείται για την διαχείριση όλων των περιπτώσεων πλην των test Pap ASCUS και το MLP μόνο για την διαχείριση των περιπτώσεων ASCUS. Η αρχιτεκτονική PNN αποτελείται από τέσσερα στρώματα : εισόδου, προτύπου το οποίο περιέχει νευρώνα για κάθε δείγμα του σετ εκπαίδευσης, άθροισης το οποίο περιέχει έναν νευρώνα για κάθε κατηγορία του εκπαιδευτικού σετ και εξόδου. Κατά την διαδικασία κατάρτισης δεν εφαρμόζεται αλγόριθμος μάθησης καθώς επιτυγχάνεται με τον καθορισμό βαρών του δικτύου. Το MLP είναι ένα νευρωνικό δίκτυο το οποίο περιλαμβάνει έναν ή περισσότερους νευρώνες με ένα ή περισσότερα κρυφά στρώματα, ένα στρώμα εισόδου και ένα στρώμα εξόδου. Στην συγκεκριμένη αρχιτεκτονική χρησιμοποιείται ο αλγόριθμος μάθησης BP στον οποίο οι πληροφορίες σχετικά με τα σφάλματα του δικτύου σε γνωστά δεδομένα πολλαπλασιάζονται προς τα πίσω, δηλαδή από το στρώμα εξόδου στο στρώμα εισόδου. Για την αποφυγή της υπερφόρτωσης εφαρμόζεται μια τεχνική εκμάθησης πρόωρης διακοπής. Για την διαδικασία εκπαίδευσης πληρούνται κριτήρια σύγκλισης με 40 μέγιστες αποτυχίες ή όταν MSE (μέσο τετραγωνικό κανονικοποιημένο σφάλμα) ≤ 0.0001 ή όταν έχει φτάσει το μέγιστο 1000 επαναλήψεις. Όμως προκειμένου να εντοπιστεί η βέλτιστη τοπολογία του δικτύου προστέθηκαν νευρώνες στα κρυμμένα στρώματα μια φορά, την φορά, αυτή η διαδικασία σταμάτησε όταν τα MLP εμφάνισαν συνεχή μειωμένη απόδοση.

Καταληκτικά, για την σωστή επιλογή των βέλτιστων PNN και MLP ρυθμίστηκαν οι παράμετροι αξιολογώντας τα ανεπτυγμένα μοντέλα στο σετ επικύρωσης και λήφθηκε υπόψη η απόδοση της ταξινόμησης στο σετ εκπαίδευσης.

Η αξιολόγηση του συστήματος έγινε με βάση την απόδοση των τελικών μοντέλων ενώ δεν χρησιμοποιήθηκαν πολύπλοκα συστήματα εκτίμησης σφαλμάτων ώστε να αξιολογηθεί το σύστημα σε ατομικό επίπεδο.

Classifier	k-NN	NB	CART	MLP	RBF	PNN
Optimal parameters	k = 5	—	(i) Pruning level = 7/11 (ii) Number of terminal nodes = 8	2 hidden layers 18 × 18	(i) 486 neurons in the hidden layer (all samples of training set) (ii) Sigma = 0.6	Sigma = 0.4
Training set	78.6%	76.8%	77.6%	78.4%	87.6%	87.6%
Validation set	79.4%	80.9%	78.6%	77.0%	77.7%	80.2%
Test set	82.8%	82.8%	75.8%	80.5%	70.3%	80.0%

k-NN: k-nearest neighbours classifier, NB: naïve Bayesian classifier, CART: classification and regression tree, MLP: multilayer perceptron network, RBF: radial basis function network, PNN: probabilistic neural network.

Πίνακας 1.5 : Ταξινόμηση Ακρίβειας στα σετ εκπαίδευσης, επικύρωσης και στα σετ δοκιμής για τους 6 ταξινομητές που δημιουργήθηκαν (προσέγγιση μεμονωμένου ταξινομητή)

Τα αποτελέσματα που προέκυψαν είναι ότι η συνολική ακρίβεια ταξινόμησης του PNN για τα σύνολα εκπαίδευσης, επικύρωσης και δοκιμών είναι 90.0%, 82.0% και 84.0% αντίστοιχα. Το CDSS έδειξε υψηλή ευαισθησία (89,4%), υψηλή εξειδίκευση (97,1%), υψηλό PPV (89,4%) και υψηλό NPV (97,1%), για ανίχνευση CIN2 +. Επιπλέον, παρήγαγε πιο ισορροπημένα αποτελέσματα από τις ιατρικές εξετάσεις όσον αφορά την ευαισθησία, την εξειδίκευση, το PVV και το NPV καθώς και κατά την ταξινόμηση των δοκιμών με τον μέγιστο δείκτη Youden (το CDSS κατέλαβε την υψηλότερη βαθμολογία με δείκτη Youden 0,87). Η τελευταία έκδοση του διαγράμματος ροής καλύπτει 51 σενάρια αποφάσεων που προσδιορίζονται στον πληθυσμό της πρακτικής ενώ το σύστημα είχε ακρίβεια 87% σε 333 περιπτώσεις επικύρωσης το οποίο βελτιώθηκε μετά την αποκατάσταση των σφαλμάτων που εντοπίστηκαν κατά την επικύρωση.

1.6 Σύστημα υποστήριξης αποφάσεων με χρήση NLP για τον έλεγχο και την επιτήρηση του καρκίνου του τραχήλου της μήτρας⁹

Η συγκεκριμένη μελέτη επικεντρώνεται στην δημιουργία ενός συστήματος υποστήριξης αποφάσεων με χρήση NLP (Natural Language Processing) για τον έλεγχο του καρκίνου του τραχήλου της μήτρας. Η μελέτη διεξήχθη σε τρεις περιοχές πρωτοβάθμιας φροντίδας στην τοποθεσία Baldwin. Όσον αφορά τους συμμετέχοντες συνολικά 25.500 γυναίκες ηλικίας 18-65 ετών λαμβάνουν φροντίδα με 63% στο Baldwin, 22% σε NE (Northeast Clinic) και 15% σε NW (Northwest Clinic).

Το CDSS έχει 54 κλινικές οδούς με 13 μονοπάτια για συστηματική διαλογή και 41 μονοπάτια για ασθενείς υψηλού κινδύνου. Η λειτουργία του περιλαμβάνει τρία στάδια και

⁹ Rajeev Chaudhry, Kathy L. MacLaughli, Ravikumar Komandur Elayavili, Maya E. Kessler, Hongfang Liu, Kavishwar B. Waghlikar et al., “NLP-enabled decision support for cervical cancer screening and surveillance”, (2014 to 2016)

είναι η εξής. Αρχικά, γίνεται εξαγωγή των πρωτογενών στοιχείων και δεδομένων από τα κλινικά αρχεία, έπειτα επανασυναρμολογούνται τα στοιχεία των δεδομένων που εξάγονται από πολλαπλές αναφορές του ασθενούς και τέλος παρέχονται συστάσεις περί φροντίδας βάσει των οδηγιών της αμερικανικής εταιρείας για κολποσκόπηση (ASCCP). Το σύστημα εφαρμόστηκε σε όλους τους ασθενείς πρωτοβάθμιας φροντίδας στην κλινική Mayo Clinic Rochester. Επίσης, πραγματοποιήθηκε δειγματοληψία από περίπου 4000 ασθενείς για δοκιμή ακρίβειας του CDSS τόσο για φυσιολογικά όσο και για μη φυσιολογικά αποτελέσματα. Συγκεκριμένα περίπου το 10% (393 ασθενείς) επιλέχθηκε για εξέταση. Η αξιολόγηση πραγματοποιήθηκε από έναν εμπειρογνώμονα στον τομέα και έναν γιατρό ο οποίος ήταν εξοικειωμένος με τις οδηγίες της ASCCP. Στη συνέχεια πραγματοποιήθηκε αντιμετώπιση των σφαλμάτων προκειμένου να γίνει εφικτή η εφαρμογή του συστήματος. Παρόλο που το σύστημα αρχικά χρειαζόταν κατά μέσο όρο 54 δευτερόλεπτα για την επεξεργασία των δεδομένων και τον υπολογισμό των συστάσεων για τον ασθενή, πραγματοποιήθηκαν ορισμένες βελτιώσεις με την συνεργασία μιας ομάδας πληροφορικής στην υποδομή Big Data. Επιπλέον, βάσει του συστήματος οι γυναίκες με αυξημένο κίνδυνο εμφάνισης καρκίνου του τραχήλου της μήτρας ήταν κυρίως αυτές με ιστορικό μη φυσιολογικού τεστ Παπ, HPV ή κολποσκοπικής βιοψίας που είχαν καθυστερήσει την παρακολούθηση, γυναίκες με ιστορικό in utero DES έκθεσης, CIN 2-3 ή αδενοκαρκίνωμα in situ τα προηγούμενα 20 χρόνια ή καρκίνο του τραχήλου της μήτρας, μεταμόσχευση συμπαγών οργάνων, μόλυνση από τον ιο HIV ή με χρόνια ανοσοκατασταλτική φαρμακευτική αγωγή. Περιπτώσεις οι οποίες είχαν ήδη κλινική συνάντηση για τεστ Παπανικολάου, αποκλείστηκαν. Οι γυναίκες που προσδιορίστηκαν ως οι κατάλληλοι υποψήφιοι για την παρέμβαση έστειλαν ηλεκτρονικές υπενθυμίσεις για παρακολούθηση αν ήταν καταχωρημένες στην πύλη της κλινικής ασθενούς ή επιστολές υπενθυμίσεων αν δεν ήταν. Οι υπενθυμίσεις που στάλθηκαν, εξηγούσαν ότι ο ασθενής είχε εντοπιστεί ως υψηλότερος κίνδυνος ή είχε καθυστερήσει για παρακολούθηση ενός παρελθόντος μη φυσιολογικού αποτελέσματος. Ακόμη, ορισμένες πληροφορίες που συγκεντρώθηκαν αφορούσαν την ηλικία, την εθνικότητα, την ασφάλιση και την αγγλική επάρκεια από το δημογραφικό τμήμα του EHR.

Αξιοσημείωτο ωστόσο κρίνεται το γεγονός ότι υπήρξαν ορισμένοι περιορισμοί όπως η αδυναμία του CDSS στον εντοπισμό ασθενών που είχαν ήδη προγραμματίσει τεστ Παπ, ασθενών που ολοκλήρωσαν την παρακολούθηση των δοκιμών Παπ σε μία εξωτερική μονάδα ή είχαν εξαντλήσει τον έλεγχο λόγω περιορισμένου προσδόκιμου ζωής. Αυτός ο περιορισμός ξεπεράστηκε με ηλεκτρονική ιατρική ανασκόπηση από μια ομάδα μελέτης (KLM, MEK, MRS) ή από έναν κλινικό γιατρό πριν την αποστολή υπενθύμισης στην ασθενή.

Τα αποτελέσματα που προέκυψαν από την μελέτη είναι αναλυτικά τα εξής : από τους 393 ασθενείς που αξιολογήθηκαν το σύστημα έκανε σωστές συστάσεις για τους 369 ασθενείς με ακρίβεια 93,4% , 307 ασθενείς είχαν φυσιολογικά αποτελέσματα προηγούμενων εξετάσεων ενώ 87 είχαν μη φυσιολογικά. Συγκεκριμένα, η ακρίβεια του συστήματος ήταν υψηλότερη σε ασθενείς με φυσιολογικά αποτελέσματα του τεστ Παπ (96,7%) σε αντίθεση με ασθενείς με μη φυσιολογικά αποτελέσματα (83,7%). Ωστόσο, τα σφάλματα του συστήματος ήταν κυρίως σφάλματα προέλευσης δεδομένων, μοντελοποίησης, προγραμματισμού και αξιολόγησης. Σημαντική βελτίωση παρουσιάστηκε επίσης για την ομάδα παρέμβασης που έλαβε μια

υπενθύμιση σε σύγκριση με τον ιστορικό έλεγχο (από 5,7% σε 23,7%) ενώ τα αποτελέσματα της παρέμβασης του σε πραγματικό χρόνο στους παρόχους σε σύγκριση με την περίοδο ελέγχου παρουσίασαν βελτίωση, καθώς τα ποσοστά εξέτασης αυξήθηκαν από 32,6% σε 60,8%. Τα ποσοστά επιτήρησης δεν παρουσίασαν σημαντική διαφορά. Εν κατακλείδι το CDSS ήταν ακριβές από τους παρόχους που το χρησιμοποίησαν σε πραγματικό χρόνο κατά την διάρκεια συνάντησης ασθενών από τον Μάρτιο του 2016 και μετά.

1.7 Εφαρμογή της βαθιάς μάθησης για την ταξινόμηση εικόνων από κολποσκόπηση¹⁰

Στην συγκεκριμένη μελέτη χρησιμοποιείται η τεχνική της βαθιάς εκμάθησης (Deep Learning) η οποία αποτελεί μέρος των μεθόδων της μηχανικής μάθησης. Η μελέτη εγκρίθηκε από την επιτροπή θεωρητικής ηθικής του κέντρου καρκίνου saitama .

Τα δεδομένα που συλλέχθηκαν ανήκουν σε 158 ασθενείς ηλικίας 21-63 ετών με καρκίνο του τραχήλου της μήτρας ενώ χρησιμοποιήθηκαν και εικόνες από την κολποσκόπηση ως δεδομένα εισόδου. Οι διαγνώσεις ήταν 49 ασθενείς με σοβαρή δυσπλασία, 78 με carcinoma in situ (CIS), 21 με επιθετικό καρκίνο (IC) , 10 ασθενείς με αδενοκαρκίνωμα in situ και διηθητικό αδενοκαρκίνωμα κ.α. Η βαθιά εκμάθηση που χρησιμοποιήθηκε βασίστηκε στις βιβλιοθήκες Keras Neural Network 1.1.0 και TensorFlow 0.8.0 (βιβλιοθήκη ανοικτού κώδικα από την google) καθώς και σε κώδικα γραμμένο σε python. Οι ασθενείς χάρη στην χρήση Deep Learning κατατάχθηκαν σε 3 ομάδες οι οποίες ήταν σοβαρή δυσπλασία, carcinoma in situ (CIS) και διηθητικό καρκίνωμα (IC). Οι εικόνες που χρησιμοποιήθηκαν ήταν συνολικά 485 ενώ μόνο οι 142 από αυτές είχαν σοβαρή δυσπλασία (2,9 εικόνες / ασθενή) , οι 257 CIS (3,3 εικόνες / ασθενή) και 86 IC (4,1 εικόνες / ασθενή) . Οι 233 καταγράφηκαν με πράσινο φίλτρο ενώ οι υπόλοιπες 252 χωρίς . Η νομιμοποίηση L1 , η ρύθμιση L2 , η εγκατάλειψη και η αύξηση των δεδομένων επέφερε ακρίβεια στο σύνολο δεδομένων περίπου 50 %. Ωστόσο, οι προεγχειρητικές εικόνες χρησιμοποιήθηκαν ως είσοδος στην βαθιά εκμάθηση ενώ, έπειτα από επεξεργασία και προσθήκη πράσινου φίλτρου αποθηκεύτηκαν σε μορφή png (640x480) οι οποίες χρησιμοποιήθηκαν στις διαγνώσεις. Όσον αφορά το σύνολο επικύρωσης αποτελούνταν συνολικά από 75 εικόνες και συγκεκριμένα 25 τυχαίες για κάθε διάγνωση ενώ δεν χρησιμοποιήθηκε για εκπαίδευση. Ακόμη, οι εικόνες IC είχαν πράσινο φίλτρο και έτειναν να έχουν πιο μεγενθυμένες βλαβες συγκριτικά με εικόνες σοβαρής δυσπλασίας και CIS.

Η ακρίβεια ήταν μικρότερη από 33% πράγμα που σημαίνει ότι η εκμάθηση δεν συνέβη γιατί ο αριθμός εικόνων για κάθε διάγνωση αντιστοιχίστηκε στο σύνολο δεδομένων επικύρωσης.

Υπήρξαν περιπτώσεις όπου δημιουργήθηκε υπερφόρτωση με την ακρίβεια επικύρωσης να αγγίζει το 40 ~50% . Τέτοιες περιπτώσεις ήταν όταν τα πυκνά στρώματα συντονίστηκαν με τα στρώματα συνέλιξης και επηρέασαν τα ποσοστά εκμάθησης, με την ακρίβεια της εκπαίδευσης να ξεπεράσει το 90%. Προκειμένου να αποφευχθεί η υπερφόρτωση το σύνολο

¹⁰ Masakazu Sato, Koji Horie, Aki Hara, Yuichiro Miyamoto, Kazuko Kurihara et al., “*Application of deep learning to the classification of images from colposcopy*”, Oncology Letter Spandidos Publications (2018)

των εικόνων περιελάμβανε εικόνες με όχι τόσο πράσινο φίλτρο που χρησιμοποιήθηκαν για μάθηση, ενώ το σύνολο δεδομένων επικύρωσης επιλέχθηκε εκ νέου (10 εικόνες για κάθε διάγνωση, 30 συνολικά) με αποτέλεσμα την μείωση της ακρίβεια επικύρωσης ανεξαρτήτως της παρουσίας ή όχι πράσινου φίλτρου. Συνεπώς, ο συνολικός αριθμός εικόνων είναι πιο σημαντικός από την διαίρεση των δεδομένων εισόδου (παρουσία ή όχι πράσινου φίλτρου) προκειμένου να αυξηθεί η ακρίβεια επικύρωσης και να αποφευχθεί η υπερφόρτωση. Επιπλέον, η εφαρμογή κανονικοποίησης L2 συνέβαλε στην αποφυγή υπερφόρτωσης και η L1 προκάλεσε αποτυχία μάθησης στο πρώτο στρώμα εισόδου ενώ η εφαρμογή εγκατάλειψης (ρυθμός 0,5) στα υπόλοιπα επίπεδα μείωσε την αποτελεσματικότητα μάθησης. Στη συνέχεια ελήφθησαν 20 στιγμιότυπα (π.χ μεγεθύνσεις, περιστροφή κ.α) από μόνο μια εικόνα και χρησιμοποιήθηκαν ως είσοδος με σκοπό την αύξηση των δεδομένων ώστε να διερευνηθεί αν βελτιώνονται τα ποσοστά ακρίβειας. Όπως προέκυψε η αύξηση των δεδομένων επιδείνωσε τον περιορισμό υπερφόρτωσης ενώ η εφαρμογή L2, η νομιμοποίηση L1 και η απόρριψη βελτίωσε την υπερφόρτωση με αποτέλεσμα η ακρίβεια επικύρωσης να σταθεροποιηθεί στο ~50%.

Τα συμπεράσματα της μελέτης είναι ότι η υπερφόρτωση αποτελεί μια κατάσταση μη ανταπόκρισης του μοντέλου και παρατηρήθηκε πιθανότατα λόγω ελάχιστων δεδομένων, συνήθως χρησιμοποιούνται 500-1000 εικόνες . Επίσης, στόχος της μελέτης όπως αναφέρεται είναι να αποδείξει ότι οι ιατροί μη έχοντας γνώση της τεχνητής νοημοσύνης μπορούν να χρησιμοποιήσουν deep learning στην κλινική πρακτική. Παρόλο που το ποσοστό 50% φαίνεται μη ικανοποιητικό δεν είναι δυνατό να κατηγοριοποιηθεί αν είναι καλό ή κακό αφού δεν υπάρχουν παρόμοιες μελέτες.

1.8 Προσδιορισμός της πιθανότητας καρκίνου του τραχήλου της μήτρας με χρήση μεθόδων μηχανικής μάθησης¹¹

Στόχος της παρούσας μελέτης είναι η πρόληψη του καρκίνου του τραχήλου της μήτρας χρησιμοποιώντας αλγόριθμους μηχανικής μάθησης.

Το σύνολο δεδομένων συλλέχθηκε από το νοσοκομείο της Βενεζουέλας και περιελάμβανε 35 χαρακτηριστικά τα οποία αφορούσαν δημογραφικά στοιχεία, συνήθειες καθώς και ιστορικά αρχεία από 858 ασθενείς (π.χ ηλικία, αριθμός από εγκυμοσύνες, καπνιστές ή μη κ.α). Όπως προέκυψε 55 ασθενείς διαγνώστηκαν με καρκίνο του τραχήλου της μήτρας και 803 ήταν υγιείς. Τα δεδομένα κατηγοριοποιήθηκαν τυχαία σε 2 ομάδες ως σύνολο εκπαίδευσης το οποίο περιελάμβανε 66% του συνόλου δεδομένων (566 εγγραφές) και σύνολο δοκιμής (292 εγγραφές). Όσον αφορά τους δείκτες απόδοσης που χρησιμοποιήθηκαν είναι True Negative Class 0 (TNC0), False Positive Class 0 (FPC0) ,False Negative Class 1 (FNC1) και True Positive Class 1 (TPC1). Η ταξινόμηση των δεδομένων πραγματοποιήθηκε με μεθόδους μηχανικής μάθησης ώστε τα αποτελέσματα που θα προκύψουν να είναι αντικειμενικά. Οι

¹¹ Munhammed Fahri Unlesern, Kadir Sabanci, Muciz Özcan, “*Determining Cervical Cancer Possibility by Using Machine Learning Methods*”, International Journal of Latest Research in Engineering and Technology (IJLRET) (2017)

μέθοδοι που χρησιμοποιήθηκαν ήταν ο Multilayer Perceptron (MLP) , ο BayesNet και ο k-Nearest Neighbor (kNN).

Στην περίπτωση του k-Nearest Neighbor το κάθε χαρακτηριστικό ταξινομείται με βάση την πιο κοντινή απόσταση από τα άλλα χαρακτηριστικά. Στην παρούσα μελέτη ο αλγόριθμος k-NN γίνεται με πειράματα. Ο αλγόριθμος MLP έχει πολλά στρώματα κόμβων τα οποία συνδέονται μεταξύ τους ενώ χρησιμοποιείται κυρίως για την εκπαίδευση του δικτύου και τέλος ο αλγόριθμος BayesNet ο οποίος είναι πιθανοτικός αντιπροσωπεύει τις σχέσεις όπως μεταξύ ασθενειών και συμπτωμάτων.

Τα αποτελέσματα που προέκυψαν είναι ότι ο k-NN έχει αλλάξει μεταξύ 1 και 90 ενώ όσες περιπτώσεις δεν έχουν διαγνωστεί με καρκίνο ταξινομήθηκαν ως TNC0, έχουν διαγνωστεί με καρκίνο (FNC1) , δεν έχουν διαγνωστεί με καρκίνο ταξινομήθηκαν ως θετικές (FPC0) και τέλος όσες έχουν καρκίνο (TPC1). Καλύτερα αποτελέσματα παρουσίασε ο αλγόριθμος MLP ο οποίος έχει αλλάξει 1 σε 50 ενώ τα ίδια περίπου αποτελέσματα είχε ο και ο BayesNet.

Συγκεκριμένα τα ποσοστά ορθής ταξινόμησης για κάθε μια από τις μεθόδους ήταν 95,89% για τον k-NN, 95,89% για τον MLP και 97,26% για τον Bayes Net. Τα ψευδή ταξινομημένα τα ποσοστά ήταν 1,37%, 1,71% και 2,05% αντίστοιχα. Συνεπώς, παρόλο που το καλύτερο αποτέλεσμα φαίνεται να έχει ο BayesNet το ποσοστό των ψευδών ποσοστών είναι μεγαλύτερο, οπότε οι ειδικοί υποστηρίζουν ότι ο καλύτερος αλγόριθμος ταξινόμησης στην παρούσα μελέτη είναι ο k-NN με 86 γείτονες.

1.9 Εύρεση συμπτωμάτων καρκίνου του τραχήλου της μήτρας στο σουηδικό κλινικό κείμενο χρησιμοποιώντας μια προσέγγιση μηχανικής μάθησης και NegEx¹²

Όπως έχει προαναφερθεί και σε προηγούμενες μελέτες ο καρκίνος του τραχήλου της μήτρας επηρεάζει γυναίκες κάτω των 40 ετών με τα ποσοστά να παραμένουν υψηλά ακόμη και στις αναπτυσσόμενες χώρες.

Στόχος αυτής της μελέτης είναι η αξιολόγηση αναπτυσσόμενων εργαλείων εξόρυξης δεδομένων σχετικά με τα αρχεία των ασθενών που πάσχουν από καρκίνο του τραχήλου της μήτρας προκειμένου να γίνει ανίχνευση των πρώιμων συμπτωμάτων. Συγκεκριμένα όπως αναφέρεται η εξόρυξη δεδομένων που έχει πραγματοποιηθεί σε προηγούμενη έρευνα αφορά έγγραφα ογκολογικού περιεχομένου και παθολογικές αναφορές. Αρχικά, γίνεται αναφορά σε δύο άρθρα σχετικά με τον καρκίνο του τραχήλου της μήτρας εκ των οποίων στο πρώτο συλλέχθηκαν 250 αρχεία για πειράματα με 15 παραμέτρους για ταξινόμηση και ακρίβεια 0,73 μέσω νευρωνικού δικτύου. Ωστόσο, στο δεύτερο άρθρο χρησιμοποιήθηκαν 221 περιπτώσεις τραχηλικού ασθενούς ως είσοδο σε ένα σύστημα σταδιοποίησης (στάδια I-IV) με την χρήση νευρωνικών/υπολογιστικών δικτύων. Έπειτα ταξινομήθηκαν με βαθμολογία 0,79 στα δεδομένα δοκιμών καθώς έγινε εφαρμογή και του αλγορίθμου C4.5 με βαθμολογία

¹² Rebecka Weegar, Maria Kvist, Karin Sundström, Søren Brunak, Hercules Dalianis, “Finding Cervical Cancer Symptoms in Swedish Clinical Text using a Machine Learning Approach and NegEx”, AMIA Annual Symposium Proceedings Archive (2015)

0,80. Σε παρόμοια μελέτη χρησιμοποιήθηκαν οι αλγόριθμοι Naïve Bayes, Bayesian Network, Υποστηρικτικοί Μηχανισμοί Υποστήριξης και Αλγόριθμοι Τυχαίου Δάσους σε αναφορές παθολογίας παρουσιάζοντας μείωση επιδόσεων τουλάχιστον 25%.

Η αναγνώριση των οντοτήτων έγινε με τον Κλινικό Εντοπισμό Οντότητας (CEF) ένα εργαλείο βασισμένο στον αλγόριθμο μηχανικής μάθησης CRF++a εκπαιδευμένο σε σουηδικά κείμενα. Όμως, όπως αποκαλύπτει προηγούμενη προσέγγιση το συγκεκριμένο εργαλείο δεν διέθετε ανίχνευση άρνησης και για τον λόγο αυτό χρησιμοποιήθηκε το NegEx το οποίο βασίζεται σε κανόνες. Τα δεδομένα που χρησιμοποιήθηκαν, ήταν από τα τμήματα μαιευτικής/ γυναικολογίας και ογκολογίας από τα έτη 2009-2010 του πανεπιστημιακού νοσοκομείου Karolinska με κωδικό διάγνωσης κακοήγη νεοπλάσματος στον τράχηλο της μήτρας, συνολικά 646 εγγραφές ασθενών.

Οι ειδικοί χειροκίνητα έκαναν την διαδικασία σήμανσης των σχετικών οντοτήτων ή αλλιώς σχολιασμό που ακολουθεί μια σειρά κατευθυντήριων οδηγιών βασιζόμενες σε εκείνες που χρησιμοποιούνται κατά την ανάπτυξη του Clinical Entity Finder. Όλες οι σημειώσεις των γιατρών από το Cervical Cancer Corpus προ σχολιάστηκαν με το Clinical Finder και με την χρήση του εργαλείου NegEx ταξινομήθηκαν οι οντότητες ως αρνητικές ή μη. Χαρακτηριστικό παράδειγμα αποτελεί η έκφραση “Ο ασθενής δεν έχει πόνο” η οποία δηλώνει άρνηση μέσω μιας λέξης. Σχολιάστηκαν τουλάχιστον 100 σημειώσεις χειροκίνητα με δύο γιατρούς (A και B) όπου ο A ήταν ειδικός στον τομέα εξόρυξης κειμένου και στον σχολιασμό και ο B στον τομέα του καρκίνου της μήτρας. Ουσιαστικά πραγματοποιήθηκε αφαίρεση ελαττωματικών σχολιασμών, πρόσθεση σχολίων κ.α. Στη συνέχεια όλα τα ευρήματα που βρέθηκαν από τον Clinical Entity Finder διερευνήθηκαν από το NegEx για τυχόν αρνητικές αντιδράσεις.

Συμπερασματικά, σε ποσοστό 54% ο σχολιαστής A και τα εργαλεία συμφωνούσαν πλήρως. Επιπλέον, υπήρξαν 500 περιπτώσεις κατά τις οποίες η απόδοση του NegEx απομονώθηκε και από αυτές οι 57 επισημάνθηκαν ως αρνητικές τόσο από τον σχολιαστή A όσο και από το NegEx. Η ακρίβεια, η ανάκληση και η βαθμολογία F1 του NegEx για σωστό προσδιορισμό των αρνητικών τιμών ήταν 0,78 , 0,75 και 0,76 αντίστοιχα. Ακόμη, 333 στιγμιότυπα ήταν ψευδώς αρνητικά δηλαδή σχολιασμοί που επισημάνθηκαν από τον A αλλά χάθηκαν από τα εργαλεία και 71 ψευδώς θετικά δηλαδή σχολιασμοί που εκτελούνταν μόνο από τα εργαλεία. Ωστόσο, υπήρξαν περιπτώσεις κατά τις οποίες ο Clinical Entity Finder δεν εντόπισε λέξεις και εκφράσεις που σχετίζονται με τον καρκίνο του τραχήλου της μήτρας λόγω συντομογραφιών ή ορθογραφικών λαθών και θεωρήθηκαν ψευδώς αρνητικές όπως cervixcancer. Οι false positive περιπτώσεις προέκυψαν από ονομασίες φαρμάκων ή μέρη του σώματος. Συνολικά, το NegEx οδήγησε σε 16 false positives τα οποία προκλήθηκαν λόγω των διαταραχών που εμφανίζονται κοντά σε ανεξάρτητες αιτίες άρνησης και 19 false negatives λόγω του ότι δεν συμπεριλήφθηκαν αρνητικές εκφράσεις στους καταλόγους των αρνητικών ενεργοποιήσεων του NegEx από 500 περιπτώσεις. Καταληκτικά, από τις 776.719 σημειώσεις ο Clinical Entity Finder εντόπισε 58.366 διαταραχές και μέρη του σώματος από τις οποίες 43.334 ταξινομήθηκαν ως ευρήματα ή διαταραχές και 12% προσδιορίστηκαν να αφαιρεθούν από το NegEx.

Είναι σημαντικό να τονιστεί ότι η μελλοντική δουλειά αφορά την επέκταση των δεδομένων εκπαίδευσης του Clinical Entity Finder με εγχειρίδια μη αυτόματης επισήμανσης του καρκίνου του τραχήλου της μήτρας πράγμα το οποίο θα μπορούσε να βελτιώσει τα αποτελέσματα μειώνοντας τον αριθμό των false negatives.

1.10 Ένα σύστημα υποστήριξης αποφάσεων βασισμένο σε ένα σύνολο Random Forests για τη βελτίωση της διαχείρισης των γυναικών με μη φυσιολογικά ευρήματα στην εξέταση του καρκίνου του τραχήλου της μήτρας¹³

Η παρούσα μελέτη αποτελεί συνέχεια προηγούμενων μελετών οι οποίες είχαν επικεντρωθεί στην δημιουργία ενός CDSS βασιζόμενου σε 6 ταξινομητές όπως k-NN, CART, NB κ.τλ του οποίου τα αποτελέσματα δεν ήταν ικανοποιητικά. Συνεπώς, έγινε προσπάθεια να συνδυαστεί ένα PNN για όλες τις ταξινομήσεις πλην των περιπτώσεων με κυτταρολογία ASCUS και MLP για ταξινόμηση μόνο των περιπτώσεων με κυτταρολογία ASCUS.

Το νέο CDSS αποτελείται από ένα σύνολο RFs(Random Forests) όπου κάθε RF χρησιμοποιείται για τη διαχείριση διαφορετικών περιπτώσεων κυτταρολογίας. Τα δεδομένα συλλέχθηκαν τυχαία και ανώνυμα από γυναίκες εγγεγραμμένες στην μελέτη του Τμήματος Κυτταρολογίας της Ιατρικής Σχολής του Πανεπιστημίου Αθηνών (Πανεπιστημιακό Νοσοκομείο ΑΤΤΙΚΟΝ) . Όσον αφορά το περιεχόμενο τους ήταν κυρίως γενετικές εξετάσεις HPV, κυτταρολογικές διαγνώσεις, βιοψίες, ηλικία και άλλα στοιχεία τα οποία αποθηκεύτηκαν σε βάση δεδομένων και επιλέχθηκαν 740 περιπτώσεις με πλήρη σειρά εξετάσεων. Παρακάτω αναλύεται ο τρόπος λειτουργίας του RF το οποίο ουσιαστικά είναι σύνολο χωρίς κλάδεμα δέντρων ταξινόμησης που δημιουργήθηκαν από δείγματα εκκίνησης των δεδομένων εκπαίδευσης χρησιμοποιώντας επιλογή τυχαίων στοιχείων στη διαδικασία επαγωγής των δέντρων. Τα δέντρα αυτά εκπαιδεύονται ανεξάρτητα και οι προβλέψεις συνδυάζονται μέσω της ομαδοποίησης προκειμένου να παραχθεί η πρόβλεψη.

Τα βήματα κατασκευής όπως περιγράφονται σε κάθε δέντρο RF είναι τα εξής:

Ως σύνολο κατάρτισης χρησιμοποιείται ένα δείγμα εκκίνησης διαφορετικό από τα αρχικά δεδομένα για την ανάπτυξη του κάθε δέντρου. Το ένα τρίτο των περιπτώσεων παραμέτρου εκτός του δείγματος εκκίνησης δεν χρησιμοποιούνται στην κατασκευή του k δέντρου. Έπειτα, σε περίπτωση που υπάρχουν μεταβλητές εισόδου M , όπου $m < M$, οι μεταβλητές m ορίζονται με τέτοιο τρόπο ώστε σε κάθε κόμβο του δέντρου να επιλέγονται τυχαία από το M και ο καλύτερος διαχωρισμός σε αυτές τις m μεταβλητές να χρησιμοποιείται για τον διαίρεση του κόμβου. Καταληκτικά, κάθε δέντρο αναπτύσσεται όσο τον δυνατόν περισσότερο. Υπογραμμίζεται ότι, η τιμή του m διατηρείται σταθερή κατά της διάρκεια της ανάπτυξης του δέντρου ενώ στον συγκεκριμένο αλγόριθμο όπως έχει προαναφερθεί δεν χρησιμοποιείται το

¹³ Panagiotis Bountris, Maria Hartiou, Abraham Pouliakis, Petros Karakitsos, Dimitrios Koutsouris, “A decision support system based on an ensemble of random forests for improving the management of women with abnormal findings at cervical cancer screening”, Conf Proc IEEE Eng Med Biol Soc. (2015)

κλάδεμα. Το CDSS που αναπτύχθηκε χρησιμοποιεί τέσσερα WRF (αλγόριθμος ο οποίος προσαρμόζει τα σταθμικά βάρη προκειμένου να επιτευχθούν τα επιθυμητά αποτελέσματα) για την ταξινόμηση των γυναικών με κυτταρολογία WLN και HPV θετική, με κυτταρολογία ASCUS, LSIL και υποθέσεων HSIL. Τονίζεται, ότι περιπτώσεις κατά τις οποίες το τεστ Pap ήταν θετικό ήταν ελάχιστες (15) με αποτέλεσμα να μη δημιουργηθεί ταξινομητής ενώ κάθε WRF λαμβάνει ως είσοδο 17 χαρακτηριστικά (HPV16, HPV8, NASBA16, NASBA18 κ.α) και ταξινομεί σε 4 ομάδες : αρνητική CIN1,CIN2/3 και CxCa. Οι περιπτώσεις κάθε ομάδας χρησιμοποιήθηκαν για την δημιουργία του σετ εκπαίδευσης καθενός από τα τέσσερα WRF αντίστοιχα. Έπειτα, για κάθε WRF έγινε εκτίμηση των παραμέτρων παρατηρώντας το σφάλμα oob κατά την διάρκεια της κατάρτισης, του ελάχιστου αριθμού δεντρών που περιλαμβάνεται σε κάθε WRF και της βέλτιστης τιμής m των μεταβλητών που χρησιμοποιήθηκαν για την κατάρτιση κάθε δέντρου. Ο υπολογισμός του πίνακα κόστους του κάθε WRF πραγματοποιήθηκε χρησιμοποιώντας βάρος ίσο με το ποσοστό του για κάθε κλάση. Παρατηρήθηκαν λανθασμένες ταξινομημένες out-of-bag περιπτώσεις, προσαρμόστηκαν τα βάρη για μείωση λανθασμένων περιπτώσεων ASCUS ή LSIL με ιστολογία CIN2+ και περιπτώσεων ψευδών CIN2+. Για να υπάρξει σωστή εκτίμηση του σφάλματος κάθε WRF, το ολοκληρωμένο σύστημα εκπαιδεύτηκε με τον αλγόριθμο 10 fold cross-validation σε όλο το σύνολο δεδομένων. Όμως, το σύστημα για την ανίχνευση HPV έπρεπε κάθε εκπαιδευμένο WRF να ταξινομήσει όλες τις περιπτώσεις του συνόλου δεδομένων ενώ για κάθε περίπτωση να εξάγει το αποτέλεσμα μαζί με την πιθανότητα της πιο επιτυχημένης κλάσης. Τα αποτελέσματα μέσω ενός πλειοψηφικού συστήματος ψηφοφορίας ταξινομήθηκαν σε CIN1 ή CIN2+.

Συμπερασματικά, τα αποτελέσματα της μελέτης παρουσιάζονται στον παρακάτω πίνακα.

	SN%	SP%	PPV%	NPV%	OA%
Pap test (cut-off ASCUS+)	98.1	45.3	33.3	98.9	56.8
Pap test (cut-off LSIL+)	89.4	67.0	43.0	96.0	71.9
Pap test (cut-off HSIL+)	71.4	95.3	81.0	92.3	90.1
HPV DNA test	91.9	61.5	39.9	96.5	68.1
HR-HRV DNA	89.4	67.4	43.2	95.8	72.2
NASBA test	77.0	90.2	68.5	93.4	87.3
FLOW	93.2	81.9	58.8	97.7	84.3
p16	58.4	92.9	69.6	88.9	85.4
WRF (single classifier approach)	86.3	89.5	70.6	96.0	88.8
Proposed System (Fig. 1)	86.1	93.3	77.7	96.1	91.8
Modified System for HPV-based Screening (Fig.2)	73.9	93.7	77.3	92.8	89.6

Histology endpoint is CIN2+ for all cases. CIN2+: Cervical intraepithelial neoplasia grade 2 or worse. HR-HPV: High-risk human papillomavirus. NASBA: Nucleic Acid Sequence Based Amplification for the identification of E6/E7 mRNA of the HPV types: 16, 18, 31, 33 and 45. FLOW: Flow cytometric E6/E7 HPV mRNA assay. p16: Immunocytochemical expression of p16. SN: Sensitivity. SP: Specificity. PPV: Positive Predictive Value. NPV: Negative Predictive Value. OA: Overall Accuracy.

Εικόνα 1.10.α : Αποτελέσματα Μελέτης (SN, SP, PPV, NPV, OA)

Σημειώνεται ότι τα μέτρα απόδοσης είναι μέσες τιμές ενώ κάθε WRF έχει εκπαιδευτεί με τις περιπτώσεις της αντίστοιχης κυτταρολογικής ομάδας ενώ τα περιστατικά άλλων ομάδων είναι άγνωστα για αυτά. Το WRF αποτελείται από 80 δέντρα όπου κάθε δέντρο χρησιμοποιεί 5 μεταβλητές εισόδου και συγκριτικά με ιατρικά τεστ και την ενιαία προσέγγιση WRF παρήγαγε τα πιο ισορροπημένα αποτελέσματα όσον αφορά την ευαισθησία, την εξειδίκευση, το PPV, NPV και της συνολικής ακρίβειας. Αντίθετα, για την περίπτωση του HPV το σύστημα παρουσίασε καλύτερη εξειδίκευση, το καλύτερο PPV και την καλύτερη ακρίβεια σε σύγκριση με του βιοδείκτες. Εν κατακλείδι, το CDSS παρέχει πιθανότητες αξιολόγησης κινδύνου μαζί με το αποτέλεσμα της ταξινόμησης, χειρισμό ελλείψεων δοκιμών/τιμών χωρίς την ανάγκη προεπεξεργασίας δεδομένων αφού τα RF χειρίζονται ελλιπή δεδομένα. Επιπλέον, μπορεί να χειριστεί περιστατικά με ανεπαρκή κυτταρολογική έκβαση συνεπώς μπορεί να παράγει ακριβή αποτελέσματα στην περίπτωση αυτόνομου διαγνωστικού ελέγχου με βάσει τον HPV όπου δεν εφαρμόζεται δοκιμασία Pap.

1.11 Αξιολόγηση της απόδοσης των τεχνικών μηχανικής μάθησης για τον έλεγχο του καρκίνου του τραχήλου της μήτρας¹⁴

Ο καρκίνος του τραχήλου της μήτρας προκαλείται συνήθως από τον ιο HPV, τα είδη του οποίου είναι σε αριθμό πάνω από 120. Στην παρούσα μελέτη γίνεται αξιολόγηση των μεθόδων μηχανικής μάθησης χρησιμοποιώντας δεδομένα από την βάση δεδομένων Benchmark Pap smear η οποία περιέχει 1417 εικόνες από Pap smear ενώ δημιουργήθηκε από το πανεπιστημιακό νοσοκομείο Herlev της Δανίας.

Από τις 1417 εικόνες οι 500 είναι από παλιά βάση δεδομένων του 2003 και οι 917 από μία νέα του 2005 ενώ κατατάσσονται άνισα σε 7 κατηγορίες τα δείγματα των οποίων χρησιμοποιούν 20 χαρακτηριστικά που λαμβάνονται από τις εικόνες του κάθε κελιού. Τα χαρακτηριστικά περιέχουν την περιοχή του πυρήνα, του κυτταροπλάσματος, τον λόγο πυρήνων/κυτταροπλάσματος, την φωτεινότητα των πυρήνων, την φωτεινότητα του κυτταροπλάσματος, την μικρότερη διάμετρο του πυρήνα, την μεγαλύτερη διάμετρο του κυτταροπλάσματος, την επιμήκυνση του πυρήνα, την στρογγυλότητα του πυρήνα, την μικρότερη διάμετρο του κυτταροπλάσματος, την μεγαλύτερη διάμετρο του κυτταροπλάσματος, την επιμήκυνση του κυτταροπλάσματος, την στρογγυλότητα κυτταροπλάσματος, την θέση του πυρήνα, τα μέγιστα και ελάχιστα στον πυρήνα και τα μέγιστα και ελάχιστα στο κυτταρόπλασμα. Ο πίνακας που ακολουθεί περιέχει τον συνολικό αριθμό των διαφόρων τύπων εικόνων Παπανικολάου από κάθε κατηγορία, η κατάταξη τους έγινε χειροκίνητα σε 7 κατηγορίες με αποτέλεσμα η τελική βάση δεδομένων να περιέχει πιο ακριβείς διαγνώσεις.

¹⁴ Abid Sarwar, Mehbob Ali, Jyotsna Suri, Vinod Sharma, “*Performance Evaluation of Machine Learning Techniques for Screening of Cervical Cancer*”, 2nd International Conference on Computing for Sustainable Global Development (2015)

Class	Category	Type of cell	Number of Cell images	Sub total
1	Normal	Superficial squamous epithelial	124	392 instances of normal cells
2		Intermediate squamous epithelial	120	
3		Columnar epithelial	148	
4	Abnormal	Mild squamous non-keratinizing dysplasia	232	1025 instances of abnormal cells
5		Moderate squamous non-keratinizing dysplasia	246	
6		Severe squamous non-keratinizing dysplasia	297	
7		Squamous carcinoma in situ intermediate	250	

Πίνακας 1.11.α : Συνολικός αριθμός τύπων εικόνων Παπανικολάου για φυσιολογικά και μη κύτταρα

Η βάση δεδομένων έπειτα διαιρείται με κώδικα σε Matlab σε 2 σύνολα εκ των οποίων το πρώτο γίνεται με τυχαία επιλογή του 75% των πλειάδων, ομοιόμορφα κατανεμημένα μεταξύ των 7 κατηγοριών για εκπαίδευση ενώ το 25% είναι για έλεγχο. Στο δεύτερο σύνολο δεδομένων οι περιπτώσεις εκπαίδευσης δημιουργήθηκαν από το σετ δοκιμών του πρώτου και περίπου 66% των στιγμιότυπων εκπαίδευσης του πρώτου συνόλου με αποτέλεσμα τα δεδομένα εκπαίδευσης να είναι περίπου το 75% της συνολικής βάσης δεδομένων το ίδιο δηλαδή με το μέγεθος συνόλου των περιπτώσεων κατάρτισης του πρώτου συνόλου. Το υπόλοιπο σχεδόν 33% των περιπτώσεων εκπαίδευσης του πρώτου συνόλου διατηρήθηκε ως έλεγχος για το δεύτερο σύνολο, που είναι περίπου το 25% της συνολικής βάσης δεδομένων. Στην συνέχεια αναφέρονται και αναλύονται οι αλγόριθμοι που χρησιμοποιήθηκαν, τα αποτελέσματα των οποίων παρουσιάζονται συγκεντρωμένα στον παρακάτω πίνακα.

S No.	Prediction model	First Dataset			Second Dataset		
		Correctly classified (%)	Incorrectly classified (%)	ROC Area	Correctly classified (%)	Incorrectly classified (%)	ROC Area
1.	Bagging	69.9717%	30.0283%	0.943	67.4286%	32.5714%	0.939
2	PART	62.6062%	37.3938%	0.818	61.4286%	38.5714%	0.929
3	Decision Table	66.0057%	33.9943%	0.912	62.00%	38.00%	0.903
4	Decorate	74.2210%	25.7790%	0.943	66.5714%	33.4286%	0.935
5	END	77.3881%	28.6119%	0.943	72.2857%	27.7143%	0.944
6	Filtered Classifier	62.3229%	37.6771%	0.834	63.4286%	36.5714%	0.85
7	J48 Graft	64.5892%	35.4108%	0.823	67.7143%	32.2857%	0.83
8	Multi-Class Classifier	70.5382%	429.4618%	0.935	68.8571%	31.1429%	0.933
9	MBP ANN	67.5141%	32.4859%	0.921	62.8571%	37.1429%	0.905
10	Naive Bayes	50.1416%	49.8584%	0.896	60.8571%	39.1429%	0.907
11	Radial Basis Function	71.3881%	28.6119%	0.938	68.8571%	31.1429%	0.935
12	Random Committee	70.8215%	29.1785%	0.936	65.4286%	34.5714%	0.921
13	Random Forest	67.7054%	32.2946%	0.937	68.00%	32.00%	0.92
14	Random Sub Space	77.0538%	22.9462%	0.959	71.1429%	28.8571%	0.947
15	Rotation Forest	54.5706%	45.4294%	0.94	67.7143%	32.2857%	0.934

Πίνακας 1.11.β : Αποτελέσματα Αλγορίθμων

Επισημαίνεται ακόμη ότι τα εργαλεία που χρησιμοποιήθηκαν ήταν το WEKA (έκδοση 3.6.10) το οποίο περιέχει μία συλλογή αλγορίθμων και MBP (Multiple Back-Propagation έκδοση 2.2.4) λογισμικό που χρησιμοποιείται κυρίως για νευρωνικά δίκτυα. Τα δεδομένα μετατράπηκαν στην κατάλληλη μορφή αρχείου(.arff) ώστε να καθίσταται δυνατή η χρήση τους στο weka αντίθετα για το MBP δίκτυα δύο κρυμμένων στρώσεων 20 και 3 νευρώνων αντίστοιχα, εκτός από 20 εισόδους και έναν νευρώνα εξόδου. Τα δίκτυα αυτά εκπαιδεύτηκαν με 1063 στιγμιότυπα εικόνων και για έλεγχο 354 περιπτώσεις αντίστοιχων συνόλων δεδομένων

Τα αποτελέσματα που προέκυψαν είναι ότι ο αλγόριθμος END είχε την μέγιστη απόδοση και για τα δύο σύνολα δεδομένων (77% και 72% αντίστοιχα). Για το πρώτο σύνολο που είναι multiclass (πρόβλημα τάξης 7) ο END το μετατρέπει σε κατηγορία 2 ωστόσο για το δεύτερο σύνολο τα δεδομένα ήταν ήδη στην κατάλληλη μορφή για τον END. Εν αντιθέση με τον Naive Bayes που είχε την μικρότερη ακρίβεια προβλέψεων (περίπου 50% στο πρώτο σύνολο και 60% στο δεύτερο σύνολο δεδομένων). Συνεπώς, ο αλγόριθμος END βοηθάει του ιατρούς στην ακριβή διάγνωση για τον καρκίνο του τραχήλου της μήτρας.

1.12 Παρακολούθηση ενσωματώσεων βαθιάς μάθησης για την πρόβλεψη του καρκίνου του τραχήλου της μήτρας¹⁵

Η παρούσα μελέτη βασίζεται στα τεχνητά νευρωνικά δίκτυα (Artificial Intelligence Networks) τα οποία αποτελούν μία μέθοδο μηχανικής μάθησης προκειμένου να γίνει διάγνωση του καρκίνου του τραχήλου της μήτρας.

Αρχικά, επισημαίνεται ότι χρησιμοποιήθηκαν δεδομένα μικρότερης διάστασης σε σχέση με την αρχική καθώς τα δεδομένα μεγάλης διάστασης προκαλούν αρκετά προβλήματα όπως το υψηλό υπολογιστικό κόστος. Οι προτεινόμενες τεχνικές μείωσης των διαστάσεων που περιγράφηκαν ήταν οι εξής:

- Μείωση και ταξινόμηση των κοινών διαστάσεων (*Joint dimensionality reduction and classification*)
- Πλήρως επιτηρούμενες ενσωματώσεις (*Fully supervised embeddings*)
- Αυτοέλεγχος με βαθιά επίβλεψη (*Deep supervised autoencoders*)

Όσον αφορά τα δεδομένα που χρησιμοποιήθηκαν, ήταν ιατρικά αρχεία 858 ασθενών ηλικίας 13 με 84 ετών, μεταξύ του 2012-2014 του Πανεπιστημιακού Νοσοκομείου στο Καρακας της Βενεζουέλας. Συγκεκριμένα, στα ιατρικά αρχεία συμπεριλαμβάνεται εκτός από την ηλικία της ασθενούς, η σεξουαλική δραστηριότητα όπου στην προκειμένη περίπτωση όλες οι ασθενείς ήταν σεξουαλικά ενεργές, ο αριθμός των κυήσεων (98% ήταν έγκυες τουλάχιστον μία φορά), το κάπνισμα, η χρήση αντισυλληπτικών και ιστορικά αρχεία σχετικά με τα σεξουαλικά μεταδιδόμενα νοσήματα. Τα χαρακτηριστικά αυτά μετατράπηκαν σε boolean μεταβλητές είτε η ασθενής απάντησε ή όχι στην ερώτηση, είτε αν απάντησε ποιά ήταν η

¹⁵ Kelwin Fernandes, Davide Chicco, James S. Cardoso, Jessica Fernandes, “*Supervised deep learning embeddings for the prediction of cervical cancer diagnosis*”. PeerJ Comput. Sci 4:3154; DOI 10.7717/peerj-cs.154 (2018)

απάντηση. Όμως, κάποιες ασθενείς δεν απάντησαν λόγω του ότι ανησύχησαν για την ιδιωτικότητα τους, σε τέτοιες περιπτώσεις προστέθηκε ένα νέο γνώρισμα το οποίο επέτρεψε στο μοντέλο να αναγνωρίσει αν η ερώτηση είχε απαντηθεί. Έπειτα, κωδικοποιήθηκαν τα χαρακτηριστικά χρησιμοποιώντας το σχήμα one-of-K και όλα τα αρχεία ανωνυμοποιήθηκαν από το νοσοκομείο πριν δημιουργήσει το σύνολο δεδομένων και το δημοσιοποιήσει στο διαδικτυακό τόπο Machine Learning Repository του Πανεπιστημίου της Καλιφόρνιας. Στη συνέχεια έγινε κλιμάκωση των χαρακτηριστικών με κανονικοποίηση [0,1] και εισαγωγή των δεδομένων που έλειπαν με μέση τιμή. Όπως αναφέρεται χαρακτηριστικά δεν χρησιμοποιήθηκε ο αλγόριθμος k-nearest neighbor για λόγους πολυπλοκότητας. Η τιμή 0 υποδηλώνει έλλειψη του κινδύνου ενώ η τιμή 1 τον μέγιστο κίνδυνο.

Η αξιολόγηση των μεθόδων έγινε με 10 fold cross-validation ενώ παράλληλα βελτιστοποιήθηκαν τα νευρωνικά δίκτυα με την στρατηγική RMSProp για μέγιστο αριθμό 500 εποχών και πρόωρης διακοπής μετά από 100 επαναλήψεις χωρίς βελτίωση και μέγεθος παρτίδας 32. Επιπλέον, αναμορφώθηκε το μοντέλο με ολόκληρο το σετ εκπαίδευσης.

Τα αποτελέσματα που προέκυψαν, φανερώνουν ότι συγκριτικά με άλλα νευρωνικά δίκτυα ήταν καλύτερα. Το σύστημα χρησιμοποιεί συνάρτηση απώλειας και επιτρέπει την βελτιστοποίηση της μείωσης των διαστάσεων ενώ έχει μεγάλη ακρίβεια ακόμη και αν δεν υπάρχουν στοιχεία σχετικά με την βιοψία. Εν κατακλείδι, αναφέρεται ότι στο μέλλον επιδιώκεται να χρησιμοποιηθεί ο k-nearest neighbor για τα ελλιπή δεδομένα καθώς και να εμπλουτιστεί το σύστημα ώστε να αναφέρει τα πιο συναφή χαρακτηριστικά στο σύνολο δεδομένων.

1.13 Πρόβλεψη του καρκίνου του τραχήλου της μήτρας με χρήση υβριδικής τεχνικής επαγωγής: Μια λύση για τα μοντέλα κληρονομικής νόσου του ανθρώπου¹⁶

Τα δεδομένα που συλλέχθηκαν ήταν στο NCBI (National Center for Bio-technology Information), αποτελούμενα από 500 εγγραφές και 61 μεταβλητές, όπως η αριθμητική τιμή της βιοψίας. Ο πρώτος αλγόριθμος που χρησιμοποιήθηκε ήταν ο Regression Tree τα αποτελέσματα του οποίου όσον αφορά την ακρίβεια ήταν 83,87% με έξοδο CART TREE. Αντίθετα, ο αλγόριθμος Random Forest Tree (RFT) πέτυχε με κωδικοποίηση MATLAB ακρίβεια 93,54% με έξοδο RTF TREE. Ωστόσο, προκειμένου να υπάρξει μεγαλύτερη ακρίβεια έγινε συνδυασμός δύο αλγορίθμων, συγκεκριμένα του k-means και RTF η ακρίβεια των οποίων ήταν μεγαλύτερη από των προηγούμενων (96,77%) με έξοδο RFT-KMean Learning Tree.

Ο αλγόριθμος CART προβλέπει την πιο πιθανή τιμή της ανεξάρτητης μεταβλητής, διαιρεί όλο το σύνολο δεδομένων σε υποσύνολα. Χρησιμοποιεί τον δείκτη GINI για να αποφασίσει ποια χαρακτηριστικά θα επιλεγούν ενώ τα αποτελέσματα θα πρέπει να είναι η επιλογή του

¹⁶ R. Vidya, G. M. Narisa, “Prediction of Cervical Cancer using Hybrid Induction Technique: A Solution for Human Hereditary Disease Patterns”, Indian Journal of Science and Technology, Vol 9(30), DOI: 10.17485/ijst/2016/v9i30/82085 (2016)

χαρακτηριστικού με τον μικρότερο δείκτη GINI. Στην παρούσα μελέτη το δέντρο CART περιέχει 500 εγγραφές όπου 100 εγγραφές με 60 training set και 40 testing set τοποθετούνται στην πρώτη θέση. Για κάθε χαρακτηριστικό ο αλγόριθμος εφαρμόζεται στο σύνολο δεδομένων NCBI για να βρεθεί η ρίζα και το φύλλο του δέντρου μέχρις ότου δημιουργηθεί το δέντρο απόφασης. Επίσης, για επίτευξη μεγαλύτερης ακρίβειας χρησιμοποιείται κλάδεμα, με τον αλγόριθμο να επιτυγχάνει ποσοστό πρόβλεψης 83,37%.

Ο αλγόριθμος RFT χειρίζεται μεγάλο σύνολο δεδομένων εισόδου και επιλέγει την σωστή μεταβλητή ώστε να παράγει την επιθυμητή έξοδο. Χρησιμοποιεί δύο τρόπους τυχαιοποίησης οι οποίοι είναι είτε τυχαία δειγματοληψία δεδομένων είτε τυχαία επιλογή χαρακτηριστικών εισόδου για δημιουργία διαφορετικών δέντρων απόφασης. Τα συμπεράσματα που προέκυψαν έπειτα από την χρήση του κώδικα στο MATLAB με δείγματα 100 εγγραφών πέτυχε ακρίβεια 93,54%.

Ο τελευταίος αλγόριθμος που αναλύθηκε ήταν ο RTF σε συνδυασμό με τον K-means. Ουσιαστικά, η προεπεξεργασία των δεδομένων έγινε με τον k-means και τα αποτελέσματα εισήχθησαν στον RTF. Η ακρίβεια του συγκεκριμένου μοντέλου ήταν 96,77% στην πρόβλεψη, πράγμα το οποίο σήμαινει ότι μπορούν να δημιουργηθούν καλύτερα μοντέλα πρόβλεψης καθώς και τα συστήματα υποστήριξης αποφάσεων να καθιστούν πιο αποτελεσματικά.

Παρακάτω παρουσιάζεται πίνακας με τα αποτελέσματα πρόβλεψης κάθε αλγορίθμου, όπου χρησιμοποιήθηκαν 100 δεδομένα δοκιμών και 60 δεδομένα εκπαίδευσης.

No.	Methods	Prediction in percentage
1	CART	83.87%
2	RFT	93.54%
3	RFT with K-mean	96.77%

Πίνακας 1.13 : Αποτελέσματα πρόβλεψης αλγορίθμων

Το σημείο υπεροχής στην παρούσα μελέτη σε σχέση με άλλες είναι το γεγονός ότι δεν χρησιμοποιούνται εικόνες αλλά μόνο αριθμητικά δεδομένα. Όπως άλλωστε επισημαίνεται η πρόβλεψη εμφάνισης του καρκίνου της μήτρας αποτελεί δύσκολο έργο λόγω των μη διαθέσιμων και κατάλληλων συνόλων δεδομένων.

1.14 Έλεγχος του καρκίνου του τραχήλου της μήτρας με χρήση τεχνητής νοημοσύνης βασισμένη στην ανάλυση των ψηφιοποιημένων εικόνων κηλίδας Παπανικολάου¹⁷

Στην παρούσα μελέτη αναλύονται διάφορες τεχνικές τεχνητής νοημοσύνης προκειμένου να διαγνωστεί ο καρκίνος του τραχήλου της μήτρας με την χρήση εικόνων.

Τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση των αλγορίθμων αποτελούνταν από 8091 τραχηλικά κύτταρα τα οποία είχαν διαμορφωθεί βασισμένα σε 40 μορφολογικές ιδιότητες. Από αυτές τις ιδιότητες 19 χαρακτηριστικά ανήκουν στο κυτταρόπλασμα, 19 στον πυρήνα, 1 αφορά την περιοχή του πυρήνα και του κυτταροπλάσματος και 1 τελευταίο περιέχει την επιγραφή που προσδιορίζει σε ποια κατηγορία Bethesda ανήκει το κύτταρο. Η λήψη των δεδομένων πραγματοποιήθηκε από τρία ιδρύματα υγειονομικής περίθαλψης της Ινδίας.

Έπειτα, οι εικόνες που είχαν την περιοχή ενδιαφέροντος μεγενθύνθηκαν (40x) και επεξεργάστηκαν ώστε να είναι πλέον ορατά τα κύτταρα. Στη συνέχεια, τα μεμονωμένα κύτταρα αποκομίστηκαν από τις ομάδες κυττάρων και δόθηκαν μοναδικά αναγνωριστικά ώστε να μπορεί κάθε κύτταρο να αναγνωριστεί στην βάση δεδομένων. Η βάση δεδομένων ελέγχθηκε από παθολόγους προκειμένου τα δεδομένα να είναι ακριβεί και έγκυρα. Επισημαίνεται επίσης, ότι χρησιμοποιήθηκε το ανοικτό λογισμικό “CellProfiler” ώστε να αξιολογήσει τα κύτταρα με βάση 39 μορφολογικά χαρακτηριστικά. Όσον αφορά τους αλγόριθμους που χρησιμοποιήθηκαν ήταν οι εξής 15:

- **Random subset space** : Τα δέντρα απόφασης χρησιμοποιούνται ως ταξινομητές βάσης.
- **Decorate** : Προσθέτει σε κάθε επανάληψη στο τρέχον σύνολο έναν νέο μαθηματικό ταξινομητή. Ο ταξινομητής προστίθεται μόνο εάν το συνολικό σφάλμα μειώνεται.
- **Ensemble of Nested dichotomies (END)** : Μετασχηματίζει ένα πρόβλημα κατηγοριοποίησης σε ένα πιο εύκολο.
- **Back propagation artificial neural network** : Αποτελείται από δύο υποδίκτυα (main και space network).
- **J48 graft** : Δημιουργεί ένα ειδικό δέντρο αποφάσεων προσθέτοντας νέους κόμβους.
- **Radial basis function network (RBF)** : Χρησιμοποιεί την μη γραμμική και την γραμμική λειτουργία ενεργοποίησης.
- **Random Committee** : Συναθροίζει μεμονωμένους ταξινομητές πολλαπλών δέντρων οι οποίοι εκπαιδεύονται και δοκιμάζονται στο ίδιο σύνολο δεδομένων.
- **Decision Table** : Μετατρέπει το σύνολο δεδομένων εκπαίδευσης σε πίνακα αποφάσεων.
- **PART** : Παράγει ένα σύνολο κανόνων (λίστες αποφάσεων) ενώ αποτελεί συνδυασμό του κανόνα RIPPER και του αλγορίθμου C4.5.

¹⁷ Rajeev Gupta, Abid Sarwa, Vinod Sharma, “Screening of Cervical Cancer by Artificial Intelligence based Analysis of Digitized Papanicolaou-Smear Images”, International Journal of Contemporary Medical Research Volume 4 | Issue 5 (2017)

- **Naïve Bayes** : Δημιουργεί ταξινομητές χρησιμοποιώντας χαρακτηριστικά ανεξάρτητα μεταξύ τους.
- **Multi-class classifier** : Μετατρέπει ένα σύνθετο πρόβλημα σε απλό πρόβλημα δύο τάξεων.
- **Filtered Classifier (FC)**
- **Rotation Forest** : Δημιουργεί ένα σύνολο ή ταξινομητές χρησιμοποιώντας διαφορετικό υποσύνολο χαρακτηριστικών από το κύριο σύνολο δεδομένων.
- **Random forest** : Δημιουργεί πολλαπλά δέντρα απόφασης.
- **Bagging** : Ο αλγόριθμος χωρίζει το σύνολο δεδομένων σε διάφορα υποσύνολα τα οποία χρησιμοποιούνται για την εκπαίδευση μοντέλων.

Τα δεδομένα μετατράπηκαν σε μορφή ARFF προκειμένου να χρησιμοποιηθούν στο WEKA ενώ τα δεδομένα εκπαίδευσης αποτελούσαν το 70% (5664 κυτταρικές εικόνες) και το 30% (2427 κυτταρικές εικόνες) για δοκιμή. Στο πείραμα που πραγματοποιήθηκε στην βάση δεδομένων χρησιμοποιήθηκαν νευρωνικά δίκτυα όπου στην πρώτη έγινε δοκιμή ώστε το νευρωνικό δίκτυο να αναγνωρίσει την περίπτωση ως φυσιολογική (τάξης 1,2,3,4) ή μη (τάξης 5,6,7) ενώ στην δεύτερη περίπτωση ελέγχθηκε ο διαχωρισμός των κυττάρων στην αντίστοιχη τάξη μεταξύ των επτά τάξεων. Επιπλέον, και στις δύο περιπτώσεις χρησιμοποιήθηκε 10 fold cross validation. Τα αποτελέσματα που προέκυψαν για το πρόβλημα δεύτερης κατηγορίας σε σύγκριση με το πρόβλημα έβδομης κατηγορίας παρουσιάζονται στον παρακάτω πίνακα.

S No.	Algorithms	Binary classification problem		Multiclass classification problem	
		Correctly classified (%)	Sensitivity	Correctly classified (%)	Sensitivity
1.	Bagging	95.2536%	0.953	76.4629%	0.765
2	PART	94.4083%	0.944	73.5587%	0.736
3	Decision Table	94.5384%	0.945	74.0355%	0.74
4	Decorate	95.0585%	0.951	76.4629%	0.769
5	END	93.9965%	0.940	75.8778%	0.759
6	Filtered Classifier	94.7334%	0.947	74.5340%	0.745
7	J48 Graft	94.4083%	0.944	73.6238%	0.736
8	Multi-Class Classifier	94.9502%	0.950	76.2029%	0.762
9	Rotation Forest	95.4270%	0.954	75.1409%	0.751
10	Naïve Bayes	93.6064%	0.936	69.7226%	0.697
11	Radial Basis Function	94.1482%	0.941	74.7941%	0.748
12	Random Committee	95.3836%	0.954	76.2029%	0.762
13	Random Forest	95.2536%	0.953	75.3576%	0.754
14	Random Sub Space	95.2536%	0.953	76.9181%	0.766
15	Multiple Backpropagation ANN	95.6220%	0.956	78.0668%	0.781

Πίνακας 1.14 : Αποτελέσματα για 10 folds cross επικύρωση για διαφορετικούς αλγορίθμους

Καταληκτικά, το σύστημα που αναπτύχθηκε βασίζεται στα νευρωνικά δίκτυα Multiple Back Propagation με αρχιτεκτονική MLP. Τα τυχαία αρχικοποιημένα βάρη στο γράφημα το οποίο βασίζεται σε στρώματα, ενημερώνονται και ρυθμίζονται με τις τιμές των σφαλμάτων μεταξύ της εξόδου του δικτύου και της πραγματικής εξόδου. Επιπλέον, χρησιμοποιήθηκε η λειτουργία sigmoid ως λειτουργία ενεργοποίησης στους κόμβους του δικτύου. Το δίκτυο αναπτύχθηκε βασιζόμενο στο MATLAB και παρουσίασε τα καλύτερα αποτελέσματα για 21 νευρώνες στο κρυμμένο στρώμα με τον αριθμό των νευρώνων στο επίπεδο εισόδου να είναι 40, στο κρυφό 21 και εξόδου μόνο 1. Ακόμη, αξιολογήθηκε η απόδοση του ANN

χρησιμοποιώντας την ανάλυση των λειτουργικών χαρακτηριστικών του δέκτη (ROC) με το ποσοστό της ευαισθησίας να είναι 87,5% και της εξειδίκευσης 83,3%.

1.15 Ένα κβαντικό υβριδικό PSO σε συνδυασμό με την ασαφή προσέγγιση k-NN για την επιλογή χαρακτηριστικών και την ταξινόμηση κυττάρων στην ανίχνευση του καρκίνου του τραχήλου της μήτρας¹⁸

Το παρόν άρθρο πραγματεύεται την δημιουργία ενός κβαντικού υβριδικού μοντέλου σε συνδυασμό με τον αλγόριθμο k-NN για την ανίχνευση του καρκίνου του τραχήλου της μήτρας. Τονίζεται ότι, τα υβριδικά ευφυή συστήματα (Hybrid Intelligent systems) ενσωματώνουν ταυτόχρονα δύο ή περισσότερες προσεγγίσεις όπως γενετικούς αλγόριθμους ή νευρωνικά δίκτυα προκειμένου να αντιμετωπιστούν σύνθετα προβλήματα.

Σκοπός της συγκεκριμένης μελέτης είναι η ενσωμάτωση κβαντικής μηχανικής μάθησης ή αλλιώς QML (Quantum Machine Learning) σε άλλους αλγορίθμους μηχανικής μάθησης ώστε να ανιχνευθεί ο καρκίνος του τραχήλου της μήτρας.

Ο αλγόριθμος Fuzzy k-NN δέχεται ως είσοδο τις εικόνες της τραχηλικής κηλίδας και εξάγει χαρακτηριστικά όπως το χρώμα, το σχήμα ή την υφή των εικόνων που έχουν μολυνθεί. Έπειτα, ο Quantum Hybrid (QPSO και Fuzzy k-NN) επιλέγει τα χαρακτηριστικά και ταξινομεί τα κύτταρα. Προκειμένου να γίνει σωστή ταξινόμηση, στοχεύει σε 17 χαρακτηριστικά που αφορούν το σχήμα, το χρώμα και την υφή των CS (cervical smear) εικόνων εισόδου. Η εξαγωγή των υφών από τις εικόνες CS έγινε με βάση το πρωτόκολλο Fourier (Local Binary Pattern Histogram).Ο αλγόριθμος QPSO που προαναφέρθηκε είναι βελτιστοποίηση του αλγορίθμου PSO ο οποίος παίρνει ως είσοδο τυχαίες λύσεις και προσδιορίζει την βέλτιστη λύση αναβαθμίζοντας κάθε γενιά.

Τα δεδομένα που χρησιμοποιήθηκαν ήταν 917 CS εικόνες από το σύνολο δεδομένων του Herlev οι οποίες εγκρίθηκαν από τους ειδικούς και στην συνέχεια ταξινομήθηκαν χειροκίνητα σε επτά τάξεις. Οι τάξεις αυτές αφορούσαν τα φυσιολογικά κύτταρα στα οποία περιλαμβάνονταν τα Superficial squamous, Intermediate squamous και Columnar και στα μη φυσιολογικά κύτταρα στα οποία περιλαμβάνονταν Mild dysplasia, Moderate dysplasia, Severe dysplasia και Carcinoma in situ. Τα αποτελέσματα της ταξινόμησης ήταν 242 φυσιολογικά και 675 μη φυσιολογικά κύτταρα. Με το λογισμικό CHAMP έγινε κατακερματισμός των εικόνων στο κυτταρόπλασμα, στον πυρήνα και στο υπόβαθρο.

Η αξιολόγηση της απόδοσης έγινε με βάση την ακρίβεια (Precision), την ανάκληση (Recall) και την ανάλυση “F1 Score” ή αλλιώς PRS. Αντίστοιχα και με προηγούμενες μελέτες χρησιμοποιήθηκε r-fold cross validation για να γίνει επικύρωση των αποτελεσμάτων, όπου r=5 λόγω του μεγέθους των δεδομένων.

Στη συνέχεια μειώθηκαν των χαρακτηριστικών του διανύσματος από 54 εγγραφές σε επτά χαρακτηριστικά 32 εγγραφών. Τα χαρακτηριστικά αυτά αφορούσαν την περιοχή του πυρήνα, την στρογγυλότητα, την φωτεινότητα, την φωτεινότητα του κυτταροπλάσματος, την περιοχή

¹⁸ Abdullah M Ilyasu, Chastine Fatichah, “A Quantum Hybrid PSO Combined with Fuzzy k-NN Approach to Feature Selection and Cell Classification in Cervical Cancer Detection”, Sensors (2017)

του κυττάρου, τον λόγο του πυρήνα και του κυτταροπλάσματος και το τοπικό δυαδικό σχέδιο με Fourier ιστόγραμμα. Για το LBR-HF επιλέχθηκαν 25 από τους 38 φορείς. Ωστόσο, παρατηρήθηκε ότι στον Fuzzy k-NN με προσέγγιση Q-Fuzzy για $k=4$ η ακρίβεια μειώθηκε με αποτέλεσμα το ανώτατο όριο ήταν $k=7$.

Συνεπώς, δημιουργήθηκαν τρία σενάρια προκειμένου να διαπιστωθεί η χρησιμότητα της προτεινόμενης προσέγγισης. Στο πρώτο σενάριο έγινε σύγκριση των αποτελεσμάτων P-Fuzzy και Q-Fuzzy με εκείνες με όλα τα χαρακτηριστικά χωρίς προηγούμενη επιλογή. Στο δεύτερο σενάριο έγινε σύγκριση με τα αποτελέσματα ταξινόμησης των P-Fuzzy, Q-Fuzzy με τεχνικές με και χωρίς προηγούμενη επιλογή χαρακτηριστικών χρησιμοποιώντας όλα τα χαρακτηριστικά. Ενώ στην τελευταία έγινε σύγκριση μεταξύ άλλων αλγορίθμων όπως Naïve Bayes, Υποστήριξη Vector Machines (SVM), Random Forest (RF) και Fuzzy k-NN.

Συμπερασματικά, διαπιστώθηκε ότι η χρήση τεχνικών QH βελτιστοποιεί τα αποτελέσματα των αλγορίθμων με χαρακτηριστικό παράδειγμα την περίπτωση του Fuzzy k-NN καθώς όμως και των υπολοίπων όπου η βελτιστοποίηση ήταν μέτρια.

1.16 Βελτίωση της ακρίβειας ενός συστήματος υποστήριξης κλινικών αποφάσεων για την εξέταση και επιτήρηση του καρκίνου του τραχήλου της μήτρας¹⁹

Η αλλαγή των οδηγιών με την προσθήκη νέων ενημερώσεων της Αμερικανικής Εταιρείας Κολποσκόπησης και Παθολογίας του καρκίνου της μήτρας (ASCCP) στο πρωτότυπο σύστημα (*CDSS with automated text processing for cervical cancer screening*) οδήγησε στην βελτίωση του.

Το CDSS εξάγει τα πρωτογενή δεδομένα από τα κλινικά αρχεία, υπολογίζει την σύσταση βάση των κατευθυντηρίων γραμμών ASCCP και τέλος παράγει τις απαραίτητες συστάσεις για την φροντίδα της ασθενούς. Τα δεδομένα εξάγονται με βάσει τις δοκιμές κυτταρολογίας, HPV, ιστολογίας και κολποσκόπησης. Επιπλέον, χρησιμοποιείται ο αλγόριθμος NLP για την εξαγωγή στοιχείων μέσω κανονικών εκφράσεων. Οι αναφορές της βιοψίας χρησιμοποιούνται από το σύστημα υποστήριξης κλινικών αποφάσεων προκειμένου να εξαχθεί η ιστολογική κατάσταση CIN (CIN2 και CIN3) ενώ εξάγονται και ανεξάρτητες μεταβλητές που αφορούν τον καρκίνο, την υστερεκτομή, τους παράγοντες κινδύνου, την ανοσοανεπάρκεια και τον ιο HIV από δομημένα δεδομένα όπως λίστες προβλημάτων, PPI κ.α. Έπειτα, εξάγονται τα χρονικά χαρακτηριστικά σχετικά με τα CIN2, CIN3 και την υστερεκτομή από τα τελευταία 25 χρόνια.

Στη συνέχεια γίνεται αυτοματοποίηση των κανόνων απόφασης στις οδηγίες ASCCP και τίθενται σε εφαρμογή με το μηχανισμό κανόνων Drools.

Όσον αφορά τις αλλαγές που πραγματοποιήθηκαν, το νέο CDSS είχε 54 κλινικές οδούς με 13 μονοπάτια για φυσιολογική ρουτίνα και 41 για μη φυσιολογική. Οι ασθενείς που βρίσκονταν

¹⁹ K.E. Ravikumar, Kathy L. MacLaughlin, Marianne R. Scheitel, Maya Kessler, Kavishwar B. Waghlikar, Hongfang Liu et al., “Improving the Accuracy of a Clinical Decision Support System for Cervical Cancer Screening and Surveillance”, Appl Clin Inform DOI: 10.1055/s-0037-1617451 (2017)

σε υψηλό κίνδυνο ήταν λόγω ιστορικού μη φυσιολογικού τεστ Pap, θετικού HPV, αναφορών κολποσκοπήσης με CIN2-CIN3 τα τελευταία 25 χρόνια, ιστορικού με καρκίνο του τραχήλου της μήτρας, έκθεσης σε DES in-utero, ανοσοκαταστολής από HIV, ιστορικού μεταμόσχευσης στερεών οργάνων ή χρόνιων ανοσοκατασταλτικών φαρμακευτικών αγωγών.

Ακόμη, η τελική έξοδος του συστήματος ήταν σύσταση για παρακολούθηση ημερομηνίας του Pap smear, κολποσκόπηση για κυτταρολογικά ευρήματα ή ενημέρωση στην ασθενή ότι δεν απαιτείται πλέον παρακολούθηση λόγω ηλικίας ή υστερεκτομής. Οι αλλαγές αφορούσαν την απομάκρυνση κριτηρίων ζώνης ενδοεμβρυϊκού μετασχηματισμού (ETZ), την αλλαγή των κατευθυντήριων οδηγιών για γυναίκες ηλικίας 21-24 ετών με μικρές ανωμαλίες Pap, την ενημέρωση για την διαχείριση άτυπων πλακωδών κυττάρων όπου σε περιπτώσεις με αρνητικό HPV συνίσταται πλέον παρακολούθηση Pap/HPV σε 3 χρόνια αντί για 5, την ενημέρωση αλγορίθμων για την παρακολούθηση συγκεκριμένων αποτελεσμάτων, την απομάκρυνση ηλικιωμένων ασθενών άνω των 65 ετών χωρίς παράγοντες υψηλού κινδύνου και τέλος την επίλυση σφαλμάτων. Τα σφάλματα αφορούσαν κυρίως μη προβλεπόμενες μεταβλητές οι οποίες οδήγησαν σε λάθος συστάσεις, μη σωστή εξέταση της ηλικίας της ασθενούς και τέλος μη σωστή ανίχνευση του αλγορίθμου NLP για τον HPV λόγω της αλλαγής γλώσσας στις εκθέσεις του εργαστηρίου.

Το ενημερωμένο CDSS έχει υλοποιηθεί σε Java χρησιμοποιώντας SQL ερωτήματα και web services για την ανάκτηση δεδομένων. Η έξοδος του περιλαμβάνει το τελικό σημείο απόφασης, το κείμενο με την κλινική σύσταση και την επομένη παρακολούθηση. Προκειμένου να διαπιστωθεί η ακρίβεια του συγκεκριμένου συστήματος ελήφθησαν τυχαία δείγματα από το 10% των ασθενών, δηλαδή από συνολικά 393 ασθενείς.

Τα αποτελέσματα σχετικά με την ακρίβεια του συστήματος παρατίθενται στον παρακάτω πίνακα.

End point type	Total patients	Total correct	Accuracy
Routine/normal	307	297	96.7%
Abnormal	86	72	83.7%
Total	393	369	93.4%

Πίνακας 1.16 : Αποτελέσματα ακρίβειας συστήματος

Καταληκτικά, παρήχθει επόμενη ημερομηνία παρακολούθησης για 281 ασθενείς από τις 393, από τις 307 γυναίκες με συστάσεις ρουτίνας στις 200 καθορίστηκε ημερομηνία εξέτασης καθώς και για τις 81 από τις 86 ασθενείς με παθολογικό ιστορικό ημερομηνία παρακολούθησης. Ωστόσο, για τις 107 από τις 307 ασθενείς δεν καθορίστηκε ημερομηνία παρακολούθησης καθώς όπως προέκυψε και από την διαπίστωση του ιατρού οι γυναίκες δεν χρειάζονταν διαλογή λόγω ιστορικού υστερεκτομής χωρίς ιστορικό CIN 2-3 ή καρκίνου του τραχήλου της μήτρας (70), ή ηλικίας μεγαλύτερη από 65. Υπήρχαν 7 γυναίκες στην ομάδα ελέγχου ρουτίνας με προηγούμενη διάγνωση σε εξωτερική μονάδα με μη προσβάσιμο αρχείο

Παρ από το CDSS με αποτέλεσμα να καθίσταται αδύνατη η ακριβής ημερομηνία της επόμενης παρακολούθησης. Από τις 81 γυναίκες που είχαν μη φυσιολογικό παρελθόν, το CDSS αναγνώρισε σχεδόν 67% (54 ασθενείς) οι οποίες καθυστέρησαν την επόμενη εξέταση. Μεταξύ των γυναικών με φυσιολογικό ιστορικό, το 21% καθυστέρησε για την επόμενη παρακολούθηση. Το CDSS δεν μπόρεσε να καθορίσει μια ακριβή ημερομηνία παρακολούθησης για πέντε γυναίκες που είχαν παθολογικό ιστορικό παρελθόν λόγω εσφαλμένων συστάσεων του συστήματος.

Όσον αφορά τα λάθη ήταν κυρίως λόγω έλλειψης πρόσβασης σε ακριβείς πληροφορίες όπως λάθη στην βάση δεδομένων, σφάλματα μοντελοποίησης κυρίως σε περιπτώσεις που δεν τηρήθηκαν οι κατευθυντήριες οδηγίες της ASCCP, προγραμματισμού και αξιολόγησης. Σε μία αξιολόγηση που έγινε παρατηρήθηκε ότι οι πάροχοι χωρίς να γνωρίζουν τα αποτελέσματα του CDSS διαφώνησαν σε 75 από 169 (44%) σενάρια ασθενών αλλά σύμφωνα με τους εμπειρογνώμονες η σύσταση του συστήματος ήταν καταλληλότερη στις περισσότερες περιπτώσεις, 53/75 (71%) έναντι 22/75 (29%) για τους παρόχους. Κατά μέσο όρο το σύστημα λειτουργεί 72 δευτερόλεπτα για τον υπολογισμό μιας σύστασης με μέγιστο χρόνο 2 λεπτά και ελάχιστο 9 δευτερόλεπτα. Μελλοντικά, θα γίνει επανεξέταση του συστήματος χρησιμοποιώντας το SMART-on-FHIR με αρχιτεκτονική REST-API προς αποφυγή ορισμένων σφαλμάτων και χρήση καλύτερης διαλειτουργικότητας.

ΚΕΦΑΛΑΙΟ 2^ο : Περιγραφή Συνόλου Δεδομένων, Μεθοδολογίας και Αλγορίθμων

Στο παρόν κεφάλαιο αναλύεται το σύνολο δεδομένων που χρησιμοποιήθηκε καθώς και η μεθοδολογία που ακολουθήθηκε χρησιμοποιώντας το WEKA.

2.1 Σύνολο Δεδομένων

Το σύνολο δεδομένων²⁰ που χρησιμοποιήθηκε, προέρχεται από το Πανεπιστήμιο του Καράκας στην Βενεζουέλα το 2017. Συνολικά περιλαμβάνει 858 εγγραφές γυναικών ηλικίας 13 έως 84 ετών. Τα χαρακτηριστικά που έχουν συλλεχθεί αφορούν δημογραφικά χαρακτηριστικά όπως την ηλικία, συνήθειες όπως το κάπνισμα και ιστορικά δεδομένα υγείας σχετικά με την πρώτη σεξουαλική επαφή, το Schiller test κατά το οποίο γίνεται με χρήση ιωδιούχου διαλύματος στην εσωτερική περιοχή του κόλπου κατά την διάρκεια της κολποσκόπησης με σκοπό να διερευνηθεί αν υπάρχει κάποιο παθολογικό αποτέλεσμα που υποδηλώνει συνήθως καρκίνο του τραχήλου της μήτρας²¹, την δυσπλασία και με διάφορες ασθένειες όπως την Ηπατίτιδα, το AIDS, τη σύφιλη κτλ που συμβάλλουν στην εμφάνιση του. Συγκεκριμένα, αποτελείται από 36 χαρακτηριστικά (ποσοτικά και ποιοτικά) εκ των οποίων τα 16 είναι τύπου boolean όπου για την τιμή true χρησιμοποιείται το 1 και για την τιμή false χρησιμοποιείται το 0. Επιπλέον, επισημαίνεται ότι ορισμένες τιμές είναι ελλιπείς καθώς οι ασθενείς επέλεξαν να μην απαντήσουν λόγω ανησυχίας για την προστασία της ιδιωτικότητας τους ενώ ορισμένα χαρακτηριστικά είχαν μηδενικές τιμές για όλες τις εγγραφές. Εν κατακλείδι, για περαιτέρω λεπτομέρειες παρατίθεται στην βιβλιογραφία η πηγή του εν λόγω συνόλου.

2.2 Μεθοδολογία

Η μεθοδολογία που ακολουθήθηκε είναι η εξής. Το σύνολο δεδομένων που προαναφέρθηκε μετατράπηκε σε μορφή .arff με την χρήση του εργαλείου WEKA. Χαρακτηριστικά των οποίων οι τιμές ήταν μηδενικές για όλες τις εγγραφές του συνόλου δεδομένων αφαιρέθηκαν με αποτέλεσμα την δημιουργία ενός νέου συνόλου δεδομένων με λιγότερα χαρακτηριστικά. Επιπλέον, ο τύπος των χαρακτηριστικών από numeric μετατράπηκε σε {0,1 κτλ} προκειμένου να είναι πιο εμφανής οι τιμές των χαρακτηριστικών. Έπειτα, διασπάστηκε σε περεταίρω σύνολα δεδομένων ώστε να εφαρμοστούν οι απαραίτητοι αλγόριθμοι. Αναλυτικότερα, τα παθολογικά δεδομένα, δηλαδή τα δεδομένα των οποίων η τιμή της μεταβλητής Biopsy (βιοψία) ισούται με 1 και ο συνολικός αριθμός τους είναι 55 με χρήση κώδικα σε Python αποσπάστηκαν από το αρχικό σύνολο και αποθηκεύτηκαν σε ξεχωριστό αρχείο όπως αντίστοιχα και τα μη παθολογικά. Τονίζεται ότι, η επιλογή των δεδομένων έγινε τυχαία μέσω κώδικα σε Python. Στη συνέχεια, δημιουργήθηκαν φάκελοι 15 αρχείων τα οποία αποτελούνταν από 55 παθολογικά και 55 φυσιολογικά, 55 παθολογικά και 165

²⁰ Cervical cancer (Risk Factors) Data Set

²¹ Π. Σταματόπουλος Καθηγητής Μαιευτικής-Γυναικολογίας, HPV Λοίμωξη & Πρόληψη Ca Τραχήλου

φυσιολογικά, 55 παθολογικά και 220 φυσιολογικά και τέλος 55 παθολογικά και 250 φυσιολογικά. Στην πρώτη περίπτωση στα δεδομένα δεν εφαρμόστηκε SMOTE ενώ στην δεύτερη εφαρμόστηκε προκειμένου ο αριθμός των μη φυσιολογικών να συνάδει με τον αριθμό των φυσιολογικών. Η τεχνική SMOTE (Synthetic Minority Over-sampling Technique) χρησιμοποιείται για την επίτευξη της ισορροπίας στο σύνολο δεδομένων δημιουργώντας συνθετικά μοτίβα στην κατηγορία των μειονοτήτων επιτυγχάνοντας έτσι την μη απώλεια πληροφοριών.

Η τεχνική undersampling έχει ως στόχο την μείωση των δεδομένων πλειοψηφίας και χρησιμοποιείται κυρίως σε περιπτώσεις όπως αυτή κατά την οποία η ποσότητα των δεδομένων είναι αρκετά μεγάλη προκειμένου τα αποτελέσματα να είναι πιο ακριβή. Όμως, η υποδειγματοληψία μπορεί να οδηγήσει στην αφαίρεση σημαντικών δεδομένων.

Ο λόγος που επιλέχθηκαν να χρησιμοποιηθούν οι παραπάνω τεχνικές oversampling και undersampling είναι προκειμένου να αντιμετωπιστεί η ανισορροπία των δεδομένων διότι είναι δύσκολα διαχειρίσιμη.

Η εφαρμογή των αλγορίθμων Naive Bayes, J48, K-NN και Random Forest έγινε και στις δύο περιπτώσεις ενώ τα αποτελέσματα που αφορούσαν την ακρίβεια, την ευαισθησία και την εξειδίκευση αποθηκεύτηκαν σε αρχεία της μορφής .xlsx.

2.3 Περιγραφή Αλγορίθμων

2.3.1 J48/ C4.5

Ο αλγόριθμος J48 ή C4.5 δημιουργεί ένα δέντρο απόφασης με βάση τον λόγο κέρδους της πληροφορίας ο οποίος δημιουργήθηκε από τον Ross Quinlan²² ενώ αφορά αριθμητικές τιμές και αποτελεί βελτιστοποίηση του ID3. Τονίζεται ότι τα ελλιπή δεδομένα αγνοούνται ενώ μπορεί να εφαρμοστεί και κλάδεμα. Ο τρόπος που λειτουργεί είναι ο εξής:

Αρχικά, δημιουργείται ο αρχικός κόμβος για ολόκληρο το δείγμα στον οποίο τοποθετείται το καταλληλότερο χαρακτηριστικό το οποίο διαιρεί πιο αποτελεσματικά το σύνολο δεδομένων σε υποσύνολα. Το μέτρο επιλογής του καλύτερου διαχωρισμού ή αλλιώς ο λόγος κέρδους

προκύπτει από τον τύπο $Gain Ratio = \frac{\Delta_{info}}{Split Info}$ όπου $Split Info = - \sum_{i=1}^k P(u_i) \log_2 P(u_i)$

και k ο συνολικός αριθμός των κατηγοριών²³.

Σε περίπτωση που τα γνωρίσματα ανήκουν στην ίδια κλάση τότε ο κόμβος μετατρέπεται σε φύλλο, αντίθετα αν δεν ανήκουν στην ίδια κλάση τότε επιλέγεται το γνώρισμα που διαχωρίζει καλύτερα τα δεδομένα.

Η διαδικασία επαναλαμβάνεται για όλους του κόμβους μέχρις ότου ο κάθε κόμβος να γίνει φύλλο. Συγκεκριμένα, οι περιπτώσεις κατά τις οποίες τερματίζεται η διαδικασία είναι όταν όλα τα δεδομένα ανήκουν σε μία κλάση ή όταν δεν υπάρχουν άλλα γνωρίσματα για έλεγχο.

²² Ερευνητής στον τομέα της πληροφορικής και ειδικότερα στην εξόρυξη δεδομένων.

²³ Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Εισαγωγή στην Εξόρυξη Δεδομένων, Εξαγωγή Δέντρου Απόφασης σελ 175-178

Η διαχείριση μη αριθμητικών δεδομένων δεν είναι εύκολη και απαιτείται κλάδεμα καθώς το δέντρο που δημιουργείται είναι αρκετά μεγάλο. Από το δέντρο απόφασης που προκύπτει εξάγονται και οι σχετικοί κανόνες.

2.3.2 NaiveBayes

Αποτελεί έναν πιθανοτικό αλγόριθμο ο οποίος είναι βασισμένος στο θεώρημα του Bayes και λειτουργεί με βάση υποθέσεις ανεξαρτησίας ανάμεσα στις προβλέψεις. Είναι ένας από τους πιο διαδεδομένους ταξινομητές λόγω της εκπληκτικής λειτουργίας και των αποτελεσμάτων του.

Ο αλγόριθμος λειτουργεί με βάση την δεσμευμένη πιθανότητα, δηλαδή την πιθανότητα να συμβεί ένα γεγονός δεδομένου ότι έχει ήδη συμβεί ένα άλλο (εκ των υστέρων πιθανότητα).

$$p(C_k | \mathbf{x}) = \frac{p(C_k) p(\mathbf{x} | C_k)}{p(\mathbf{x})}$$

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

Εικόνα 2.3.2 Τύποι Δεσμευμένης Πιθανότητας

Ο ταξινομητής προβλέπει την πιθανότητα ένα σημείο δεδομένων να ανήκει σε μία συγκεκριμένη κλάση. Η υψηλότερη πιθανότητα θεωρείται η πιο πιθανή κατηγορία. Η απουσία ή η παρουσία ενός χαρακτηριστικού δεν επηρεάζει την παρουσία ενός άλλου ενώ αποτελεί έναν απλό αλγόριθμο ο οποίος μπορεί να εκπαιδεύσει εύκολα ένα μικρό σύνολο δεδομένων.

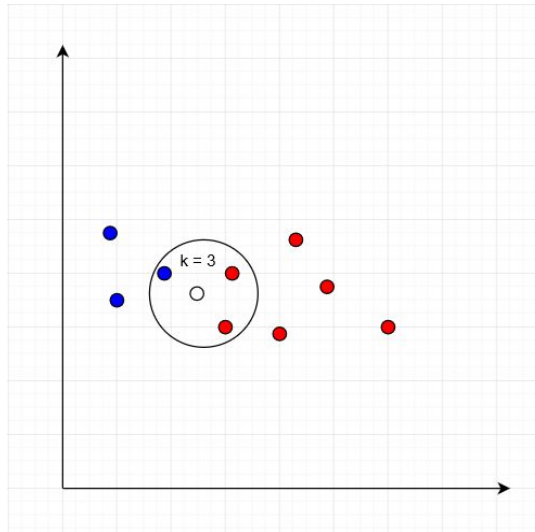
Κάθε στιγμιότυπο του προβλήματος αποτελείται από ένα σύνολο γνωρισμάτων. Αν το πρόβλημα έχει m κλάσεις τότε δοθέντος ενός αγνώστου στιγμιότυπου του προβλήματος ο ταξινομητής προβλέπει την κλάση στην οποία ανήκει με βάση την μεγαλύτερη εκ των υστέρων πιθανότητα. Ωστόσο, προκειμένου να μειωθεί το υπολογιστικό κόστος υπάρχει ανεξαρτησία ως προς τις τιμές των γνωρισμάτων του στιγμιότυπου.

Τα μειονεκτήματα που διέπουν τον συγκεκριμένο αλγόριθμο είναι η απαίτηση μεγάλου αριθμού δεδομένων ενώ σε συγκεκριμένες περιπτώσεις τα αποτελέσματα είναι μεροληπτικά.

2.3.3 K-NN

Ο αλγόριθμος k πλησιέστερων γειτόνων ανήκει στους κατηγοριοποιητές βασισμένοι σε παραδείγματα οι οποίοι δεν παράγουν κάποιο μοντέλο γενίκευσης.

Ο αλγόριθμος αναζητά μέσα στον χώρο τα σημεία που βρίσκονται πλησιέστερα στη νέα παρατήρηση. Αν για παράδειγμα $k=5$ τότε εντοπίζονται τα 5 πλησιέστερα σημεία. Έπειτα αν τρία σημεία px είναι κόκκινα και δύο μπλε τότε θα θεωρηθεί και αυτό κόκκινο.



Εικόνα 2.3.3 Διάγραμμα για $k=3$ πλησιέστερους γείτονες

Εντάσσεται στους αλγορίθμους Instance Based Classifiers. Όσον αφορά την λειτουργία του είναι η εξής.

Υπολογίζεται η απόσταση ενός νέου σημείου από κάθε σημείο που υπάρχει ως παράδειγμα εκπαίδευσης. Η συγκεκριμένη απόσταση υπολογίζεται είτε με την Ευκλείδεια απόσταση είτε με κανονικοποίηση τιμών για ετερόκλητα χαρακτηριστικά.

Αφού γίνει υπολογισμός των αποστάσεων, μετά η μικρότερη απόσταση και ο αριθμός των γειτόνων από το σημείο υποδηλώνει την κατηγορία στην οποία θα καταχωρηθεί με βάση την συχνότερη κατηγορία.

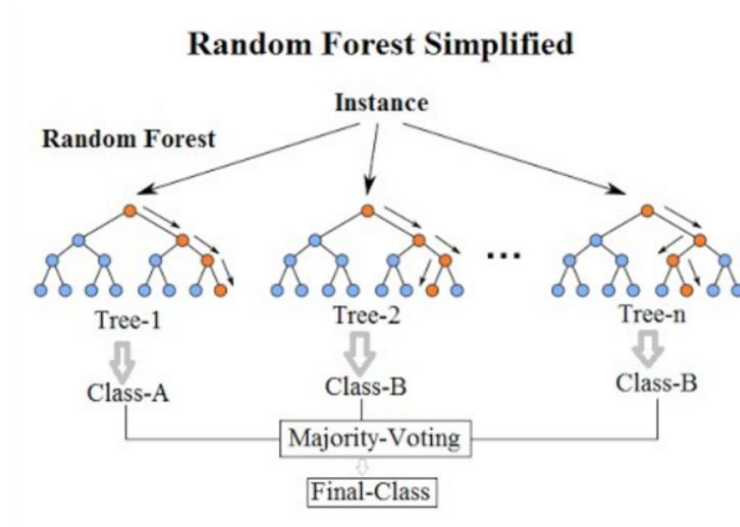
Ωστόσο, ορισμένα από τα μειονεκτήματα που παρουσιάζει ο συγκεκριμένος αλγόριθμος είναι η μεγάλη απαίτηση σε αποθηκευτικό χώρο, το μεγάλο υπολογιστικό κόστος στην κλήση καθώς και ο επηρεασμός των αποτελεσμάτων βάσει του αριθμού των γειτόνων.

2.3.4 RandomForest

Ο συγκεκριμένος αλγόριθμος αναπτύχθηκε από τους Leo Breiman και Adele Cutler, αποτελεί μια επέκταση του bagging και είναι από τους δημοφιλέστερους αλγορίθμους λόγω της μεγάλης ακρίβειας που παρουσιάζει καθώς και της ταχύτητας του.

Η μέθοδος cross validation δεν είναι απαραίτητη για τον έλεγχο ακρίβειας και η αποδοτικότητα του παραμένει ακόμη και σε περιπτώσεις ελλιπών δεδομένων και μεγάλου αριθμού χαρακτηριστικών.

Αρχικά, δημιουργούνται πολλά σύνολα δεδομένων εκπαίδευσης με την μέθοδο bootstrap και αντίστοιχα για κάθε σύνολο δημιουργείται το δέντρο απόφασης. Για την κατασκευή του κάθε δέντρου δεν λαμβάνονται υπόψη όλα τα χαρακτηριστικά. Στο τέλος επιλέγεται ο καλύτερος ταξινομητής από το σύνολο των δέντρων απόφασης που δημιουργούνται.

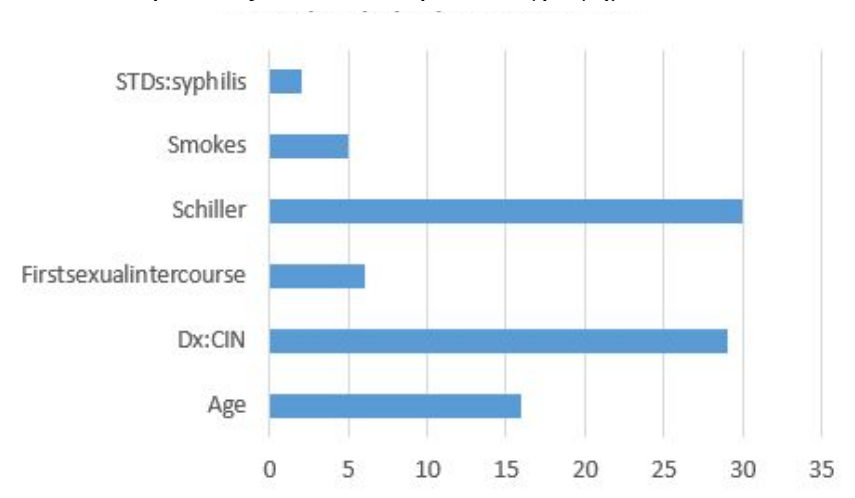


2.3.4 Random Forest²⁴

2.4 Αποτελέσματα

Όσον αφορά τα αποτελέσματα που προέκυψαν είναι τα εξής.

Τα πιο συχνά χαρακτηριστικά έπειτα από την χρήση του αλγορίθμου C4.5 (J48) στα δεδομένα με SMOTE παρουσιάζονται στο παρακάτω γράφημα.

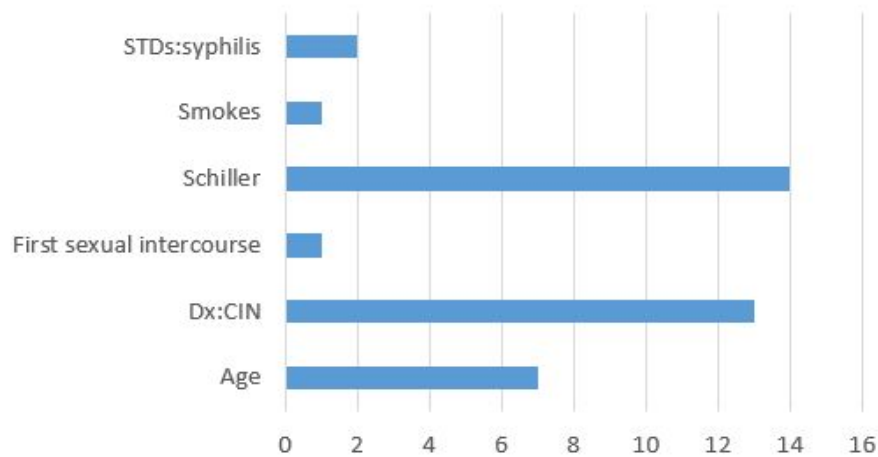


Εικόνα 2.4.1 Τα πιο συχνά χαρακτηριστικά δεδομένων με εφαρμογή SMOTE

Συγκεκριμένα, όπως παρατηρείται η πρώτη αιτία εμφάνισης καρκίνου του τραχήλου της μήτρας είναι το Schiller ενώ σε περίπτωση που κάποια ασθενής έχει διαγνωστεί στο παρελθόν με νεοπλασία τραχηλικής ενδοεπιθηλίου (Dx:CIN) αποτελεί τον δεύτερο σημαντικότερο παράγοντα εμφάνισης καρκίνου του τραχήλου της μήτρας. Η σύφιλη είναι η λιγότερο σημαντική αιτία καθώς επηρεάζει την εξωτερική επιφάνεια του σώματος και των γεννητικών οργάνων ενώ σε σοβαρές περιπτώσεις μπορούν να επηρεαστούν και τα εσωτερικά όργανα.

²⁴ Will Koehrsen, Random Forest Simple Explanation, Dec 27, 2017, A Medium Corporation

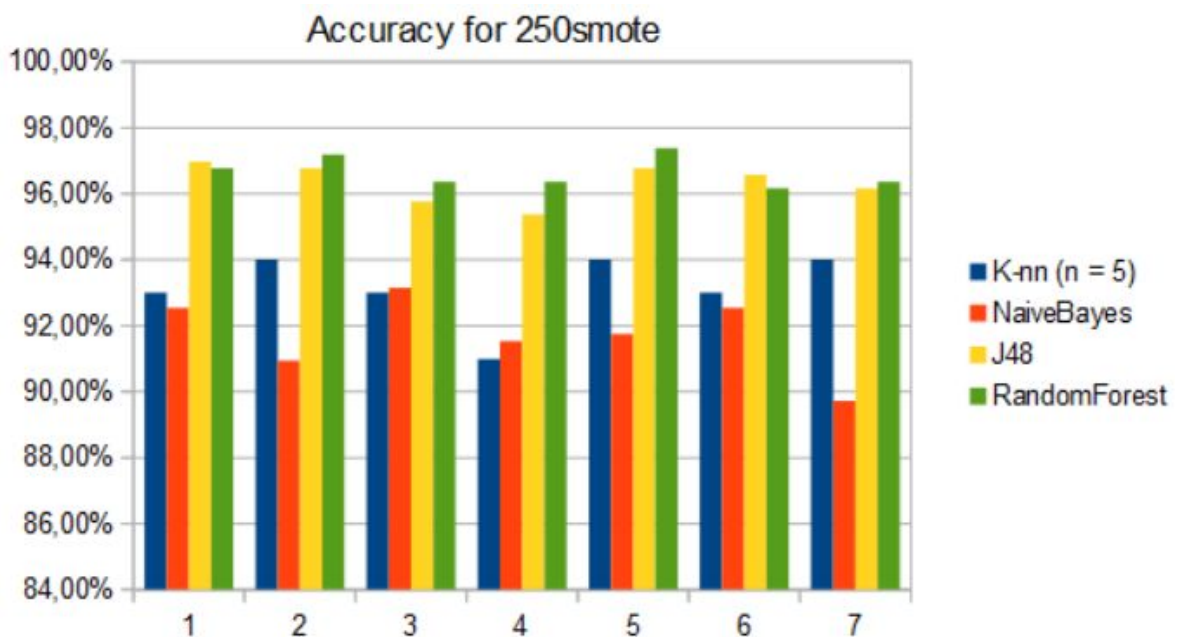
Στην περίπτωση που υπάρχει ταξική ανισορροπία, δηλαδή δεν έχει εφαρμοστεί το SMOTE, τότε η κύρια αιτία εμφάνισης καρκίνου του τραχήλου της μήτρας είναι το Schiller ενώ η λιγότερο σημαντική το κάπνισμα και η πρώτη σεξουαλική επαφή.



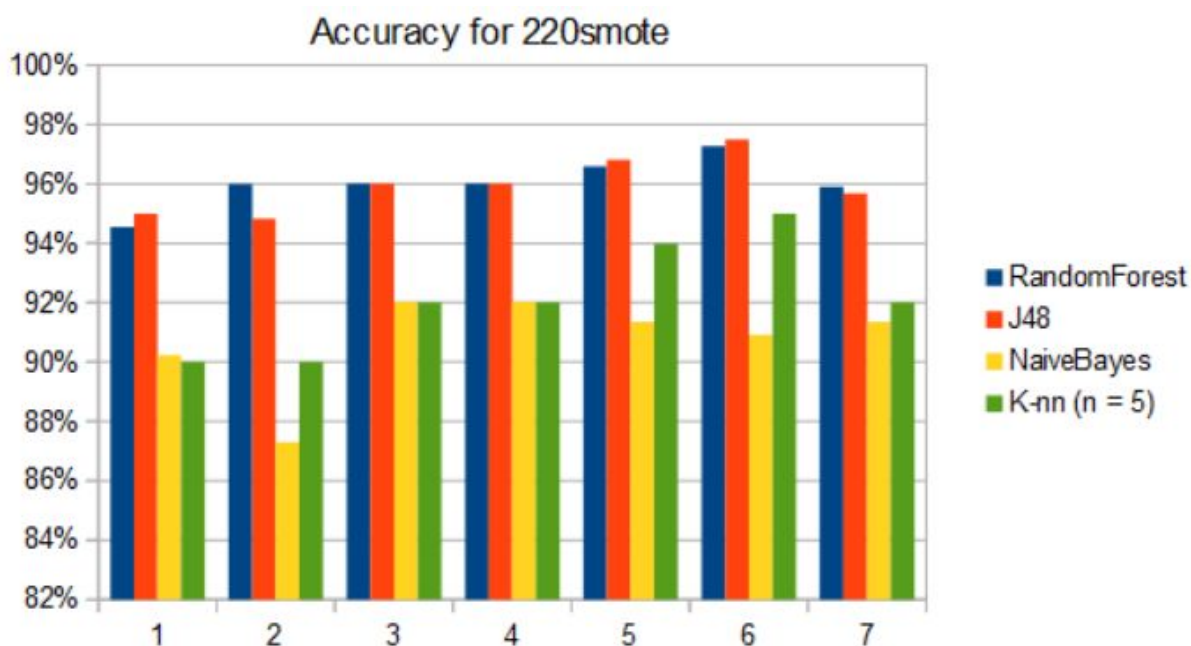
Εικόνα 2.4.2 Τα πιο συχνά χαρακτηριστικά των δεδομένων χωρίς SMOTE

Η ακρίβεια των αλγορίθμων διαχωρίζεται με βάση το γεγονός αν έχει εφαρμοστεί SMOTE ή όχι στα δεδομένα. Τονίζεται ότι χρησιμοποιήθηκαν συνολικά 7 αρχεία για κάθε αλγόριθμο. Στην πρώτη περίπτωση η ακρίβεια των αλγορίθμων κυμαίνεται από 83% μέχρι 97% με τα υψηλότερα αποτελέσματα να ανήκουν στον J48 και Random Forest ενώ ακολουθούν ο K-NN και ο Naive Bayes.

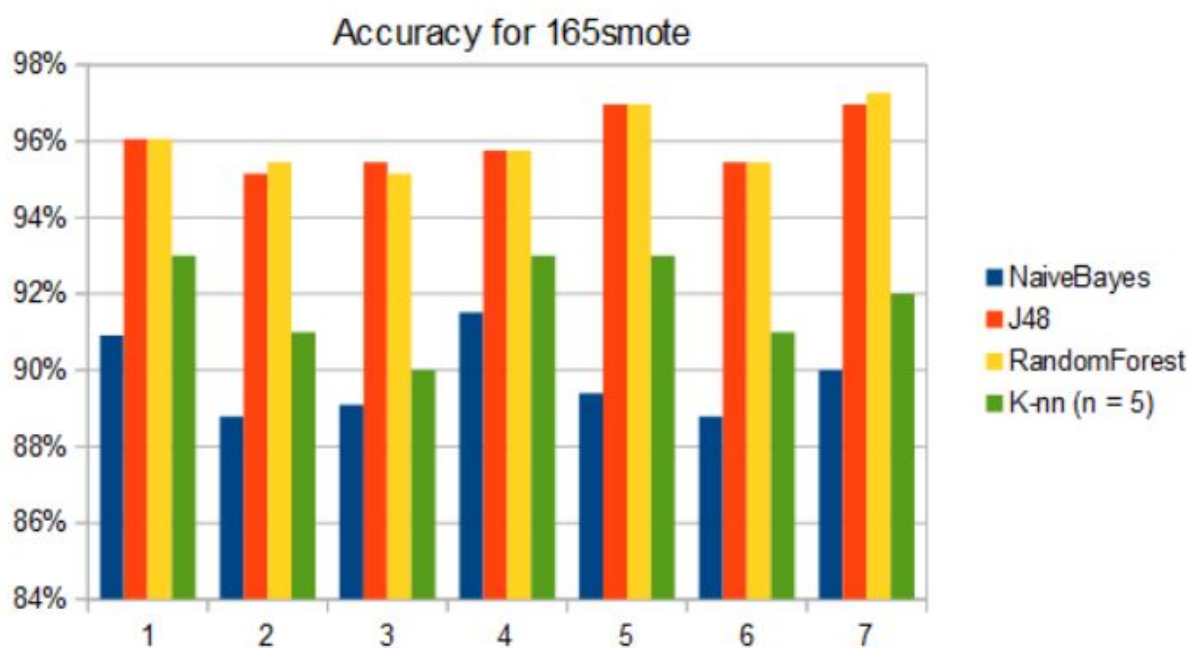
Παρακάτω παρουσιάζονται τα σχετικά διαγράμματα.



Εικόνα 2.4.3 Ακρίβεια αλγορίθμων για 250 φυσιολογικά και 250 παθολογικά



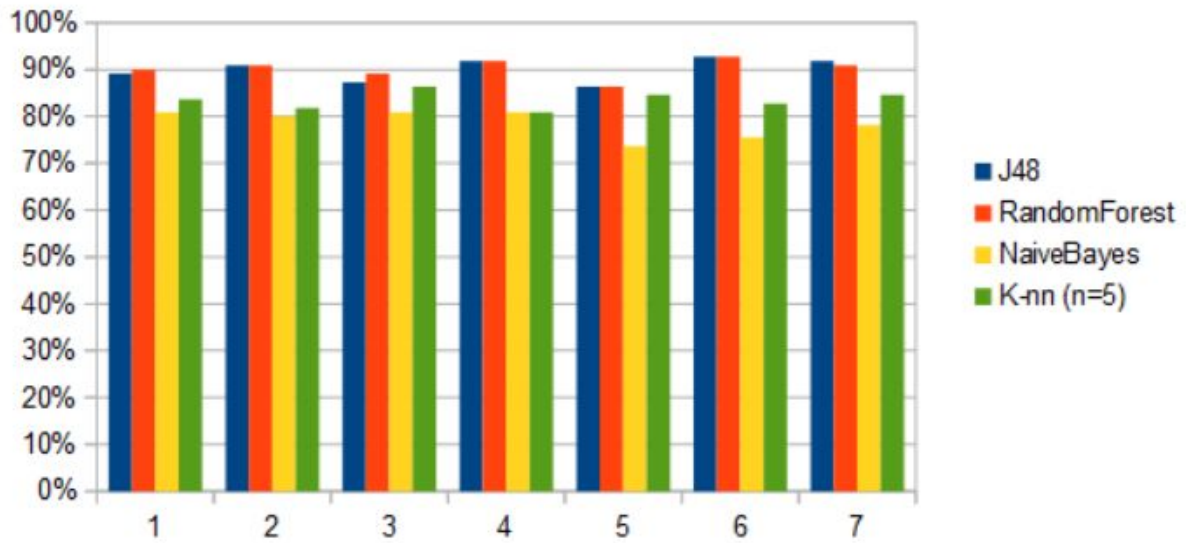
Εικόνα 2.4.4 Ακρίβεια αλγορίθμων για 220 φυσιολογικά και 220 παθολογικά



Εικόνα 2.4.5 Ακρίβεια αλγορίθμων για 165 φυσιολογικά και 165 παθολογικά

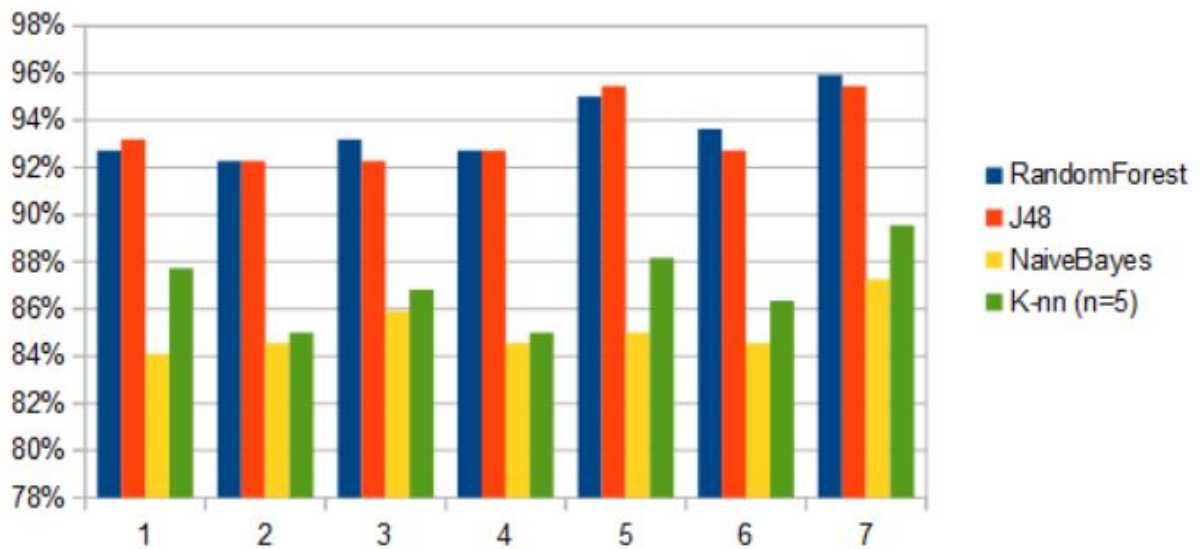
Στην δεύτερη περίπτωση παρατηρείται ότι οι αλγόριθμοι J48 και Random Forest εξακολουθούν να έχουν τα καλύτερα αποτελέσματα εν αντιθέση με τους υπόλοιπους αλγορίθμους. Αναλυτικότερα τα ποσοστά κυμαίνονται από 80% μέχρι 95% με τον Naive Bayes να παρουσιάζει τα χαμηλότερα ποσοστά ακρίβειας.

Accuracy of Algorithms for 55_55



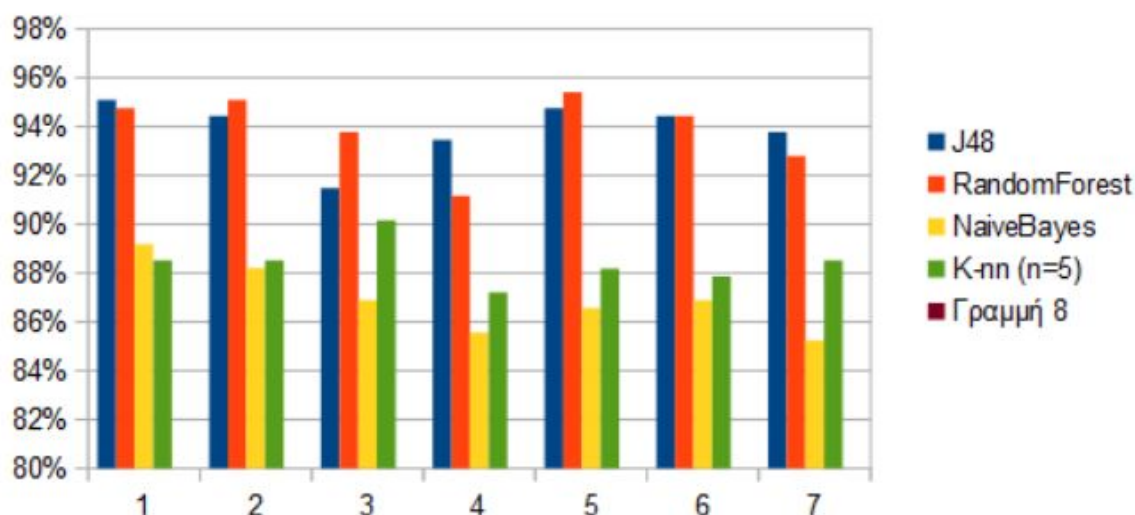
Εικόνα 2.4.6 Ακρίβεια αλγορίθμων για 55 φυσιολογικά και 55 παθολογικά

Accuracy of Algorithms for 165_55



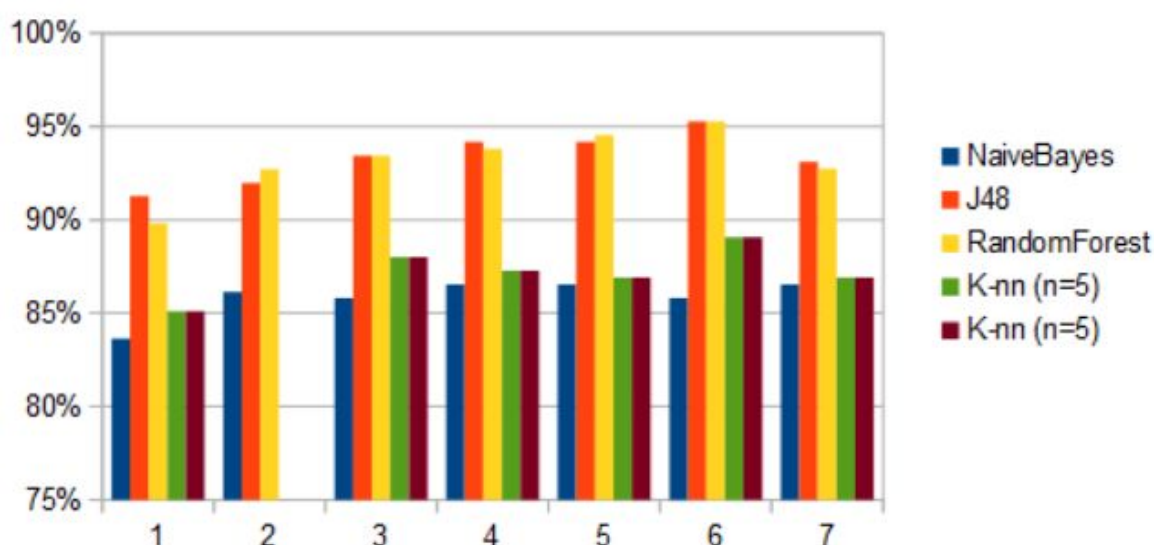
Εικόνα 2.4.7 Ακρίβεια αλγορίθμων για 165 φυσιολογικά και 55 παθολογικά

Accuracy of Algorithms for 250_55



Εικόνα 2.4.8 Ακρίβεια αλγορίθμων για 250 φυσιολογικά και 55 παθολογικά

Accuracy of Algorithms for 220_55



Εικόνα 2.4.9 Ακρίβεια αλγορίθμων για 220 φυσιολογικά και 55 παθολογικά

Επιπλέον, μελετήθηκε η ευαισθησία (recall) η οποία μετρά το ποσοστό των θετικών που έχουν σωστά αναγνωριστεί και η πιστότητα (precision == positive predictive value) η οποία είναι τα πραγματικά θετικά παραδείγματα διά των παραδειγμάτων που προβλέψαμε θετικά. Για τα δεδομένα στα οποία δεν έχει εφαρμοστεί SMOTE παρατηρείται ότι για όλους τους αλγορίθμους, η κλάση που αναφέρεται στους ασθενείς χωρίς καρκίνο παρουσιάζει μεγαλύτερη ευαισθησία και πιστότητα πλην των δεδομένων στα οποία ο αριθμός των ασθενών με καρκίνο ισούται με τον αριθμό των υγιών κατα τα οποία η πιστότητα είναι

μεγαλύτερη για τους μη υγιείς ενώ η ευαισθησία είναι μεγαλύτερη για τους υγιείς. Ακόμη, η ευαισθησία και η πιστότητα είναι χαμηλότερη για τους αλγορίθμους Naive Bayes και K-NN.

	File2		File3		File4		File5		File6		File7		File8	
Naive Bayes	Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall	
220_55	0,911	0,882	0,906	0,922	0,921	0,900	0,926	0,905	0,918	0,914	0,921	0,900	0,922	0,909
	0,581	0,655	0,667	0,618	0,633	0,691	0,650	0,709	0,661	0,673	0,633	0,691	0,655	0,691
250_55	Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall	
	0,919	0,952	0,935	0,920	0,930	0,908	0,915	0,908	0,934	0,900	0,927	0,912	0,918	0,900
	0,739	0,618	0,661	0,709	0,623	0,691	0,596	0,618	0,609	0,709	0,627	0,673	0,583	0,636
165_55	Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall	
	0,901	0,885	0,902	0,891	0,890	0,927	0,912	0,879	0,907	0,891	0,902	0,891	0,910	0,921
	0,672	0,709	0,684	0,709	0,750	0,655	0,672	0,745	0,690	0,727	0,684	0,709	0,755	0,727
55_55	Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall	
	0,774	0,873	0,770	0,855	0,766	0,891	0,766	0,891	0,710	0,800	0,694	0,909	0,712	0,945
	0,854	0,745	0,837	0,745	0,870	0,727	0,870	0,727	0,771	0,673	0,868	0,600	0,919	0,618

Εικόνα 2.4.10 Αποτελέσματα ευαισθησίας και πιστότητας του αλγορίθμου Naive Bayes

J48	Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall	
220_55	0,962	0,927	0,954	0,945	0,963	0,955	0,959	0,968	0,959	0,968	0,968	0,973	0,967	0,945
	0,746	0,855	0,789	0,818	0,825	0,855	0,868	0,836	0,868	0,836	0,889	0,873	0,800	0,873
220_55	Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall	
	0,968	0,972	0,964	0,968	0,944	0,952	0,960	0,960	0,968	0,968	0,972	0,960	0,968	0,956
	0,870	0,855	0,852	0,836	0,774	0,745	0,818	0,818	0,855	0,855	0,828	0,873	0,810	0,855
165_55	Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall	
	0,946	0,964	0,946	0,952	0,946	0,952	0,946	0,958	0,959	0,982	0,946	0,958	0,959	0,982
	0,885	0,836	0,852	0,836	0,852	0,836	0,868	0,836	0,941	0,873	0,868	0,836	0,941	0,873
55_55	Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall	
	0,877	0,909	0,881	0,945	0,873	0,873	0,883	0,964	0,833	0,909	0,885	0,982	0,883	0,964
	0,906	0,873	0,941	0,873	0,873	0,873	0,960	0,873	0,900	0,818	0,980	0,873	0,960	0,873

Εικόνα 2.4.11 Αποτελέσματα ευαισθησίας και πιστότητας του αλγορίθμου J48

	File2		File3		File4		File5		File6		File7		File8	
Random Forest	Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall	
220_55	0,932	0,941	0,954	0,954	0,955	0,964	0,951	0,973	0,964	0,968	0,964	0,977	0,950	0,959
	0,755	0,727	0,818	0,818	0,849	0,818	0,880	0,800	0,870	0,855	0,904	0,855	0,830	0,800
250_55	Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall	
	0,961	0,976	0,964	0,976	0,953	0,972	0,941	0,952	0,965	0,980	0,960	0,972	0,956	0,956
	0,882	0,818	0,885	0,836	0,860	0,782	0,769	0,727	0,902	0,836	0,865	0,818	0,800	0,800
165_55	Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall	
	0,946	0,958	0,940	0,958	0,957	0,952	0,946	0,958	0,953	0,982	0,947	0,970	0,959	0,988
	0,868	0,836	0,865	0,818	0,857	0,873	0,868	0,836	0,940	0,855	0,902	0,836	0,960	0,873
55_55	Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall	
	0,855	0,964	0,881	0,945	0,877	0,909	0,883	0,964	0,845	0,891	0,885	0,982	0,881	0,945
	0,958	0,836	0,941	0,873	0,906	0,873	0,960	0,873	0,885	0,836	0,980	0,873	0,941	0,873

Εικόνα 2.4.12 Αποτελέσματα ευαισθησίας και πιστότητας του αλγορίθμου Random Forest

K-NN (N=5)	Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall	
220_55	0,878	0,945	0,882	0,959	0,882	0,982	0,884	0,968	0,877	0,973	0,896	0,977	0,880	0,968
	0,684	0,473	0,750	0,491	0,867	0,473	0,794	0,491	0,806	0,455	0,857	0,545	0,788	0,473
250_55	Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall	
	0,891	0,980	0,885	0,988	0,904	0,984	0,889	0,964	0,896	0,968	0,884	0,980	0,894	0,976
	0,833	0,455	0,885	0,418	0,879	0,527	0,735	0,455	0,771	0,491	0,821	0,418	0,813	0,473
165_55	Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall	
	0,888	0,958	0,859	0,958	0,878	0,958	0,863	0,952	0,876	0,982	0,869	0,964	0,882	0,994
	0,833	0,636	0,806	0,527	0,825	0,600	0,789	0,545	0,914	0,582	0,838	0,564	0,971	0,600
55_55	Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall	
	0,803	0,891	0,778	0,891	0,813	0,945	0,758	0,909	0,817	0,891	0,781	0,909	0,806	0,909
	0,878	0,782	0,872	0,745	0,935	0,782	0,886	0,709	0,880	0,800	0,891	0,745	0,896	0,782

Εικόνα 2.4.13 Αποτελέσματα ευαισθησίας και πιστότητας του αλγορίθμου K-NN

Στα δεδομένα με SMOTE τα αποτελέσματα των αλγορίθμων διαφέρουν με την ανάκληση να είναι σε ορισμένες περιπτώσεις μεγαλύτερη από την πιστότητα και το αντίστροφο. Αυτό συμβαίνει διότι η βελτίωση της πιστότητας μειώνει την ευαισθησία και το αντίστροφο. Όταν ένα μοντέλο παρουσιάζει μεγαλύτερη πιστότητα και ευαισθησία τότε είναι καλύτερο από ένα μοντέλο με μικρή ευαισθησία και πιστότητα.

	File2		File3		File4		File5		File6		File7		File8	
NaiveBayes	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
165	0,919	0,897	0,881	0,897	0,869	0,921	0,931	0,897	0,887	0,903	0,868	0,915	0,898	0,903
	0,899	0,921	0,895	0,879	0,916	0,861	0,901	0,933	0,901	0,885	0,910	0,861	0,902	0,897
220	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
	0,897	0,909	0,851	0,913	0,919	0,927	0,935	0,918	0,910	0,918	0,910	0,918	0,917	0,909
	0,908	0,895	0,900	0,830	0,922	0,913	0,914	0,932	0,917	0,909	0,917	0,909	0,910	0,918
250	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
	0,917	0,936	0,912	0,908	0,935	0,928	0,926	0,904	0,937	0,896	0,935	0,916	0,906	0,888
	0,934	0,915	0,907	0,911	0,928	0,935	0,905	0,927	0,899	0,939	0,917	0,935	0,889	0,907

Εικόνα 2.4.14 Αποτελέσματα ευαισθησίας και πιστότητας του αλγορίθμου Naive Bayes

J48	Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall	
165	0,958	0,964	0,957	0,945	0,957	0,952	0,958	0,958	0,959	0,982	0,957	0,952	0,959	0,982
	0,963	0,958	0,946	0,958	0,952	0,958	0,958	0,958	0,981	0,958	0,952	0,958	0,981	0,958
220	Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall	
	0,967	0,932	0,958	0,941	0,968	0,955	0,968	0,959	0,968	0,968	0,968	0,968	0,967	0,945
	0,934	0,968	0,938	0,956	0,952	0,966	0,957	0,966	0,968	0,968	0,968	0,968	0,947	0,968
250	Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall	
	0,972	0,968	0,972	0,964	0,964	0,952	0,963	0,944	0,972	0,964	0,972	0,960	0,968	0,956
	0,968	0,972	0,964	0,972	0,952	0,964	0,944	0,964	0,964	0,972	0,960	0,972	0,956	0,968

Εικόνα 2.4.15 Αποτελέσματα ευαισθησίας και πιστότητας του αλγορίθμου J48

RandomForest	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
165	0,958	0,964	0,957	0,952	0,946	0,958	0,958	0,958	0,959	0,982	0,952	0,958	0,959	0,988
	0,963	0,958	0,952	0,958	0,957	0,945	0,958	0,958	0,981	0,958	0,957	0,952	0,988	0,958
220	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
	0,954	0,936	0,963	0,959	0,963	0,959	0,964	0,968	0,964	0,968	0,964	0,968	0,963	0,955
	0,938	0,955	0,957	0,961	0,957	0,961	0,966	0,961	0,968	0,964	0,968	0,964	0,955	0,964
250	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
	0,964	0,972	0,972	0,972	0,972	0,956	0,960	0,968	0,968	0,980	0,964	0,960	0,964	0,964
	0,971	0,964	0,972	0,972	0,956	0,972	0,960	0,968	0,980	0,968	0,960	0,964	0,964	0,964

Εικόνα 2.4.16 Αποτελέσματα ευαισθησίας και πιστότητας του αλγορίθμου Random Forest

K-NN (N=5)	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
165	0,944	0,915	0,915	0,915	0,904	0,915	0,939	0,939	0,929	0,952	0,915	0,915	0,922	0,933
	0,918	0,945	0,915	0,915	0,914	0,903	0,939	0,939	0,929	0,952	0,915	0,915	0,933	0,921
220	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
	0,916	0,895	0,908	0,904	0,928	0,936	0,919	0,932	0,938	0,955	0,950	0,955	0,927	0,923
	0,898	0,918	0,899	0,903	0,931	0,922	0,926	0,913	0,954	0,936	0,954	0,950	0,923	0,927
250	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
	0,939	0,924	0,951	0,940	0,940	0,936	0,923	0,908	0,959	0,940	0,943	0,924	0,959	0,932
	0,924	0,939	0,940	0,951	0,935	0,939	0,908	0,923	0,940	0,960	0,925	0,943	0,933	0,960

Εικόνα 2.4.17 Αποτελέσματα ευαισθησίας και πιστότητας του αλγορίθμου K-NN

Εν κατακλείδι μελετήθηκε η εξειδίκευση η οποία παρουσιάζει αυξημένες τιμές για όλους τους αλγορίθμους πλην του K-nn (n=5) για τα δεδομένα που δεν έχει εφαρμοστεί SMOTE.

55_55				165_55			
J48	0,87	0,87	0,87	J48	0,84	0,84	0,84
	0,96	0,91	0,95		0,96	0,95	0,95
RF	0,87	0,84	0,87	RF	0,84	0,82	0,87
	0,95	0,96	0,95		0,96	0,96	0,95
KNN	0,78	0,78	0,75	KNN	0,64	0,53	0,6
	0,91	0,89	0,86		0,96	0,96	0,96
NB	0,71	0,75	0,75	NB	0,71	0,71	0,66
	0,84	0,87	0,86		0,89	0,89	0,93

Εικόνα 2.4.18 Αποτελέσματα εξειδίκευσης αλγορίθμων για τα δεδομένα χωρίς SMOTE

	File1	File2	File3		File1	File2	File3
220_55				250_55			
J48	0,86	0,86	0,82	J48	0,86	0,84	0,75
	0,93	0,96	0,95		0,97	0,97	0,95
RF	0,73	0,82	0,82	RF	0,82	0,84	0,78
	0,94	0,96	0,95		0,98	0,98	0,97
KNN	0,47	0,47	0,49	KNN	0,46	0,64	0,53
	0,95	0,98	0,96		0,98	0,96	0,98
NB	0,66	0,69	0,62	NB	0,62	0,71	0,78
	0,88	0,9	0,92		0,95	0,92	0,97

Εικόνα 2.4.19 Αποτελέσματα εξειδίκευσης αλγορίθμων για τα δεδομένα χωρίς SMOTE

	File1	File2	File3		File1	File2	File3		File1	File2	File3
165_55				220_55				250_55			
J48	0,96	0,96	0,96 J48		0,97	0,96	0,96 J48		0,97	0,97	0,96
	0,94	0,95	0,95		0,93	0,94	0,96		0,97	0,96	0,95
RF	0,96	0,96	0,95 RF		0,96	0,96	0,96 RF		0,96	0,97	0,97
	0,96	0,95	0,96		0,94	0,96	0,96		0,97	0,97	0,96
KNN	0,9	0,92	0,09 KNN		0,92	0,9	0,92 KNN		0,94	0,95	0,94
	0,9	0,99	0,9		0,9	0,9	0,94		0,92	0,94	0,94
NB	0,83	0,88	0,83 NB		0,9	0,83	0,91 NB		0,92	0,91	0,94
	0,91	0,9	0,91		0,1	0,91	0,93		1	0,91	0,93

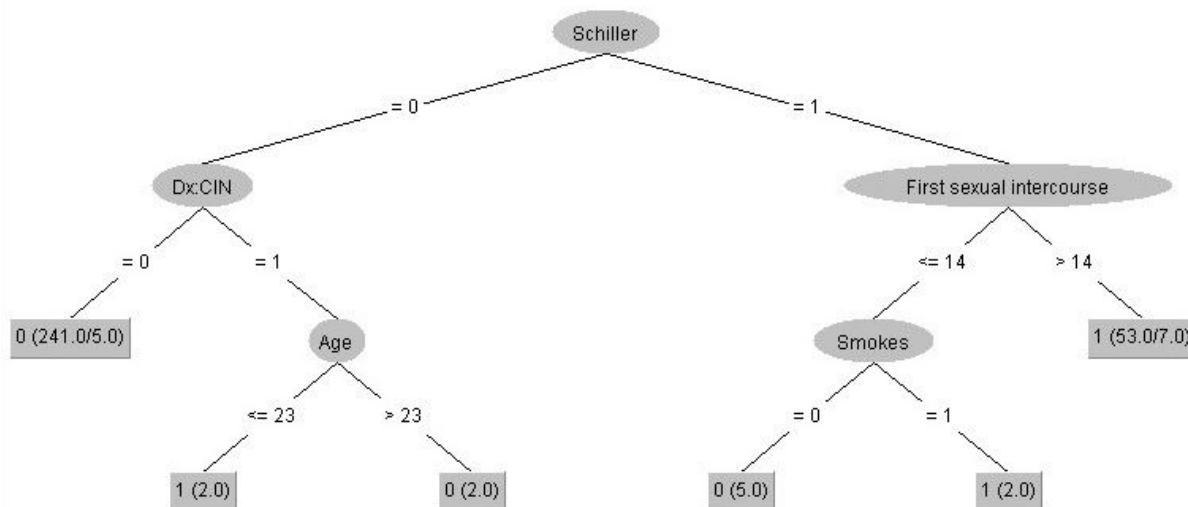
Εικόνα 2.4.20 Αποτελέσματα εξειδίκευσης αλγορίθμων για τα δεδομένα με SMOTE

Προκειμένου να επιλεγεί ο κατάλληλος αλγόριθμος για την υλοποίηση του συστήματος υποστήριξης αποφάσεων υπολογίστηκε ο μέσος όρος της ευαισθησίας, της εξειδίκευσης και της πιστοτητας. Από τα αποτελέσματα που προέκυψαν ο αλγόριθμος που τελικά επιλέγεται είναι ο J48 καθώς εν συγκρίση με τους άλλους αλγορίθμους έχει μεγαλύτερες τιμές και για τις τρεις παραμέτρους.

	Sensitivity	Precision	Recall
J48	0,9	0,95	0,95
	0,95	0,89	0,89
RF	0,88	0,94	0,96
	0,82	0,91	0,88
KNN	0,71	0,89	0,94
	0,94	0,87	0,74
NB	0,78	0,89	0,91
	0,87	0,79	0,78

Εικόνα 2.4.21 Συνοπτικά αποτελέσματα Αλγορίθμων

Έπειτα, υπολογίζοντας τον μέσο όρο εξειδίκευσης και ευαισθησίας παρόλο που όλα τα αρχεία παρουσιάζουν υψηλό μέσο όρο για την υλοποίηση του συστήματος επιλέχθηκε το αρχείο με μέσο όρο 89%, συνολικά 6 φύλλα και χωρίς SMOTE . Με βάση το δέντρο που παρουσιάζεται, έγινε και η ανάπτυξη του κώδικα σε Javascript.



Εικόνα 2.4.22 Δέντρο J48

ΚΕΦΑΛΑΙΟ 3^ο: Ολοκληρωμένο Σύστημα

Έπειτα από την ανάλυση που προηγήθηκε το σύστημα που αναπτύχθηκε είναι το εξής.

Κλινικό Σύστημα Υποστήριξης Αποφάσεων για τον Καρκίνο του τραχήλου της μήτρας

Ο καρκίνος του τραχήλου της μήτρας αποτελεί έναν από τους συχνότερους τύπους καρκίνου ο οποίος προκαλείται κυρίως από τον ιό του ανθρώπινου θηλώματος (HPV). Σκοπός του συστήματος υποστήριξης κλινικών αποφάσεων είναι η αύξηση της ακρίβειας στην διάγνωση γυναικών που διατρέχουν υψηλό κίνδυνο εμφάνισης καρκίνου.

Προσοχή! Τα κλινικά συστήματα υποστήριξης αποφάσεων δεν αντικαθιστούν τον γιατρό αλλά αποτελούν ένα βοηθητικό εργαλείο.

Αρχική Στατιστικά

Είναι θετικό το Schiller Test? (Is Schiller Test positive?)

- ☐ Ναι(Yes)
☐ Όχι(No)

Υπάρχει δυσπλασία στην περιοχή του τραχήλου?(Is there any dysplasia in Cervix?)

- ☐ Ναι(Yes)
☐ Όχι(No)

Ηλικία (Age)

- ☐ 23 ή μικρότερη(23 or Less)
☐ Μεγαλύτερη από 23(More than 23)

Σε ποιά ηλικία ήταν η πρώτη σεξουαλική επαφή? (First Sexual Intercourse)

- ☐ 14 ή μικρότερη (14 or Less)
☐ Μεγαλύτερη από 14 (More than 14)

Κάπνισμα (Smoking)

- ☐ Ναι(Yes)
☐ Όχι(No)

Submit

3.1 Αρχική σελίδα συστήματος

Αφού ο χρήστης απαντήσει σε όλες τις ερωτήσεις τότε εμφανίζεται ένα αναδυόμενο παράθυρο με τα σχετικά αποτελέσματα είτε είναι θετικά είτε αρνητικά.

Κλινικό Σύστημα Υποστήριξης Αποφάσεων για τον Καρκίνο του τραχήλου της μήτρας

Ο καρκίνος του τραχήλου της μήτρας αποτελεί έναν από τους συχνότερους τύπους καρκίνου ο οποίος προκαλείται κυρίως από τον ιό του ανθρώπινου θηλώματος (HPV). Σκοπός του συστήματος υποστήριξης κλινικών αποφάσεων είναι η αύξηση της ακρίβειας στην διάγνωση γυναικών που διατρέχουν υψηλό κίνδυνο εμφάνισης καρκίνου.

Προσοχή! Τα κλινικά συστήματα υποστήριξης αποφάσεων δεν αντικαθιστούν τον γιατρό αλλά αποτελούν ένα βοηθητικό εργαλείο.

Αρχική Στατιστικά

Είναι θετικό το Schiller Test? (Is Schiller Test positive?)
☐ Ναι(Yes)
☒ Όχι(No)

Υπάρχει δυσπλασία στην περιοχή του τραχήλου?(Is there any dysplasia in Cervix?)
☒ Ναι(Yes)
☐ Όχι(No)

Ηλικία (Age)
☐ 23 ή μικρότερη(23 or Less)
☒ Μεγαλύτερη από 23(More than 23)

Σε ποιά ηλικία ήταν η πρώτη σεξουαλική επαφή? (First Sexual Intercourse)
☐ 14 ή μικρότερη (14 or Less)
☒ Μεγαλύτερη από 14 (More than 14)

Κάπνισμα (Smoking)
☐ Ναι(Yes)
☒ Όχι(No)

Submit

Αυτή η τοποθεσία αναφέρει...
Η πιθανότητα εμφάνισης Καρκίνου είναι μικρή. The possibility of Cancer is small
☐ Να μην επιτραπεί σε αυτήν τη σελίδα να δημιουργήσει περισσότερα μηνύματα
OK

3.2 Μικρή πιθανότητα εμφάνισης καρκίνου

Κλινικό Σύστημα Υποστήριξης Αποφάσεων για τον Καρκίνο του τραχήλου της μήτρας

Ο καρκίνος του τραχήλου της μήτρας αποτελεί έναν από τους συχνότερους τύπους καρκίνου ο οποίος προκαλείται κυρίως από τον ιό του ανθρώπινου θηλώματος (HPV). Σκοπός του συστήματος υποστήριξης κλινικών αποφάσεων είναι η αύξηση της ακρίβειας στην διάγνωση γυναικών που διατρέχουν υψηλό κίνδυνο εμφάνισης καρκίνου.

Προσοχή! Τα κλινικά συστήματα υποστήριξης αποφάσεων δεν αντικαθιστούν τον γιατρό αλλά αποτελούν ένα βοηθητικό εργαλείο.

Αρχική

Στατιστικά

Είναι θετικό το Schiller Test? (Is Schiller Test positive?)

☒ Ναι(Yes)

☐ Όχι(No)

Υπάρχει δυσπλασία στην περιοχή του τραχήλου?(Is there any dysplasia in the area of the cervix?)

☐ Ναι(Yes)

☒ Όχι(No)

Ηλικία (Age)

☒ 23 ή μικρότερη(23 or Less)

☐ Μεγαλύτερη από 23(More than 23)

Σε ποιά ηλικία ήταν η πρώτη σεξουαλική επαφή? (First Sexual Intercourse)

☐ 14 ή μικρότερη (14 or Less)

☒ Μεγαλύτερη από 14 (More than 14)

Κάπνισμα (Smoking)

☐ Ναι(Yes)

☒ Όχι(No)

Submit

Αυτή η τοποθεσία αναφέρει...

Η πιθανότητα εμφάνισης καρκίνου είναι μεγάλη. The possibility of Cancer is high

☐ Να μην επιτραπεί σε αυτήν τη σελίδα να δημιουργήσει περισσότερα μηνύματα

OK

3.3 Μεγάλη πιθανότητα εμφάνισης καρκίνου

Σε περίπτωση που ο χρήστης δεν απαντήσει σε όλες τις ερωτήσεις τότε εμφανίζεται ένα αναδυόμενο παράθυρο με αντίστοιχο μήνυμα.

Κλινικό Σύστημα Υποστήριξης Αποφάσεων για τον Καρκίνο του τραχήλου της μήτρας

Ο καρκίνος του τραχήλου της μήτρας αποτελεί έναν από τους συχνότερους τύπους καρκίνου ο οποίος προκαλείται κυρίως από τον ιό του ανθρώπινου θηλώματος (HPV). Σκοπός του συστήματος υποστήριξης κλινικών αποφάσεων είναι η αύξηση της ακρίβειας στην διάγνωση γυναικών που διατρέχουν υψηλό κίνδυνο εμφάνισης καρκίνου.

Προσοχή! Τα κλινικά συστήματα υποστήριξης αποφάσεων δεν αντικαθιστούν τον γιατρό αλλά αποτελούν ένα βοηθητικό εργαλείο.

Αρχική

Στατιστικά

Είναι θετικό το Schiller Test? (Is Schiller Test positive?)

☐ Ναι(Yes)

☐ Όχι(No)

Υπάρχει δυσπλασία στην περιοχή του τραχήλου?(Is there any dysplasia in the area of the cervix?)

☐ Ναι(Yes)

☐ Όχι(No)

Ηλικία (Age)

☐ 23 ή μικρότερη(23 or Less)

☐ Μεγαλύτερη από 23(More than 23)

Σε ποιά ηλικία ήταν η πρώτη σεξουαλική επαφή? (First Sexual Intercourse)

☐ 14 ή μικρότερη (14 or Less)

☒ Μεγαλύτερη από 14 (More than 14)

Κάπνισμα (Smoking)

☐ Ναι(Yes)

☒ Όχι(No)

Submit

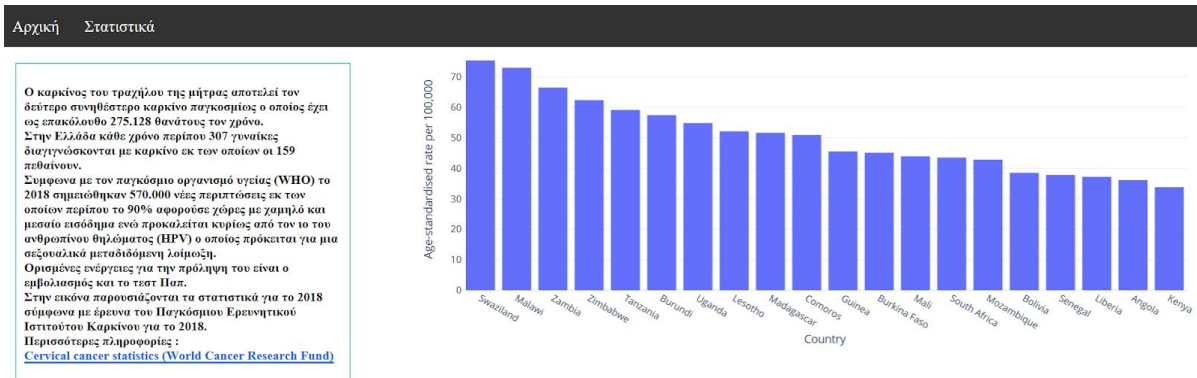
Αυτή η τοποθεσία αναφέρει...

Παρακαλώ απαντήστε σε όλες τις ερωτήσεις.Please answer all the questions

OK

3.4 Αναδυόμενο παράθυρο συμπλήρωσης όλων των ερωτήσεων από τον χρήστη

Επιλέγοντας το κουμπί “Στατιστικά” εμφανίζονται σχετικές πληροφορίες καθώς και ένα διάγραμμα με τα ποσοστά καρκίνου της μήτρας σύμφωνα με το World Cancer Research Fund.



3.5 Σελίδα με τα στατιστικά στοιχεία

Συζήτηση/Συμπεράσματα

Συμπερασματικά, η χρήση της τεχνητής νοημοσύνης και γενικότερα της τεχνολογίας αποτελεί ένα εργαλείο το οποίο μπορεί να συμβάλλει στην έγκαιρη διάγνωση του καρκίνου χρησιμοποιώντας διάφορες τεχνικές και αλγορίθμους.

Όσον αφορά την διεπαφή του προγράμματος WEKA δεν παρουσιάστηκε κάποιο πρόβλημα καθώς είναι αρκετά κατανοητό και εύκολο στην χρήση ενώ παράλληλα αποτέλεσε ένα σημαντικό εργαλείο προκειμένου να επιλεγεί ο κατάλληλος αλγόριθμος και να εξαχθούν τα αποτελέσματα σχετικά με την ακρίβεια, εξειδίκευση και ευαισθησία του κάθε αλγορίθμου.

Το σύστημα που αναπτύχθηκε είναι απλό στην χρήση και αποτελεί συμβουλευτικό εργαλείο για τους γιατρούς ελέγχοντας συγκεκριμένες παραμέτρους. Η διαδικασία που ακολουθήθηκε μέχρι την ολοκλήρωση του συστήματος είχε ως επακόλουθο την καλύτερη κατανόηση των εννοιών σχετικά με την εξόρυξη δεδομένων.

Συγκριτικά με τα αποτελέσματα που προέκυψαν από τις μελέτες που παρουσιάστηκαν στο δεύτερο κεφάλαιο οι τιμές των μέτρων είναι εξίσου υψηλές ενώ παρατηρήθηκε ότι χρησιμοποιήθηκαν παρόμοιοι αλγόριθμοι (π.χ K-NN). Ακόμη, διαπιστώθηκε ότι η ανάπτυξη κλινικών συστημάτων τα τελευταία χρόνια έχει αυξηθεί κατακόρυφα καθώς παρουσιάζεται αρκετό ενδιαφέρον για τον τομέα της τεχνητής νοημοσύνης.

Συνοψίζοντας, αποδεικνύεται ότι η αξιοποίηση των συστημάτων συμβάλλει στην πιο έγκαιρη διάγνωση των ασθενών. Όσο η τεχνολογία εξελίσσεται τόσο περισσότερο θα αναζητούνται τρόποι ώστε να επιτυγχάνεται η καλύτερη διάγνωση των ασθενών.

Μελλοντικά το σύστημα θα μπορούσε να χρησιμοποιηθεί από τους γιατρούς προκειμένου να αξιολογηθεί αλλά και από ασθενείς ώστε να διερευνηθεί κατά πόσο εμπιστεύονται τέτοιου είδους συστήματα.

Παράρτημα Κώδικας

Κώδικας για την “Αρχική” σελίδα

```
<!DOCTYPE html>
<html>
<head>
<meta name="viewport"content="width=device-width,initial-scale=1">
<title>Clinical Decision Support System for Cervical Cancer</title>
<link href="style.css" rel="stylesheet">
</head>
<body>
<section class="dss">
<div class="cdss">
<h1> Κλινικό Σύστημα Υποστήριξης Αποφάσεων για τον Καρκίνο του τραχήλου της
μήτρας</h1>
<p> Ο καρκίνος του τραχήλου της μήτρας αποτελεί έναν από τους συνηθέστερους τύπους
καρκίνου ο οποίος προκαλείται κυρίως από τον ιό του ανθρώπινου θηλώματος (HPV).
<br>Σκοπός του συστήματος υποστήριξης κλινικών αποφάσεων είναι η αύξηση της
ακρίβειας στην διάγνωση γυναικών που διατρέχουν υψηλό κίνδυνο εμφάνισης
καρκίνου.</br></p>
<p><b>Προσοχή! Τα κλινικά συστήματα υποστήριξης αποφάσεων δεν αντικαθιστούν τον
γιατρό αλλά αποτελούν ένα βοηθητικό εργαλείο.</b></p>
</div>
</section>
<div class="topnavmenu">
<a class="active" href="file:///C:/Users/rania/Desktop/Cdss/index.html">Αρχική</a>
<a href="file:///C:/Users/rania/Desktop/Cdss/index2.html">Στατιστικά</a>
</div>
<div>
<h2>&nbsp;Είναι θετικό το Schiller Test? (Is Schiller Test positive?)</h2>

<label>
<input type="radio" class="schiller" name="quest1" value="yes"
onclick="getSchiller();" />&nbsp;Ναι(Yes)</br>
</label>
<label>
<input type="radio" class="schiller" name="quest1" value="no"
onclick="getSchiller();" />&nbsp;Όχι(No)</br>
</label>

</div>
<div>
```

<h2> Υπάρχει δυσπλασία στην περιοχή του τραχήλου?(Is there any dysplasia in Cervix?)</h2>

<label>

<input type="radio" class="DxCin" name="quest2" value="yes" onclick="getDxCin();" /> Ναι(Yes)</br>

</label>

<label>

<input type="radio" class="DxCin" name="quest2" value="no" onclick="getDxCin();" /> Όχι(No)</br>

</label>

</div>

<div>

<h2> Ηλικία (Age)</h2>

<label>

<input type="radio" class="age" name="quest3" value="0" onclick="getAge();" /> 23 ή μικρότερη(23 or Less)</br>

</label>

<label>

<input type="radio" class="age" name="quest3" value="1" onclick="getAge();" /> Μεγαλύτερη από 23(More than 23)</br>

</label>

</div>

<div>

<h2> Σε ποιά ηλικία ήταν η πρώτη σεξουαλική επαφή? (First Sexual Intercourse)</h2>

<label>

<input type="radio" class="firstsexualintercourse" name="quest4" value="0" onclick="getFirstSexualIntercourse();" /> 14 ή μικρότερη (14 or Less)</br> <!--Using class because its better than id -->

</label>

<label>

<input type="radio" class="firstsexualintercourse" name="quest4" value="1" onclick="getFirstSexualIntercourse();" /> Μεγαλύτερη από 14 (More than 14)</br>

</label>

</div>

<div>

<h2> Κάπνισμα (Smoking)</h2>

<label> <!--Click the letters -->

<input type="radio" class="smokes" name="quest5" value="yes" onclick="getSmokes();" /> Ναι(Yes)</br>

</label>

<label>

```

        <input      type="radio"      class="smokes"      name="quest5"      value="no"
onclick="getSmokes();" />&nbsp; Όχι(No)</br>
    </label>
</div>
<div class="submitresults">
    <input id="rslt" type="button" value="Submit" onclick="result();" />
<!--<div id="results"><br>&nbsp; Αποτελέσματα(Results)<br></div>-->
</div>

<style>
body { margin:0; background-color:white;}
.dss h1 {

    font-size:4em;
    margin-top: 0;
    margin-bottom: 0;
    text-align: center;
    }
.dss {

    background-size: cover;
    background-attachment: fixed;
    background-color:lightSeaGreen;
    padding:40px;
    color: white;
    background-repeat: no-repeat;
    text-align:center;
    }
.cdss { font-size:18px;}
.topnavmenu {

    overflow:hidden;
    background-color:#333;

    }
.topnavmenu a {
    float: left;
    color: white;
    text-align: center;
    padding: 15px 16px;
    text-decoration: none;
    font-size: 20px;

```

```

}
.topnavmenu a:hover {
  background-color: lightSeaGreen;
  color: white;
}
input[type=button] {
  background-color: lightSeaGreen;
  border: none;
  color: white;
  padding 20px 40px;
  cursor: pointer;
  margin: 4px 2px;
  text-align: center;
  font-size: 16px;
  text-decoration: none;

}
</style>
<script>
var schiller;
var DxCin;
var age;
var firstsexualintercourse;
var smokes;
var answer;

function getSchiller() { //check answers

var sch = document.getElementsByClassName("schiller");
if (sch[0].checked) {

  schiller = sch[0].value;

}
else if (sch[1].checked) {

  schiller = sch[1].value;

}
}

function getDxCin() {

```

```

var dx = document.getElementsByClassName("DxCin");
if (dx[0].checked) {

    DxCin = dx[0].value;

}
else if (dx[1].checked) {

    DxCin = dx[1].value;

}
}

function getAge() {
var ag = document.getElementsByClassName("age");
if (ag[0].checked) {
    age = ag[0].value;
}
else if (ag[1].checked) {
    age = ag[1].value;
}
}

function getFirstSexualIntercourse(){

var fsc = document.getElementsByClassName("firstsexualintercourse");
if (fsc[0].checked) {
    firstsexualintercourse = fsc[0].value;
}
else if (fsc[1].checked) {
    firstsexualintercourse = fsc[1].value;
}
}

function getSmokes() {

var smk = document.getElementsByClassName("smokes");
if (smk[0].checked ) {

    smokes = smk[0].value;

```

```

}
else if (smk[1].checked) {

    smokes = smk[1].value;

}

}

function result() {

    if(schiller === "no") {
        if(DxCin === "no") {
            answer = "no cancer";
        }
        else if(DxCin === "yes") {
            if(age === "0") { //if age is less than 23
                answer = "cancer";
            }
            else if(age === "1") { //if age is under 23

                answer = "no cancer";
            }
        }
    } else if(schiller === "yes") {
        if(firstsexualintercourse === "0") { //if firstsexualintercourse is above 14
            if(smokes === "no") {
                answer = "no cancer";

            }
            else if(smokes === "yes") {
                answer = "cancer";
            }
        }
        else if(firstsexualintercourse === "1") { //if firstsexualintercourse is under 14
            answer = "cancer";
        }
    }

    if (answer === "no cancer") {
        alert("Η πιθανότητα εμφάνισης Καρκίνου είναι μικρή, The possibility of Cancer is small");
    }
}

```

```

}
else if (answer === "cancer") {
    alert("Η πιθανότητα εμφάνισης καρκίνου είναι μεγάλη, The possibility of Cancer is high");
}
else {
    alert("Παρακαλώ απαντήστε σε όλες τις ερωτήσεις, Please answer all the questions");
}
}
}
</script>
</body>
</html>

```

Κώδικας για την σελίδα “Στατιστικά”

```

<!DOCTYPE html>
<html>
<head>
<meta name="viewport"content="width=device-width;initial-scale=1">
<title>Clinical Decision Support System for Cervical Cancer</title>
<link href="style.css" rel="stylesheet">
</head>
<body>
<section class="dss">
<div class="cdss">
    <h1> Κλινικό Σύστημα Υποστήριξης Αποφάσεων για τον Καρκίνο του τραχήλου της
    μήτρας</h1>
    <p> Ο καρκίνος του τραχήλου της μήτρας αποτελεί έναν από τους συνηθέστερους τύπους
    καρκίνου ο οποίος προκαλείται κυρίως από τον ιό του ανθρωπίνου θηλώματος (HPV).
    <br>Σκοπός του συστήματος υποστήριξης κλινικών αποφάσεων είναι η αύξηση της
    ακρίβειας στην διάγνωση γυναικών που διατρέχουν υψηλό κίνδυνο εμφάνισης
    καρκίνου.</br></p>
    <p><b>Προσοχή! Τα κλινικά συστήματα υποστήριξης αποφάσεων δεν αντικαθιστούν τον
    γιατρό αλλά αποτελούν ένα βοηθητικό εργαλείο.</b></p>
</div>
</section>
<div class="topnavmenu">
    <a class="active" href="file:///C:/Users/rania/Desktop/Cdss/index.html">Αρχική</a>
    <a href="file:///C:/Users/rania/Desktop/Cdss/index2.html">Στατιστικά</a>
</div>

<div class="statistics">
<p><b> Ο καρκίνος του τραχήλου της μήτρας αποτελεί τον δεύτερο συνηθέστερο καρκίνο
    παγκοσμίως ο οποίος έχει ως επακόλουθο 275.128 θανάτους τον χρόνο.

```


Στην Ελλάδα κάθε χρόνο περίπου 307 γυναίκες διαγιγνώσκονται με καρκίνο εκ των οποίων οι 159 πεθαίνουν.

Συμφωνα με τον παγκόσμιο οργανισμό υγείας (WHO) το 2018 σημειώθηκαν 570.000 νέες περιπτώσεις

εκ των οποίων περίπου το 90% αφορούσε χώρες με χαμηλό και μεσαίο εισόδημα ενώ προκαλείται κυρίως από τον ιο του ανθρωπίνου θηλώματος (HPV) ο οποίος πρόκειται για μια σεξουαλικά μεταδιδόμενη λοίμωξη.

Ορισμένες ενέργειες για την πρόληψη του είναι ο εμβολιασμός και το τεστ Παπ.</br>

Στην εικόνα παρουσιάζονται τα στατιστικά για το 2018 σύμφωνα με έρευνα του Παγκόσμιου Ερευνητικού Ινστιτούτου Καρκίνου για το 2018.

Περισσότερες πληροφορίες :</br> Cervical cancer statistics (World Cancer Research Fund)</p>

</div>

<style>

body { margin:0; background-color:white;background-repeat:no-repeat;}

.dss h1 {

font-size:4em;

margin-top: 0;

margin-bottom: 0;

text-align: center;

}

.dss {

background-size: cover;

background-attachment: fixed;

background-color:lightSeaGreen;

padding:40px;

color: white;

background-repeat: no-repeat;

text-align:center;

}

.cdss { font-size:18px;}

.topnavmenu {

overflow:hidden;

background-color:#333;

}

.topnavmenu a {

float: left;

```
color: white;
text-align: center;
padding: 15px 16px;
text-decoration: none;
font-size: 20px;
}
.topnavmenu a:hover {
background-color: lightSeaGreen;
color: white;
}
.statistics {
background-color:white;
padding:10px;
border: 1px solid lightSeaGreen;
width:400px;
height:150px;
word-wrap: break-word;
margin:20px;
}
img{
max-width:100%;
height:auto;
float:right;
}
</style>

</body>
</html>
```

Βιβλιογραφία

1. Weka, “Machine Learning Software in Java”,
<https://www.cs.waikato.ac.nz/~ml/weka/index.html>
2. World Health Organization, “Cervical Cancer, Early diagnosis and screening” ,
<https://www.who.int/cancer/prevention/diagnosis-screening/cervical-cancer/en/>
3. World Cancer Research Fund, “Cervical Cancer Statistics”,
<https://www.wcrf.org/dietandcancer/cancer-trends/cervical-cancer-statistics>
4. Priyanka K Malli, Dr. Suvarna Nandyal, “*Machine Learning Technique for detection of Cervical Cancer using k-NN and Artificial Neural Network*” ,International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) (2017),
<http://www.ijettcs.org/Volume6Issue4/IJETTCS-2017-08-20-62.pdf>
5. Dipti N. Punjani, Dr. Kishor H. Atkotiya, “*Cervical Cancer Prediction using Data Mining*”, International Journal for Research in Applied Science & Engineering Technology (IJRASET) (2017), <https://www.ijraset.com/fileserve.php?FID=12245>
6. Kavishwar B. Waghlikar, Kathy L. MacLaughlin, Michael R. Henry, Robert A. Greenes, Ronald A. Hankey, Hongfang Liu et al., “*Clinical decision support with automated text processing for cervical cancer screening*”, Journal of American Medical Informatics Association (2012),
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3422840>
7. Kavishwar B. Waghlikar, Kathy L. MacLaughlin, Petra M. Casey, Thomas M. Kastner, Michael R. Henry, Ronald A. Hankey et al., “*Automated Recommendation for Cervical Cancer Screening and Surveillance*”, Cancer Informatics (2014),
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4214690/>
8. Panagiotis Bountris, Maria Haritou, Abraham Pouliakis, Niki Margari, Maria Kyrgiou, Aris Spathis et al., “*An Intelligent Clinical Decision Support System for Detection*”, BioMed Research International (2014),
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4000928/>
9. Rajeev Chaudhry, Kathy L. MacLaughlin, Ravikumar Komandur Elayavili, Maya E. Kessler, Hongfang Liu, Kavishwar B. Waghlikar et al., “*NLP-enabled decision support for cervical cancer screening and surveillance*”, (2014 to 2016),
<https://healthit.ahrq.gov/site/default/files/docs/citation/r2hs022911-chaudhry-final-report-2017.pdf>
10. Masakazu Sato, Koji Horie, Aki Hara, Yuichiro Miyamoto, Kazuko Kurihara et al., “*Application of deep learning to the classification of images from colposcopy*”, Oncology Letter Spandidos Publications (2018),
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5795879/>
11. Munhammed Fahri Unlesern, Kadir Sabanci, Muciz Özean, “*Determining Cervical Cancer Possibility by Using Machine Learning Methods*”, International Journal of Latest Research in Engineering and Technology (IJLRET) (2017),
<https://www.researchgate.net/publication/322233711>

12. Rebecka Weegar, Maria Kvist, Karin Sundström, Søren Brunak, Hercules Dalianis, “*Finding Cervical Cancer Symptoms in Swedish Clinical Text using a Machine Learning Approach and NegEx*”, AMIA Annual Symposium Proceedings Archive (2015), <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4765575>
13. Panagiotis Bountris, Maria Hartiou, Abraham Pouliakis, Petros Karakitsos, Dimitrios Koutsouris, “*A decision support system based on an ensemble of random forests for improving the management of women with abnormal findings at cervical cancer screening*”, Conf Proc IEEE Enq Med Biol Soc. (2015), <https://ieeexplore.ieee.org/document/7320286>
14. Abid Sarwar, Mehbob Ali, Jyotsna Suri, Vinod Sharma, “*Performance Evaluation of Machine Learning Techniques for Screening of Cervical Cancer*”, 2nd International Conference on Computing for Sustainable Global Development (2015). <https://ieeexplore.ieee.org/document/7100375>
15. Kelwin Fernandes, Davide Chicco, James S. Cardoso, Jessica Fernandes, “*Supervised deep learning embeddings for the prediction of cervical cancer diagnosis*”. PeerJ Comput. Sci 4:3154; DOI 10.7717/peerj-cs.154 (2018), <https://peerj.com/articles/cs-154.pdf>
16. R. Vidya, G. M. Narisa, “*Prediction of Cervical Cancer using Hybrid Induction Technique: A Solution for Human Hereditary Disease Patterns*”, Indian Journal of Science and Technology, Vol 9(30), DOI: 10.17485/ijst/2016/v9i30/82085 (2016), <http://www.indjst.org/index.php/indjst/article/viewFile/82085/72318>
17. Rajeev Gupta, Abid Sarwa, Vinod Sharma, “*Screening of Cervical Cancer by Artificial Intelligence based Analysis of Digitized Papanicolaou-Smear Images*”, International Journal of Contemporary Medical Research Volume 4 | Issue 5 (2017), https://www.ijcmr.com/uploads/7/7/4/77464738/ijcmr_1457_june_26.pdf
18. Abdullah M Iliyasu, Chastine Fatichah, “*A Quantum Hybrid PSO Combined with Fuzzy k-NN Approach to Feature Selection and Cell Classification in Cervical Cancer Detection*”, Sensors (2017), <https://ncbi.nlm.nih.gov/pmc/articles/PMC5750777/>
19. K.E. Ravikumar. Kathy L. MacLaughlin, Marianne R. Scheitel, Maya Kessler, Kavishwar B. Waghlikar, Hongfang Liu et al., “*Improving the Accuracy of a Clinical Decision Support System for Cervical Cancer Screening and Surveillance*”, Appl Clin Inform DOI: 10.1055/s-0037-1617451 (2017), <https://www.thieme-connect.com/products/ejournals/html/10.1055/s-0037-1617451>
20. Kelwin Fernandes, Jaime S. Cardoso, Jessica Fernandes, Cervical cancer (Risk Factors) Data Set (2017), <https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29>
21. HPV ΛΟΙΜΩΞΗ & ΠΡΟΛΗΨΗ Ca TPAXHΛOY
www.med.auth.gr/depts/amg/docs/Stamatopoulos_-_HPV_Loimoksi_&_Prolipsi_Ca_Traxilou.pdf

22. Pang-Ning Tan, Michael Steinbach, Vipin Kumar, “Εισαγωγή στην Εξόρυξη Δεδομένων, Εξαγωγή Δέντρου Απόφασης” (σελ 175-178), Εκδόσεις Τζιόλα, Μετάφραση: Σταύρος Σουραβλάς, ISBN 978-960-162-9
23. Will Koehrsen, Random Forest Simple Explanation, Dec 27, 2017, A Medium Corporation,
<https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d>