ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Εξόρυξη δεδομένων και αναγνώριση προτύπων με χρήση γενετικών αλγόριθμων και τεχνικών θεωρίας πληροφορίας για τη βελτιστοποιημένη ταξινόμηση περιστατικών τραχηλικής ενδοεπιθηλιακής νεοπλασίας

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Παναγιώτης Π. Κοσσιώρης

Επιβλέπων : Δημήτριος-Διονύσιος Κουτσούρης

Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2012

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Εξόρυξη δεδομένων και αναγνώριση προτύπων με χρήση γενετικών αλγόριθμων και τεχνικών θεωρίας πληροφορίας για τη βελτιστοποιημένη ταξινόμηση περιστατικών τραχηλικής ενδοεπιθηλιακής νεοπλασίας

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Παναγιώτης Π. Κοσσιώρης

Επιβλέπων : Δημήτριος-Διονύσιος Κουτσούρης

Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 23/03/2012

.....	.....	.....
Δ.-Δ. Κουτσούρης	Κ. Νικήτα	Γ. Ματσόπουλος

Αθήνα, Μάρτιος 2012

.....

Παναγιώτης Π. Κοσσιώρης

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

«Το στυλ και η δομή είναι τα βασικά σε ένα βιβλίο.

Οι μεγάλες ιδέες είναι μπαρούφες.»

Vladimir Nabokov

ΠΕΡΙΛΗΨΗ

Ο καρκίνος του τραχήλου της μήτρας αποτελεί έναν από τους πλέον θανατηφόρους καρκίνους των γυναικών και οφείλεται κατά κύριο λόγο στη λοίμωξη από τον ιό των ανθρώπινων θηλωμάτων (HPV). Είναι πολλοί οι τύποι του HPV που ενοχοποιούνται (άλλοι σε μεγαλύτερο βαθμό και άλλοι σε μικρότερο) για την πρόκληση προκαρκινικών αλλοιώσεων στον τράχηλο της μήτρας.

Από την πρώτη στιγμή που έγινε γνωστή η βασική αιτία των αλλοιώσεων του τραχήλου της μήτρας, έγινε προσπάθεια για να εξευρεθούν τεχνικές οι οποίες θα προσέφεραν έγκυρη και όσο το δυνατόν πιο έγκαιρη ανίχνευση, με απώτερο σκοπό τον αποτελεσματικότερο προληπτικό πληθυσμιακό έλεγχο της νόσου. Σημαντικότατο ρόλο προς την κατεύθυνση αυτή διαδραμάτισε ο Γεώργιος Παπανικολάου, ο οποίος ανέπτυξε το 1943 το γνωστό test Pap το οποίο αποτελεί ακόμα και σήμερα το πιο διαδεδομένο και παράλληλα το αποδοτικότερο test για τη διάγνωση του καρκίνου του τραχήλου της μήτρας. Ωστόσο το test Pap εμφανίζει ένα σημαντικό ποσοστό εσφαλμένης πρόβλεψης. Η λανθασμένη αυτή πρόβλεψη μπορεί να οδηγήσει την ασθενή είτε σε μία σειρά περιττών εξετάσεων και θεραπειών σε περίπτωση υπερεκτίμησης της πραγματικής κατάστασης, είτε σε έναν λανθασμένο καθησυχασμό σε περίπτωση υποεκτίμησης της πραγματικής κλινικής κατάστασης.

Για τον λόγο αυτό, αναπτύχθηκαν τα τελευταία χρόνια ορισμένες νέες τεχνικές για την ανίχνευση των τραχηλικών αλλοιώσεων, που έχουν σαν στόχο να βοηθήσουν τους γιατρούς όσον αφορά την αναγνώριση της κλινικής κατάστασης και εν τέλει να μειώσουν, με τον τρόπο αυτό, το ποσοστό λάθους του test Pap. Οι πιο σημαντικές τεχνικές, που μελετώνται σήμερα σε ολόκληρο τον ερευνητικό κόσμο, είναι το HPV DNA, το NASBA mRNA, το p16 και η κυτταρομετρία ροής. Με τις τεχνικές αυτές γίνεται μία προσπάθεια ανίχνευσης της πορείας της έκφρασης του γενετικού υλικού του HPV στα τραχηλικά κύτταρα. Γίνεται έτσι κατανοητό πως είναι ιδιαίτερα χρήσιμο να κατανοήσουμε πώς αυτές οι τεχνικές μπορούν να συνδυαστούν και κυρίως ποιες από αυτές μας παρέχουν πληροφορίες για την εκτίμηση της κλινικής κατάστασης.

Στην έρευνα που διεξάγαμε εφαρμόσαμε τεχνικές επιλογής χαρακτηριστικών, θεωρίας πληροφορίας, γενετικών αλγορίθμων και αναγνώρισης προτύπων με σκοπό την εξόρυξη δεδομένων από ένα δείγμα 212 γυναικών, οι οποίες εμφάνιζαν αποτέλεσμα test Pap χαμηλού βαθμού (LgSIL) ή υψηλού βαθμού (HgSIL) πλακώδης ενδοεπιθηλιακή βλάβη και η αντίστοιχη ιστολογική εξέταση έδειχνε κάποιας μορφής τραχηλική ενδοεπιθηλιακή νεοπλασία. Τα 212 αυτά περιστατικά ανήκαν σε μία ευρύτερη βάση 500 δειγμάτων, που περιλαμβάνουν τα αποτελέσματα των εξετάσεων των πέντε τεχνικών ανίχνευσης που προαναφέρθηκαν και συγκεντρώθηκαν από το Αττικό Νοσοκομείο και από το Νοσοκομείο Ιωαννίνων. Με βάση την εργασία μας καταλήξαμε στην ανάπτυξη ενός συστήματος που προβλέπει την κλινική κατάσταση και κάναμε σύγκριση με την ιστολογική βιοψία, η οποία αποτέλεσε το χρυσό κανόνα στην έρευνά μας. Συνολικά, από τις πέντε ξεχωριστές τεχνικές

ανίχνευσης είχαμε μία βάση με 50 ξεχωριστές τιμές-χαρακτηριστικά για κάθε μία ασθενή. Με χρήση προηγμένων τεχνικών υπολογιστικής νοημοσύνης καταφέραμε να μειώσουμε σημαντικά τον αριθμό των χαρακτηριστικών που συμμετέχουν ουσιαστικά στη διαμόρφωση της κλινικής κατάστασης και να αυξήσουμε σε μεγάλο βαθμό τα στατιστικά μέτρα απόδοσης που περιγράφουν την εκτίμηση για την πρόβλεψη της ιστολογίας σε σύγκριση με το test Pap. Ακόμα, βασιζόμενοι σε αυτό, κατασκευάσαμε ένα αρχικό μοντέλο για την εκτίμηση της πραγματικής κλινικής κατάστασης. Τα συμπεράσματά μας μπορούν να εκληφθούν ως μία προκαταρκτική ερευνητική εργασία και εφόσον επιβεβαιωθούν από μία μεγαλύτερη βάση δεδομένων θα μπορούσαν να αποτελέσουν τη βάση για τη δημιουργία ενός συστήματος για την κατάλληλη ιατρική διαλογή ασθενών με τραχηλική ενδοεπιθηλιακή νεοπλασία.

Λέξεις-κλειδιά: Καρκίνος τραχήλου μήτρας, Pap test, Προληπτικός πληθυσμιακός έλεγχος, HPV DNA test, NASBA mRNA test, p16 test, Κυτταρομετρία φθορισμού, Επιλογή Χαρακτηριστικών, Θεωρία Πληροφορίας, Γενετικοί Αλγόριθμοι, Αναγνώριση Προτύπων, Εξόρυξη δεδομένων, Στατιστικά μέτρα απόδοσης

ABSTRACT

Cervical cancer is one of the most common and fatal cancers among women, and it is mainly due to infection by Human Papillomavirus (HPV). There are many types of HPV which are implicated (to a bigger or lesser extent) for the cause of precancerous lesions in cervix uteri.

From the first moment that the main cause of cervical lesions became clear, there was an effort made to develop techniques which would lead to a valid and as timely as possible detection, in order to provide a more efficient screening for the disease. To this direction a very important role was played by George Papanicolaou, who developed the well-known test Pap which is still today the most effective test for the diagnosis of cervical cancer. Nevertheless, the test Pap shows an important percentage of incorrect prediction. Such an incorrect prediction could lead a patient either to having a series of unnecessary medical tests in case of an overestimated prediction or to a false reassurance in case of an underestimation of the true clinical state.

For this reason, some new techniques have been developed the last years for the detection of cervical lesions, which aim to support physicians in the detection of the clinical state and, hence, to decrease the error rate of the test Pap. The most important techniques are HPV DNA, NASBA mRNA, p16 and flow cytometry. These techniques try to detect the route of gene expression of HPV in cervical cells. It is clear, therefore, that it is very useful to understand in which way all these tests could be combined and mainly which of these tests provide us with information for better assessing the clinical state.

In this work, we applied techniques of feature selection, information theory, genetic algorithms and pattern recognition to mine data from a sample of 212 women which had a test Pap result of Low Grade (LgSIL) or High Grade (HgSIL) Squamous Intraepitheliel Lesion and their comprehensive biopsy showed some sort of a Cervical Intraepithelial Neoplasia. These 212 incidents belonged to a wider database of 500 samples, consisting of the results of the five screening tests which were compiled in Attikon Hospital and Hospital of Ioannina, Greece. Our work resulted in the development of a system which predicts the clinical state. Furthermore, we compared the outcomes of the system with the histological biopsy result, which was the golden standard in our research. Including all of the 5 separate detection techniques we developed a database of separate values-features for each one of the patients. With the use of advanced computational intelligence techniques we managed to decrease the number of the features which participate in the form of the final clinical state. Moreover, we managed in this way to increase the statistical measures for the prediction of the histology compared to the test Pap. Moreover, based on all these we developed an initial risk model for the estimation of the true clinical state. Our conclusions can be regarded as a preliminary work which, if confirmed on a larger database, could lead to the development of a system for the proper triage of patients with cervical intraepithelial neoaplasia.

Key words: Cervical cancer, Pap test, Screening, HPV DNA test, NASBA mRNA test, p16 test, Flow cytometry, Feature selection, Information theory, Genetic algorithms, Pattern recognition, Data mining, Statistical measures

ΠΡΟΛΟΓΟΣ

Η συγκεκριμένη διπλωματική εργασία εκπονήθηκε στο Εργαστήριο Βιοϊατρικής Τεχνολογίας της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου, σε συνεργασία με το Αττικό Νοσοκομείο και το Νοσοκομείο Ιωαννίνων.

Για την πραγματοποίησή της θα ήθελα να εκφράσω τις θερμές ευχαριστίες μου προς τον επιβλέποντα καθηγητή, κ. Δημήτριο Κουτσούρη, που μου έδωσε τη δυνατότητα να ασχοληθώ με ένα τόσο ενδιαφέρον και σύγχρονο θέμα καθώς επίσης και για τη συνολική υποστήριξή του. Τη Δρα. Μαρία Χαρίτου για τις χρήσιμες συμβουλές της και την υπομονή της κατά τη διόρθωση της εργασίας μου. Ακόμα, θα ήθελα να ευχαριστήσω τον καθηγητή Κυτταρολογίας του Ε.Κ.Π.Α, Διευθυντή του Εργαστηρίου Διαγνωστικής Κυτταρολογίας της Ιατρικής Σχολής του Ε.Κ.Π.Α., Π.Γ.Ν. «Αττικόν», κ. Πέτρο Καρακίτσο για τη βάση δεδομένων που μας παρείχε, χωρίς την οποία η παρούσα εργασία δε θα μπορούσε να είχε εκπονηθεί. Επίσης, θα ήθελα να ευχαριστήσω θερμότατα τον υποψήφιο Δρα. Παναγιώτη Μπούντρη, για την αδιάκοπη καθοδήγησή του και την καταλυτική του συνεργασία σε όλα τα προβλήματα που συνάντησα. Χωρίς την ουσιαστική αλλά και ψυχολογική του στήριξη, η εργασία αυτή δεν θα μπορούσε να είχε υλοποιηθεί τόσο αποτελεσματικά. Επίσης, ευχαριστώ τα υπόλοιπα μέλη της τριμελούς επιτροπής, κυρία Κωνσταντίνα Νικήτα και κύριο Γιώργο Ματσόπουλο, καθηγητές της Σχολής των Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών. Τέλος, θα ήθελα να ευχαριστήσω όλους εκείνους που έβαλαν έστω και ένα μικρό λιθαράκι στην περάτωση της διπλωματικής εργασίας, και ιδιαιτέρως τον Κοράκη Βασίλειο για τη διαρκή υπενθύμιση του ότι κάνει Master και τη Βραδή Μαρία-Στέλλα για τις αγχολυτικές περιγραφές της σχετικά με το εργασιακό της περιβάλλον.

Θέμα της διπλωματικής εργασίας είναι η εξόρυξη δεδομένων και αναγνώριση προτύπων με χρήση γενετικών αλγόριθμων και τεχνικών θεωρίας πληροφορίας για τη βελτιστοποιημένη ταξινόμηση περιστατικών τραχηλικής ενδοεπιθηλιακής νεοπλασίας. Στόχος της εργασίας είναι η εξαγωγή συμπερασμάτων για το ρόλο των επιμέρους τεχνικών ανίχνευσης του καρκίνου του τραχήλου της μήτρας και πιο συγκεκριμένα για τις προκαρκινικές αλλοιώσεις, δηλαδή αυτές των τραχηλικών ενδοεπιθηλιακών νεοπλασιών. Ακόμα, φιλοδοξία της εργασίας αποτελεί η εξαγωγή ενός συνόλου κανόνων για την ανίχνευση των τραχηλικών ενδοεπιθηλικακών νεοπλασιών.

Στο πρώτο κεφάλαιο γίνεται η παρουσίαση της ιατρικής πλευράς του θέματος από την οποία φαίνεται και η σημαντικότητα του προβλήματος. Στο δεύτερο κεφάλαιο γίνεται μια εισαγωγή στις τεχνικές τις οποίες χρησιμοποιήσαμε στα πλαίσια της έρευνάς μας. Στο τρίτο κεφάλαιο παρουσιάζεται εκτενώς ολόκληρη η μεθοδολογία που ακολουθήσαμε καθώς επίσης και τα συμπεράσματα στα οποία καταλήξαμε. Τέλος, στο τέταρτο κεφάλαιο, αφού έχει προηγηθεί η αποτίμηση των συμπερασμάτων, θα παρατεθούν οι προτάσεις μας για τη μελλοντική έρευνα επί του συγκεκριμένου θέματος.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Κεφάλαιο 1

Ο Καρκίνος του Τραχήλου της Μήτρας

1.1 Εισαγωγή

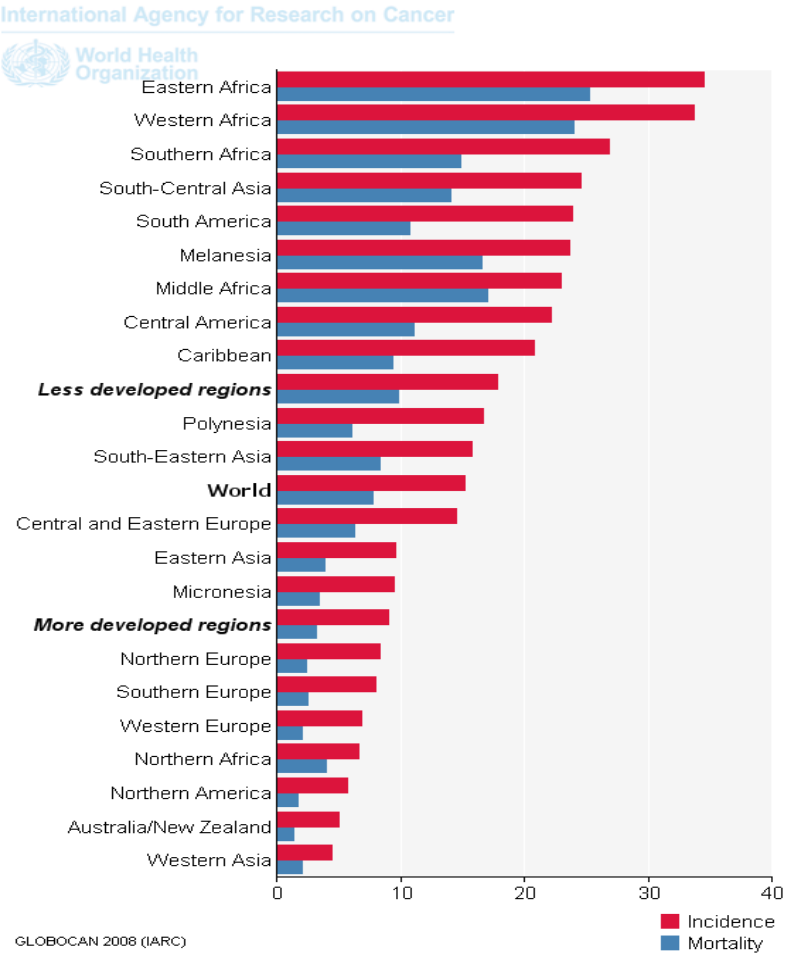
Ο καρκίνος του τραχήλου της μήτρας είναι ο όρος που περιγράφει το κακόηθες νεόπλασμα που προκαλείται στα κύτταρα του τραχήλου της μήτρας. Ένα από τα πιο κοινά συμπτώματα του καρκίνου του τραχήλου της μήτρας είναι η κολπική αιμορραγία, αλλά σε αρκετές περιπτώσεις μπορεί να μην εμφανιστεί κανένα προφανές σύμπτωμα μέχρις ότου ο καρκίνος βρεθεί σε προχωρημένο στάδιο. Η χειρουργική επέμβαση συνιστά στις περισσότερες περιπτώσεις την ενδεδειγμένη θεραπεία στα αρχικά στάδια της ασθένειας, ενώ σε πιο προχωρημένα στάδια η αντιμετώπιση της ασθένειας γίνεται με χημειοθεραπεία ή ραδιοθεραπεία.

Ο προληπτικός έλεγχος γίνεται με το τεστ Παπανικολάου ή πιο σύντομα test Pap, το οποίο εντοπίζει προκαρκινικές ή δυνητικά προκαρκινικές αλλοιώσεις στα κύτταρα του τραχήλου και του ιστού. Η θεραπεία των υψηλού βαθμού (*high-grade*) αλλοιώσεων μπορούν να προλάβουν την ανάπτυξη καρκίνου σε πολλές ασθενείς. Στις ανεπτυγμένες χώρες, η ευρέως διαδεδομένη χρήση των προγραμμάτων προληπτικού ελέγχου έχει ελαττώσει την εμφάνιση του καρκίνου του τραχήλου της μήτρας κατά 50% ή ακόμα και περισσότερο.

Η παρουσία του ιού των ανθρώπινων θηλωμάτων, ή πιο σύντομα HPV, αποτελεί έναν αναγκαίο παράγοντα για την ανάπτυξη σχεδόν όλων των περιπτώσεων του καρκίνου του τραχήλου της μήτρας. Μάλιστα, η παρουσία μερικών τύπων του HPV αποτελεί και τον μεγαλύτερο παράγοντα κινδύνου για την ανάπτυξη του καρκίνου του τραχήλου της μήτρας και ακολουθείται από το κάπνισμα. Μπορεί να μην είναι γνωστοί όλοι οι παράγοντες που οδηγούν στην ανάπτυξη του καρκίνου του τραχήλου της μήτρας, ωστόσο έχουν ήδη ενοχοποιηθεί και άλλοι παράγοντες πέρα από τον HPV που συμβάλλουν στην ανάπτυξη του.

1.1.1 Επιδημιολογία

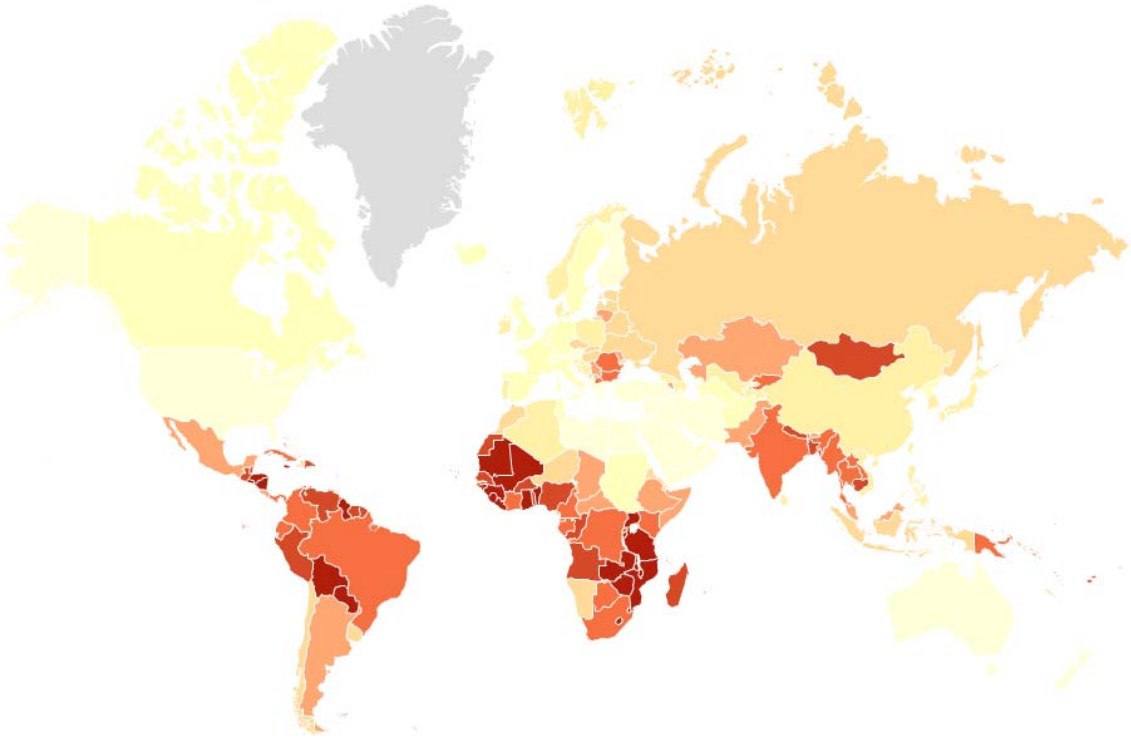
Παγκοσμίως, ο καρκίνος του τραχήλου της μήτρας είναι ο έκτος πιο θανατηφόρος καρκίνος, σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας (*WHO*) [2]. Προσβάλλει περίπου 16 γυναίκες στις 100.000 κάθε χρόνο και από τις 16 οι 9 τελικά πεθαίνουν, πράγμα που απεικονίζεται στην παρακάτω στατιστική εικόνα [3].



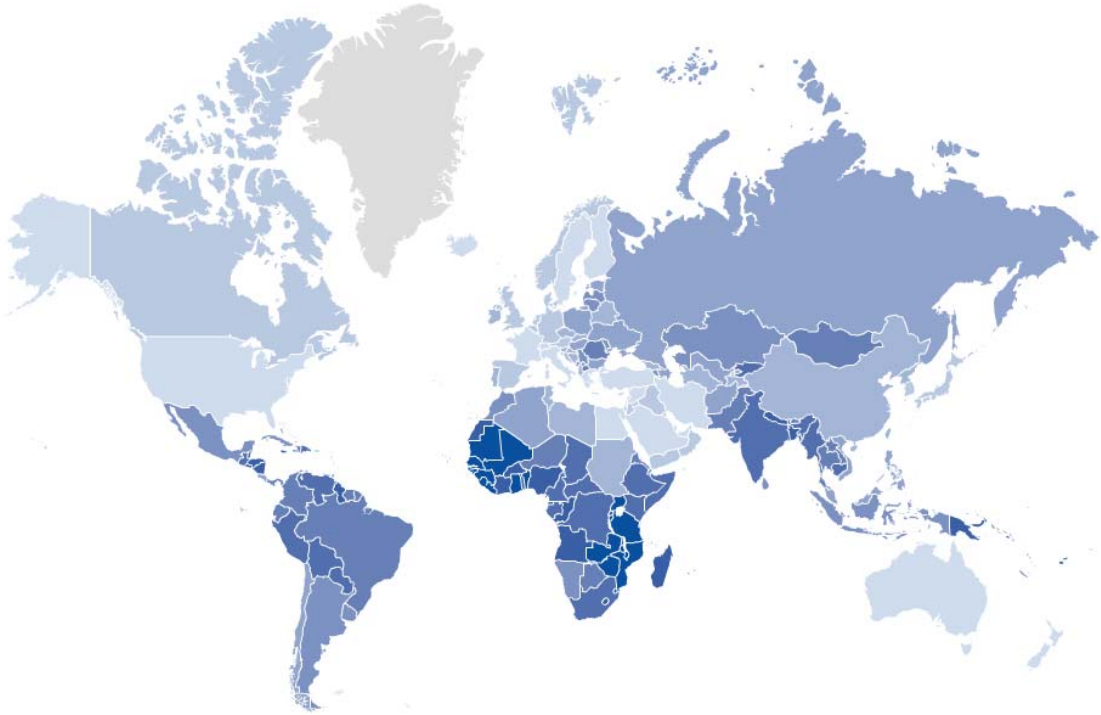
Εικόνα 1.1*: Εκτίμηση εμφάνισης περιστατικών και θνησιμότητας σε διάφορες γεωγραφικές περιοχές ανά 100.000 γυναίκες*

Από την παραπάνω εικόνα καθίσταται επίσης σαφές ότι τα ποσοστά εμφάνισης του καρκίνου και η θνησιμότητα σχετίζονται άμεσα με το βιοτικό επίπεδο, καθώς σε προηγμένες χώρες ο προληπτικός έλεγχος γίνεται με τρόπο συστηματικό και έτσι πολλές καταστάσεις προλαμβάνονται. Συνολικά, τα περιστατικά καρκίνου του τραχήλου της μήτρας ανά έτος ανέρχονται παγκοσμίως σε περίπου 530.000 εκ των οποίων τα 275.000 περίπου είναι θανατηφόρα [3].

Πιο συγκεκριμένα, η εμφάνιση περιστατικών και η θνησιμότητα ποικίλει από χώρα σε χώρα (εικόνα 1.2). Η αιτία γι αυτό, όπως αναφέρθηκε και προηγουμένως, εντοπίζεται στις κοινωνικοοικονομικές συνθήκες που επικρατούν ανά χώρα, με υψηλότερη επίπτωση σε χώρες όπου η δυνατότητα οικογενειακού προγραμματισμού, η μαιευτική και η γυναικολογική δημόσια υγεία είναι μηδαμινές και η πρόνοια για τη δημιουργία προληπτικών προγραμμάτων ελέγχου για τον καρκίνο της μήτρας δεν υφίσταται. Περιοχές όπως η Αφρική, η Κεντρική Αμερική, η Νότια Αμερική και η Καραϊβική παρουσιάζουν τη μεγαλύτερη εμφάνιση κρουσμάτων αλλά και θνησιμότητας από αυτά. Αντίθετα, σε χώρες αναπτυγμένες και ο αριθμός των κρουσμάτων και ο αντίστοιχος αριθμός θυμάτων είναι εμφανώς μικρότερος, πράγμα το οποίο οφείλεται στον συστηματικό προληπτικό έλεγχο [3].



***Εικόνα 1.2**: Παγκόσμιος χάρτης εμφάνισης κρουσμάτων καρκίνου του τραχήλου της μήτρας*



***Εικονα 1.3**: Παγκόσμιος χάρτης θνησιμότητας από καρκίνο του τραχήλου της μήτρας*

Στις Ηνωμένες Πολιτείες, ο καρκίνος του τραχήλου της μήτρας αποτελεί μόλις τον όγδοο συχνότερο καρκίνο που εμφανίζεται στις γυναίκες, ενώ μεταξύ των γυναικολογικών καρκίνων κατατάσσεται μόλις τρίτος πίσω από τον καρκίνο του ενδομητρίου και των ωοθηκών [1]. Η συχνότητα εμφάνισης και η θνησιμότητα στις ΗΠΑ είναι περίπου το ήμισυ αυτών για τον υπόλοιπο κόσμο και οφείλεται κυρίως στην επιτυχία του προληπτικού πληθυσμιακού ελέγχου με το test Pap[4]. Η συχνότητα εμφάνισης νέων κρουσμάτων στις ΗΠΑ ανερχόταν το 2008 περίπου στα 6 ανά 100.000 γυναίκες, εκ των οποίων μόλις τα 2 περίπου ήταν θανατηφόρα.

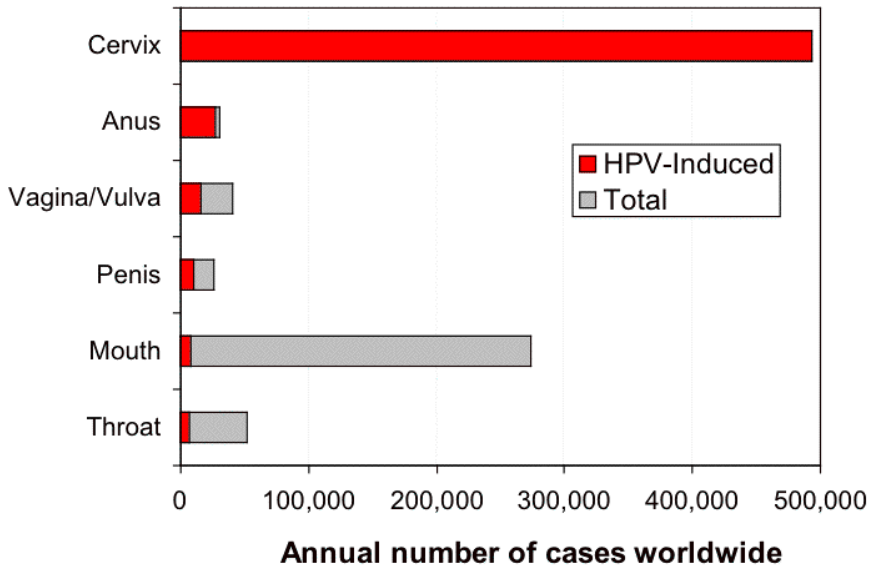
Στο Ηνωμένο Βασίλειο τα κρούσματα ανέρχονταν περίπου σε 7 ανά 100.000 το 2008 [3]. Ο καρκίνος του τραχήλου της μήτρας είναι ο δωδέκατος συχνότερος που εμφανίζεται στις γυναίκες του Ηνωμένου Βασιλείου και ο δεύτερος πιο κοινός καρκίνος στις γυναίκες ηλικίας κάτω των 35 ετών [1]. Παρόλα αυτά, η θνησιμότητα το 2008 ήταν σχετικά μικρή, καθώς ανερχόταν σε 2 ανά 100.000 γυναίκες.

Στον Καναδά, το 2008, 1.300 γυναίκες διαγνώστηκαν με καρκίνο τραχήλου και 380 πέθαναν. Στην Αυστραλία, υπήρξαν 734 περιπτώσεις καρκίνου του τραχήλου το 2005. Οι γυναίκες που διαγιγνώσκονται με καρκίνο τραχήλου μειώνονται κατά μέσο όρο 5% κάθε χρόνο μετά από το πρόγραμμα πληθυσμιακού ελέγχου που ξεκίνησε το 1991 [5,6].

1.1.2 Αιτιολογία

1.1.2.1 Συσχέτιση ιού HPV με τον καρκίνο τραχήλου μήτρας

Ο καρκίνος του τραχήλου της μήτρας, όπως αναφέρθηκε και στην εισαγωγή του παρόντος κεφαλαίου, οφείλεται κυρίως στη λοίμωξη από τον **ιό των ανθρώπινων θηλωμάτων** (***HPV***). Ο HPV ανήκει στην ευρύτερη οικογένεια των ιών των θηλωμάτων που μπορούν να μολύνουν τον άνθρωπο. Είναι υπεύθυνος για τη δημιουργία λοιμώξεων, οι οποίες μπορούν να οδηγήσουν σε καρκίνου του τραχήλου, αλλά και σε άλλα είδη καρκίνου σε άνδρες και γυναίκες, όπως φαίνεται και στην παρακάτω εικόνα [7].



***Εικόνα 1.4**: Αριθμός περιστατικών καρκίνου λόγω HPV ανά έτος παγκοσμίως*

Στις Ηνωμένες Πολιτείες παρουσιάζονται ετησίως περισσότερες από 6,2 εκατομμύρια νέες λοιμώξεις από τον HPV τόσο σε άντρες όσο και σε γυναίκες. Για το λόγο αυτό ο HPV είναι γνωστός ως το «κοινό κρυολόγημα» των σεξουαλικά μεταδιδόμενων νοσημάτων. Επηρεάζει περίπου το 80% του συνόλου των σεξουαλικά ενεργών ατόμων, είτε εμφανίζουν συμπτώματα είτε όχι. Σε νέες γυναίκες, οι περισσότερες λοιμώξεις από τον HPV είναι προσωρινές, δηλαδή δεν έχουν μεγάλη μακροπρόθεσμη σημασία, μια και ο οργανισμός απαλλάσσεται από τον ιό με την πάροδο των ετών [1]. Πιο συγκεκριμένα, το 70% των λοιμώξεων έχουν φύγει εντός ενός έτους και το 90 % εντός δύο ετών [8]. Αυτό επιτυγχάνεται είτε με την οριστική απομάκρυνση του ιού από τον οργανισμό είτε με βύθιση της τιμής της λοίμωξης σε επίπεδα μη ανιχνεύσιμα. Παρόλα αυτά ο HPV πιθανότατα θα παραμείνει στα κύτταρα του μολυσμένου ατόμου για αόριστο χρονικό διάστημα. Τις περισσότερες φορές ο ιός θα βρίσκεται σε λανθάνουσα κατάσταση, με τον κίνδυνο όμως να ελλοχεύει σε περιπτώσεις εξασθένισης του ανοσοποιητικού συστήματος, κι έτσι να οδηγηθεί το άτομο αυτό σε εμφάνιση συμπτωμάτων [9].

Ο HPV συνδέεται με την ανάπτυξη καρκίνου του τραχήλου εξαιτίας των αλλοιώσεων που αυτός προκαλεί στα τραχηλικά κύτταρα. Όταν η λοίμωξη εξακολουθεί να είναι παρούσα, πράγμα που συμβαίνει περίπου σε ένα ποσοστό 5-10% των αρχικά μολυσμένων γυναικών, τότε υπάρχει μεγάλος κίνδυνος ανάπτυξης **τραχηλικής ενδοεπιθηλιακής νεοπλασίας** (*cervical intraepithelial neoplasia ή εν συντομία CIN*). Η τραχηλική ενδοεπιθηλιακή νεοπλασία αποτελεί ένα προκαρκινικό στάδιο και μπορεί να οδηγήσει σε διηθητικό καρκίνο [1]. Η διαδικασία αυτή έχει μεγάλη χρονική διάρκεια, περίπου 15-20 έτη, οπότε παρέχονται πολλές ευκαιρίες για την ανίχνευση και αντίστοιχη αντιμετώπιση των προκαρκινικών αλλοιώσεων που μπορεί να έχουν αναπτυχθεί [9]. Οι γυναίκες που έχουν πολλούς

σεξουαλικούς συντρόφους ή ακόμα και αν έχουν σεξουαλικές σχέσεις με άνδρες που είχαν πολλές συντρόφους διατρέχουν μεγαλύτερο κίνδυνο [10,11].

Περισσότεροι από 150 τύποι του HPV έχουν αναγνωρισθεί ως σήμερα, αν και κάποιοι αναφέρουν ως και πάνω από 200 υποτύπους [12,13]. Από αυτούς, δεκαπέντε χαρακτηρίζονται ως **υψηλού κινδύνου** (*16,18,31,33,35,39,45,51,52,58,59,68,73,82*), τρεις ως **πιθανώς υψηλού κινδύνου** (*26,53,66*) και δώδεκα ως **χαμηλού κινδύνου** (*6,11,40,42,43,44,54,61,70,72,81*) [14]. Εμμένουσα λοίμωξη για 10 ή περισσότερα χρόνια με υψηλού κινδύνου τύπους του HPV μπορεί να οδηγήσει σε προκαρκινικές αλλοιώσεις του τραχήλου ή ακόμα και σε διηθητικό καρκίνο [15]. Οι περισσότερες περιπτώσεις καρκίνου του τραχήλου, περίπου το 70%, οφείλεται στους τύπους 16 και 18. Οι παραπάνω δύο τύποι μαζί με τον τύπο 31 αποτελούν τους πρωταρχικούς παράγοντες κινδύνου για την ανάπτυξη καρκίνου του τραχήλου της μήτρας [16]. Παρόλα αυτά ο καρκίνος μπορεί να προκληθεί και από τύπους χαμηλού κινδύνου [1].

Το ιατρικώς αποδεκτό πρότυπο, που έχει εγκριθεί από την Αμερικανική Εταιρεία για τον Καρκίνο (*American Cancer Society*), είναι ότι η ασθενής πρέπει να έχει μολυνθεί από τον HPV για να αναπτύξει καρκίνο τραχήλου. Συνεπώς, θεωρείται ως μια σεξουαλικά μεταδιδόμενη ασθένεια. Παρ' όλα αυτά, οι περισσότερες γυναίκες που έχουν μολυνθεί με υψηλού κινδύνου HPV δεν αναπτύσσουν καρκίνο του τραχήλου της μήτρας [17]. Η χρήση των προφυλακτικών μειώνει αλλά δεν εμποδίζει πάντα τη μετάδοση. Ομοίως, ο HPV μπορεί να μεταδοθεί με την επαφή του δέρματος με μολυσμένες περιοχές.

Στους **άνδρες** δεν υπάρχει διαθέσιμο τεστ ελέγχου του HPV αν και πιστεύεται ότι ο HPV αναπτύσσεται επιλεκτικά στη βάλανο του πέους. Ως ένα είδος πρόληψης συνιστάται ο καθαρισμός της περιοχής αυτής [1]. Αξίζει να σημειωθεί ότι υπάρχουν κάποιες μελέτες [18,19] οι οποίες υποστηρίζουν ότι οι άνδρες που έχουν κάνει περιτομή έχουν λιγότερες πιθανότητες να μολυνθούν από τον HPV. Όμως, δεν υπάρχει κάποια αδιαπραγμάτευτη απόδειξη γι αυτό και δεν έχει αναγνωριστεί από τις διεθνείς αντικαρκινικές οργανώσεις.

1.1.2.2 Κύκλος ζωής και δράση του HPV

Για να γίνει κατανοητή η δράση του ιού είναι απαραίτητη η γνώση του **κύκλου ζωής του HPV**. Η λοίμωξη από τον HPV περιορίζεται στα βασικά κύτταρα (*basal cells*) του στρωματοποιημένου επιθηλίου που είναι και ο μοναδικός ιστός στον οποίο αναπαράγονται τα βασικά κύτταρα. Τα κύτταρα των επιθηλιακών ιστών μολύνονται από τον ιό συνήθως στη ζώνη μετάπλασης μέσω μικρών εκδορών ή άλλων επιθηλιακών τραυμάτων που εκθέτουν τομείς της βασικής μεμβράνης [20]. Το γονιδίωμα του ιού κατόπιν μεταφέρεται στον πυρήνα του ξενιστή και αντιγράφεται μαζί με το υπόλοιπο υγειές γονιδίωμα του ξενιστή. Μεταγράφεται και μεταφράζεται κανονικά κατά τη διαδικασία της κυτταρικής διαίρεσης του ξενιστή και διαφοροποιείται στα πάνω στρώματα του επιθηλίου. Με τον τρόπο αυτό οι

υψηλού κινδύνου τύποι του HPV έχουν την ικανότητα να διαιωνίζουν τα βασικά κύτταρα, επεκτείνοντας τη διάρκεια ζωής τους. Για την έναρξη της μεταγραφής απαιτούνται 12-24 ώρες. Τα εμπλεκόμενα αντισώματα παίζουν σημαντικό ρόλο στην εξουδετέρωση, ωστόσο οι ιϊκοί παράγοντες συνεχίζουν και παραμένουν στη βασική μεμβράνη και στην επιφάνεια των κυττάρων [9].

Οι πρωτεΐνες του HPV που συνδέονται με την ανάπτυξη καρκίνου είναι οι E6 και E7, καθώς είναι υπεύθυνες για την αναστολή της λειτουργίας των ογκοκατασταλτικών γονιδίων που εμπεριέχονται στα μεταβολικά μονοπάτια των πρωτεϊνών p53 [21] και pRb (*ρετινοβλάστωμα*) [22]. Το γονιδίωμα του HPV αποτελείται από έξι πρώιμες πρωτεΐνες (*early proteins*) που παράγονται πριν την αντιγραφή του ιού και δύο όψιμες πρωτεΐνες (*late proteins*) που παράγονται μετά την αντιγραφή του ιού (L1 και L2) [23].

Με το που προσβληθεί το κύτταρο-ξενιστής, εκφράζονται οι **E1 και E2 πρωτεΐνες**. Τα υψηλά επίπεδα της E2 είναι αυτά που αρχικά καταστέλλουν την παραγωγή των E6/E7 πρωτεϊνών. Όμως με την ενσωμάτωση του γονιδιώματος του HPV στον ξενιστή αναστέλλεται η δράση της E2. Αυτό έχει σαν αποτέλεσμα την αναστολή της καταστολής των E6/E7 πρωτεϊνών. Οι E6/E7 οδηγούν σε αδρανοποίηση δύο άλλες βασικές ογκοκατασταλτικές πρωτεΐνες. Πιο συγκεκριμένα, η E6 αναστέλλει την p53 και η E7 την pRb [24]. Τα ογκογονίδια E6 και E7 του ιού ευθύνονται για την τροποποίηση του κυτταρικού κύκλου έτσι ώστε το κύτταρο-ξενιστής να διατηρηθεί σε μια κατάσταση ευνοϊκή για την αναπαραγωγή του ιϊκου γονιδιώματος [9].

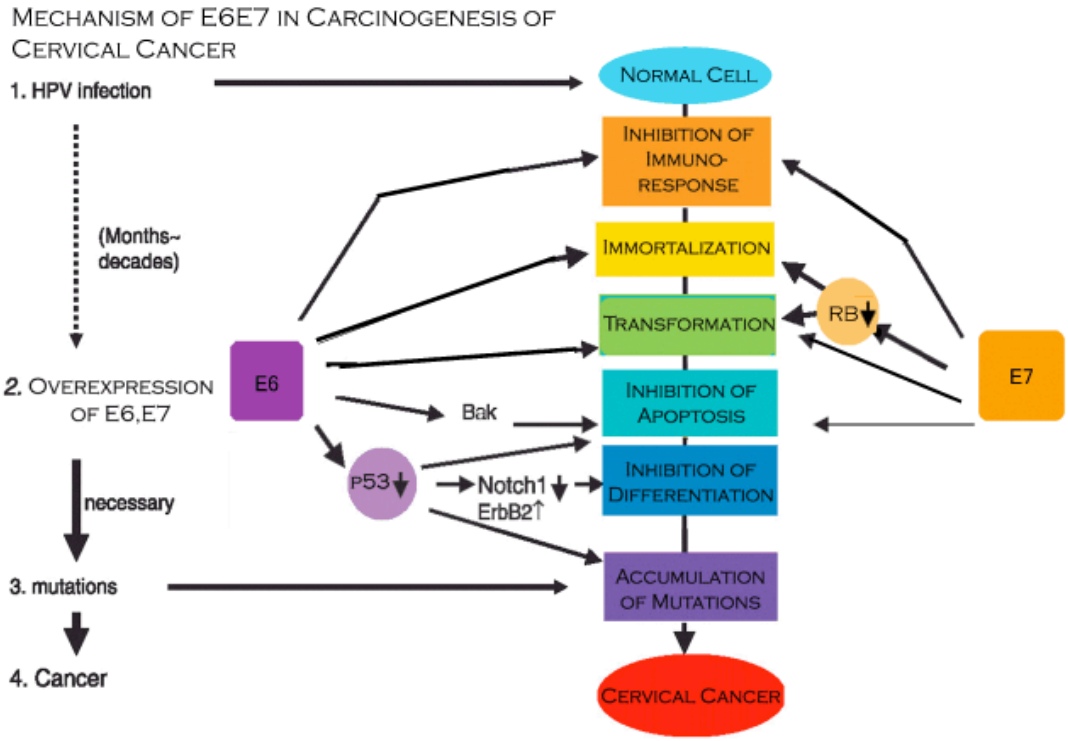
Η **E6 πρωτεΐνη** του HPV σε συνεργασία με την αντίστοιχη πρωτεΐνη του κυττάρου-ξενιστή (που σχετίζεται αποκλειστικά με την E6) επεμβαίνει ρυθμιστικά στην ογκοκατασταλτική p53 πρωτεΐνη και οδηγεί στην πρωτεοσωμική αποσύνθεσή της. Σε ένα φυσιολογικό κύτταρο (που δεν έχει μολυνθεί από τον HPV), η πρωτεΐνη p53 αποτρέπει την κυτταρική ανάπτυξη και οδηγεί το κύτταρο σε απόπτωση αν διαπιστώσει βλάβη στο DNA. Όταν όμως υπάρχει η E6 πρωτεΐνη του HPV, η p53 δεν μπορεί να εμποδίσει τον ανεξέλεγκτο πολλαπλασιασμό των κυττάρων και ευνοείται η περαιτέρω αντιγραφή, μεταγραφή και μετάφραση του HPV στα μολυσμένα κύτταρα [9,25].

Η **E7 πρωτεΐνη** του HPV ενεργεί ως πρωταρχική πρωτεΐνη μετασχηματισμού (*transforming protein*). Όταν δεν υπάρχει η πρωτεΐνη E7 (σε φυσιολογικό κύτταρο που δεν έχει μολυνθεί), τότε το ρετινοβλάστωμα pRb συνδέεται με το μεταγραφικό παράγοντα E2F και τον αποτρέπει από το να προχωρήσει τον κυτταρικό κύκλο στις επόμενες φάσεις του που περιλαμβάνουν τον πολλαπλασιασμό του κυττάρου, μέχρι το κύτταρο να είναι όντως έτοιμο να διαιρεθεί. Στην περίπτωση όμως που υπάρχει η E7 πρωτεΐνη του ιού HPV, τότε αυτή συνδέεται με την πρωτεΐνη ρετινοβλάστωμα (pRb) και δεν επιτρέπει στο μεταγραφικό παράγοντα E2F να συνδεθεί με την pRb. Έτσι, απελευθερώνεται ο μεταγραφικός παράγοντας E2F, ο οποίος στη συνέχεια επιτρέπει τον ανεξέλεγκτο πολλαπλασιασμό των μολυσμένων κυττάρων [9,25].

Στα ανώτερα στρώματα του επιθηλίου, τα όψιμα γονίδια **L1/L2** μεταγράφονται και μεταφράζονται σε δομικές πρωτεΐνες που ενθυλακώνουν το γονιδίωμα του ιού σε ένα καψίδιο για να το προστατεύσουν. Μόλις το γονιδίωμα του ιού ενθυλακωθεί, το καψίδιο

υποβάλλεται σε μια ωρίμανση χημικής οξειδοαναγωγής που σταθεροποιεί τα ιϊκά σωματίδια και αυξάνει την εξειδικευμένη μολυσματικότητά τους [26].

Όταν ένας HPV εισέρχεται σε ένα κύτταρο, το μολύνει και μπορεί να μεταδοθεί. Μπορεί να χρειαστεί να περάσουν αρκετοί μήνες ή και χρόνια για να αναπτυχθούν και να ανιχνευθούν **πλακώδεις ενδοεπιθηλιακές αλλοιώσεις (Squamous Intraepithelial Lesions ή SIL)**. Ο χρόνος (*latency period*) που μεσολαβεί από την πραγματική λοίμωξη μέχρι την κλινική ανίχνευση της ασθένειας καθιστά δύσκολο για το άτομο που έχει μολυνθεί να καθορίσει ποιος ή ποια σύντροφος ήταν η πηγή της μόλυνσης [9].



***Εικόνα 1.5**: Ρύθμιση κυτταρικού κύκλου μέσω ρυθμιστικών πρωτεϊνών σε φυσιολογικό κύτταρο*

1.1.2.3 Αντίδραση ανοσοποιητικού συστήματος στον ιό HPV

Οι περισσότερες των λοιμώξεων που προκαλούνται από τον HPV είναι παροδικές και κλινικά μη ανιχνεύσιμες, πράγμα το οποίο υποδεικνύει ότι το ανοσολογικό σύστημα του ασθενούς έχει την ικανότητα ελέγχου και αντιμετώπισης της λοίμωξης. Παρά το γεγονός όμως ότι η λοίμωξη προκαλεί ανοσολογική αντίδραση, ο HPV είναι γενικά ένας δύσκολος στόχος για το ανοσοποιητικό σύστημα και η αντίδρασή του σε γενικές γραμμές είναι

αδύναμη. Είναι αρκετοί οι μηχανισμοί που εμπλέκονται και συμβάλλουν στην ανοσολογική διαφυγή του HPV. Οι σημαντικότεροι είναι οι παρακάτω τρεις [27]:

1. Ο HPV δεν προκαλεί λύση κυττάρων, αντίθετα για παράδειγμα με τον ιό του έρπη. Προκαλεί κυτταρικό πολλαπλασιασμό παρά κυτταρική καταστροφή και για το λόγο αυτό δεν προκαλεί κάποια φλεγμονώδη αντίδραση.
2. Το γεγονός ότι ο HPV επηρεάζει μόνο τα επιθηλιακά κύτταρα διευκολύνει τη διαφυγή του από το ανοσοποιητικό σύστημα. Μόνο σε τελικώς διαφοροποιημένα πλακώδη κύτταρα βρίσκονται ολόκληρα σωματίδια του ιού. Τα τελικώς διαφοροποιημένα αυτά κύτταρα βρίσκονται αρκετά μακριά από τα λεμφοκυτταρικά βλαστικά κέντρα στον υποβλεννογόνιο που σχετίζονται με την πρόκληση ανοσολογικής αντίδρασης.
3. Ο HPV διαφεύγει της αναγνώρισης από το ενδογενές αμυντικό σύστημα μπλοκάροντας κατά αυτόν τον τρόπο την παραγωγή της ιντερφερόνης για τη διασφάλιση της αναπαραγωγής του. Αυτό επιτυγχάνεται μέσω της παραγωγής των πρωτεϊνών E6/E7 οι οποίες δεσμεύουν και αδρανοποιούν ενώσεις στο κύκλωμα της ιντερφερόνης.

1.1.2.4 Άλλοι παράγοντες κινδύνου για εμφάνιση καρκίνου τραχήλου

Εκτός από τον ιό των ανθρώπινων θηλωμάτων, η Αμερικανική Εταιρεία για τον Καρκίνο προβλέπει τον ακόλουθο κατάλογο των **συμπαραγόντων κινδύνου** για τον καρκίνο του τραχήλου: κάπνισμα, λοίμωξη από τον ιό της ανθρώπινης ανοσοαναπάρκειας HIV, μόλυνση από χλαμύδια, διαταραχές σχετιζόμενες με το άγχος, παράγοντες δίαιτας, ορμονική αντισύλληψη, πολλαπλές κυήσεις, έκθεση στο ορμονικό φάρμακο diethylstilbestrol (*DES*) καθώς και οικογενειακό ιστορικό καρκίνου του τραχήλου της μήτρας [28].

1.1.3 Ταξινόμηση προκαρκινικών αλλοιώσεων τραχήλου μήτρας

Η ονομασία και η **ταξινόμηση** των προκαρκινικών αλλοιώσεων του τραχήλου της μήτρας έχει αλλάξει πολλές φορές κατά τον 20ό αιώνα. Το σύστημα ταξινόμησης του Παγκόσμιου Οργανισμού Υγείας [29,30] ήταν περιγραφικό της αλλοίωσης, χαρακτηρίζοντάς την ως **ήπια, μέτρια ή σοβαρή δυσπλασία** (ανάλογα με το πάχος του τραχηλικού επιθηλίου που καταλαμβάνεται από νεοπλασματικά κύτταρα) **ή καρκίνωμα in situ-CIS** (αλλοιώσεις στις οποίες όλο το πάχος του επιθηλίου έχει αντικατασταθεί από αδιαφοροποίητα νεοπλασματικά κύτταρα).

Η **τραχηλική ενδοεπιθηλιακή νεοπλασία (*Cervical Intraepithelial Neoplasia ή CIN*)** αναπτύχθηκε ως όρος αργότερα με σκοπό να δώσει έμφαση στο φάσμα των ανωμαλιών και να βοηθήσει στην τυποποίηση της θεραπείας [30]. Πιο συγκεκριμένα, η ήπια δυσπλασία κατατάσσεται ως **CIN-1**, η μέτρια δυσπλασία ως **CIN-2** και τέλος η σοβαρή δυσπλασία και το καρκίνωμα in-situ ως **CIN-3**. Πιο πρόσφατα, οι αλλοιώσεις CIN-2 και CIN-3 έχουν συνδυαστεί και περιγράφονται από τον όρο CIN-2/3 [1]. Ο καρκίνος περνά συνήθως από όλα τα στάδια αλλοιώσεων (CIN-1,2,3), παρότι υπάρχουν αναφορές για καρκίνους που αναπτύχθηκαν κατευθείαν από αλλοίωση τύπου CIN-1. Στην παρακάτω εικόνα φαίνεται ένα παράδειγμα των σταδίων που περνούν οι αλλοιώσεις από ένα φυσιολογικό επιθήλιο έως τον διηθητικό καρκίνο, μέσω των ενδιάμεσων σταδίων CIN-1,2,3.

Εικόνα 1.6: *Στάδια ανάπτυξης πλακώδους καρκίνου τραχήλου σε αρχικά φυσιολογικό επιθήλιο*

1.1.4 Σταδιοποίηση και ιστολογικοί τύποι καρκίνου τραχήλου μήτρας

Ο καρκίνος του τραχήλου της μήτρας θεωρείται μια πολυδιάστατη νόσος. Από την Παγκόσμια Ομοσπονδία Μαιευτικής και Γυναικολογίας (*FIGO*) έχουν περιγραφεί πέντε κλινικά στάδια [27] ως ακολούθως:

- Τα στάδια 0 και I αποτελούν τα λεγόμενα προκλινικά στάδια του καρκίνου του τραχήλου της μήτρας. Ο προληπτικός έλεγχος παίζει πολύ σημαντικό ρόλο στην έγκαιρη διάγνωση κυρίως λόγω του γεγονότος ότι γυναίκες με καρκίνο σταδίου 0 και I δεν εμφανίζουν συμπτώματα και δεν μπορεί να εντοπιστεί με γυμνό μάτι.
- Τα στάδια II,III,IV σχετίζονται με διάφορα συμπτώματα, όπως για παράδειγμα η μετεμμηνοπαυσική αιμορραγία, και αποτελούν τη διηθητική φάση.

Η σταδιοποίηση αυτή της FIGO χρησιμοποιείται ευρέως από τους κλινικούς για το σχεδιασμό της πρέπουσας θεραπείας για την κάθε ασθενή [27].

Ο Παγκόσμιος Οργανισμός Υγείας (*WHO*) αναγνωρίζει **δύο κύριους ιστολογικούς τύπους διηθητικού καρκίνου**: τα πλακώδη καρκινώματα και τα αδενοκαρκινώματα.

Τα πλακώδη καρκινώματα, που είναι και συχνότερα (περίπου το 85% των περιπτώσεων), κατηγοριοποιούνται σε κερατινοποιούμενα και μη κερατινοποιούμενα. Τα κερατινοποιούμενα μπορεί να παρουσιάζουν υψηλή ή μέτρια διαφοροποίηση και αποτελούνται από μεγάλα καρκινικά κύτταρα. Αντίθετα τα μη κερατινοποιούμενα (χαμηλής διαφοροποίησης) μπορεί να αποτελούνται είτε από μεγάλου είτε από μικρού τύπου κύτταρα [27].

Τα αδενοκαρκινώματα είναι λιγότερο συχνά, αντιπροσωπεύουν περίπου το 10-12%. Παρότι κάθε τύπος αποτελεί ξεχωριστή ιστολογική οντότητα, δεν είναι καθόλου σπάνια η παρουσία περισσοτέρων του ενός τύπων αδενοκαρκινώματος σε έναν όγκο. Η συχνή συνύπαρξη αδενικού και πλακώδους καρκινώματος υπονοεί την πιθανή κοινή προέλευση από τα εφεδρικά κύτταρα του τραχήλου και κοινή αιτιοπαθογένεια. Ο πιο συχνός τύπος αδενοκαρκινώματος στον τράχηλο είναι το βλεννώδες ενδοτραχηλικό αδενοκαρκίνωμα. Αναγνωρίζονται τρεις βαθμοί ενδοτραχηλικού καρκινώματος –υψηλής, μέτριας και χαμηλής διαφοροποίησης– που εξαρτώνται από το βαθμό ομοιότητας του καρκινικού κυττάρου με το φυσιολογικό αδενικό κύτταρο του ενδοτραχηλικού επιθηλίου [27].

Άλλοι τύποι καρκινώματος (αδενοπλακώδες, αδενοκυστικό, μεταστατικό) αποτελούν το υπόλοιπο 3-5% όλων των περιπτώσεων [27].

1.1.5 Συμπτώματα

Στα πρώτα στάδια του καρκίνου του τραχήλου της μήτρας δεν υπάρχει εμφάνιση συμπτωμάτων. Η παρουσία κακοήθειας μπορεί να υποδειχθεί από κολπική αιμορραγία, αιμορραγία επαφής ή πιο σπάνια από μία κολπική μάζα. Ακόμα, συμπτώματα που μπορούν να παρατηρηθούν είναι ο μέτριος πόνος κατά τη σεξουαλική επαφή ή κάποιο κολπικό έκκριμα. Σε προχωρημένο στάδιο της ασθένειας μπορεί να παρατηρηθούν μεταστάσεις στην κοιλιά, στους πνεύμονες ή και αλλού. Τα συμπτώματα του προχωρημένου καρκίνου του τραχήλου μπορεί να περιλαμβάνουν: απώλεια όρεξης, απώλεια βάρους, κόπωση, πυελικό άλγος, οσφυαλγία, πόνο στα κάτω άκρα, πρησμένο πόδι, βαριά αιμορραγία από τον κόλπο, απώλεια ούρων [31] και κατάγματα οστών [1].

1.1.6 Διάγνωση

Η έγκαιρη και έγκυρη διάγνωση του καρκίνου του τραχήλου της μήτρας είναι πολύ σημαντική. Όσο νωρίτερα ανιχνευθούν οι τραχηλικές αλλοιώσεις τόσο μεγαλύτερες είναι οι πιθανότητες επιτυχούς θεραπείας. Διαφορετικά, αν ο καρκίνος αναπτυχθεί σε προχωρημένα στάδια τότε η αντιμετώπισή του γίνεται ολοένα και πιο δύσκολη.

Είναι επίσης πολύ σημαντικό σε αυτό το σημείο να καταστούν κατανοητές και να διαχωριστούν οι έννοιες της **διάγνωσης (diagnosis)** και του **προληπτικού πληθυσμιακού ελέγχου (screening)**. Στη διάγνωση περιλαμβάνονται κλασσικές μέθοδοι όπως η οπτική μακροσκοπική εξέταση του τραχήλου και η βιοψία μέσω κολποσκόπησης [27]. Αντίθετα, ο προληπτικός πληθυσμιακός έλεγχος είναι η συστηματική παρακολούθηση των γυναικών ως ένα είδος δευτερογενούς πρόληψης της ασθένειας.

Η **ιατρική διαλογή (triage)** είναι ακόμα μία έννοια που θα πρέπει να τονιστεί και να γίνει κατανοητή. Το triage συνδέεται με την ξεχωριστή αντιμετώπιση του κάθε περιστατικού ανάλογα με τα αποτελέσματα των ιατρικών εξετάσεων αλλά και με τα ξεχωριστά προσωπικά στοιχεία της καθε ασθενούς. Στα προσωπικά στοιχεία περιλαμβάνονται πληροφορίες όπως: ηλικία, οικογενειακή κατάσταση, ιατρικό ιστορικό, ψυχολογικοί παράγοντες κ.ά.. Με βάση τα παραπάνω σχηματίζεται ένας προσωπικός ιατρικός φάκελος για κάθε γυναίκα, ο οποίος σε συνδυασμό με τα αποτελέσματα των ιατρικών εξετάσεων ανίχνευσης αλλοιώσεων δίνει στο γιατρό τη δυνατότητα να πάρει την απόφαση για το τι πρέπει να πράξει και ποια είναι η διαδικασία που πρέπει να ακολουθηθεί μελλοντικά. Σύγχρονες έρευνες προσανατολίζονται στην εύρεση πρωτοκόλλων για το triage τραχηλικών αλλοιώσεων, ιδιαίτερα για τις αλλοιώσεις υψηλού κινδύνου (HgSIL), πράγμα στο οποίο προσπαθεί να συνδράμει και η συγκεκριμένη εργασία.

1.1.7 Πρόληψη

Η **πρωτογενής πρόληψη** από τον HPV είναι πολύ σημαντική και μπορεί να αποτρέψει τη μελλοντική λοίμωξη μιας γυναίκας από αυτόν. Περιλαμβάνει διάφορους τρόπους (που αναλύονται εδώ) όπως εμβολιασμός, χρήση προφυλακτικού, αποφυγή καπνίσματος και κατανάλωση φρούτων και λαχανικών. Η **δευτερογενής πρόληψη** αφορά τον συστηματικό προληπτικό πληθυσμιακό έλεγχο (*screening*) των γυναικών και αναπτύσσεται λόγω σημαντικότητας εκτενέστερα στο υποκεφάλαιο 1.2.

1.1.8 Θεραπεία

Η κλινική σταδιοποίηση του διηθητικού καρκίνου του τραχήλου της μήτρας είναι σημαντική καθώς καθορίζει την θεραπευτική αγωγή της ασθενούς. Καρκίνος που έχει εντοπιστεί στο στάδιο 0 (**προδιηθητικό στάδιο**) απαιτεί μόνο τοπική εξαίρεση του καρκινικού επιθηλίου με διαθερμία ή λέιζερ [27].

Οι καρκίνοι **σταδίου Ι** συνήθως θεραπεύονται με ριζική υστερεκτομή με ή χωρίς διατήρηση των ωοθηκών, απόφαση που εξαρτάται από την ηλικία της ασθενούς. Εντούτοις, σε γυναίκες με πρώιμο στάδιο καρκίνου τραχήλου, η τραχηλεκτομή μπορεί να χρησιμοποιηθεί ως θεραπεία διατήρησης της γονιμότητας. Η ριζική τραχηλεκτομή είναι μία χειρουργική επέμβαση κατά την οποία ο τράχηλος, το ανώτερο τμήμα του κόλπου, ο παραμητρικός ιστός (ιστός στο χαμηλότερο τμήμα της μήτρας) και οι πυελικοί λεμφαδένες εξαιρούνται. Το σώμα της μήτρας και οι ωοθήκες δεν εξαιρούνται και έτσι η δυνατότητα απόκτησης παιδιών παραμένει. Αυτό γίνεται σε πρώιμο στάδιο του καρκίνου. Σκοπός είναι η διατήρηση της γονιμότητας. Αυτή η θεραπεία έχει αναπτυχθεί τα τελευταία χρόνια από ογκολόγους γυναικολογίας σε ειδικευμένα κέντρα σε όλο τον κόσμο. Γίνεται διακολπικά και μέσω μικρών χειρουργικών τομών στην κοιλιά, χρησιμοποιώντας λαπαροσκόπιο [1,27].

Η καθιερωμένη θεραπεία για καρκίνο του τραχήλου **προχωρημένου σταδίου**, είναι η ριζική υστερεκτομή ή/και η πυελική ακτινοθεραπεία. Περισσότερο εξαπλωμένος καρκίνος ίσως απαιτεί ακτινοθεραπεία ή/και χημειοθεραπεία. Το κλινικό στάδιο τη στιγμή της διάγνωσης επηρεάζει το προσδόκιμο επιβίωσης της ασθενούς [1,27].

Η **πρόγνωση** (πρόβλεψη εξέλιξης θεραπείας και προσδόκιμο επιβίωσης) είναι πολύ καλή σε γυναίκες με καρκίνο σε πρώιμο στάδιο (Στάδιο 0 ή Στάδιο Ι) αλλά είναι φτωχή σε γυναίκες με προχωρημένο καρκίνο. Μάλιστα, η πρόγνωση σε ασθενείς σε Στάδιο 0 (προδιηθητικό στάδιο) είναι εξαιρετική με πολύ μικρό κίνδυνο υποτροπής (< 1%). Υπάρχουν κάποιες ενδείξεις ότι ο καρκίνος του τραχήλου της μήτρας είναι περισσότερο επιθετικός σε νεαρές γυναίκες παρά σε γυναίκες μεγαλύτερες των 40 ετών [27].

1.2 Προληπτικός πληθυσμιακός έλεγχος και ανίχνευση καρκίνου του τραχήλου της μήτρας

Η ανίχνευση των προκαρκινικών αλλοιώσεων του τραχήλου της μήτρας καθώς και η διάγνωση του καρκίνου πρέπει να είναι τόσο έγκαιρες όσο και έγκυρες. Το κατά πόσο σημαντική είναι η έγκαιρη ανίχνευση φαίνεται από το γεγονός ότι όσο νωρίτερα ανιχνευθούν οι τραχηλικές αλλοιώσεις τόσο οι πιθανότητες επιτυχούς αντιμετώπισης αυξάνουν. Διαφορετικά, αν ο καρκίνος αναπτυχθεί σε προχωρημένα στάδια τότε η αντιμετώπισή του γίνεται ολοένα και πιο δύσκολη. Για το λόγο αυτό, ο **συστηματικός προληπτικός έλεγχος (*screening*)** των γυναικών είναι πολύ διαδεδομένος τα τελευταία χρόνια σε όλο τον κόσμο. Πρόκειται για τη συστηματική παρακολούθηση όλων των γυναικών ως ένα είδος δευτερογενούς πρόληψης της ασθένειας. Έτσι, υπάρχει η δυνατότητα έγκαιρης ανίχνευσης των αλλοιώσεων πριν καν παρουσιαστούν τα συμπτώματα της νόσου.

Τα screening tests αυτά δεν είναι τέλεια, όπως και κάθε ιατρική εξέταση άλλωστε. Τα αποτελέσματα αυτών μπορεί να είναι εσφαλμένα θετικά για υγιείς ασθενείς (*ψευδή θετικά ή false positives*) ή εσφαλμένα αρνητικά για άτομα που έχουν τη νόσο (*ψευδή αρνητικά ή false negatives*). Μερικά από τα βασικά μειονεκτήματα των screening tests είναι [32]:

1. κόστος και χρήση ιατρικών πόρων που δεν είναι αναγκαία στην πλειονότητα των περιπτώσεων
2. δυσμενείς επιπτώσεις στο άτομο που υποβάλλεται σε αυτά (π.χ. άγχος, δυσφορία)
3. άγχος που προκαλεί ένα ψευδές θετικό αποτέλεσμα εξέτασης
4. περιττή διερεύνηση και αντιμετώπιση των ψευδώς θετικών αποτελεσμάτων
5. άγχος που προκαλείται από την παράταση της γνώσης της ύπαρξης ή μη μιας ασθένειας χωρίς καμία βελτίωση στην έκβασή της
6. λανθασμένη αίσθηση ασφάλειας που προκαλεί ένα ψευδές αρνητικό αποτέλεσμα το οποίο πιθανώς θα οδηγήσει σε καθυστέρηση της τελικής διάγνωσης

Η πιο διαδεδομένη μέθοδος που χρησιμοποιείται ακόμα και σήμερα για την ανίχνευση αλλοιώσεων στον τράχηλο της μήτρας είναι το τεστ Pap. Σε πρόσφατες μελέτες έχει αποδειχθεί ότι υπάρχουν τύποι του HPV που είναι ογκογόνοι. Για αυτό το λόγο τα τελευταία χρόνια έχουν αναπτυχθεί νέες τεχνικές για τον εντοπισμό των συγκεκριμένων τύπων υψηλού κινδύνου, με σκοπό την υποβοήθηση των γιατρών στη λήψη έγκυρης απόφασης, δηλαδή στη διενέργεια ορθής ιατρικής διαλογής (*triage*).

1.2.1 PAP test

Το Τεστ Παπανικολάου (ή πιο σύντομα test Pap) είναι ένα τεστ προσυμπτωματικού ελέγχου το οποίο χρησιμοποιείται στη γυναικολογία με σκοπό την ανίχνευση προκαρκινικών και καρκινικών (κακοηθών) διαδικασιών στο ενδοτραχηλικό κανάλι (ή αλλιώς ζώνη μετασχηματισμού). Οι αλλοιώσεις μπορούν να αντιμετωπιστούν, προλαμβάνοντας τον καρκίνο του τραχήλου της μήτρας. Το τεστ αυτό εισήχθη από τον Γεώργιο Παπανικολάου, έναν Έλληνα γιατρό από τον οποίο πήρε και το όνομά του [33]. Το 1926 ανέφερε ότι καρκινικά κύτταρα μπορούσαν να παρατηρηθούν μέσα στις κολπικές εκκρίσεις από γυναίκες με καρκίνο του τραχήλου της μήτρας (Παπανικολάου 1928). Το 1944 δημοσίευσε το άρθρο «Διάγνωση του καρκίνου του τραχήλου της μήτρας από το κολπικό επίχρισμα» (Παπανικολάου & Traut 1943). Οι παρατηρήσεις του σύντομα επιβεβαιώθηκαν και από άλλους (Ayre 1944, Meigs et al 1945). Η πρώτη κλινική προληπτικού ελέγχου του καρκίνου του τραχήλου της μήτρας λειτούργησε στη Μασαχουσέτη το 1945 [34].

Το τεστ Παπανικολάου χρησιμοποιείται παγκοσμίως ως προληπτική εξέταση για την ανίχνευση του προδιηθητικού και του μικροδιηθητικού καρκίνου του τραχήλου της μήτρας. Το τεστ περιλαμβάνει απομάκρυνση ενός δείγματος επιθηλιακών κυττάρων από την επιφάνεια του τραχήλου χρησιμοποιώντας σπάτουλα ή άλλη δειγματοληπτική συσκευή. Τα κύτταρα που συλλέγονται μεταφέρονται σε γυάλινη επιφάνεια και εν συνεχεία αποστέλλονται στο κυτταρολογικό εργαστήριο, όπου προετοιμάζονται με ειδικές χρώσεις για τη μικροσκοπική εξέταση. Η αρχική μικροσκόπηση γίνεται από κυτταρολόγο (primary screener), ο οποίος έχει εκπαιδευτεί για να είναι ικανός να ανιχνεύει άτυπα κύτταρα ανάμεσα σε χιλιάδες φυσιολογικών κυττάρων στο επίχρισμα. Οι γυναίκες που εντοπίζεται να έχουν προβληματικά επιχρίσματα παραπέμπονται για περαιτέρω έρευνα και θεραπεία [34].

Ο πρωταρχικός στόχος της εξέτασης κολποτραχηλικού επιχρίσματος Παπανικολάου είναι να προστατέψει από την ανάπτυξη διηθητικού καρκίνου του τραχήλου της μήτρας ανιχνεύοντας προκαρκινικές αλλοιώσεις (*Τραχηλική Ενδοεπιθηλιακή Νεοπλασία*) στο επιθήλιο του τραχήλου της μήτρας. Το test Pap, επίσης, είναι μία ευαίσθητη μέθοδος διάγνωσης του μικροδιηθητικού καρκίνου του τραχήλου. Οι γυναίκες με μικροδιηθητικό καρκίνο (*FIGO στάδιο 1*) συχνά δεν γνωρίζουν την ύπαρξη της νόσου γιατί συνήθως είναι ελεύθερες συμπτωμάτων. Η διάγνωση και η θεραπεία του διηθητικού καρκίνου, που βρίσκεται ακόμη στα πρώτα στάδια εξέλιξης, βελτιώνει σημαντικά την πρόγνωση (αλλαγή του προσδόκιμου επιβίωσης) τηςυ ασθενούς. Για αυτούς τους λόγους το test Pap χρησιμοποιείται κυρίως ως μέθοδος προληπτικού ελέγχου των γυναικών που είναι ελεύθερες τραχηλικής ενδοεπιθηλιακής νεοπλασίας (*CIN*) και μικροδιηθητικού καρκίνου [34].

Σε μια γενική εξέταση ή χαμηλού κινδύνου του πληθυσμού, τα περισσότερα αποτελέσματα του test Pap είναι φυσιολογικά. Η πιο πρόσφατη κατάταξη είναι η **Bethesda** που κατατάσσει όλες τις αλλοιώσεις του τραχήλου της μήτρας σε 2 ομάδες: Χαμηλού βαθμού πλακώδης ενδοεπιθηλιακή αλλοίωση (*Low-Grade Squamous Intraepithelial Lesion* ή **LgSIL**)

και υψηλού βαθμού πλακώδης ενδοεπιθηλιακή αλλοίωση (*High-Grade Squamous Intraepithelial Lesion* ή **HgSIL**). Η LgSIL αντιστοιχεί σε CIN-1 και η HgSIL περιλαμβάνει τις CIN-2 και CIN-3 αλλοιώσεις [30]. Αναλυτικότερα, όλα τα μη φυσιολογικά αποτελέσματα του test Pap ταξινομούνται σύμφωνα με το σύστημα Bethesda στις ακόλουθες κατηγορίες:

- Ανωμαλίες Πλακωδών Κυττάρων (SIL)

→ Άτυπα πλακώδη κύτταρα απροσδιορίστου σημασίας (**ASC-US**)

Ένα αποτέλεσμα ASC-US προκύπτει από αλλοίωση στα τραχηλικά κυτταρα λόγω του HPV. Στις περισσότερες περιπτώσεις οι αλλοίωσεις αυτές δεν εξελίσσονται σε καρκινώματα, αλλά απαιτούν περαιτέρω παρακολούθηση και πιθανόν θεραπευτική αγωγή για να αποφευχθεί ο κίνδυνος εμφάνισης καρκίνου [56].

→ Χαμηλού βαθμού πλακώδης ενδοεπιθηλιακή αλλοίωση (**LgSIL**)

Η χαμηλού βαθμού πλακώδης ενδοεπιθηλιακή δυσπλασία καταδεικνύει μία πιθανή τραχηλική δυσπλασία, συνήθως μίας ήπιας μορφής δυσπλασία (CIN-1). Επειδή η CIN-1 είναι μία κοινή και συχνά καλοήθης μορφή δυσπλασίας και γιατρεύεται εντός 2 ετών συνήθως, στα περιστατικά με LgSIL η αγωγή που προτείνεται συνήθως είναι η τήρηση μίας στάσης αναμονής. Καθώς το 12-16% των περιστατικών αυτών θα μπορούσαν να εξελιχθούν σε μία σοβαρότερη δυσπλασία ο κυτταρολόγος μπορεί να προβεί σε κολποσκόπηση με βιοψία [66].

→ Άτυπα πλακώδη κύτταρα - δεν μπορεί να αποκλείσει HgSIL (**ASC-H**)

Η κατάσταση ASC-H είναι πολύ πιο πιθανόν να αποτελεί μια προκαρκινική κατάσταση σε σχέση με την κατάσταση ASC-US. Απαιτεί εκτίμηση της κλινικής κατάστασης μέσω κολποσκόπησης.

→ Υψηλού βαθμού πλακώδης ενδοεπιθηλιακή αλλοίωση (**HgSIL**)

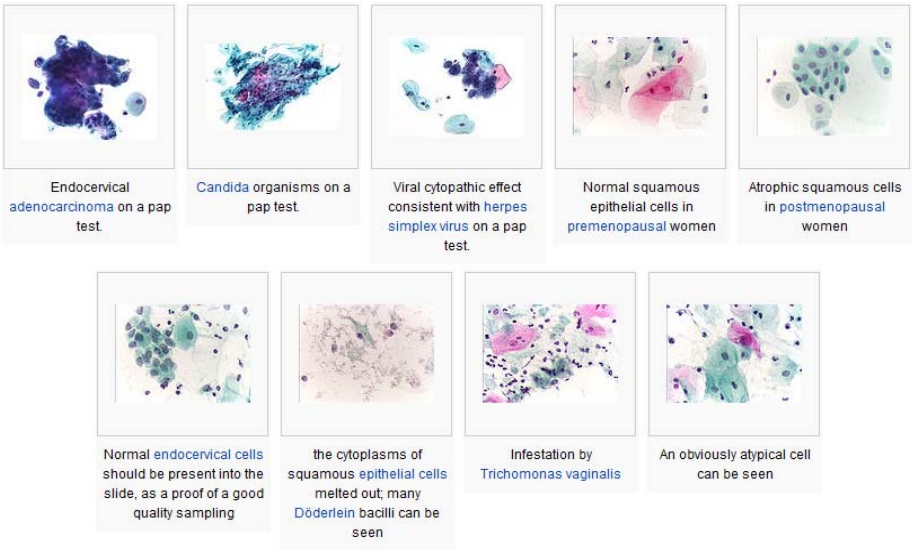
Η υψηλού βαθμού πλακώδης ενδοεπιθηλιακή αλλοίωση καταδεικνύει μια σοβαρότερη δυσπλασία ή μπορεί ακόμα και καρκίνωμα. Η HgSIL κατάσταση δεν υποδεικνύει την παρουσία καρκίνου. Από όλα τα περιστατικά μόλις το 2% έχει στην πραγματικότητα καρκίνο τη στιγμή της εξέτασης και 20% αυτών θα μπορούσαν να εξελιχθούν σε καρκίνο. Για το λόγο αυτό απαιτείται άμεση κολποσκόπηση με βιοψία σε τέτοιες περιπτώσεις. Η συνηθέστερη ταξινόμηση του HgSIL είναι σε CIN-2/3 [66].

→ Πλακώδες καρκίνωμα

Πρόκειται για ιστολογικά διακριτή μορφή καρκίνου. Προκαλείται από τον μη ελεγχόμενο πολλαπλασιασμό επιθηλιακών κυττάρων.

- Αδενικές ανωμαλίες επιθηλιακών κυττάρων

→ Άτυπα αδενικά κύτταρα που δεν προσδιορίζονται αλλιώς (AGC ή AGC-NOS)



Εικόνα 1.7*: Οπτικά αποτελέσματα των περιπτώσεων του test Pap*

Επίσης μπορούν να ανιχνευθούν ανωμαλίες του ενδοτράχηλου και του ενδομητρίου, όπως μια σειρά μολυσματικών διαδικασιών, συμπεριλαμβανομένης της μυκητίασης [35], του ιού του απλού έρπητα και τριχομονάδες. Το test Pap όμως δεν παρουσιάζει μεγάλη ανιχνευτική ευαισθησία ως προς αυτές τις κλινικές καταστάσεις, οπότε απουσία ανίχνευσης κάποιας τέτοιου είδους λοίμωξης δε συνεπάγεται και την πραγματική απουσία της λοίμωξης [33].

Αν και η πιο σημαντική χρήση του test Pap είναι ο προληπτικός έλεγχος του πληθυσμού, ώστε να εντοπίζει τις γυναίκες που παρουσιάζουν αυξημένες πιθανότητες να αναπτύξουν καρκίνο του τραχήλου της μήτρας, μπορεί επίσης να χρησιμοποιηθεί ως διαγνωστική εξέταση του καρκίνου του τραχήλου. Μπορεί να χρησιμοποιηθεί για να επιβεβαιώσει την παρουσία διηθητικού καρκίνου σε γυναίκες με συμπτώματα και/ή ενδεικτικά στοιχεία της νόσου. Εντούτοις , όταν χρησιμοποιείται κατ' αυτόν τον τρόπο, τα ευρήματα πρέπει να ερμηνεύονται με προσοχή. Μελέτες έχουν δείξει ότι το κολποτραχηλικό επίχρισμα μπορεί να είναι αρνητικό ακόμη και σε παρουσία προχωρημένου διηθητικού καρκίνου του τραχήλου. Αυτό συμβαίνει γιατί το αίμα, τα φλεγμονώδη κύτταρα και το νεκρωμένο υλικό από τη θέση του καρκίνου συχνά καλύπτουν τα άτυπα κύτταρα που υπάρχουν στο επίχρισμα. Όποιος κάνει τη λήψη του κολποτραχηλικού επιχρίσματος πρέπει να το γνωρίζει αυτό, και να χρησιμοποιεί όλες τις μεθόδους, δηλαδή κολποσκόπηση και βιοψία, για να εξακριβώσει την παρουσία ή την απουσία της κακοήθειας σε μία συμπτωματική ασθενή με αρνητική εξέταση Παπανικολάου. Το test Pap χρησιμοποιείται συχνά για **παρακολούθηση (*follow-up*)** γυναικών οι οποίες έχουν θεραπευτεί για CIN ή

διηθητικό καρκίνο. Είναι μία ευαίσθητη μέθοδος ανίχνευσης υποτροπής της νόσου σε τέτοιες περιπτώσεις. Τέλος, όπως αναφέρθηκε και προηγουμένως, αν και μια ποικιλία παθογόνων μικροοργανισμών μπορούν να διαγνωσθούν στο κολποτραχηλικό επίχρισμα, εντούτοις το τεστ Παπανικολάου δεν αποτελεί έγκυρη μέθοδο για την ανίχνευση λοιμώξεων.

1.2.2 Νέες τεχνικές ανίχνευσης καρκίνου του τραχήλου της μήτρας

1.2.2.1 HPV DNA test

Στην αρχή του παρόντος κεφαλαίου έγινε μια εισαγωγή στον ιό των ανθρώπινων θηλωμάτων (*Human Papilloma Virus ή HPV*). Οι ιοί του ανθρώπινου θηλώματος (*HPV*) είναι μέλη μίας οικογένειας ιών, που είναι γνωστοί ως ιοί Papova. Είναι επιθηλιοτρόποι ιοί, οι οποίοι προάγουν τον κυτταρικό πολλαπλασιασμό με αποτέλεσμα την ανάπτυξη υπερπλαστικών καλοηθών θηλωματωδών αλλοιώσεων του γεννητικού συστήματος, του ανωτέρου αναπνευστικού συστήματος, του πεπτικού σωλήνα και του δέρματος. Είναι πάρα πολλοί οι τύποι HPV που έχουν αναγνωριστεί ως αποτέλεσμα μοριακού υβριδισμού του εξαγόμενου DNA από επίπεδα ή εξωφυτικά κονδυλώματα και από μία ποικιλία θέσεων. Κάθε τύπος ιού έχει μια περιορισμένη περιοχή που προσβάλλει και ιοί οι οποίοι καταλαμβάνουν παρόμοιες περιοχές/εσοχές συσχετίζονται γενετικά. Μοριακός υβριδισμός των θηλωμάτων του ανώτερου γεννητικού συστήματος έδειξε ότι λίγο λιγότεροι από τους μισούς αναγνωρισμένους τύπους HPV περιορίζονται στο γυναικείο γεννητικό σύστημα. Με βάση τα παραπάνω γίνεται κατανοητό πως η παρουσία τύπων του HPV παίζει σημαντικό ρόλο για την εξέλιξη κάποιας λοίμωξης του τραχήλου της μήτρας [36].

Αν και το test Pap έχει αποδειχτεί ότι είναι ένα αξιόλογο μέσο στην πρόληψη του τραχηλικού καρκίνου, η ανάλυση των κολποτραχηλικών επιχρισμάτων είναι μία εργασία που απαιτεί κόπο και μπορούν να την αναλάβουν μόνο πολύ καλά εκπαιδευμένοι κυτταρολόγοι. Επίσης η ερμηνεία των κολποτραχηλικών επιχρισμάτων εμπεριέχει τον υποκειμενικό παράγοντα και είναι αντικείμενο διαγνωστικού λάθους. Ένα αντικειμενικό τεστ που βασίζεται στην ανίχνευση του DNA των HPV υψηλού κινδύνου φαίνεται να είναι μία πρακτική εναλλακτική λύση.

Η ανάλυση DNA έχει γίνει η μέθοδος επιλογής εξαιτίας των περιορισμών των άλλων διαγνωστικών μεθόδων. Οι πιο ικανοποιητικές εξετάσεις για HPV ανιχνεύουν το DNA του ιού. Επειδή όλοι οι τύποι του HPV έχουν στενή συγγένεια, γίνονται αναλύσεις που στοχεύουν σε συγκεκριμένες περιοχές του γενετικού υλικού για να γίνει διαχωρισμός μεταξύ των διαφόρων τύπων HPV. Η εξέταση Hybrid Capture 2 παρέχει το προφίλ ενός αριθμού διαφορετικών τύπων HPV, οι οποίοι τότε μπορούν να αναλυθούν περαιτέρω χρησιμοποιώντας την αλυσιδωτή αντίδραση πολυμεράσης (*PCR*). Η PCR χρησιμοποιείται για

να αναγνωρίζει ειδικούς τύπους HPV. Η Hybrid Capture 2 (Digene Diagnostics, Gaithersburg, MD, USA) έχει πρόσφατα εγκριθεί από την Αμερικανική Εταιρία Τροφίμων και Φαρμάκων [American Food and Drug Agency (FDA)], ως βοήθημα στον κυτταρολογικό προληπτικό έλεγχο γυναικών ηλικίας 30 ετών και πάνω. Είναι γενικά παραδεκτό ότι η PCR και HC2, μαζί, ανιχνεύουν αξιόπιστα τους υψηλού κινδύνου και άλλους τύπους HPV σε κλινικά δείγματα [38,39].

Όσον αφορά το HPV DNA test και την Hybrid Capture 2 έχουν διατυπωθεί και έχουν αξιολογηθεί τρεις τοποθετήσεις για το ρόλο και τη χρησιμότητά του για τη διάγνωση προκαρκινικών και καρκινικών καταστάσεων του τραχήλου της μήτρας. Οι τοποθετήσεις αυτές είναι οι ακόλουθες:

1. ως πρωταρχικής σημασίας προληπτικός έλεγχος
2. ως βοηθητική-επικουρική κυτταρολογική εξέταση
3. ως μέσον ιατρικής παρακολούθησης μετά από τη θεραπεία ή ως εξέταση επιβεβαίωσης της θεραπείας

Η Hybrid Capture 2 χρησιμοποιείται κυρίως ως screening test. Σύμφωνα με την 1^{η} προσέγγιση συνίσταται ο **συνδυασμός του HPV DNA test και της κυτταρολογικής εξέτασης (*1.*)** (δηλαδή του test Pap) για τον ιατρικό έλεγχο και την αξιολόγηση των κλινικών αποτελεσμάτων για γυναίκες ηλικίας άνω των 30 ετών. Ο συνδυασμός αυτών των διαφορετικών εξετάσεων χρησιμοποιείται με σκοπό την παροχή ασφαλέστερων συμπερασμάτων για τις υπό έλεγχο γυναίκες. Γυναίκες που εμφανίζουν αρνητικά αποτελέσματα και στις δύο εξετάσεις θα μπορούσαν να διαβεβαιωθούν διπλά ότι είναι ελεύθερες νόσου. Το αυξημένο κόστος της διπλής εξέτασης αντισταθμίζεται με έλεγχο σε αραιότερα χρονικά διαστήματα. Μία δεύτερη προσέγγιση αφορά γυναίκες που αρχικά ελέγχονται για HPV DNA, ενώ η κυτταρολογική εξέταση χρησιμοποιείται μόνο σε περιπτώσεις εμφάνισης θετικού HPV DNA test. Κατά την προσέγγιση αυτή μία γυναίκα παραπέμπεται για περεταίρω κλινική αξιολόγηση μόνο στην περίπτωση εμφάνισης θετικών αποτελεσμάτων και στις δύο εξετάσεις.

Το μεγαλύτερο πρόβλημα του HPV DNA test είναι η χαμηλή ειδικότητα που παρουσιάζει στον έλεγχο για την ύπαρξη CIN ή καρκίνου. Δεν υπάρχει ακόμα ξεκάθαρη άποψη για το πώς θα πρέπει να αντιμετωπίζονται γυναίκες με θετικό HPV DNA test και αρνητική κυτταρολογική εξέταση. Συνίσταται αυξημένη επίβλεψη, καθώς αυτές οι γυναίκες έχουν αυξημένες πιθανότητες εμφάνισης CIN σε βάθος χρόνου. Στο πλαίσιο της παρούσας μελέτης, οι προσπάθειες μας επικεντρώνονται στις περιπτώσεις αυτές με στόχο τη βελτίωση της εκτίμησης της κλινικής κατάστασης.

Ακόμα, όπως διατυπώθηκε και παραπάνω, το HPV DNA test χρησιμοποιείται και ως **βοηθητικό μέσο της κυτταρολογίας (*2.*)**. Ένα από τα προβλήματα στο χώρο της κυτταρολογίας του τραχήλου της μήτρας είναι η αντιμετώπιση των γυναικών με ASCUS ή LgSIL. Αρκετές μελέτες έχουν δείξει ότι οι γυναίκες, των οποίων το κολποτραχηλικό

επίχρισμα έχει δώσει ASCUS αλλοιώσεις, μπορεί να έχουν περισσότερο προχωρημένη αλλοίωση (CIN-2 ή σοβαρότερη).

Το HPV DNA test χρησιμοποιείται και ως **μέθοδος παρακολούθησης (*3.*)** μετά από θεραπεία. Οι γυναίκες οι οποίες έχουν διαγνωστεί με CIN-2 ή CIN-3 και έχουν υποβληθεί σε χειρουργική θεραπεία ή κωνοειδή εκτομή, πρέπει να παρακολουθούνται στενά για τουλάχιστον πέντε χρόνια μετά τη θεραπεία. Αν και πάνω από 90% των γυναικών θεραπεύονται, υπάρχει ο κίνδυνος της υποτροπής ή της ανάπτυξης διηθητικού καρκινώματος στο 5%-19% των περιπτώσεων. Στις περισσότερες περιπτώσεις η παρακολούθηση περιλαμβάνει και κολποσκόπηση και κυτταρολογική διερεύνηση, ανά 6 μήνες τον πρώτο χρόνο, και μία φορά το χρόνο κυτταρολογική εξέταση και/ή κολποσκόπηση για τα επόμενα 5 χρόνια.

Όπως έγινε φανερό το HPV DNA test προσπαθεί να ανιχνεύσει τους τύπους του HPV που υπάρχουν στα τραχηλικά κύτταρα της εξεταζόμενης ασθενούς. Τα αποτελέσματα που επιστρέφει αφορούν την παρουσία ή απουσία αντίστοιχα των διαφόρων τύπων. Έτσι το τεστ θεωρείται θετικό αν υπάρχει η παρουσία ογκογόνων τύπων και αντίστοιχα αρνητικό όταν δεν ανιχνεύει κάποιον ιό γενικότερα ή όταν ανιχνεύει ιούς χαμηλού κινδύνου.

Τέλος, αν και το HPV DNA τεστ από μόνο του είναι μία ευαίσθητη μέθοδος διερεύνησης των CIN αλλοιώσεων του τραχήλου της μήτρας, δεν είναι επαρκώς ειδικό για να αποτελέσει μία πρακτική μέθοδο στον πρωτογενή έλεγχο του καρκίνου του τραχήλου της μήτρας. Η σημασία ενός θετικού HPV DNA τεστ σε απουσία υποκείμενης νόσου δεν είναι ακόμη γνωστή και μπορεί να προκαλέσει περιττή ανησυχία σε γυναίκες με φυσιολογικούς τραχήλους που «φιλοξενούν» τον ιό. Ο ρόλος του HPV DNA τεστ για την περαιτέρω αντιμετώπιση των ASCUS ή LgSIL είναι επίσης αμφισβητούμενη όπως και ο ρόλος του στην παρακολούθηση γυναικών που έχουν υποβληθεί σε θεραπεία για CIN-2/3 [37].

1.2.2.2 mRNA test

Η ανίχνευση E6/E7 mRNA και η παρουσία της δραστηριότητας κάποιου ογκογονιδίου σε δείγματα του τραχήλου της μήτρας μπορεί να πραγματοποιηθεί με αντίστροφη μεταγραφάση (RT)-PCR ή με την ενίσχυση αλληλουχίας νουκλεϊκών οξέων (*NASBA*) [4]. NASBA είναι μια μέθοδος στη μοριακή βιολογία η οποία χρησιμοποιείται για να ενισχύσει RNA ακολουθίες, η οποία αναπτύχθηκε από τον J. Compton το 1991, και αρχικά χρησιμοποιήθηκε για την ανίχνευση της έκφρασης του HIV [40].

Τα επίπεδα mRNA των E6 και E7 πρωτεϊνών έχουν δείξει ότι αυξάνονται ανάλογα με τη σοβαρότητα της βλάβης, συνεπώς η ανίχνευση mRNA των E6/E7 μπορεί να έχει μεγαλύτερη προγνωστική αξία και να βελτιώσει την ειδικότητα και τη θετική προγνωστική αξία σε σύγκριση με το HPV DNA test για τον προσυμπτωματικό έλεγχο [41]. Οι

κυτταρολογικές και ιστολογικές εξετάσεις δεν δίνουν ακριβείς πληροφορίες για το ποιες γυναίκες με επιθηλιακές ανωμαλίες και ατυπίες θα αναπτύξουν καρκίνο. Το HPV DNA test δηλώνει την παρουσία του ιού αλλά δεν παρέχει στοιχεία για την εξέλιξη της μόλυνσης. Η ανίχνευση του mRNA των ογκογονιδίων E6/E7 των πιο συχνών ογκογόνων τύπων του ιού, **HPV 16, 18, 31, 33 και 45,** σε δείγματα τραχηλικού επιχρίσματος αποτελεί τον πιο σύγχρονο και άμεσο προγνωστικό δείκτη για την πιθανότητα ανάπτυξης καρκίνου του τραχήλου της μήτρας και δίνει τη δυνατότητα καλύτερης κατάταξης ασθενών με μη-φυσιολογικά κυτταρολογικά ευρήματα σε ομάδες υψηλού ή χαμηλού κινδύνου. Γίνεται κατανοητό λοιπόν πως η ανίχνευση mRNA έστω και ενός από του ιούς υψηλού κινδύνου συνεπάγεται ότι το test είναι θετικό, ενώ αντίθετα η απουσία mRNA καταδεικνύει ένα αρνητικό αποτέλεσμα. Με βάση τα παραπάνω δίνεται η δυνατότητα αποφυγής κωνοειδών εκτομών σε γυναίκες με χαμηλού κινδύνου μολύνσεις, καθώς και παρακολούθησης ασθενών που έχουν υποστεί χειρουργική επέμβαση για την αφαίρεση της ύποπτης περιοχής του τραχήλου [42].

Οι E6/E7 σχετίζονται άμεσα με την εξέλιξη προκαρκινικών αλλοιώσεων του τραχήλου. Πιο συγκεκριμένα, η σημαντικότητα του mRNA test φαίνεται από το γεγονός ότι η περαιτέρω μεταγραφή και μετάφραση των ογκογονιδίων του HPV οδηγεί με μεγαλύτερη πιθανότητα σε αυξημένη ογκογονική δραστηριότητα και κατ` επέκταση στην εξέλιξη της τραχηλικής νεοπλασίας (*CIN*). Σε πολλές μελέτες έχει εξακριβωθεί ότι το mRNA test δίνει περισσότερα θετικά αποτελέσματα σε αλλοιώσεις CIN-3 ή σοβαρότερες παρά σε αλλοιώσεις CIN-2 [50,51]. Σε περίπτωση που το mRNA test δώσει λιγότερα θετικά αποτελέσματα σε σχέση με το HPV DNA test του ίδιου δείγματος σημαίνει είτε ότι ορισμένοι τύποι του HPV δεν εκφράστηκαν είτε ότι εκφράστηκαν αλλά σε πολύ χαμηλό βαθμό.

Τέλος, έχει αποδειχθεί ότι το mRNA test σε συνδυασμό με το HPV DNA test μπορεί να αποτελέσει μια πολύ καλή επιλογή ως πρωταρχικό screening test. Αν και τα δύο test είναι θετικά τότε η ασθενής υποβάλλεται σε βιοψία για επιβεβαίωση. Αν μόνο το HPV DNA test είναι θετικό τότε η ασθενής επανεξετάζεται μετά από 6 μήνες και αν αυτό παραμένει θετικό τότε γίνεται κολποσκόπηση για επιβεβαίωση [41].

1.2.2.3 p16 test

Η p16 είναι μια ογκοκατασταλτική πρωτεΐνη που παράγουν τα φυσιολογικά κύτταρα για να σταματούν τον κυτταρικό κύκλο πριν τη φάση της σύνθεσης και της περαιτέρω διαίρεσής του [49]. Αποτελεί έναν από τους πιο σημαντικούς βιολογικούς δείκτες (*biomarkers*) των κυττάρων που είναι μολυσμένα από HPV τύπο υψηλού κινδύνου. Πιο συγκεκριμένα, η υπερέκφραση (*overexpression*) της p16 μπορεί να υποδείξει την εξέλιξη των προκαρκινικών αλλοιώσεων. [43]

Η **δράση** της p16 πρωτεΐνης είναι ογκοκατασταλτική. Στα φυσιολογικά κύτταρα, υπάρχει ένα σημείο αναμονής (*restriction point*) στο τέλος της G1 φάσης του κυτταρικού

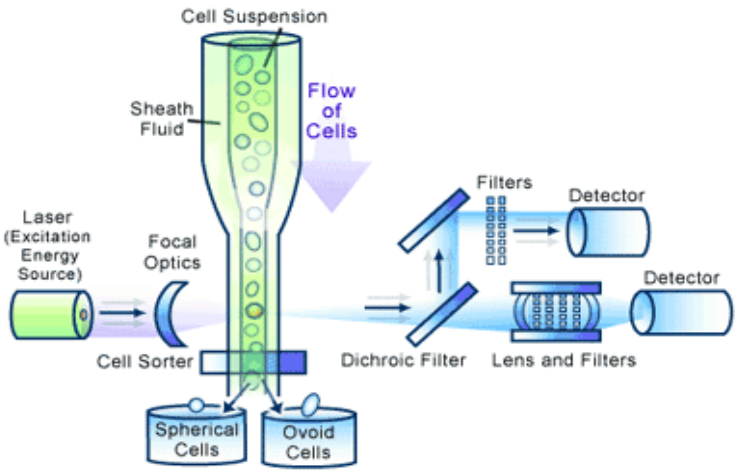
κύκλου, στο οποίο λαμβάνεται η απόφαση για το αν το κύτταρο είναι έτοιμο να προχωρήσει στις επόμενες φάσεις του [44]. Η p16 επεμβαίνει ρυθμιστικά και αποτρέπει τη σύνδεση των πρωτεϊνών cyclin-dependent kinases (*CDKs*) με τις κυκλίνες (*cyclins*). Όταν επιτευχθεί μια τέτοια σύνδεση CDK/cyclin (συγκεκριμένα είναι σύνδεση της CDK4 ή 6 με την cyclin D) τότε ενεργοποιείται η εξέλιξη του κυτταρικού κύκλου στις επόμενες φάσεις του. Εν προκειμένω, ενεργοποιούνται οι παράγοντες μεταγραφής του DNA με απώτερο στόχο την περαιτέρω διαίρεση του κυττάρου [45].

Στα κύτταρα που έχουν υποστεί HPV-λοίμωξη παρατηρείται αυξημένη έκφραση κυκλινών, οι οποίες ανταγωνίζονται την p16. Προσπαθούν δηλαδή να συνδεθούν με τις CDKs ώστε να ευνοηθεί η συνέχιση του κυτταρικού κύκλου. Στη συνέχεια, η σύνδεση CDK4-6/cyclinD φωσφορυλιώνει (απενεργοποιεί με χημικό τρόπο) το ρετινοβλάστωμα pRb. Έτσι, απελευθερώνεται ο μεταγραφικός παράγοντας E2F που επιτρέπει τη μετάβαση από τη G1 φάση του κυτταρικού κύκλου στην S (*φάση σύνθεσης*). Με αυτόν τον τρόπο γίνεται δυνατή η ανεξέλεγκτη κυτταρική ανάπτυξη που οδηγεί σε καρκίνο [46]. Οπότε γίνεται σαφές ότι η ανίχνευση υπερέκφρασης της πρωτεΐνης p16 καταδεικνύει ένα θετικό αποτέλεσμα, ενώ σε αντίθετη περίπτωση θεωρούμε ότι έχουμε ένα αρνητικό αποτέλεσμα.

Μια μελέτη [43] απέδειξε ότι όταν η p16 test συνοδεύει το HPV DNA test τότε αποτελεί ιδιαίτερα χρήσιμο βιολογικό δείκτη της **εξέλιξης των LgSIL αλλοιώσεων σε HgSIL**. Ειδικότερα, οι ασθενείς με ASCUS ή LgSIL αλλοίωση και ύπαρξη HPV τύπου υψηλού κινδύνου που έδωσαν θετική p16 είχαν πολύ μεγαλύτερη πιθανότητα να αναπτύξουν HgSIL απ’ ότι εκείνες με αρνητική p16.

1.2.2.4 Flow Cytometry test

Το Flow Cytometry test είναι ένα τεστ το οποίο χρησιμοποιείται για τη μέτρηση της ποσότητας DNA στα κύτταρα. Το τεστ αυτό μπορεί, μετρώντας την ποσότητα DNA στα κύτταρα, να προσδιορίσει το ποσοστό των κυττάρων που βρίσκονται σε διαφορετική φάση της κυτταρικής διαίρεσης. Μπορεί επίσης να ανιχνεύσει τους πληθυσμούς των κυττάρων, τα οποία εμφανίζουν ασυνήθιστα ποσά DNA, δηλαδή κύτταρα που έχουν πολλές ανωμαλίες γονιδίων [5]



Εικόνα 1.8: *Η βασική αρχή λειτουργίας του Flow Cytometry*

Η βασική αρχή, πάνω στην οποία στηρίζεται η λειτουργία του Flow Cytometry, είναι η πρόσπτωση μίας μονοχρωματικής (με ένα μόνο μήκος κύματος) δέσμης φωτός, συνήθως laser, πάνω σε μία υδροδυναμικά συγκλίνουσα ροή υγρού. Ένας αριθμός από ανιχνευτές περιβάλλουν το σημείο στο οποίο η ροή του υγρού περνά μέσα από τη δέσμη του laser: ένας κατά το διαμήκη άξονα της δέσμης, αρκετοί κάθετα προς αυτήν και ένας ή περισσότεροι ανιχνευτές φθορισμού. Κάθε σωματίδιο από 0.2 ως 150μm όταν διαπερνά τη δέσμη, προκαλεί σκέδαση της δέσμης και λόγω φθοριζουσών ουσιών που βρίσκονται στο εκάστοτε σωματίδιο προκαλείται εκπομπή φωτός σε μεγαλύτερο μήκος κύματος σε σχέση με την πηγή της δέσμης. Ο συνδυασμός της σκεδαζόμενης και της φθορίζουσας δέσμης εντοπίζεται από τους ανιχνευτές και εν συνεχεία, αναλύοντας τις διακυμάνσεις φωτεινότητας σε κάθε ανιχνευτή, είναι δυνατόν να εξαχθούν πληροφορίες για τη χημική και φυσική δομή του κάθε σωματιδίου. Η εμπρόσθια σκέδαση (FSC) σχετίζεται με τον όγκο του κυττάρου και η πλάγια σκέδαση (SSC) εξαρτάται από την εσωτερική πολυπλοκότητα του σωματιδίου (π.χ. σχήμα του πυρήνα) [47].

Οι σύγχρονες συσκευές κυτταρομετρίας ροής παρέχουν τη δυνατότητα ανάλυσης αρκετών χιλιάδων σωματιδίων το δευτερόλεπτο και παράλληλα το διαχωρισμό και την απομόνωση των σωματιδίων με βάση συγκεκριμένες ιδιότητες που αυτά παρουσιάζουν. Μια συσκευή κυτταρομετρίας ροής μοιάζει με ένα μικροσκόπιο, με τη διαφορά πως αντί να παράγει την εικόνα ενός κυττάρου, αυτή παρέχει δεδομένα για τα χαρακτηριστικά ενός μεγάλου αριθμού κυττάρων σε λίγο χρόνο. Μία τυπική συσκευή κυτταρομετρίας ροής αποτελείται από 5 κυρίως εξαρτήματα: μία ροή υγρού περιβλήματος (μεταφέρει και διευθετεί τα κύτταρα ώστε να περνούν ένα-ένα από τη δέσμη του λέιζερ), ένα οπτικό σύστημα που παράγει φωτεινά σήματα, έναν ανιχνευτή με σύστημα μετατροπής σήματος από αναλογικό σε ψηφιακό, ένα σύστημα ενίσχυσης και έναν ηλεκτρονικό υπολογιστή για την ανάλυση των σημάτων [47].

Εδώ, σε σχέση με τα προηγούμενα tests δεν είναι τόσο προφανές πότε το flow cytometry test θεωρείται θετικό και πότε αρνητικό. Θετικό θεωρείται λοιπόν σε περίπτωση που το ποσοστό των φθοριζόντων κυττάρων υπερβαίνουν το 1,5%. Αυτό το γεγονός υποδεικνύει την ενσωμάτωση του ιου. Σε αντίθετη περίπτωση το flow cytometry test

θεωρείται αρνητικό [67]. Στον καρκίνο του τραχήλου, το flow cytometry ενδείκνυται ως **screening συμπληρωματικό** του κυτταρολογικού test Papενώ τελευταίες μελέτες δείχνουν ότι ίσως μπορεί να χρησιμοποιηθεί ακόμα και ως πρωτεύον screening test. Μάλιστα, παρουσιάζει πολύ υψηλή απόδοση όταν χρησιμοποιείται ως δευτερεύων βιολογικός δείκτης για ασθενείς με θετικό HPV DNA test [48].

1.3 Πρόβλημα βελτιστοποίησης της απόδοσης των τεχνικών ανίχνευσης καρκίνου του τραχήλου της μήτρας

Στα πλαίσια του υποκεφαλαίου αυτού θα γίνει αναφορά και σύγκριση της απόδοσης των διαφόρων screening tests για τον καρκίνο του τραχήλου της μήτρας, όπως αυτά έχουν παρουσιαστεί σε διεθνώς εμπεριστατωμένες μελέτες. Με βάση αυτά, θα παρουσιαστεί σε γενικές γραμμές το θέμα της παρούσας εργασίας καθώς και τι έρχεται αυτή να προσθέσει στην ήδη υπάρχουσα βιβλιογραφία γύρω από το θέμα αυτό.

Όπως προαναφέρθηκε, ο στόχος που έχει το screening είναι ο έγκαιρος προσδιορισμός μίας ασθένειας, έτσι ώστε να υπάρχει η δυνατότητα άμεσης παρέμβασης και να μειωθεί κατά το δυνατόν η θνησιμότητα. Παρόλο που το screening μπορεί να οδηγήσει σε μία πρώιμη διάγνωση, δεν έχουν όλα τα screening tests την ίδια διαγνωστική αξία.Ακόμα, η πιθανότητα για υπερδιάγνωση ή εσφαλμένη διάγνωση είναι πάντα υπαρκτή. Επίσης, δημιουργείται συχνά και μία ψευδής αίσθηση ασφάλειας. Για όλους τους παραπάνω λόγους, ένα screening test πρέπει να παρουσιάζει υψηλή **ειδικότητα** (*specificity*) σε συνδυασμό με μία αποδεκτή **ευαισθησία** (*sensitivity*). Πιο συγκεκριμένα, θα παρατεθούν στη συνέχεια μελέτες, στις οποίες προσπαθείται, μέσω συνδυασμού του test Pap και των νέων τεχνικών που αναφέρθηκαν στο υποκεφάλαιο 1.2, να επιτευχθεί καλύτερη ειδικότητα και ευαισθησία στην ανίχνευση των κλινικών αποτελεσμάτων.

1.3.1 Σύγχρονες εμπεριστατωμένες μελέτες

Όπως αναφέρθηκε και προηγουμένως, ο συνδυασμός του test Pap με τις νεότερες τεχνικές για την ανίχνευση τραχηλικών αλλοιώσεων είναι κάτι που έχει απασχολήσει πολλούς ερευνητές. Αρχικά κρίνεται σκόπιμο να παρατεθούν μελέτες οι οποίες αφορούν την κάθε εξέταση ξεχωριστά για να δοθεί μία πρώτη εντύπωση για την ευαισθησία και την ειδικότητα που επιτυγχάνει αυτοτελώς η κάθε μια από αυτές.

PAP test

Σύμφωνα με μελέτες [1], το **PAP test** έδωσε παρόμοια αποτελέσματα στη συμβατική εκδοχή του και στην κυτταρολογία υγρής φάσης. Ιδιαίτερη σημασία έχει η υψηλή ειδικότητά του. Πιο συγκεκριμένα, οι δύο μελέτες στις οποίες αναζητήσαμε στατιστικά αποτελέσματα ασχολούνταν με το κατά πόσο το test Pap είναι σύμφωνο με την ιστολογία. Στη μία από αυτές [59] στόχος ήταν η σύγκριση του test Pap με άλλα διαγνωστικά test για τον καρκίνο του τραχήλου της μήτρας. Συνολικά στα πλαίσια αυτής της μελέτης εξετάστηκαν 1.757 γυναίκες εκ των οποίων οι 828 παραπέμφθηκαν για κολποσκόπηση λόγω μη φυσιολογικών αποτελεσμάτων του test Pap. Στη δεύτερη από αυτές [60], στόχος των συγγραφέων ήταν να αναδείξουν τη μεγάλη κλινική χρησιμότητα του HPV DNA test σε περιπτώσεις ASCUS. Για το λόγο αυτό αρχικά εξήγαγαν ορισμένα στατιστικά για το test Pap ώστε να συγκριθούν με τα τελικά τους αποτελέσματα.

Ειδικότερα, η απόδοση της **συμβατικής κυτταρολογίας** test Pap έδωσε [59]:

- *ευαισθησία*=72%
- *ειδικότητα*=94%

Αντίστοιχες μελέτες της ακρίβειας της **κυτταρολογίας υγρής φάσης** έδωσαν:

- *ευαισθησία*=από 61% [60] έως 66% [59]
- *ειδικότητα*=από 82% [60] έως 91% [59]

Τα αποτελέσματα που προέκυψαν από αυτές τις μελέτες όσον αφορά τα στατιστικά μέτρα είναι αποτέλεσμα δυαδικής διάκρισης. Θεωρούν δηλαδή ότι η CIN-3 είναι η χειρότερη κλινική κατάσταση προ του καρκίνου, οπότε θεωρείται αρνητικό ένα αποτέλεσμα μεχρι και CIN-2 και θετικό ένα αποτέλεσμα με CIN-3 ή με καρκίνωμα.

HPV DNA test

Μελέτες ως προς την ακρίβεια του **HPV DNA test** υπολόγισαν:

- *ευαισθησία* = από 88% έως 91% (για την ανίχνευση CIN-3 και πάνω) [60] και έως 97% (για την ανίχνευση CIN-2+) [61]
- *ειδικότητα* = από 73% έως 79% (για την ανίχνευση CIN-3 και πάνω) [60] και έως 93% (για την ανίχνευση CIN-2+) [61]
- *αρνητική προγνωστική αξία* = από 97% έως 100% [27]

Αξίζει να τονίσουμε στο σημείο αυτό πως οι παραπάνω μελέτες επικεντρώνονται σε περιστατικά με σοβαρής μορφής CIN. Στα πλαίσια των δύο αυτών μελετών εξετάσθηκαν 4.075 [60] και 11.085 [61] γυναίκες αντίστοιχα. Τα αποτελέσματα σε αυτές αντιμετωπίζονται και στις δύο μελέτες με τρόπο δυαδικό, διότι επικεντρώνονται κυρίως στην ανίχνευση

περιπτώσεων με σοβαρή τραχηλική ενδοεπιθηλιακή νεοπλασία (CIN-2+). Τα νούμερα που προκύτπουν για την ευασισθησία και την ειδικότητα στα δείγματά τους είναι πολύ υψηλά κυρίως χάρη σε αυτό το διαχωρισμό, καθώς, όπως προαναφέρθηκε, η παρουσία λοίμωξης από κάποιον τύπο HPV είναι αναγκαία συνθήκη για την πρόκληση αλλοιώσεων ή καρκίνου. Παρατηρούμε λοιπόν ότι σε σχέση με την κυτταρολογία PAP test, το HPV DNA test προσφέρει καλύτερη ευαισθησία αλλά και χειρότερη ειδικότητα στη διερεύνηση γυναικών της κατηγορίας υψηλού κινδύνου για ανάπτυξη καρκίνου του τραχήλου.

Ένα αρνητικό HPV DNA test έχει υψηλή αρνητική προγνωστική αξία (ικανότητα επιβεβαίωσης αρνητικού αποτελέσματος), δηλαδή η πιθανότητα μίας γυναίκας να έχει καρκίνο είναι πολύ χαμηλή. Τα αντίστοιχα αποτελέσματα για τη θετική προγνωστική αξία διαφέρουν κατά πολύ από μελέτη σε μελέτη και απαιτούν μεγαλύτερη διερεύνηση.

Μια άλλη μελέτη [62] απέδειξε ότι η ευαισθησία του PAP test μαζί με το HPV DNA test είναι μεγαλύτερη από την ευαισθησία του κάθε test ξεχωριστά, ιδιαίτερα για γυναίκες άνω των 30 ετών. Αντίθετα, βρήκε ότι η ειδικότητα του συνδυασμού των δύο tests ήταν μικρότερη από την ειδικότητα του καθενός ξεχωριστά.

mRNA test

Σε μια μετα-ανάλυση των Cuschieri, Wentzensen (*Πίνακας 1.2)* [55], συνοψίζονται άλλες σημαντικές μελέτες και παρουσιάζεται η υψηλή απόδοση του mRNA test στην ανίχνευση των E6/E7 σε αλλοιώσεις CIN-2,3 και σε διηθητικό καρκίνο τραχήλου.

Πίνακας 1.1: Απόδοση mRNA test για ανίχνευση CIN-2,3 και διηθητικού καρκίνου.

Detection rate of E6/E7 transcipts in CIN2/3 and cervical cancers

Author	CIN2/3	Invasive CA
Rose, 1995 (27)		28/28 (100%)
Nakagawa, 1995 (26)	19/19 (100%)	31/31 (100%)
Kraus, 2006 (89)		199/204 (98%)
Lie, 2005 (36)	225/291 (77%)	20/20 (100%)
Molden, 2005 (37)	13/14 (93%)	
Sotlar, 2004 (30)	95/109 (87%)	
Total	352/433 (81%)	278/283 (98%)

Μια άλλη μελέτη της απόδοσης του mRNA test έδωσε τα εξής αποτελέσματα [50]:

- *ευαισθησία* = 64% σε γυναίκες με CIN-2+ αλλοίωση

- *ειδικότητα* = 97% σε γυναίκες με φυσιολογική κυτταρολογική

Και σε αυτή τη μελέτη παρατηρούμε ότι η εκτίμηση της ευασθησίας και της ειδικότητας γίνεται με τρόπο δυαδικό και η διάκριση γίνεται μεταξύ πολύ γενικών καταστάσεων.

p16 test

Η μετα-ανάλυση των Cuschieri, Wentzensen (*Πίνακας 1.3)* [55] αναφέρεται και στην απόδοση της p16 test βάσει αποτελεσμάτων άλλων έγκυρων μελετών.

Πίνακας 1.2: Ανίχνευση της p16 στις διάφορες αλλοιώσεις του τραχήλου.

p16 staining in immunohistochemistry

Author	Nondysplastic*	CIN1	CIN2	CIN3	Invasive CA
Sano, 1998 (44)	0/15 (0%)	8/15 (53%)	16/17 (94%)	27/27 (100%)	38/39 (97%)
Klaes, 2001 (45)	1/111 (1%)	29/47 (61%)	32/32 (100%)	60/60 (100%)	58/60 (97%)
Klaes, 2002 (49)	7/58 (12%)	15/17 (88%)	10/10 (100%)	43/43 (100%)	46/46 (100%)
Agoff, 2003 (51)	30/247 (12%)	43/76 (57%)	60/80 (75%)	103/113 (91%)	47/53 (89%)
Wang, 2004 (56) †	12/179 (7%)	27/75 (36%)	12/19 (63%)	19/19 (100%)	
Hu, 2005 (54) ‡,§		20/45 (44%)	43/46 (93%)	51/51 (100%)	
Benevolo, 2006 (52) ‡	0/17 (0%)	17/54 (31%)	9/10 (90%)	11/11 (100%)	8/8 (100%)
Ishikawa, 2006 (55) ‡	0/7 (0%)	13/53 (25%)	32/40 (80%)	45/48 (94%)	16/16 (100%)
Focchi, 2007(53) ‡	0/114 (0%)	80/88 (91%)	33/33 (100%)	32/32 (100%)	44/47 (94%)
Total	50/748 (7%)	252/470 (54%)	247/287 (86%)	391/404 (96%)	257/269 (96%)

*Includes normal, reactive, inflammation, hyperplasia, atypical, equivocal.
† Population-based study.
‡ Criteria modified from Horn et al. (49).
§ Adolescents.

Η μεγάλη χρησιμότητα του p16 ως συνοδευτικό του HPV DNA test φαίνεται και στη μετα-ανάλυση των Παρασκευαΐδη, Κολιόπουλου (*Πίνακας 1.4)* [63], όπου παρουσιάζεται η ανίχνευση της p16 στα διάφορα στάδια τραχηλικής αλλοίωσης.

Πίνακας 1.3: Ανίχνευση της p16 στα διάφορα στάδια τραχηλικής αλλοίωσης.

P16 Positivity in cervical cell samples and percentage of p16-positive samples per cytological category.

Study	p16 Positivity		% of p16 Positive samples			
	n or c	cut-off (+)	WNL	ASCUS	LSIL	HSIL
Bibbo 2002	n+c	>10 AC	0%	NA	74%	96%
Saqi 2002	n±c	>10 cells	21%	50%	74%	90%
Bibbo 2003	n+c	>10 AC	NA	NA	NA	63%
Murphy 2003	ND	ND	0%	100%	86%	100%
Pientong 2003	n+c	⩾3 cells	0%	57%	33%	93%
Akpolat 2004	n±c	⩾10% cells	0%	10%	NA	42%
Guo 2004	n±c	⩾5 AC	NA	NA	58%	97%
Pientong 2004	n+c	⩾10 cells	0%	53%	54%	100%
Sahebali 2004	n+c	ND	NA	NA	NA	N A
Trunk 2004	ND	ND	11%	67%	79%	100%
Yoshida 2004	n+c	⩾1 AC	NA	13%	58%	100%
Bose 2005	n+c	>10 cells	13%	29%	21%	81%
Filho 2005	ND	Any staining	49%	56%	NA	NA
Moore 2005	ND	>30% AC	0%	14%	44%	100%
Nieh 2005	ND	Any staining	NA	61%	NA	NA
Wentzensen 2005	n±c	NS>2	1%	NA	10%	98%
Ekalaksananan 2006	ND	⩾10 cells	7%	63%	60%	100%
Holladay 2006	n or c	⩾1 AC	0%	35%	42%	80%
Negri 2006	n or c	ND	13%	56%	67%	100%
Sahebali 2006	n+c	ND	NA	NA	NA	NA
Bollanca 2007	n±c	>1% cells	2%	NA	35%	96%
Carozzi 2007	ND	ND	35%	50%	40%	80%
Carydis 2007	ND	>10 cells	10%	NA	NA	NA
Liu 2007	n±c	>10 cells	20%	NA	50%	82%
Meyer 2007	n±c	⩾1 AC	9%	18%	42%	81%
Monsonego 2007	n or c	Any staining	36%	55%	28%	88%
Wentzensen 2007	n	NS > 2	NA	27%	24%	NA
Pooled value (random effect model) (95% CI)			12% (7–17%)	45% (35–54%)	45% (37–57%)	89% (84–95%)

AC: atypical cells, ASCUS: atypical squamous cells of undetermined significance, c: cytoplasmic staining, cut-off (+): number or percentage of cells stained positive for p16, HSIL: high-grade squamous intraepithelial lesion, CI: confidence interval, LSIL: low-grade squamous intraepithelial lesion, n: nuclear staining, NA: not-applicable, ND: not-documented, NS: nuclear scoring, WNL: within normal limits.

Στον πίνακα 1.3 παρουσιάζονται συγκεντρωτικά αποτελέσματα μελετών που πραγματέυονται το ίδιο θέμα. Γίνεται φανερό πως η παρουσία της p16 έχει μεγαλύτερο αντίκτυπο σε περιπτώσεις που το test Pap είναι μη φυσιολογικό. Από την ανάλυση όλων των μελετών κατέληξαν λοιπόν στο γεγονός πως με χρήση διαστήματος εμπιστοσύνης 95% έχουμε θετικό p16 test σε ποσοστό 7-17% για περιπτώσεις φυσιολογικής κυτταρολογίας, 35-54% σε περιπτώσεις ASCUS, 37-57% σε περιπτώσεις LgSIL και 84-95% σε περιπτώσεις HgSIL. Το γεγονός ότι το ποσοστό αυξάνει ανάλογα με τη σοβαρότητα του περιστατικού έρχεται να επιβεβαιώσει όσα έχουν ειπωθεί για το p16 test.

Flow Cytometry test

Η μελέτη των Narimatsu και Patterson (*Πίνακας 1.5*) [57] για την ανίχνευση μέσω flow cytometry έδωσε τα εξής αποτελέσματα για HgSIL αλλοιώσεις:

- *ευαισθησία* = 83,3%
- *ειδικότητα* = 91,3%

Μάλιστα, όπως φαίνεται και στον Πίνακα 1.5, το flow cytometry έδωσε υψηλότερη ευαισθησία αλλά και ειδικότητα από το HPV DNA test (Hybrid Capture2). Και σε αυτή την περίπτωση η εκτίμηση για την ειδικότητα και την ευαισθησία γίνεται με δυαδική διάκριση των περιστατικών και συγκεκριμένα κατηγοριοποιούνται τα περιστατικά με βάση το αν είναι καλύτερα ή χειρότερα από HgSIL.

Πίνακας 1.4: Σύγκριση απόδοσης του Flow Cytometry (E6/E7 mRNA) και του HPV DNA test (Hybrid Capture II).

Comparison of HPV FISH *E6* and *E7* mRNA and HCII in a Low-Risk HPV Cohort (n = 149)*

Papanicolaou Result	Intracellular *E6* and *E7* mRNA	Hybrid Capture II
WNL (n = 109)	10	13
ASCUS (n = 21)	14	9
LSIL (n = 5)	4	3
HSIL (n = 12)	8	8
Squamous cell carcinoma (n = 2)	2	1

ASCUS, atypical squamous cells of undetermined significance; FISH, fluorescence in situ hybridization; HCII, Hybrid Capture II; HPV, human papillomavirus; HSIL, high-grade squamous intraepithelial lesion; LSIL, low-grade squamous intraepithelial lesion; mRNA, messenger RNA; WNL, within normal limits.

* For intracellular *E6* and *E7* mRNA and Hybrid Capture II, respectively, sensitivity, ≥HSIL, 71.4% and 64.2%, and specificity, 91% and 88%.

Αφού μπορέσαμε να πάρουμε μια πρώτη ιδέα για το κάθε test ξεχωριστά, τίθεται τώρα το ζήτημα τι συμβαίνει εάν αντί για μία εξέταση χρησιμοποιήσουμε περισσότερες της μίας εξετάσεις. Για το λόγο αυτό θα παρατεθούν στο σημείο αυτό μερικές μελέτες που πραγματεύονται ένα τέτοιο θέμα συνδυασμού εξετάσεων καθώς και τα αποτελέσματά τους. Μία ενδιαφέρουσα μελέτη, η οποία προσπάθησε να προβεί σε μία συγκριση τεχνικών ώστε να επιτύχουμε καλύτερα αποτελέσματα, είναι η μελέτη των H. De Vuyst και P. Claeys με τίτλο "*Comparison of pap smear, visual inspection with acetic acid, human papillomavirus DNA-PCR testing and cervicography*". Σκοπός της μελέτης αυτής ήταν η αξιολόγηση τεσσάρων μεθόδων ανίχνευσης τραχηλικών ενδοεπιθηλιακών αλλοιώσεων σε περιοχή της Αφρικής, μεταξύ των οποίων ήταν το τεστ Pap και το HPV DNA test. Τα αποτελέσματά της για τα δύο προαναφερθέντα tests ήταν [52]:

- Τεστ Pap

  Ειδικότητα: 83,3 %
  Ευαισθησία: 94,6 %

- HPV DNA test

  Ειδικότητα: 94,4 %
  Ευαισθησία: 73,9 %

Το συμπέρασμα της μελέτης αυτής ήταν ότι το test Pap παρουσίαζε μεγαλύτερη ειδικότητα στις περιπτώσεις των νεοπλασιών, ενώ αντίθετα το HPV DNA test μεγαλύτερη ευαισθησία.

Μία ακόμη μελέτη που κινείται προς αυτή την κατεύθυνση είναι η μελέτη των T. Ekalaksananan και C. Pientong με τίτλο "*Usefulness of combining testing for p16 protein and human papillomavirus (HPV) in cervical carcinoma screening*". Στο πλαίσιο αυτής της μελέτης συμμετείχαν 186 ασθενείς με μη φυσιολογικό αποτέλεσμα στο τεστ Pap. Στόχος ήταν η εκτίμηση της p16 και του HPV DNA test κατά το screening. Τα αποτελέσματα συγκρίνονταν πάντα με τη βιοψία, η οποία θεωρείται ότι είναι το αποτέλεσμα το οποίο τυγχάνει της μεγαλύτερης εμπιστοσύνης από τους γιατρούς. Τα αποτελέσματα στα οποία κατέληξαν έδειξαν ότι θετικές εμφανίζονται οι εξετάσεις για το p16 test όσο και το HPV DNA test στις περιπτώσεις σοβαρής ενδοτραχηλικής νεοπλασίας και καρκινώματος. Επίσης, η παρουσία των p16 και HPV μπορεί να αποτελεί κάποια ένδειξη για την περεταίρω πορεία της ασθένειας, ακόμα και σε περιπτώσεις φυσιολογικές ή ήπιας τραχηλικής αλλοίωσης. Το συμπέρασμα στο οποίο κατέληξαν στην μελέτη αυτή αυτή ήταν ότι συνδυάζοντας το p16 test και το HPV DNA test έχει μεγάλη χρησιμότητα για την ανίχνευση ασθενειών που εμφανίζουν μεγάλη προδιάθεση για την ανάπτυξη καρκίνου [53]. Τα αποτελέσματα της μελέτης αυτής φαίνονται στον παρακάτω πίνακα:

Πίνακας 1.5: Αποτελέσματα από μη φυσιολογικά τραχηλικά κύτταρα με χρήση Pap test, p16 test και HPV DNA test

Conventional Pap test	Number	P16 protein detection		HPV DNA detection	
		Positive (%)	Negative (%)	Positive (%)	Negative (%)
Normal	148	11 (7.4)	137 (92.6)	5 (3.4)	143 (96.6)
ASCUS	13	8 (61.5)	5 (38.5)	0 (0)	13 (100)
Low-grade dysplasia					
ASC-H	6	4 (66.7)	2 (33.3)	2 (33.3)	4 (66.7)
LSIL	5	3 (60.0)	2 (40.0)	1 (20.0)	4 (80.0)
High-grade dysplasia					
HSIL	12	12 (100)	0 (0)	12 (100)	0 (0)
SCC	2	2 (100)	0 (0)	2 (100)	0 (0)
Total	186	40 (21.5)	146 (78.5)	22 (11.8)	164 (88.2)

ASCUS, atypical squamous cells of undetermined significance; ASC-H, atypical squamous cells cannot exclude HSIL; LSIL, low squamous intraepithelial lesion; HSIL, high squamous intraepithelial lesion; SCC, squamous cervical carcinoma.

Στη μελέτη της F.M. Carozzi με τίτλο "*Combined Analysis of HPV DNA and p16INK4a Expression to Predict Prognosis in ASCUS and LSIL Pap Smears*" οι συγγραφείς συνδυάζουν το HPV DNA test με το p16 test με στόχο τη σωστότερη πρόγνωση της κλινικής κατάστασης σε περιπτώσεις που το τεστ Pap είναι είτε ASCUS είτε LgSIL. Τα αποτελέσματα της μελέτης αυτής φαίνονται στους δύο πίνακες που ακολουθούν [43].

Πίνακας 1.6: Αποτελέσματα του HPV DNA test σε αναλογία με την κυτταρολογία

	N (%) of HR-HPV negative	N (%) of HR-HPV positive	Total – N (%)
Normal	100 (92.6)	89 (83.2)	189 (87.9%)
ASCUS	3 (2.8)	2 (1.9)	5 (2.3%)
LSIL	3 (2.8)	10 (9.3)	13 (6.04%)
HSIL	2 (1.8)	6 (5.6)	8 (3.7%)
Total	108 (50.2%)	107 (49.8%)	215

Πίνακας 1.7: Αποτελέσματα του p16 test σε αναλογία με την κυτταρολογία

	N (%) of $p16^{INK4a}$ negative	N (%) of $p16^{INK4a}$ positive	Total – N (%)
Normal	58 (87.9)	31 (77.5)	89 (84.0)
ASCUS	1 (1.5)	1 (2.5)	2 (1.9)
LSIL	6 (9.1)	4 (10)	10 (9.4)
HSIL	1 (1.5)	4 (10)	5 (4.7)
Total	66 (62.3)	40 (37.7)	106

Το συμπέρασμα στο οποίο κατέληξε η μελέτη αυτή ήταν ότι το HPV DNA test παρουσιάζει μεγάλη χρησιμότητα σε αμφιλεγόμενες περιπτώσεις του test Pap. Παρόλα αυτά όμως δεν προέκυψε μεγάλη ειδικότητα για το HPV DNA test σε περιπτώσεις σοβαρής CIN. Αντίθετα το p16 test εμφανίζει μεγαλύτερη χρησιμότητα σαν εργαλείο ιατρικής διαλογής για ασθενείς με LgSIL.

Ενδιαφέρον παρουσιάζει ακόμα η μελέτη "*The Sensitivity and Specificity of p16 Cytology vs HPV Testing for Detecting High-Grade Cervical Disease in the Triage of ASC-US and LSIL Pap Cytology Results*" των K. Denton και C. Bergeron. Στο πλαίσιο αυτής, αναλύθηκε η απόδοση του p16 test σε μια σειρά από 810 γυναίκες με test Pap ASC-US ή LgSIL. Ακόμα, για τις συγκεκριμένες γυναίκες ήταν διαθέσιμο και το HPV DNA test καθώς και το αποτέλεσμα της βιοψίας. Τα αποτελέσματα της μελέτης αυτής παρουσιάζονται στον παρακάτω πίνακα.

Πίνακας 1.8: Αποτελέσματα της εκτίμησης ιστολογίας με p16 test και HPV DNA test σε 810 περιπτώσεις ASCUS και LgSIL

	Sensitivity (%)	95% CI (%)	Specificity (%)	95% CI (%)
ASC-US, CIN 2+				
p16				
Path, morphology+	76.5	65.8-85.2	71.1	65.6-76.1
Path, score 2+	78.8	68.2-87.1	65.5	59.8-70.8
CT, morphology+	92.6	84.6-97.2	63.2	57.5-68.6
HPV	90.1	81.5-95.6	37.8	32.4-43.5
ASC-US, CIN 3				
p16				
Path, morphology+	75.9	62.4-86.5	67.1	61.7-72.1
Path, score 2+	83.0	70.2-91.9	62.5	57.1-67.8
CT, morphology+	92.6	82.1-97.9	58.6	53.1-64.0
HPV	88.9	77.4-95.8	35.3	30.2-40.8
LSIL, CIN 2+				
p16				
Path, morphology+	80.1	72.6-86.4	47.0	41.1-53.0
Path, score 2+	76.4	68.5-83.2	53.3	47.4-59.2
CT, morphology+	92.2	86.5-96.0	37.3	31.7-43.2
HPV	95.7	91.0-98.4	18.5	14.2-23.5
LSIL, CIN 3				
p16				
Path, morphology+	81.1	70.3-89.3	42.0	36.8-47.4
Path, score 2+	81.1	70.3-89.3	48.7	43.4-54.1
CT, morphology+	94.6	86.7-98.5	32.2	27.3-37.4
HPV	95.9	88.6-99.2	15.9	12.2-20.1

Από τον παραπάνω πίνακα βλέπουμε τα αποτελέσματα για την ευαισθησία και την ειδικότητα που προέκυψαν. Παρατηρήθηκαν πολύ υψηλά νούμερα στην ευαισθησία περιστατικών με βιοψία CIN-2/3, πιο συγκεκριμένα 92,6% για περιπτώσεις ASC-US και 92,2% για περιπτώσεις LSIL. Η εκτίμηση στην οποία κατέληξαν ήταν ότι η παρουσία της p16 επηρεάζει σημαντικά το *triage* των περιπτώσεων ASC-US και LgSIL.

Μία άλλη μελέτη [58], αυτή των M. Bevenolo και A. Vocaturo με τίτλο "*Sensitivity, Specificity and Clinical Value of Human Papillomavirus (HPV) E6/E7 mRNA Assay as a Triage Test for Cervical Cytology and HPV DNA Test*", πραγματεύεται το θέμα του συνδυασμού του HPV DNA test και του mRNA test. Στο πλαίσιο αυτής έγινε μια επισκόπηση

με στόχο να εκτιμηθεί η απόδοση ενός εργαλείου, του PreTect HPV-Proofer E6/E7 mRNA, ως μέσο για το triage για την κυτταρολογία και το HPV DNA test. Αναλύθηκαν 1.201 γυναίκες, εκ των οποίων οι 688 είχαν υποβληθεί σε κολποσκόπηση και οι 195 είχαν διαγνωσθεί με υψηλού βαθμού τραχηλική ενδοεπιθηλιακή νεοπλασία (CIN-2+). Τα αποτελέσματα στα οποία κατέληξαν φαίνονται στον ακόλουθο πίνακα.

Πινακας 1.9: Ακρίβεια του HPV, mRNA και DNA test σε 912 γυναίκες με μη φυσιολογική ιστολογία

Cytology	No. of women tested	No. of women with colposcopy follow-up	No. of CIN2+ women	Test	Test positivity[b]		Sensitivity[c]		Specificity[d]		PPV		NPV	
					%	95% CI	%	95% CI	%	95% CI	%	95% CI	%	95% CI
ASC-US	238	136	26	HPV mRNA	21	17–26	83	63–94	82	73–89	46	31–63	94	87–98
				HPV DNA	62	56–68	99	83–100	29	21–39	22	15–32	99	87–100
L-SIL	472	289	51	HPV mRNA	31	26–35	62	47–75	76	70–81	36	26–47	91	86–94
				HPV DNA	85	81–88	91	80–97	13	9–18	19	14–24	93	80–99
ASC-US and L-SIL	755	425	77	HPV mRNA	27	24–30	67	52–75	45	73–82	80	31–48	31	88–95
				HPV DNA	76	73–79	93	85–98	18	14–22	20	16–25	97	90–100
H-SIL+	157	138	105	HPV mRNA	67	59–74	67	57–76	45	45–84	80	70–88	31	18–45
				HPV DNA	95	90–98	96	91–99	4	0–21	77	68–83	33	4–78

[a] 95% CI, 95% confidence interval; ASC-US,atypical squamous cells of undetermined significance; L-SIL, low-grade squamous intraepithelial lesion; H-SIL, high-grade squamous intraepithelial lesion; CIN2+, cervical intraepithelial neoplasia grade 2 or more severe diagnosis; PPV, positive predictive value; NPV, negative predictive value.
[b] Test positivity and its 95% CI are calculated based on the total number of women tested.
[c] Sensitivity and its 95% CI are calculated based on the number of CIN2+ women, adjusted by follow-up completeness.
[d] Specificity and its 95% CI are calculated based on the number of CIN2− women, i.e., (the number of women with colposcopy follow-up) − (the number of CIN2+ women), adjusted by follow-up completeness.

Το συμπέρασμα στο οποίο κατέληξαν ήταν ότι το συγκεκριμένο εργαλείο μπορεί να προσφέρει καλύτερες πληροφορίες για το triage σε σχέση με το HPV DNA test στις περιπτώσεις ASC-US και LgSIL. Ακόμα θεωρείται πιο αποδοτικό από την κυτταρολογική εξέταση για το triage γυναικών με θετικό αποτέλεσμα του HPV DNA test. Παρόλα αυτά όμως παρουσιάζει χαμηλή ευαισθησία, πράγμα που επιβάλλει και προσεκτική και αυστηρή παρακολούθηση των ασθενών με θετικό HPV DNA test και αρνητικό mRNA test.

Ακόμα μια μελέτη [65] που ασχολείται με το θέμα του συνδυασμού του τεστ Pap με το HPV DNA test είναι αυτή του S. Costa et al. με τίτλο "*Human papillomavirus (HPV) test and Pap smear as predictor of outcome in conservatively treated adenocarcinoma in situ (AIS) of the uterine cervix*". Σκοπός της μελέτης αυτής είναι η αξιολόγηση της κλινικής κατάστασης ασθενών που έχουν ακολουθήσει συντηρητική αγωγή για in situ αδενοκαρκίνωμα, καθώς και η απόδοση της κυτταρολογίας και του HPV DNA test στην ανίχνευση της νόσου μετά από θεραπεία. Τα αποτελέσματα στα οποία κατέληξαν φαίνονται στον παρακάτω πίνακα.

Πίνακας 1.10: Απόδοση του test Pap και του HPV DNA test σε ασθενείς που έχουν λάβει αγωγή για ενδοτραχηλικό καρκίνωμα in situ

Test/FU visit	Sensitivity[a]	Specificity	PPV	NPV	OR (95% CI)
First FU visit					
PAP smear	60.0 (26.2–87.8)	68.7 (41.3–88.9)	54.6 (23.4–83.2)	73.3 (44.9–92.2)	3.3 (0.63–17.16)
HPV test	90.0 (55.5–99.7)	58.3 (27.7–84.8)	64.3 (35.1–87.2)	87.5 (47.3–99.6)	12.6 (1.18–133.89)
Second FU visit					
PAP smear	66.7 (22.2–95.6)	73.7 (48.8–90.8)	44.4 (13.7–78.8)	87.5 (61.6–98.4)	5.6 (0.77–40.59)
HPV test	83.3 (35.8–99.5)	58.8 (32.9–81.5)	41.7 (15.1–72.3)	90.9 (58.7–99.7)	7.1 (0.67–75.21)
Third FU visit					
PAP smear	0.0 (0.0–84.2)	95.4 (77.2–99.8)	0.0 (0.0–97.5)	91.3 (71.9–98.9)	1.1 (0.96–1.24)
HPV test	0.0 (0.0–84.3)	91.7 (73.0–98.9)	0.0 (0.0–84.2)	91.7 (73.0–98.9)	1.1 (0.97–1.23)
Fourth FU visit	NC	NC	NC	NC	NC
Fifth FU visit	NC	NC	NC	NC	NC
Sixth FU visit	NC	NC	NC	NC	NC

Τα αποτελέσματα έδειξαν ότι το HPV DNA test σε συνδυασμό με την κυτταρολογία προσφέρει σημαντικά πλεονεκτήματα σε σχέση με τη χρησιμοποίηση μόνο ενός τεστ για την εξαγωγή συμπερασμάτων σε ό,τι αφορά τη διάγνωση της κλινικής κατάστασης στις περιπτώσεις αυτές.

Λαμβάνοντας υπόψη τον αριθμό των μελετών αλλά και τα αποτελέσματα αυτών, σε συνδυασμό με την πρόσφατη παραγωγή εμβολίου για τη νόσο, καταδεικνύεται η σοβαρότητα του προβλήματος του καρκίνου του τραχήλου της μήτρας. Στη συνέχεια θα δώσουμε ένα περιληπτικό πλάνο στο οποίο συμπυκνώνεται η γενικότερη ιδέα της παρούσας διπλωματικής εργασίας αλλά και ο τρόπος με τον οποίο χειριστήκαμε το θέμα αυτό.

1.3.2 Γενικό πλάνο της διπλωματικής εργασίας

Ο **σκοπός** της συγκεκριμένης διπλωματικής εργασίας και η **καινοτομία** που προσδοκούμε να εισάγει, είναι το γεγονός ότι χρησιμοποιούνται **εξελιγμένα μαθηματικά εργαλεία βασισμένα σε τεχνικές υπολογιστικής νοημοσύνης** για την εξαγωγή χρήσιμης πληροφορίας από τον συνδυασμό όλων των διαθέσιμων τεχνικών ανίχνευσης με σκοπό να οδηγηθούμε σε καλύτερα αποτελέσματα για τη διάγνωση ασθενών με τραχηλικές ενδοεπιθηλιακές αλλοιώσεις. Μ’ αυτόν τον τρόπο, θέλουμε να βοηθήσουμε τους γιατρούς να πάρουν μια έγκυρη απόφαση για το πώς θα αντιμετωπίσουν εξατομικευμένα την καθεμία ασθενή (*triage*). Η διαφορά της συγκεκριμένης μελέτης από άλλες μελέτες είναι ότι αφενός χρησιμοποιούνται εξελιγμένα μαθηματικά εργαλεία για τον συνδυασμό των διαφόρων εξετάσεων, αφετέρου χρησιμοποιούνται και τα τέσσερα νέα screening tests (HPV DNA, mRNA, p16 και Flow Cytometry) και όχι μόνο ένα ή δύο, όπως συμβαίνει στις μελέτες που ανεφέρθηκαν προηγουμένως αλλά και στις περισσότερες μελέτες που έχουν δημοσιευθεί. Πιο συγκεκριμένα, λάβαμε υπ’ όψιν μας ένα σύνολο από 50 συνολικά αποτελέσματα των 5 διαφορετικών εξετάσεων (Pap test, HPV DNA test, mRNA test, p16 test και Flow

Cytometry). Τελικός στόχος της παρούσας διπλωματικής είναι η πρόταση ενός υποσυνόλου χαρακτηριστικών (*feature subset*) και η δημιουργία ενός μοντέλου εκτίμησης κινδύνου , για την αντιμετώπιση αλλά και το *triage* ασθενών με τραχηλικές ενδοεπιθηλιακές αλλοιώσεις.

Προς την κατεύθυνση αυτή, εφαρμόσαμε μια σειρά από **προηγμένες τεχνικές μείωσης των χαρακτηριστικών, θεωρίας πληροφορίας, βελτιστοποίησης και ταξινόμησης** σε μια βάση δεδομένων ενός δείγματος 212 γυναικών από το Αττικό Νοσοκομείο και από το Νοσοκομείο Ιωαννίνων, οι οποίες διαγνώστηκαν με κάποιας μορφής νεοπλασία (CIN-1,CIN-2/3) και ταυτόχρονα το test Pap ήταν είτε LgSIL είτε HgSIL. Σαν εξέταση αναφοράς (reference test) για την εξακρίβωση της ορθότητας των αποτελεσμάτων,χρησιμοποιήθηκε η ιστολογική βιοψία. Το κυτταρολογικό επίχρισμα λήφθηκε σε LBC μορφή. Σύμφωνα με τις μεθόδους που χρησιμοποιήθηκαν, έχουμε τυποποίηση 35 HPV τύπων (υψηλού και χαμηλού κινδύνου), ανίχνευση του mRNA των τύπων 16,18,31,33,45, ανίχνευση της πρωτεΐνης p16 που υπερεκφράζεται στον καρκίνο του τραχήλου και κυτταρομετρία ροής για ανίχνευση του mRNA των HPV τύπων υψηλού κινδύνου. Στο 2^{ο} Κεφάλαιο που ακολουθεί παρουσιάζονται οι τεχνικές που χρησιμοποιήσαμε για την επίλυση του προβλήματος μας. Στο 3^{ο} Κεφάλαιο περιγράφεται λεπτομερώς ο τρόπος εφαρμογής των τεχνικών του 2^{ου} Κεφαλαίου και παρουσιάζονται τα αποτελέσματα και τα συμπεράσματα στα οποία καταλήξαμε.

Κεφάλαιο 2

Τεχνικές Επιλογής Χαρακτηριστικών και Ταξινόμησης

Στο κεφάλαιο αυτό, όπως υποδηλώνεται και από τον τίτλο του, θα αναπτυχθεί η θεωρία για τεχνικές επιλογής χαρακτηριστικών και ταξινόμησης. Οι τεχνικές αυτές αποτελούν τα εργαλεία τα οποία χρησιμοποιήσαμε για την επίλυση του προβλήματός μας. Πρόκειται για τεχνικές τόσο μαθηματικά όσο και πρακτικά τεκμηριωμένες, καθώς έχουν εφαρμοστεί σε μελέτες διαφόρων επιστημονικών τομέων με επιτυχή αποτελέσματα. Πιο συγκεκριμένα θα αναπτυχθούν θέματα που άπτονται των παρακάτω τομέων:

- Επιλογή Χαρακτηριστικών
- Θεωρία Πληροφορίας
- Γενετικοί Αλγόριθμοι, και
- Μέθοδοι Αναγνώρισης Προτύπων

Αυτές είναι οι τεχνικές που θα μας οδηγήσουν, όπως προαναφέρθηκε, στην εξόρυξη πληροφορίας και οπότε στην πρόταση ενός υποσυνόλου των διαθέσιμων χαρακτηριστικών καθώς και στη δημιουργία ενός μοντέλου εκτίμησης κινδύνου για την αντιμετώπιση των περιστατικών τραχηλικής ενδοεπιθηλιακής νεοπλασίας. Πιο συγκεκριμένα, με χρήση τεχνικών επιλογής χαρακτηριστικών με κριτήρια από τη θεωρία πληροφορίας και χρήση γενετικών αλγορίθμων για την εξασφάλιση της βελτιστότητας θα γίνει η μείωση των χαρακτηριστικών και έτσι θα μπορέσουμε να προτείνουμε ένα πολύ μικρότερο σύνολο χαρακτηριστικών σε σχέση με πλήρες σύνολο των 50 που έχουν στη διάθεσή τους οι γιατροί. Επιπρόσθετα, με χρήση τεχνικών αναγνώρισης προτύπων θα επιβεβαιώσουμε ότι το σύνολο που θα παραχθεί από τα προηγούμενα έχει τα αναμενόμενα αποτελέσματα, αλλά επίσης θα μπορέσουμε να προχωρήσουμε και στην παραγωγή των ζητούμενων κανόνων εκτίμησης κινδύνου για την αντιμετώπιση της νόσου.

2.1 Επιλογή Χαρακτηριστικών (Feature Selection)

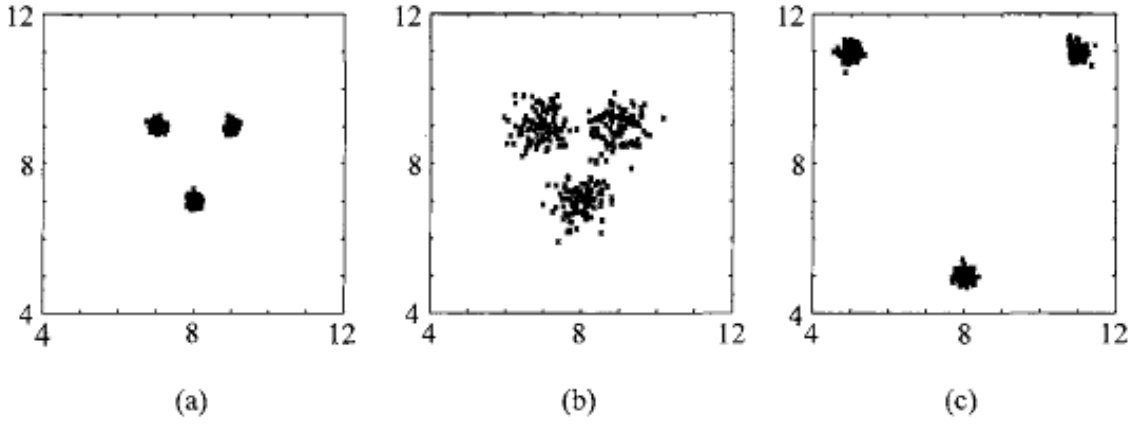
Ένα μείζον πρόβλημα στη σχεδίαση συστημάτων αναγνώρισης προτύπων είναι η λεγόμενη μάστιγα των διαστασιμότητας (*curse of dimensionality*). Ο αριθμός των χαρακτηριστικών που είναι στη διάθεση των σχεδιαστών τέτοιων συστημάτων είναι συνήθως αρκετά μεγάλος. Οι λόγοι για τους οποίους επιθυμούμε να μειώσουμε τον αριθμό των χαρακτηριστικών σε έναν επαρκή ελάχιστο αριθμό είναι πολλοί. Ο προφανής λόγος είναι η υπολογιστική πολυπλοκότητα. Ένας άλλος παρεμφερής λόγος είναι ότι ο συνδυασμός κάποιων χαρακτηριστικών μεταξύ τους σε ένα κοινό διάνυσμα χαρακτηριστικών λόγω υψηλής αμοιβαίας συσχέτισης, μπορεί να οδηγήσει σε κάποια κέρδη σε ό,τι αφορά την ταξινόμηση με τίμημα όμως τη μεγαλύτερη υπολογιστική πολυπλοκότητα. Τέλος, όπως θα δούμε, η μείωση του αριθμού των χαρακτηριστικών επιβάλλεται και από τις απαραίτητες ιδιότητες γενίκευσης που θα πρέπει να έχει το σύστημα ταξινόμησής μας [68].

Ένας μεγάλος αριθμός χαρακτηριστικών μεταφράζεται ουσιαστικά και σε έναν μεγάλο αριθμό παραμέτρων του συστήματος ταξινόμησης. Για το λόγο αυτό, δεδομένου ενός αριθμού *N* μοτίβων εκπαίδευσης (*training patterns*), η προσπάθεια για να επιλέξουμε έναν όσο το δυνατόν μικρότερο αριθμό χαρακτηριστικών συμβαδίζει με την επιθυμία μας για σχεδίαση συστημάτων με καλές ιδιότητες γενίκευσης. Ακόμα, σημαντικό βήμα για τη σχεδίαση ενός συστήματος ταξινόμησης είναι το στάδιο της εκτίμησης της απόδοσής του (*performance evaluation stage*), καθώς δεν αρκεί μόνο η σχεδίαση του συστήματος αλλά πρέπει επίσης να έχουμε αξιολογήσει και την απόδοσή του.

Η ουσία και η χρησιμότητα της επιλογής χαρακτηριστικών μπορεί να συνοψισθεί στην ακόλουθη φράση:

"*Δεδομένου ενός αριθμού χαρακτηριστικών, με ποιον τρόπο μπορεί κάποιος να επιλέξει τα πιο σημαντικά από αυτά, ώστε να μειώσει τον αριθμό τους και ταυτόχρονα να διατηρεί όσο το δυνατόν περισσότερη από την πληροφορία που αυτά φέρουν για τη διάκριση μεταξύ των κλάσεων*;"

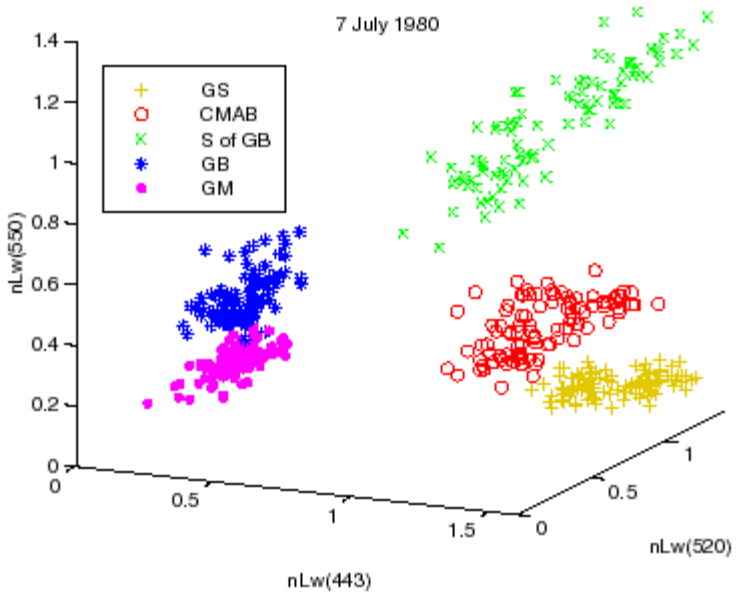
Η διαδικασία αυτή είναι γνωστή ως **επιλογή ή μείωση χαρακτηριστικών** (*feature selection/reduction*), και είναι εξαιρετικά σημαντική και κρίσιμη για ένα σύστημα ταξινόμησης. Εάν επιλέξουμε χαρακτηριστικά με μικρή διακριτική ισχύ/ικανότητα (*discrimination power*), η σχεδίαση θα οδηγήσει σε ένα σύστημα με πολύ χαμηλή απόδοση, ενώ αν επιλεγούν χαρακτηριστικά που φέρουν μεγάλη ποσότητα πληροφορίας η σχεδίαση απλοποιείται σε μεγάλο βαθμό. Τελικά, σε μια πιο ποιοτική περιγραφή του προβλήματος, θα μπορούσαμε να πούμε πως επιθυμούμε την επιλογή χαρακτηριστικών που οδηγούν σε μεγάλη απόσταση μεταξύ των κλάσεων (*large between-class distance*) και μικρή διακύμανση εντός των κλάσεων (*small within-class variation*) [68].



Εικόνα 2.1*: κλάσεις με (a) μικρή within-class διακύμανση και μικρές between-class αποστάσεις (b) μεγάλη within-class διακύμανση και μικρές between-class αποστάσεις (c) μικρή within-class διακύμανση και μεγάλες between-class αποστάσεις*

2.1.1 Χώρος Χαρακτηριστικών (Feature Space)

Πριν προχωρήσουμε στην ανάλυση μεθόδων επιλογής χαρακτηριστικών (*feature selection*) θα πρέπει να δοθεί ένας σύντομος ορισμός για το τι είναι ο χώρος χαρακτηριστικών. Ο **χώρος χαρακτηριστικών** (*feature space*) στην αναγνώριση προτύπων είναι ο χώρος όπου κάθε δείγμα προτύπου αναπαριστάται σαν ένα σημείο ενός *N*-διάστατου χώρου. Η διάσταση *N* συνδέεται με τον αριθμό των χαρακτηριστικών που χρησιμοποιούνται για την περιγραφή του μοντέλου μας. Παρόμοια δείγματα συγκεντρώνονται μεταξύ τους, πράγμα το οποίο επιτρέπει τη χρήση της εκτίμησης πυκνότητας για την αναγνώριση των μοντέλων. Η επιλογή χαρακτηριστικών είναι μια τεχνική επιλογής ενός υποσυνόλου σχετικών χαρακτηριστικών για τη δημιουργία εύρωστων μοντέλων εκμάθησης, τα οποία βρίσκουν εφαρμογή κυρίως στην εκμάθηση μηχανών. Με την απομάκρυνση άσχετων και πλεοναζόντων χαρακτηριστικών από τα δεδομένα μας επιτυγχάνουμε τη βελτίωση της απόδοσης του συστήματός μας, και βοηθούμαστε να κατανοήσουμε καλύτερα τα δεδομένα, καταλαβαίνοντας ποια είναι τα σημαντικά χαρακτηριστικά και πώς αυτά συσχετίζονται μεταξύ τους.



***Εικόνα 2.2**: Παράδειγμα 3-διάστατου χώρου χαρακτηριστικών*

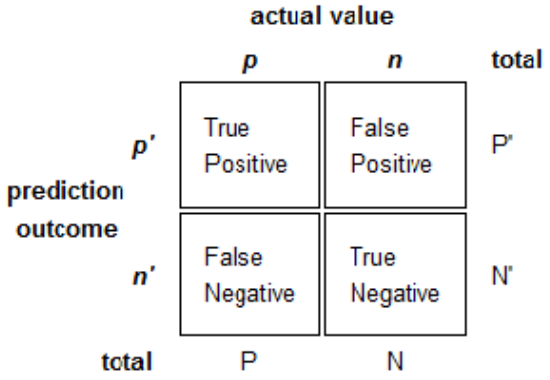
2.1.2 Καμπύλες ROC

Στη θεωρία ανίχνευσης σημάτων μια ROC (Receiver Operating Characteristic) καμπύλη αποτελεί τη γραφική αναπαράσταση της ευαισθησίας (s*ensitivity*) με την (1-ειδικότητα (*specificity*)) για έναν δυαδικό ταξινομητή του οποίου το κατώφλι ποικίλει. Η ευαισθησία και η ειδικότητα είναι στατιστικές μετρήσεις της επίδοσης μίας δοκιμής δυαδικής ταξινόμησης. Η ευαισθησία μετρά την αναλογία θετικών δειγμάτων που αναγνωρίστηκαν ορθά, ενώ η ειδικότητα μετρά την αναλογία αρνητικών δειγμάτων, τα οποία αναγνωρίστηκαν ορθά. Η ROC ανάλυση μας παρέχει τα εργαλεία ώστε να επιλέξουμε πιθανά βέλτιστα μοντέλα και ταυτόχρονα να απορρίψουμε τα υποβέλτιστα ανεξαρτήτως κόστους ή της κατανομής των κλάσεων. Έτσι γίνεται κατανοητό πως η ROC ανάλυση συνδέεται άμεσα με την ανάλυση κόστους-οφέλους σε ό,τι αφορά τη λήψη διαγνωστικής απόφασης (*diagnostic decision making*) [68].

Υποθέτουμε τώρα ένα δυαδικό πρόβλημα, του οποίου τα αποτελέσματα χαρακτηρίζονται είτε ως θετικά (Positive, p) είτε ως αρνητικά (Negative, n). Οι πιθανές καταστάσεις που μπορούν να προκύψουν από έναν τέτοιο ταξινομητή είναι 4:

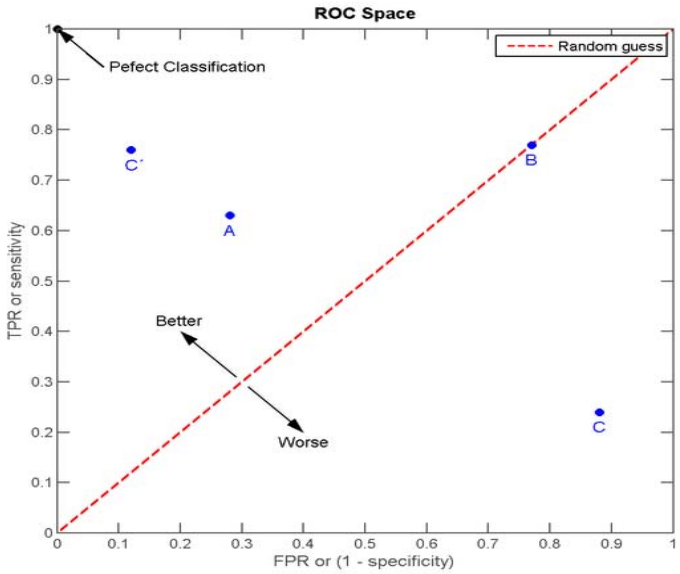
- ✓ True Positive (TP)
- ✓ True Negative (TN)
- ✓ False Positive (FP)
- ✓ False Negative (FN)

Με βάση τα παραπάνω θα μπορούσαμε να σχηματίσουμε έναν 2x2 πίνακα, ένα λεγόμενο **πίνακα σύγχυσης (*Confusion Matrix*)**. Ο πίνακας σύγχυσης αποτελεί ένα απεικονιστικό εργαλείο που χρησιμοποιείται κατά βάση στην εκμάθηση μηχανών με επίβλεψη. Κάθε γραμμή του αντιπροσωπεύει τις καταστάσεις στην προβλεπόμενη κλάση ενώ κάθε στήλη αντιπροσωπεύει τις καταστάσεις στην πραγματική κλάση, δηλαδή τις καταστάσεις που προέκυψαν από την αξιοποίηση των δεδομένων μας. Στην περίπτωση του δυαδικού ταξινομητή ο πίνακας σύγχυσης θα είχε τη μορφή που φαίνεται στην ακόλουθη εικόνα:



***Εικόνα 2.3**: Πίνακας σύγχυσης (Confusion Matrix)*

Για το σχεδιασμό μίας ROC καμπύλης χρειαζόμαστε μόνο τις τιμές του True Positive Rate (TPR) και False Positive Rate (FPR), καθώς αυτές οι τιμές είναι που καθορίζουν την επίδοση του συστήματος ταξινόμησης. Το TPR είναι ισοδύναμο με την ευαισθησία, ενώ το FPR είναι ισοδύναμο του παράγοντα (1 – ειδικότητα). Ο ROC χώρος έχει τη μορφή που φαίνεται στην παρακάτω εικόνα [68].



***Εικόνα 2.4**: ROC χώρος*

Η κόκκινη γραμμή στο σχήμα 2.4 απεικονίζει την περιοχή της τυχαίας επιλογής καθώς κατά μήκος αυτής το TPR και το FPR είναι ίσα, κατά συνέπεια δεν μπορούμε να κλίνουμε προς τη μία ή την άλλη κατάσταση. Όσο πιο κάτω βρισκόμαστε από την κόκκινη γραμμή τόσο χειρότερη είναι η επίδοση της ταξινόμησής μας. Αντίθετα όσο πιο πάνω, τόσο

καλύτερη ταξινόμηση έχουμε, με το σημείο (0,1) να αντιπροσωπεύει την *τέλεια ταξινόμηση* (*perfect classification*).

2.1.3 *Επιλογή χαρακτηριστικών με βάση τον στατιστικό έλεγχο υποθέσεων*

Ένα πρώτο βήμα για την επιλογή χαρακτηριστικών είναι να παρατηρήσουμε κάθε ένα χαρακτηριστικό ανεξάρτητα και να δούμε στην πράξη τη διαχωριστική ικανότητα που μπορεί να προσφέρει το καθένα. Το να εξετάζεται κάθε ένα χαρακτηριστικό μόνο του κάθε άλλο παρά βέλτιστο μπορεί να χαρακτηριστεί, παρ' όλα αυτά όμως μας βοηθάει στο να απορρίψουμε κάποια χαρακτηριστικά μειώνοντας το υπολογιστικό κόστος. Ο στατιστικός έλεγχος βασίζεται στην ύπαρξη μιας «μηδενικής» υπόθεσης H_0 και της «εναλλακτικής» υπόθεσης H_1. Θα προσπαθήσουμε να δείξουμε ποια από τις παρακάτω 2 υποθέσεις είναι σωστή.

H_0 : Οι τιμές του χαρακτηριστικού διαφέρουν σημαντικά

H_1 : Οι τιμές του χαρακτηριστικού δε διαφέρουν σημαντικά

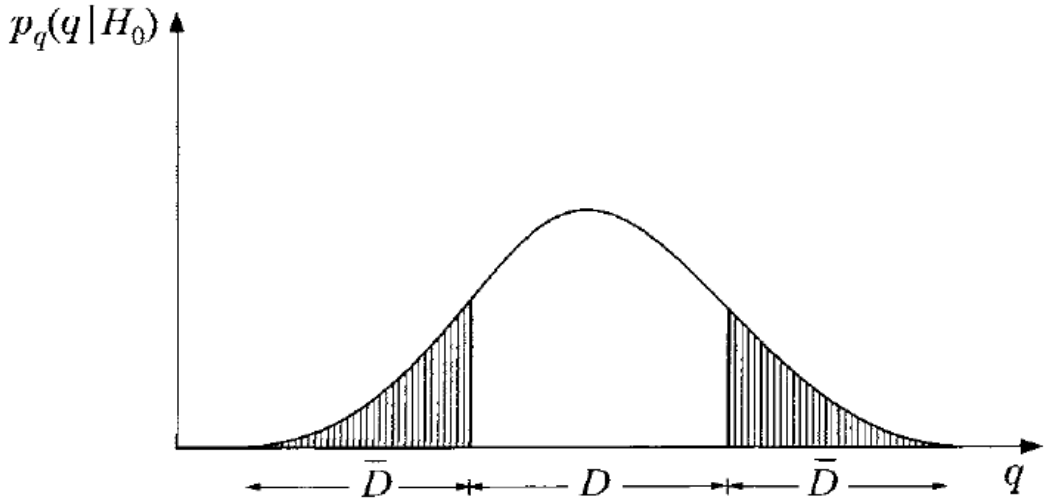
Θα προσεγγίσουμε το πρόβλημα υπολογίζοντας τις διαφορές από τις μέσες τιμές του κάθε χαρακτηριστικού στις διάφορες κλάσεις και στη συνέχεια θα ελέγξουμε το πόσο αυτές διαφέρουν από το μηδέν. Σε αυτό το σημείο πρέπει να υπενθυμίσουμε μερικές βασικές αρχές του ελέγχου υποθέσεων [68].

Έστω x μια τυχαία μεταβλητή με συνάρτηση πυκνότητας πιθανότητας (σππ), την οποία θεωρούμε γνωστή, με μια άγνωστη παράμετρο θ. Στην περίπτωση της γκαουσιανής κατανομής ως άγνωστη παράμετρος θ μπορεί να θεωρηθεί είτε η μέση τιμή είτε η διακύμανση. Η υπόθεση μας διατυπώνεται ως εξής:

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0$$

Η διαδικασία της απόφασης έχει ως εξής:

Έστω x_i, i=1,2,...,N τα πειραματικά δείγματα της τυχαίας μεταβλητής x. Με βάση το υπό εξέταση πρόβλημα διαλέγουμε μια συνάρτηση $f(\cdot,...,\cdot)$ και υποθέτουμε $q = f(x_1, x_2, ..., x_N)$. Η συνάρτηση επιλέγεται κατά τέτοιον τρόπο έτσι ώστε η σππ του q να μπορεί να παραμετροποιηθεί με βάση το θ, δηλαδή να είναι $p_q = (q, \theta)$. Θεωρούμε τώρα ως D το διάστημα στο οποίο θα εμφανίζεται μεγάλη πιθανότητα εκπλήρωσης της μηδενικής υπόθεσης H_0. Σαν $\bar{D}$ θεωρούμε το διάστημα με μικρή πιθανότητα εκπλήρωσης της H_0. Αν η τιμή του q, που προκύπτει από τα διαθέσιμα δείγματα, βρίσκεται εντός του διαστήματος D αποδεχόμαστε την H_0. Σε αντίθετη περίπτωση την απορρίπτουμε.



***Εικόνα 2.5** : Διάγραμμα στο οποίο απεικονίζονται σχηματικά τα διαστήματα αποδοχής (D) και απόρριψης ($\bar{D}$)*

Το προφανές ερώτημα που προκύπτει πλέον είναι αυτό το οποίο αφορά τη λανθασμένη απόφαση:

Έστω ότι η H_0 ισχύει. Η πιθανότητα λάθους στην απόφασή μας είναι $P(q \in \bar{D} | H_0) \equiv \rho$. Η τιμή του ρ καλείται «επίπεδο σημαντικότητας» (significance level) και την προκαθορίζουμε με βάση το ζητούμενο πρόβλημα στο οποίο εργαζόμαστε.

Ένα πρόβλημα ελέγχου υποθέσεων μπορεί να κατηγοριοποιηθεί με βάση τις 2 ακόλουθες περιπτώσεις:

- ✓ Γνωστή Διακύμανση
- ✓ Άγνωστη Διακύμανση

Στην περίπτωση γνωστής διακύμανσης μπορούμε να συνοψίσουμε τη διαδικασία λήψης απόφασης στα εξής βήματα:

- Υπολογισμός $\bar{x}$ και q δεδομένου του αριθμού N των δειγμάτων
- Επιλογή επιπέδου σημαντικότητας ρ
- Υπολογισμός μέσω των σχετικών πινάκων για $N(0,1)$ του διαστήματος αποδοχής $D=[-x_\rho, x_\rho]$ σχετικού με την πιθανότητα $1-\rho$
- Αν $q \in D$ αποφασίζεται το H_0, σε αντίθετη περίπτωση το H_1

Στην περίπτωση άγνωστης διακύμανσης υπολογίζουμε το q και παρατηρούμε πως αυτό δεν ακολουθεί πλέον γκαουσιανή κατανομή, αλλά ακολουθεί τη λεγόμενη t-κατανομή με $N-1$ βαθμούς ελευθερίας. Όπως γίνεται κατανοητό, η επιλογή χαρακτηριστικών με βάση το στατιστικό έλεγχο υποθέσεων μπορεί να αποτελέσει ένα πολύ σημαντικό εργαλείο για την εξόρυξη δεδομένων (*data mining*).

2.1.4 Μέτρα Διαχωρισιμότητας των Κλάσεων

Μέχρι στιγμής οι τεχνικές στις οποίες αναφερθήκαμε λαμβάνουν υπ' όψιν τους μόνο ένα χαρακτηριστικό. Αυτές οι τεχνικές όμως δεν συνυπολογίζουν το γεγονός της συσχέτισης που υπάρχει μεταξύ των διάφορων χαρακτηριστικών, η οποία επηρεάζει την ικανότητα διαχωρισμού των διανυσμάτων χαρακτηριστικών που έχουν σχηματιστεί. Στην παράγραφο αυτή θα επεκταθούμε στην αποδοτικότητα διαχωρισμού διανυσμάτων χαρακτηριστικών. Για τη μελέτη του προβλήματος αυτού μπορούμε να κινηθούμε με δύο τρόπους. Ο πρώτος τρόπος είναι να συνδυάσουμε χαρακτηριστικά με τρόπο που να παράγουμε εν τέλει το «καλύτερο» - αποδοτικότερο διάνυσμα χαρακτηριστικών δεδομένης μιας διάστασης l. Ο δεύτερος τρόπος είναι να μετασχηματίσουμε τα δεδομένα μας με βάση ένα κριτήριο βελτιστότητας με σκοπό να αποκτήσουμε χαρακτηριστικά τα οποία να μας προσφέρουν μεγάλη διαχωριστική ισχύ [68].

Μερικά μέτρα διαχωρισιμότητας των κλάσεων είναι :

- *<u>Απόκλιση</u> (Divergence)*

Η απόκλιση είναι ένας τελεστής που υπολογίζει το μέτρο της πηγής ή της καταβόθρας σε ένα συγκεκριμένο σημείο ενός διανυσματικού πεδίου ως ένα προσημασμένο διανυσματικό

μέγεθος. Δεδομένων 2 κλάσεων $\boldsymbol{\omega_1,\omega_2}$ και ενός διανύσματος χαρακτηριστικών $\boldsymbol{x}$, διαλέγουμε ένα χαρακτηριστικό αν ισχύει

$$P(\omega_1|x) > P(\omega_2|x)$$

Ο λόγος $\ln\frac{p(x|\omega_1)}{p(x|\omega_2)} \equiv D_{12}(x)$ μπορεί να χρησιμοποιηθεί ως ένα μέτρο διαχωριστικής ικανότητας των κλάσεων $\boldsymbol{\omega_1}$ *και* $\boldsymbol{\omega_2}$. Για πλήρως επικαλυπτόμενες κλάσεις ισχύει $D_{12}=0$. Εφόσον το x παίρνει διάφορες τιμές είναι λογικό να χρησιμοποιείται η μέση τιμή αυτού στις 2 κλάσεις, η οποία για κάθε κλάση δίνεται από τα ολοκληρώματα

$$D_{12} = \int_{-\infty}^{+\infty} p(x|\omega_1)\ln\frac{p(x|\omega_1)}{p(x|\omega_2)}dx$$

$$D_{21} = \int_{-\infty}^{+\infty} p(x|\omega_2)\ln\frac{p(x|\omega_2)}{p(x|\omega_1)}dx$$

Το άθροισμα $d_{12} = D_{12} + D_{21}$ είναι γνωστό ως απόκλιση και χρησιμοποιείται για το διαχωρισμό των κλάσεων $\boldsymbol{\omega_1}$ *και* $\boldsymbol{\omega_2}$. Το αποτέλεσμα γενικεύεται και για προβλήμα με περισσότερες των 2 κλάσεις.

- *Όριο Chernoff και απόσταση Bhattacharyya*

Το ελάχιστο σφάλμα ταξινόμησης ενός Bayesian ταξινομητή 2 κλάσεων μπορεί να γραφεί ως

$$P_e = \int_{-\infty}^{+\infty} \min[P(\omega_i)p(x|\omega_i), P(\omega_j)p(x|\omega_j)]dx$$

Ο αναλυτικός υπολογισμός ενός τέτοιου ολοκληρώματος δεν είναι δυνατός. Μπορούμε όμως να θέσουμε ένα πάνω όριο, του οποίου ο υπολογισμός βασίζεται στην ακόλουθη ανισότητα

$$\min[a,b] \le a^s b^{1-s} \quad \text{για } \alpha,\beta > 0 \text{ και } 0 \le s \le 1$$

Συνδυάζοντας τις 2 τελευταίες σχέσεις παίρνουμε

$$P_e \le P(\omega_i)^s P(\omega_j)^{1-s} \int_{-\infty}^{+\infty} p(x|\omega_i)^s p(x|\omega_j)^{1-s} dx \equiv \varepsilon_{CB}$$

Όπου το ε_{CB} είναι γνωστό ως όριο Chernoff (Chernoff bound). Το ελάχιστο όριο μπορεί να υπολογιστεί ελαχιστοποιώντας το ε_{CB} ως προς s. Για s=1/2 έχουμε μια ειδική μορφή του ορίου

$$P_e \le \varepsilon_{CB} = \sqrt{P(\omega_i)P(\omega_j)} \int_{-\infty}^{+\infty} \sqrt{p(x|\omega_i)p(x|\omega_j)}dx$$

Για γκαουσιανές κατανομές $N(\mu_i, \Sigma_i)$, $N(\mu_j, \Sigma_j)$ καταλήγουμε στην ακόλουθη σχέση

$$\varepsilon_{CB} = \sqrt{P(\omega_i)P(\omega_j)}\exp(-B)$$

$$\text{με } B = \frac{1}{8}(\mu_i - \mu_j)^T\left(\frac{\Sigma_i + \Sigma_j}{2}\right)(\mu_i - \mu_j) + \frac{1}{2}\ln\left(\frac{\left|\frac{\Sigma_i + \Sigma_j}{2}\right|}{\sqrt{|\Sigma_i||\Sigma_j|}}\right)$$

Ο όρος Β είναι γνωστός ως απόσταση *Bhattacharyya* και χρησιμοποιείται ως μέτρο διαχωρισμού κλάσεων.

- <u>*Πίνακες Διασποράς*</u> *(Scatter Matrices)*

Ορίζουμε τους ακόλουθους πίνακες

- *Within-class scatter matrix*

$$S_w = \sum_{i=1}^{M} P_i S_i$$

Όπου S_i είναι η μήτρα συμμεταβλητότητας για την κλάση $\boldsymbol{\omega_i}$

$$S_i = \mathrm{E}\{(x - \mu_i)(x - \mu_i)^T\}$$

και P_i η *a priori* πιθανότητα της κλάσης $\boldsymbol{\omega_i}$.

- *Between-class scatter matrix*

$$S_b = \sum_{i=1}^{M} P_i(\mu_i - \mu_0)(\mu_i - \mu_0)^{\mathrm{T}}$$

με μ_0 να είναι το ολικό μέσο διάνυσμα

$$\mu_0 = \sum_i^M P_i \mu_i$$

- *Mixture scatter matrix*

$$S_m = \mathrm{E}\{(x-\mu_0)(x-\mu_0)^{\mathrm{T}}\}$$

Η S_m είναι η μήτρα συμμεταβλητότητας του διανύσματος χαρακτηριστικών ως προς την ολική μέση μέση τιμή και ισχύει ότι

$$S_m = S_w + S_b$$

Με βάση τους παραπάνω ορισμούς παρατηρούμε ότι το κριτήριο $J_1 = \dfrac{trace(S_m)}{trace(S_w)}$ παίρνει μεγάλες τιμές όταν τα δείγματα στον l-διάστατο χώρο είναι ομαλά κατανεμημένα γύρω από τη μέση τιμή κάθε κλάσης και οι διαμερίσεις κάθε κλάσης είναι ευδιάκριτες. Αντίστοιχα αποτελέσματα θα μπορούσαμε να πάρουμε αν αντί για το ίχνος των πινάκων χρησιμοποιούσαμε τις ορίζουσες ($J_2 = \dfrac{|S_m|}{|S_w|} = |S_w^{-1} S_m|$). Ένα τρίτο κριτήριο θα μπορούσε να θεωρηθεί το $J_3 = trace\{S_w^{-1} S_m\}$. Τα κριτήρια J_2 και J_3 έχουν το πλεονέκτημα του να μην διαφοροποιούνται όταν υπόκεινται σε γραμμικούς μετασχηματισμούς, πράγμα πολύ χρήσιμο για την παραγωγή βέλτιστων χαρακτηριστικών.

Τέλος, ένας σημαντικός δείκτης για το κατά πόσο δύο κλάσεις διαχωρίζονται κατά επαρκή τρόπο είναι ο λεγόμενος *Fischer's discriminant ratio* (FDR) που ορίζεται ως

$$FDR = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

Ο FDR χρησιμοποιείται για να ποσοτικοποιήσει τις διαχωριστικές ικανότητες του κάθε χαρακτηριστικού. Ο FDR μπορεί να επεκταθεί και σε περιπτώσεις με περισσότερα των 2 χαρακτηριστικά και μια μορφή του κριτηρίου αυτού θα μπορούσε να είναι

$$FDR_{multicase} = \sum_i^M \sum_{j \neq i}^M \frac{(\mu_i - \mu_j)^2}{\sigma_i^2 + \sigma_j^2} \quad [68]$$

2.1.5 Επιλογή Υποσυνόλου Χαρακτηριστικών

Προηγουμένως ορίσαμε μερικά κριτήρια τα οποία μετρούν την αποδοτικότητα του κάθε ενός χαρακτηριστικού ξεχωριστά ή ενός διανύσματος χαρακτηριστικών. Βασιζόμενοι λοιπόν σε αυτά ήρθε η ώρα να περάσουμε στην καρδιά του προβήματος. Αυτή είναι η επιλογή ενός υποσυνόλου μήκους l χαρακτηριστικών από έναν αριθμό m διαθέσιμων. Οι βασικοί τρόποι για να κινηθούμε είναι δύο και θα αναλυθούν αντίστοιχα στα 2 υποκεφάλαια που ακολουθούν.

2.1.5.1 Βαθμωτή Επιλογή Χαρακτηριστικών

Τα χαρακτηριστικά αντιμετωπίζονται ξεχωριστά σε αυτή την περίπτωση. Μπορούν να υιοθετηθούν οποιαδήποτε από τα μέτρα διαχωρισιμότητας κλάσεων που αναφέρθηκαν προηγουμένως (ROC, απόκλιση, κ.ά.). Υπολογίζουμε την τιμή ενός κριτηρίου $C(k)$ για κάθε ένα χαρακτηριστικό $k = 1,2, \dots, m$. Εν συνεχεία, τα χαρακτηριστικά κατατάσσονται σε σειρά φθίνουσα. Τα l χαρακτηριστικά, που τελικά επιλέγονται, αντιστοιχούν στις l καλύτερες τιμές του κριτηρίου $C(k)$ και έπειτα προχωρούμε στο σχηματισμό ενός διανύσματος χαρακτηριστικών με αυτά [68].

Το κυριότερο πλεονέκτημα της ξεχωριστής αντιμετώπισης του κάθε ενός χαρακτηριστικού είναι η υπολογιστική απλότητα. Παρόλα αυτά όμως, αυτές οι προσεγγίσεις δεν λαμβάνουν υπ' όψιν τους τις υπάρχουσες συσχετίσεις (*correlations*) μεταξύ των χαρακτηριστικών [67]. Μερικά μέτρα τα οποία θα μπορούσαμε να χρησιμοποιήσουμε για την ιεράρχηση χαρακτηριστικών είναι:

- t-test
- Εντροπία
- Καμπύλη ROC
- Απόσταση bhattacharyya

2.1.5.2 Επιλογή Διανύσματος Χαρακτηριστικών

Η ξεχωριστή αντιμετώπιση των χαρακτηριστικών έχει όπως ειπώθηκε το πλεονέκτημα της υπολογιστικής απλότητας. Σε πολύπλοκα προβλήματα και σε περιπτώσεις όπου τα χαρακτηριστικά εμφανίζουν μεγάλες συσχετίσεις μεταξύ τους η τεχνική αυτή δεν θα ήταν αποδοτική. Για το λόγο αυτό θα εστιάσουμε στο σημείο αυτό σε τεχνικές οι οποίες βαθμονομούν τις ταξινομητικές ικανότητες διανυσμάτων χαρακτηριστικών. Γίνεται εύκολα αντιληπτό πως στην περίπτωση αυτή η υπολογιστική πολυπλοκότητα είναι ένας σοβαρός περιοριστικός παράγοντας. Έαν θέλαμε να είμαστε πιστοί στην έννοια της βελτιστοποίησης θα έπρεπε να σχηματίζαμε όλους του δυνατούς συνδυασμούς διανυσμάτων l χαρακτηριστικών. Ανάλογα με τον κανόνα βελτιστοποίησης με τον οποίο εργαζόμαστε μπορούμε να χωρίσουμε την επιλογή διανύσματος χαρακτηριστικών σε 2 κατηγορίες-προσεγγίσεις [68].

Η πρώτη προσέγγιση είναι η **προσέγγιση φίλτρου** (*filter approach*). Στην προσέγγιση αυτή, ο κανόνας για την επιλογή των χαρακτηριστικών είναι ανεξάρτητος του τύπου του ταξινομητή που θα χρησιμοποιήσουμε στη σχεδίαση του συστήματός μας. Για κάθε έναν συνδυασμό πρέπει να χρησιμοποιήσουμε ένα από τα μέτρα διαχωρισιμότητας κλάσεων και να διαλέξουμε με βάση αυτό τον κάλυτερο από τους συνδυασμούς. Συνολικά ο αριθμός των διαφορετικών συνδυασμών με l χαρακτηριστικά σε ένα σύνολο m χαρακτηριστικών $\binom{m}{l} = \frac{m!}{l!(m-l)!}$. Για ένα μεγάλο σύνολο χαρακτηριστικών γίνεται αντιληπτό πως ο αριθμός αυτός είναι πρακτικά πολύ μεγάλος. Ακόμα στις περισσότερες περιπτώσεις δε γνωρίζουμε ποιος είναι ο βέλτιστος αριθμός των χαρακτηριστικών [68].

Οι σημαντικότερες τεχνικές προσέγγισης φίλτου είναι η *Sequential Backward Selection* και η *Sequential Forward Selection*. Η παρουσίαση της πρώτης μεθόδου θα γίνει με ένα παράδειγμα. Θεωρούμε ένα σύνολο με $m = 4$ χαρακτηστικά $[x_1, x_2, x_3, x_4]$ εκ των οποίων ζητούμε να γίνει επιλογή του βέλτιστου διανύσματος με 2 χαρακτηριστικά. Η διαδικασία αποτελείται από τα ακόλουθα βήματα:

1. Υιοθετούμε ένα κριτήριο διαχωρισιμότητας C και υπολογίζουμε τις επιμέρους τιμές για κάθε χαρακτηριστικό του διανύσματος $[x_1, x_2, x_3, x_4]^T$
2. Απομακρύνουμε ένα χαρακτηριστικό και για κάθε έναν από τους συνδυασμούς που θα προκύψουν, $[x_1, x_2, x_3]^T$, $[x_1, x_2, x_4]^T$, $[x_1, x_3, x_4]^T$, $[x_2, x_3, x_4]^T$, υπολογίζουμε την αντίστοιχη τιμή του κριτηρίου μας. Επιλέγουμε το συνδυασμό με την καλύτερη τιμή, π.χ. το $[x_1, x_2, x_3]^T$
3. Από το 3-διάστατο διάνυσμα που προέκυψε απομακρύνουμε ένα χαρακτηριστικό για κάθε έναν συνδυασμό $[x_1, x_2]^T$, $[x_1, x_3]^T$, $[x_2, x_3]^T$, και υπολογίζουμε πάλι την τιμή του κριτηρίου για κάθε ένα από αυτά και επιλέγουμε αυτό με την καλύτερη τιμή κριτηρίου.

Έτσι, ξεκινώντας από έναν αριθμό m αρχικών χαρακτηριστικών, σε κάθε βήμα απομακρύνουμε ένα χαρακτηριστικό μέχρι να καταλήξουμε σε ένα διάνυσμα l χαρακτηριστικών.

Στην ίδια λογική είναι και η τεχνική *Sequential Forward Selection*, μόνο που η διαδικασία για την επιλογή γίνεται ανάποδα.

1. Για κάθε χαρακτηριστικό υπολογίζουμε την τιμή του κριτηρίου. Θεωρούμε ότι το χαρακτηριστικό με την καλύτερη τιμή είναι το x_1
2. Σχηματίζουμε λοιπόν όλα τα διδιάστατα διανύσματα που χρησιμοποιούν το βέλτιστο χαρακτηριστικό του πρώτου βήματος $[x_1, x_4]^T$, $[x_1, x_2]^T$, $[x_1, x_3]^T$. Υπολογίζουμε και πάλι την τιμή του κριτηρίου μας και συνεχίζουμε στο ίδιο μοτίβο μέχρι τον ζητούμενο αριθμό χαρακτηριστικών l.

Και οι δύο τεχνικές αυτές προφανώς είναι υποβέλτιστες μια και δεν μπορεί να δειχθεί ότι το παραχθέν υποσύνολο είναι το βέλτιστο [68].

Η δεύτερη προσέγγιση ονομάζεται **προσέγγιση περιτυλίγματος** (*wrapper approach*). Η διαφορά σε σχέση με την προηγούμενη προσέγγιση έγκειται στο γεγονός ότι αυτή τη φορά αντί να θεωρούμε σαν κριτήριο κάποιο από τα μέτρα διαχωρισιμότητας κλάσεων, θεωρούμε σαν κριτήριο την απόδοση αυτού καθεαυτού του ταξινομητή μας. Ο συνδυασμός χαρακτηριστικών που τελικά επιλέγεται είναι αυτός που παρουσιάζει το μικρότερο σφάλμα ταξινόμησης. Η υπολογιστική πολυπλοκότητα ανεβαίνει ανάλογα με το είδος του ταξινομητή που χρησιμοποιείται [68].

Για τις δύο προσεγγίσεις που παρατέθηκαν παραπάνω έχουν προταθεί μια σειρά από αποδοτικές τεχνικές. Μερικές από αυτές είναι υποβέλτιστες και μερικές βέλτιστες. Στην παρούσα εργασία κατασκευάζουμε μια νέα τεχνική επιλογής χαρακτηριστικών προσέγγισης wrapper με τη χρήση τεχνικών θεωρίας πληροφορίας και γενετικών αλγόριθμων με σκοπό την εξόρυξη δεδομένων από μία βάση δεδομένων για τη βελτίωση της ταξινόμησης των περιστατικών τραχηλικής ενδοεπιθηλιακής νεοπλασίας.

2.2 Επιλογή Χαρακτηριστικών με χρήση τεχνικών Θεωρίας Πληροφορίας

Στο πλαίσιο του υποκεφαλαίου αυτού θα γίνει μια εισαγωγή των βασικών ορισμών που απαιτούνται για την επακόλουθη ανάπτυξη της θεωρίας πληροφορίας. Μετά τον ορισμό της εντροπίας και της αμοιβαίας πληροφορίας, θα αναλυθούν οι κανόνες αλυσίδας, η μη αρνητικότητα της αμοιβαίας πληροφορίας, η ανισότητα των υπό επεξεργασία δεδομένων και

προχωρούμε σε περεταίρω επεξήγηση των παραπάνω εννοιών με την εξέταση των απαιτούμενων στατιστικών. Αφού γίνει εισαγωγή στη θεωρία πληροφορίας θα παρουσιάσουμε το πώς μπορεί να γίνει επιλογή χαρακτηριστικών με βάση τη θεωρία πληροφορίας.

Η έννοια της πληροφορίας είναι πολύ ευρεία για να αποκρυσταλλωθεί ικανοποιητικά από έναν μόνο ορισμό. Παρόλα αυτά, για κάθε πιθανοτική κατανομή ορίζουμε μία ποσότητα που αποκαλείται **εντροπία** *(entropy)*, η οποία έχει πολλές ιδιότητες που έρχονται σε συμφωνία με τη διαισθητική αντίληψή μας για το τι θα μπορούσε να αποτελεί ένα μέτρο πληροφορίας. Η έννοια αυτή επεκτείνεται για να ορίσουμε την **αμοιβαία πληροφορία** *(mutual information)*, η οποία ορίζει ένα μέτρο για το ποσό πληροφορίας που εμπεριέχεται σε μια τυχαία μεταβλητή για κάποια άλλη. Η εντροπία γίνεται εν συνεχεία η ίδια πληροφορία μίας τυχαίας μεταβλητής. Η αμοιβαία πληροφορία είναι μία ειδική περίπτωση μίας γενικότερης ποσότητας που αποκαλείται **σχετική εντροπία** *(relative entropy)*, η οποία είναι μία μετρική σχέση για την απόσταση μεταξύ δύο πιθανοτικών κατανομών. Όλες οι παραπάνω ποσότητες που αναφέρθηκαν είναι στενά συνυφασμένες μεταξύ τους και μοιράζονται ένα μεγάλο αριθμό ιδιοτήτων [71]. Τέλος αξίζει να σημειωθεί ότι μέσω της θεωρίας πληροφορίας μπορούμε να ποσοτικοποιήσουμε την πληροφορία που κρύβεται σε κάποιες μεταβλητές, πράγμα που την καθιστά ένα πολύ σημαντικό εργαλείο για την εξόρυξη δεδομένων.

2.2.1 Θεωρία Πληροφορίας

Η θεωρία πληροφορίας είναι ένας κλάδος των εφαρμοσμένων μαθηματικών που εμπεριέχει την ποσοτικοποίηση της πληροφορίας. Η θεωρία πληροφορίας αναπτύχθηκε από τον Claude E. Shannon με σκοπό την αναζήτηση θεμελιωδών ορίων σε διαδικασίες επεξεργασίας σήματος, όπως η συμπίεση δεδομένων, η αξιόπιστη αποθήκευση και επικοινωνία δεδομένων. Επίσης, η θεωρία πληροφορίας βρίσκει εφαρμογή σε ένα ευρύ φάσμα επιστημονικών τομέων, όπως η στατιστική συμπερασματολογία, η κρυπτογραφία και τα δίκτυα επικοινωνιών [71].

Σημαντική έννοια της θεωρίας πληροφορίας είναι η πηγή πληροφορίας. Μία πηγή πληροφορίας είναι ένα μαθηματικό μοντέλο για μία φυσική οντότητα, η οποία παράγει μία ακολουθία συμβόλων κατά τρόπο τυχαίο που αποκαλούνται «έξοδοι». Τα σύμβολα αυτά που παράγονται μπορεί να είναι πραγματικοί αριθμοί (π.χ. μετρήσεις τάσεων σε αισθητήρες), δυαδικά δεδομένα (π.χ. δεδομένα Ηλεκτρονικού Υπολογιστή), διδιάστατη κατανομή πεδίων (π.χ. ψηφιακές εικόνες), συνεχείς ή διακριτές κυματομορφές κ.ά.. Το διάστημα που περικλείει όλες τις δυνατές συμβολοσειρές εξόδου αποκαλείται ***αλφάβητο*** (***alphabet***) της πηγής και μια πηγή είναι ουσιαστικά η εκχώρηση (απόδοση/ανάθεση) ενός μέτρου πιθανότητας σε ένα γεγονός που περιέχει σύνολα συμβολοσειρών του αλφαβήτου. Παρόλα αυτά, είναι χρήσιμο να ερμηνευθεί η έννοια του χρόνου σαν ένας μετασχηματισμός ακολουθιών παραγόμενων από την πηγή [70].

2.2.2 Εντροπία

Πρώτα από όλα πρέπει να εισάγουμε την έννοια της **εντροπίας**, που αποτελεί ένα μέτρο της αβεβαιότητας μίας τυχαίας μεταβλητής (π.χ. στην περίπτωσή μας μία από τις διαθέσιμες εξετάσεις που έχουμε στη βάση δεδομένων μας). Έστω X μία διακριτή τυχαία μεταβλητή με αλφάβητο Μ και κατανομή πιθανότητας $p(x) = \Pr\{\mathrm{X} = \mathrm{x}\}, \mathrm{x} \in \mathrm{X}$. Αναφέρουμε σε αυτή τη φάση τη συνάρτηση πιθανότητας με $p(x)$ και όχι με $p_X(x)$ χάριν ευκολίας. Οπότε, οι $p(x)$ και $p(y)$ αναφέρονται σε δύο διαφορετικές τυχαίες μεταβλητές [72].

<u>*Ορισμός*</u> Η εντροπία $H(X)$ μίας διακριτής τυχαίας μεταβλητής X ορίζεται ως

$$H(X) = -\sum_{x \in \mathrm{X}} p(x) log p(x)$$

Ο λογάριθμος στην παραπάνω σχέση έχει βάση το 2 και η εντροπία εκφράζεται σε *bits*. Για παράδειγμα, η εντροπία ενός δίκαιου ζαριού είναι *1 bit*. Θα χρησιμοποιήσουμε ακόμα τη σύμβαση ότι η ποσότητα $0 log 0 = 0$, η οποία δικαιολογείται από τη συνέχεια της συνάρτησης, καθώς $\lim_{x \to 0\_} x log x \to 0$. Η πρόσθεση όρων μηδενικής πιθανότητας οπότε δεν μεταβάλλει την εντροπία [72].

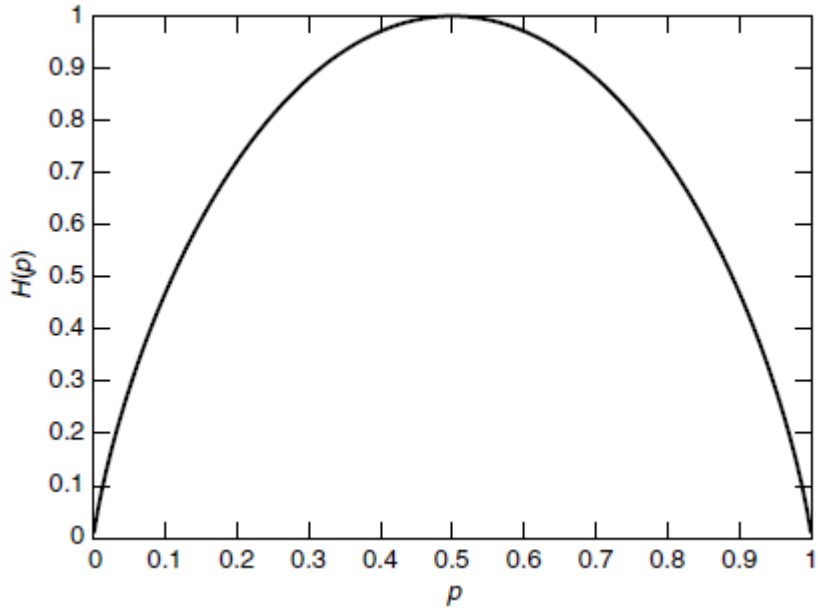
Σε αυτή τη φάση πρέπει να ορίσουμε την έννοια της **προσδοκίας** *(expectation)*. Αν $X \sim p(x)$, η προσδοκώμενη τιμή μίας τυχαίας μεταβλητής $g(X)$ δίνεται από την ακόλουθη σχέση

$$E_p g(X) = \sum_{x \in \mathrm{X}} g(x) p(x)$$

η οποία υποδηλώνεται πιο απλά και ως $Eg(X)$ αν η συνάρτηση πιθανότητας είναι προφανής με βάση τα συμφραζόμενα. Η εντροπία, λοιπόν, μπορεί να παρουσιαστεί ως η προσδοκώμενη τιμή της τυχαίας μεταβλητής $log \frac{1}{p(x)}$, με το X να ακολουθεί την πιθανοτική κατανομή $p(x)$. Άρα ισχύει:

$$H(X) = E_p log \frac{1}{p(X)}$$

Ο ορισμός αυτός της εντροπίας σχετίζεται με τον ορισμό της εντροπίας στη θερμοδυναμική.



Εικόνα 2.6: *Δυαδική εντροπία συναρτήσει της πιθανότητας εμφάνισης*

2.2.3 *Κοινή Εντροπία και υπό συνθήκη εντροπία*

Στο προηγούμενο υποκεφάλαιο ορίστηκε η εντροπία μίας τυχαίας μεταβλητής. Στα πλαίσια του υποκεφαλαίου θα επεκτείνουμε τον ορισμό σε ένα ζεύγος τυχαίων μεταβλητών. Στην πραγματικότητα δεν εισάγεται καμία νέα έννοια, αφού οι δύο τυχαίες μεταβλητές (X, Y) μπορούν να αντιμετωπιστούν ως διάνυσμα [72].

<u>*Ορισμός*</u> Η κοινή εντροπία $H(X, Y)$ ενός ζεύγους διακριτών τυχαίων μεταβλητών (X, Y) με κοινή κατανομή $p(x, y)$ ορίζεται ως

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in y} p(x, y) log p(x, y),$$

η οποία μπορεί επίσης να εκφραστεί και ακολούθως

$$H(X, Y) = -E log p(x, y).$$

Ακόμα, ορίζουμε την υπό συνθήκη εντροπία μίας τυχαίας μεταβλητής

Ορισμός Αν $(X,Y) \sim p(x,y)$, η υπό συνθήκη εντροπία $H(Y|X)$ ορίζεται ως:

$$\begin{aligned} H(Y|X) &= \sum_{x \in \mathrm{X}} p(x) H(Y|X=x) \\ &= -\sum_{x \in \mathrm{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) logp(y|x) \\ &= -\sum_{x \in \mathrm{X}} \sum_{y \in \mathcal{Y}} p(x,y) logp(y|x) \\ &= -E\ logp(Y|X) \end{aligned}$$

Η υπό συνθήκη εντροπία αποτελεί ένα μέτρο της μέσης αβεβαιότητας του εξαγόμενου δείγματος Y δεδομένου του δείγματος εισόδου X. Επίσης αποκαλείται και αμφιβολία (equivocation) [73]. Η φυσικότητα του ορισμού της κοινής εντροπίας και της υπό συνθήκη εντροπίας διαφαίνεται από το γεγονός ότι η εντροπία ενός ζεύγους τυχαίων μεταβλητών είναι το άθροισμα της εντροπίας της μίας μεταβλητής συν την υπό συνθήκη εντροπία της άλλης, πράγμα το οποίο αποδεικνύεται με το ακόλουθο θεώρημα

Θεώρημα Κανόνας Αλυσίδας

$$H(X,Y) = H(X) + H(Y|X)$$

και ισοδύναμα θα ισχύει

$$logp(X,Y) = logp(X) + logp(Y|X)$$

Πόρισμα

$$H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$$

2.2.4 Σχετική Εντροπία και αμοιβαία πληροφορία

Η εντροπία μίας τυχαίας μεταβλητής είναι ένα μέτρο της αβεβαιότητας της τυχαίας αυτής μεταβλητής. Εκφράζει, επίσης, ένα μέτρο για την ποσότητα πληροφορίας που απαιτείται κατά μέσο όρο για να περιγραφεί η τυχαία μεταβλητή. Στο υποκεφάλαιο αυτό θα παρουσιαστούν δύο σχετικές έννοιες: η **σχετική εντροπία** *(relative entropy)* και η **αμοιβαία πληροφορία** *(mutual information)* [72].

Η σχετική εντροπία αποτελεί ένα μέτρο της απόστασης μεταξύ δύο κατανομών. Εκφράζει επίσης ένα μέτρο της αδυναμίας να αποφασίσουμε εάν μια κατανομή είναι q ενώ στην πραγματικότητα είναι p.

<u>*Ορισμός*</u> Η σχετική εντροπία (ή απόσταση Kullback Leibler) μεταξύ δύο πιθανοτικών κατανομών $p(x)$ και $q(x)$ ορίζεται ως

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) log \frac{p(x)}{q(x)}$$
$$= E_p \, log \frac{p(X)}{q(X)}$$

Στον παραπάνω ορισμό χρησιμοποιούμε χάριν ευκολίας τη σύμβαση ότι $0 \, log \frac{0}{0} = 0$ και ακόμη $0 \, log \frac{0}{q} = 0$ και $0 \, log \frac{p}{0} = \infty$. Ως εκ τούτου, αν υπάρχει ένα σύμβολο $x \in \mathcal{X}$ τέτοιο ώστε $p(x) > 0$ και $q(x) = 0$, τότε $D(p||q) = \infty$.

Σε αυτό το σημείο θα ορίσουμε την αμοιβαία πληροφορία, η οποία είναι ένα μέτρο της ποσότητας πληροφορίας, που περικλείεται σε μία τυχαία μεταβλητή για κάποια άλλη τυχαία μεταβλητή. Ουσιαστικά εκφράζει τη μείωση της αβεβαιότητας μίας τυχαίας μεταβλητής λόγω της γνώσης της άλλης [72].

<u>*Ορισμός*</u> Θεωρούμε δύο τυχαίες μεταβλητές X και Y , από κοινού πιθανοτική κατανομή $p(x, y)$ και marginal πιθανοτικές κατανομές $p(x)$ και $p(y)$. Η αμοιβαία πληροφορία $I(X; Y)$ είναι η σχετική εντροπία μεταξύ των κοινών κατανομών και του γινομένου των κατανομών $p(x)p(y)$:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log \frac{p(x,y)}{p(X)p(Y)}$$
$$= D(p(x,y)||p(x)p(y))$$
$$= E_{p(x,y)} log \frac{p(X,Y)}{p(X)p(Y)}$$

2.2.5 Σχέση μεταξύ εντροπίας και αμοιβαίας πληροφορίας

Ξαναγράφοντας τον ορισμό της αμοιβαίας πληροφορίας $I(X;Y)$ έχουμε

$$I(X;Y) = \sum_{x,y} p(x,y) log \frac{p(x,y)}{p(x)p(y)}$$
$$= \sum_{x,y} p(x,y) log \frac{p(x|y)}{p(x)}$$
$$= \sum_{x,y} p(x,y) log p(x) + \sum_{x,y} p(x,y) log p(x|y)$$
$$= \sum_{x} p(x,y) log p(x) - \left(- \sum_{x,y} p(x,y) \log p(x|y) \right)$$
$$= H(X) - H(X|Y)$$

Από την τελευταία σχέση γίνεται κατανοητό γιατί η αμοιβαία πληροφορία εκφράζει τη μείωση της αβεβαιότητας του X λόγω γνώσης του Y. Επίσης, λόγω συμμετρίας ισχύει ότι

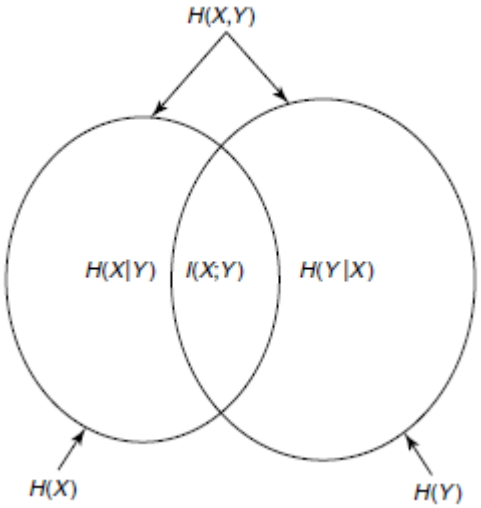
$$I(X;Y) = H(y) - H(Y|X).$$

Εύκολα συμπεραίνουμε λοιπόν ότι το X «λέει» τόσα για το Y , όσα και το Y για το X.
Με βάση όσα έχουν αναφερθεί μέχρι εδώ καταλήγουμε στο επόμενο θεώρημα [72]:

<u>*Θεώρημα*</u> (Αμοιβαία πληροφορία και εντροπία)

$$I(X;Y) = H(X) + H(X|Y)$$
$$I(X;Y) = H(Y) - H(Y|X)$$
$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$
$$I(X;Y) = I(Y;X)$$
$$I(X;X) = H(X)$$

Τέλος, η σχέση μεταξύ $H(X), H(Y), H(X,Y), H(Y|X)$ και $I(X;Y)$ φαίνεται καθαρά στο παρακάτω διάγραμμα Venn [72]



***Εικόνα 2.7**: Σχέση εντροπίας και αμοιβαίας πληροφορίας*

2.2.6 Τεχνική mRMR

Η επιλογή χαρακτηριστικών μέσω της τεχνικής **mRMR** (*minimum Redundancy Maximum Relevance*) γίνεται με χρήση είτε της αμοιβαίας πληροφορίας, είτε της συσχέτισης ή με κάποια μέτρα απόστασης/ομοιότητας. Για παράδειγμα, μέσω της αμοιβαίας πληροφορίας μπορούμε να εντοπίσουμε ταυτόχρονα ποια χαρακτηριστικά σχετίζονται μεταξύ τους (*relevant*) και ποιά μπορούν να θεωρηθούν περιττά/πλεονάζοντα (*redundant*). Η συνάφεια (*relevance*) ενός συνόλου χαρακτηριστικών (*feature set*) S με την κλάση c ορίζεται ως ο μέσος όρος των τιμών της αμοιβαίας πληροφορίας μεταξύ των επιμέρους χαρακτηριστικών f_i και της κλάσης c:

$$D(S,c) = \frac{1}{|S|} \sum_{f_i \in S} I(f_i; c)$$

όπου $|S|$ το πλήθος των χαρακτηριστικών που ανήκουν στο σύνολο. Ο πλεονασμός (*redundancy*) όλων των χαρακτηριστικών του συνόλου S ορίζεται αντίστοιχα ως ο μέσος όρος της αμοιβαίας πληροφορίας μεταξύ των χαρακτηριστικών f_i και f_j :

$$R(S) = \frac{1}{|S|^2} \sum_{f_i, f_j \in S} I(f_i ; f_j)$$

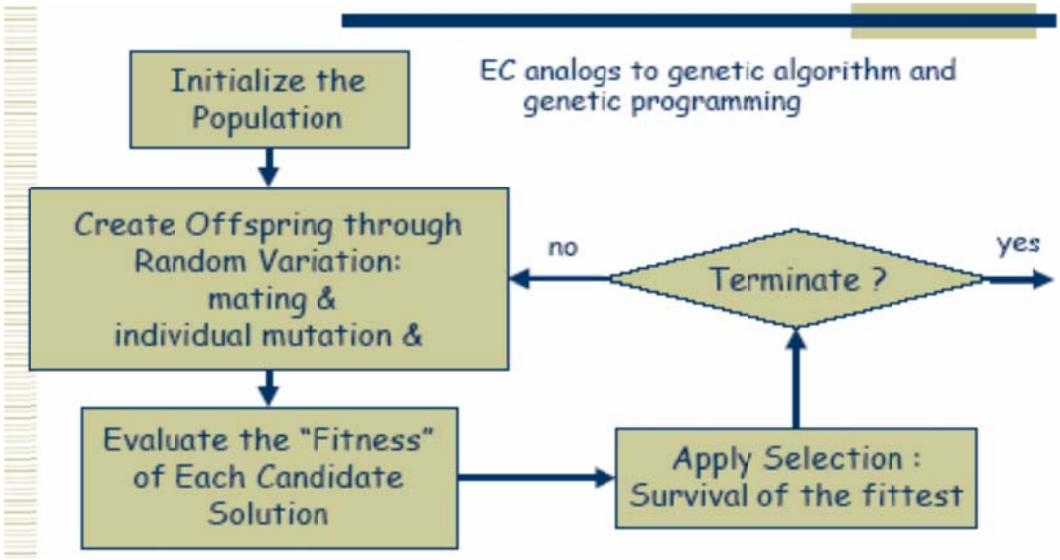
Το κριτήριο που χρησιμοποιεί η τεχνική mRMR είναι ένας συνδυασμός των 2 μέτρων που ορίστηκαν παραπάνω και πιο συγκεκριμένα περιγράφεται από την ακόλουθη σχέση:

$$\max_S [\frac{1}{|S|} \sum_{f_i \in S} I(f_i; c) - \frac{1}{|S|^2} \sum_{f_i, f_j \in S} I(f_i ; f_j)]$$

Η επιλογή χαρακτηριστικών με την τεχνική mRMR είναι μια προσέγγιση της θεωρητικά βέλτιστης τεχνικής επιλογής χαρακτηριστικών "maximum-dependency", η οποία μεγιστοποιεί την αμοιβαία πληροφορία μεταξύ της από κοινού κατανομής των επιλεγμένων χαρακτηριστικών και της μεταβλητής της ταξινόμησης. Σε γενικές γραμμές ο αλγόριθμος αυτός είναι πολύ πιο αποδοτικός σε σχέση με τον θεωρητικά βέλτιστο, όντας ταυτόχρονα πιο εύρωστος στην επιλογή χρήσιμων χαρακτηριστικών [69].

2.3 Γενετικοί Αλγόριθμοι

Ένας **γενετικός αλγόριθμος** (*Genetic Algorithm ή GA*) είναι μια τεχνική **ευρεστικής αναζήτησης (*search heuristic*)** η οποία μιμείται τη διαδικασία της φυσικής εξέλιξης. Αυτή η ευρεστική μέθοδος χρησιμοποιείται συνήθως για να εξαχθούν χρήσιμες λύσεις σε προβλήματα βελτιστοποίησης και αναζήτησης. Οι γενετικοί αλγόριθμοι ανήκουν στην ευρύτερη κατηγορία των εξελικτικών αλγορίθμων (*Evolutionary Algorithm ή EA*), οι οποίοι παράγουν λύσεις σε προβλήματα βελτιστοποίησης με χρήση τεχνικών που εμπνέονται από τη φυσική εξέλιξη όπως η κληροδότηση (*inheritance*), η μετάλλαξη (*mutation*), η επιλογή (*selection*) και η διασταυρωση (crossover) [74]. Ο γενετικός αλγόριθμος είναι μια παραλλαγή της στοχαστικής ακτινικής αναζήτησης, όπου οι διαδοχικές καταστάσεις παράγονται με το συνδυασμό δύο γονικών καταστάσεων και όχι με την τροποποίηση μιας μεμονωμένης κατάστασης. Η αναλογία με τη φυσική επιλογή είναι η ίδια όπως με αυτή της στοχαστικής ακτινικής αναζήτησης, με τη διαφορά ότι τώρα έχουμε να κάνουμε με γενετήσια και όχι άφυλη αναπαραγωγή [75].



***Εικόνα 2.8**: Διάγραμμα ροής της λειτουργίας ενός γενετικού αλγορίθμου*

2.3.1 Μεθοδολογία

Οι γενετικοί αλγόριθμοι (*GA*) ξεκινούν με ένα σύνολο *k* τυχαία παραγόμενων καταστάσεων που ονομάζονται **πληθυσμός** (*population*). Κάθε κατάσταση ή **άτομο** (*individual*), αναπαρίσταται με μία συμβολοσειρά από ένα πεπερασμένο αλφάβητο [75]. Τα άτομα αναπαρίστανται σαν ακολουθίες 0 και 1, χωρίς όμως αυτό να αποκλείει και άλλες μορφές κωδικοποίησης [1], με την κάθε κωδικοποίηση να συμπεριφέρεται διαφορετικά.

Η εξέλιξη (*evolution*) αρχίζει συνήθως από έναν πληθυσμό τυχαία παραγμένων ατόμων (individuals) και λαμβάνει χώρα με την πάροδο των γενεών (*generations*). Σε κάθε γενιά αξιολογείται η **καταλληλότητα** (*fitness*) του κάθε ατόμου στον πληθυσμό, τα πολλαπλά άτομα επιλέγονται με τρόπο στοχαστικό από τον τρέχοντα πληθυσμό, βάσει της καταλληλότητάς (*fitness*) τους και τροποποιούνται για να διαμορφώσουν έναν νέο πληθυσμό. Ο νέος αυτός πληθυσμός χρησιμοποιείται στη συνέχεια στην επόμενη επανάληψη του αλγορίθμου [74]. Η καταλληλότητα ενός ατόμου «βαθμονομείται» από μία **συνάρτηση καταλληλότητας (***fitness function***).** Η **συνάρτηση καταλληλότητας** ορίζεται μέσω της γενετικής αναπαράστασης και ποσοτικοποιεί την ποιότητα της αναπαριστώμενης λύσης. Η συνάρτηση αυτή εξαρτάται πάντοτε από το πρόβλημα του οποίου την επίλυση επιδιώκουμε. Η ιδιότητα της εν λόγω συνάρτησης θα πρέπει να είναι η επιστροφή υψηλότερων τιμών στις περιπτώσεις καλύτερων καταστάσεων [75]. Ο αλγόριθμος ολοκληρώνεται όταν είτε παραχθεί ένας μέγιστος αριθμός γενιών είτε έχει επιτευχθεί ένα ικανοποιητικό επίπεδο καταλληλότητας (*fitness level*) για τον πληθυσμό. Σε περίπτωση που ο αλγόριθμος ολοκληρωθεί λόγω μεγίστου αριθμού γενιών, μια ικανοποιητική λύση μπορεί να έχει ή να μην έχει επιτευχθεί. Μία τυπική αναπαράσταση της λύσης είναι μία ακολουθία από bits. Ακολουθίες

διαφορετικών τύπων και δομών μπορούν να χρησιμοποιηθούν ουσιαστικά με τον ίδιο τρόπο.

Η βασική ιδιότητα, η οποία καθιστά αυτές τις γενετικές αναπαραστάσεις βολικές, είναι το γεγονός ότι τα μέρη τους ευθυγραμμίζονται εύκολα λόγω του σταθερού μεγέθους τους, το οποίο διευκολύνει τις διαδικασίες διασταυρώσεων (*crossover operations*). Μπορούν ακόμη να χρησιμοποιηθούν αναπαραστάσεις μεταβλητού μεγέθους με το μειονέκτημα όμως ότι η εφαρμογή διασταυρώσεων καταλήγει να είναι περισσότερο περίπλοκη. Αναπαραστάσεις σε μορφή δέντρου εξετάζονται από τον γενετικό προγραμματισμό, ενώ αναπαραστάσεις σε μορφή γραφικών παραστάσεων εξετάζονται στον εξελικτικό προγραμματισμό.

Μόλις η γενετική αναπαράσταση και η συνάρτηση καταλληλότητας οριστούν, ο γενετικός αλγόριθμος προχωρά στην αρχικοποίηση του πληθυσμού των λύσεων, συνήθως κατά τρόπο τυχαίο, και εν συνεχεία στη βελτίωση μέσω εφαρμογής επαναλαμβανόμενων γενετικών χειριστών (*genetic operators*), **μετάλλαξης** (*mutation*), **διασταύρωσης** (*crossover*) και **επιλογής** (*selection*) [74].

2.3.2 Αρχικοποίηση (Initialization)

Αρχικά παράγονται πολλές μεμονωμένες λύσεις, κατά τρόπο τυχαίο, ώστε να διαμορφώσουν έναν αρχικό πληθυσμό. Το μέγεθος του πληθυσμού αυτού εξαρτάται από τη φύση του προβλήματος, αλλά τυπικά περιέχει έναν μεγάλο αριθμό πιθανών λύσεων. Ο πληθυσμός παράγεται τυχαία, επιτρέποντας ολόκληρο το εύρος των πιθανών λύσεων (Χώρος Αναζήτησης / Search Space). Περιστασιακά, οι λύσεις μπορούν να αναζητηθούν στις περιοχές, στις οποίες οι βέλτιστες λύσεις είναι πιθανό να βρεθούν.

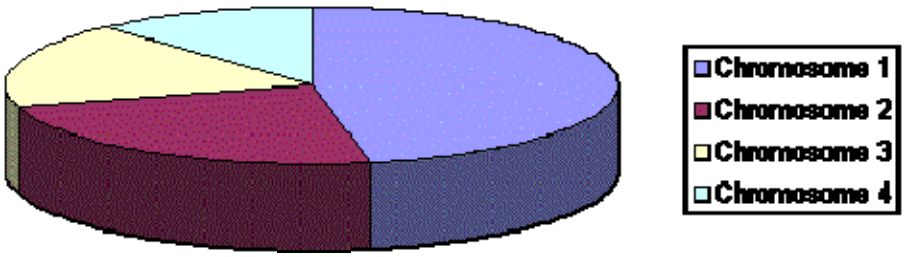
2.3.3 Επιλογή (Selection)

Η **επιλογή** (*selection*) είναι το στάδιο ενός γενετικού αλγορίθμου στο οποίο τα **άτομα** (*individuals*) επιλέγονται από έναν πληθυσμό για περαιτέρω αναπαραγωγή [74]. Τα **άτομα** (*individuals*) επιλέγονται από τον πληθυσμό να γίνουν γονείς για τη διαδικασία της **διασταύρωσης** (*crossover*). Το πρόβλημα είναι το πώς θα επιλεγούν τα άτομα αυτά. Σύμφωνα με τη δαρβινική θεωρία τα καλύτερα-ισχυρότερα άτομα θα επιζήσουν και θα δημιουργήσουν τον νέο απόγονο. Υπάρχουν πολλές μέθοδοι για την επιλογή των ατόμων-

χρωμοσωμάτων και οι βασικότερες από αυτές θα παρουσιαστούν στα πλαίσια του υποκεφαλαίου αυτού [76].

2.3.3.1 Roulette wheel selection

Η Roulette wheel selection είναι ένας γενετικός χειριστής για την επιλογή δυνητικά χρήσιμων λύσεων για επανασυνδυασμό. Στην περίπτωση του χειριστή αυτού η **συνάρτηση καταλληλότητας** (*fitness function*) αποδίδει σε κάθε πιθανή λύση / χρωμόσωμα μία καταλληλότητα. Το επίπεδο καταλληλότητας (*fitness level*) χρησιμοποιείται για να συσχετίσει κάθε χρωμόσωμα με μία πιθανότητα επιλογής. Η τιμή που παίρνει η πιθανότητα αυτή είναι $p_i = \frac{f_i}{\sum_{j=1}^{N} f_j}$, όπου f_i είναι η καταλληλότητα του κάθε ατόμου *i* και *N* το μέγεθος του πληθυσμού [68]. Οι γονείς λοιπόν επιλέγονται με βάση την καταλληλότητά τους. Όσο καλύτερα είναι τα χρωμοσώματα τόσο μεγαλύτερη πιθανότητα έχουν να επιλεγούν. Η διαδικασία θα μπορούσε να παραλληλιστεί με έναν τροχό ρουλέτας (*roulette wheel*), στον οποίο έχουν τοποθετηθεί όλα τα χρωμοσώματα του πληθυσμού. Ο χώρος τον οποίο καταλαμβάνει το κάθε χρωμόσωμα στον «τροχό» αυτό είναι ανάλογος της καταλληλότητας που εμφανίζει το κάθε χρωμόσωμα.

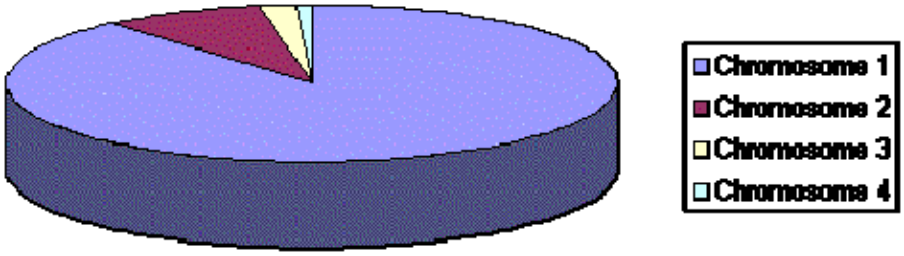


***Εικόνα 2.9**: Κυκλικό διάγραμμα στο οποίο κάθε χρωμόσωμα καταλαμβάνει ποσοστό ανάλογο της καταλληλότητας*

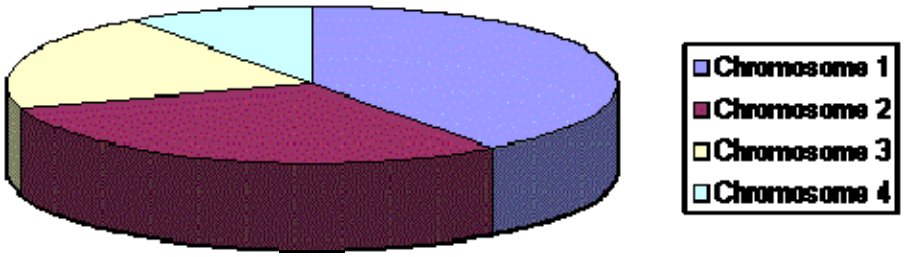
Εν συνεχεία ρίχνεται η μπίλια και επιλέγεται ένα χρωμόσωμα. Γίνεται κατανοητό πως χρωμοσώματα με μεγαλύτερη καταλληλότητα έχουν μεγαλύτερη πιθανότητα επιλογής οπότε θα επιλεγούνε περισσότερες φορές [76].

2.3.3.2 Rank Selection

Σε περιπτώσεις κατά τις οποίες η καταλληλότητα ανάμεσα στα χρωμοσώματα παρουσιάζει μεγάλες διαφορές η Roulette wheel selection θα εμφανίσει προβλήματα, καθώς χρωμοσώματα με μικρή καταλληλότητα έχουν ελάχιστες πιθανότητες να επιλεγούν. Η Rank Selection σε αντίθεση, πρώτα βαθμονομεί τον πληθυσμό και κάθε χρωμόσωμα λαμβάνει ένα βαθμό καταλληλότητας από αυτή την κατάταξη.



***Εικόνα 2.10**: Κυκλικό διάγραμμα πληθυσμού με χρωμοσώματα που εμφανίζουν μεγάλες διαφορές καταλληλότητας*



***Εικόνα 2.11**: Κυκλικό διάγραμμα μετά το Ranking*

Μετά από αυτή τη διαδικασία της βαθμονόμησης όλα τα χρωμοσώματα έχουν πιθανότητες να επιλεχθούν. Παρόλα αυτά η μέθοδος αυτή μπορεί να οδηγήσει σε πιο αργή σύγκλιση, λόγω του ότι τα χρωμοσώματα με τη μεγαλύτερη καταλληλότητα αντιμετωπίζονται με τον ίδιο τρόπο όπως τα υπόλοιπα [76].

2.3.3.3 Steady-State Selection

Η μέθοδος αυτή δεν αποτελεί μια μέθοδο επιλογής γονέων. Η κύρια ιδέα της επιλογής αυτής είναι ότι μεγάλο μέρος των χρωμοσωμάτων πρέπει να επιζήσει στην επόμενη γενιά. Εν συνεχεία ο **γενετικός αλγόριθμος** (*GA*) λειτουργεί με τον ακόλουθο τρόπο. Σε κάθε γενιά επιλέγονται μερικά χρωμοσώματα(με βάση την καταλληλότητα τους) για τη δημιουργία ενός νέου απογόνου. Εν συνεχεία τα χρωμοσώματα με τη χαμηλή καταλληλότητα αφαιρούνται και ο νέος απόγονος τοποθετείται στη θέση τους. Το υπόλοιπο του πληθυσμού επιζεί και στην επόμενη γενιά. [76].

2.3.4 Αναπαραγωγή (Reproduction)

Το επόμενο βήμα μετά την αρχικοποίηση είναι να παραχθεί ένας πληθυσμός δεύτερης γενιάς λύσεων από εκείνους μέσω των γενετικών χειριστών (*genetic operators*): **διασταύρωση** (*crossover*) ή/και **μετάλλαξη** (*mutation*). Για κάθε νέα λύση που παράγεται, ένα ζευγάρι λύσεων «γονέων» επιλέγεται για την περαιτέρω αναπαραγωγή από τη «δεξαμενή» λύσεων που επιλέχθηκε προηγουμένως. Με την παραγωγή μίας λύσης «παιδιού», για την οποία χρησιμοποιήθηκε μία από τις παραπάνω μεθόδους της διασταύρωσης και της μετάλλαξης, δημιουργείται μια νέα λύση η οποία τυπικά περιλαμβάνει πολλές από τις ιδιότητες των «γονέων». Για κάθε νέο «παιδί» επιλέγονται νέοι «γονείς» και η διαδικασία συνεχίζεται έως ότου παραχθεί ένας νέος πληθυσμός λύσεων κατάλληλου μεγέθους.

Οι διαδικασίες αυτές οδηγούν τελικά στον πληθυσμό επόμενης γενιάς των χρωμοσωμάτων, η οποία είναι διαφορετική από την αρχική γενιά. Η μέση καταλληλότητα θα παρουσιάζει αύξηση με αυτή τη διαδικασία στην οποία υπόκειται ο πληθυσμός, δεδομένου πως μόνο οι καλύτεροι οργανισμοί από την πρώτη γενιά επιλέγονται για την αναπαραγωγή σε συνδυασμό με ένα μικρό ποσοστό λιγότερο κατάλληλων λύσεων, για λόγους που έχουν αναφερθεί προηγουμένως. Πέραν των δύο πιο διαδεδομένων **γενετικών χειριστών** (*genetic operators*), δηλαδή της **διασταύρωσης** (*crossover*) και της **μετάλλαξης** (*mutation*), υπάρχει και δυνατότητα να χρησιμοποιηθούν και άλλοι τελεστές όπως η ανασυγκρότηση (*regrouping*), η αποίκηση-εξάλειψη (*colonization-extinction*) ή η μετανάστευση (*migration*) [74].

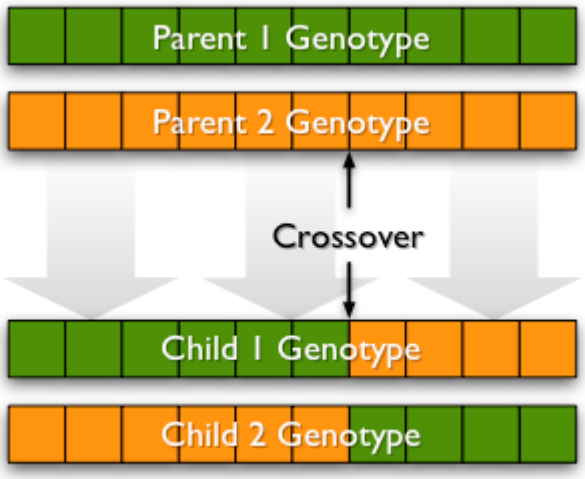
2.3.4.1 Διασταύρωση (Crossover)

Η διασταύρωση, όπως αναφέρθηκε παραπάνω, είναι ένας γενετικός τελεστής. Χρησιμοποιείται στον προγραμματισμό ενός ή περισσοτέρων χρωμοσωμάτων με σκοπό αυτά να ποικίλουν από τη μία γενιά στην επόμενη. Η όλη διαδικασία είναι ανάλογη της βιολογικής αναπαραγωγής και διασταύρωσης, στην οποία είναι βασισμένη και η θεωρία των γενετικών αλγορίθμων. Η διαδικασία της διασταύρωσης συνίσταται στην επιλογή παραπάνω από μίας λύσης «γονείς» και εν συνεχεία με βάση αυτές στην παραγωγή μιας λύσης «παιδί» [74].

Υπάρχουν διάφορες μέθοδοι διασταύρωσης. Θα παρουσιαστούν συνοπτικά οι ακόλουθες:

- Διασταύρωση ενός σημείου (*Single point crossover*) [75]

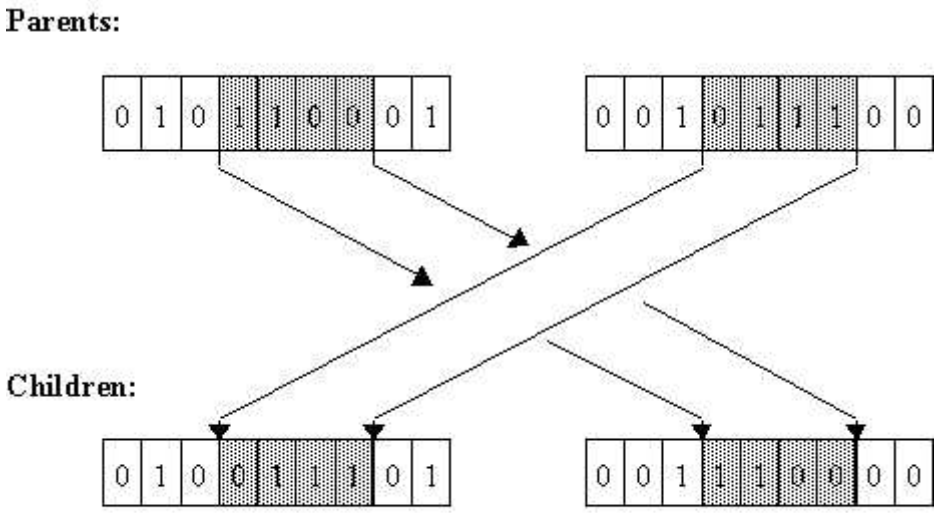
Επιλέγεται ένα σημείο διασταύρωσης και η ακολουθία του ατόμου (*individual*) από το σημείο έναρξης της μέχρι το σημείο που επιλέχθηκε αντιγράφεται από τον πρώτο γονιό και το υπόλοιπο από τον δεύτερο γονιό.



***Εικόνα 2.12**: Παράδειγμα διασταύρωσης ενός σημείου*

- Διασταύρωση δύο σημείων (*Two point crossover*) [75]

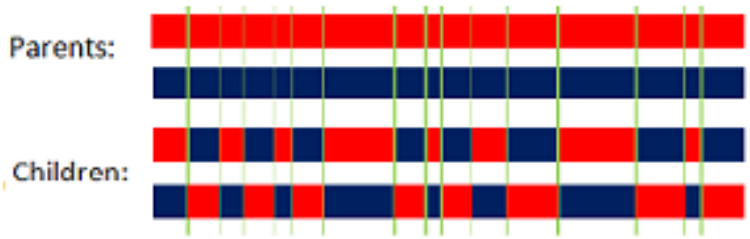
Στην περίπτωση αυτή επιλέγονται δύο σημεία και η ακολουθία του ατόμου (*individual*) μέχρι το πρώτο σημείο διασταύρωσης αντιγράφεται από τον πρώτο γονιό, η ακολουθία από το πρώτο σημείο ως το δεύτερο σημείο αντιγράφεται από το δεύτερο γονιό και τέλος το υπόλοιπο κομμάτι του νέου ατόμου που θα προκύψει αντιγράφεται από τον πρώτο γονιό.



***Εικόνα 2.13**: Παράδειγμα διασταύρωσης δύο σημείων*

- Ομοιόμορφη Διασταύρωση (*Uniform Distribution*) [75]

Τα μπιτ της ακολουθίας ενός ατόμου αντιγράφονται κατά τρόπο τυχαίο είτε από τον πρώτο είτε από το δεύτερο γονιό.



***Εικόνα 2.14**: Παράδειγμα ομοιόμορφης διασταύρωσης*

- Διασταύρωση τριών γονέων (*Three Parent Crossover*) [77]

Με την τεχνική αυτή, ένας απόγονος παράγεται από τρεις γονείς, οι οποίοι επιλέγονται τυχαία. Κάθε κομμάτι του πρώτου γονιού ελέγχεται με κάποιο κομμάτι του δεύτερου γονιού με σκοπό να ανιχνευθεί κατά πόσο είναι ίδια. Εάν είναι ίδια η ακολουθία αυτή επιλέγεται για την γέννηση του απογόνου αλλιώς επιλέγεται το αντίστοιχο κομμάτι από τον τρίτο γονιό.

parent1	**1**	**1**	**0**	**1**	**0**	**0**	**0**	**1**	**0**
parent2	**0**	**1**	**1**	**0**	**0**	**1**	**0**	**0**	**1**
parent3	**1**	**1**	**0**	**1**	**1**	**0**	**1**	**0**	**1**
offspring	**1**	**1**	**0**	**1**	**0**	**0**	**0**	**0**	**1**

Εικόνα 2.15: *Παράδειγμα διασταύρωσης τριών γονέων*

2.3.4.2 Μετάλλαξη (Mutation)

Η μετάλλαξη (*mutation*) είναι ένας γενετικός χειριστής που χρησιμοποιείται ώστε να επιτευχθεί γενετική ποικιλία από τη μία γενιά ενός πληθυσμού στην επόμενη. Η μετάλλαξη αλλάζει την τιμή σε ένα ή περισσότερα μπιτ στην ακολουθία ενός ατόμου σε σχέση με την αρχική του κατάσταση. Με την εφαρμογή μετάλλαξης, η λύση μπορεί να αλλάξει εντελώς σε σχέση με την προηγούμενη. Οπότε γίνεται κατανοητό πως ένας γενετικός αλγόριθμος μπορεί να καταλήξει σε καλύτερα αποτελέσματα με χρήση του τελεστή της μετάλλαξης.

Η κλασική εκδοχή ενός τελεστή μετάλλαξης περιλαμβάνει την πιθανότητα αλλαγής της τιμής ενός αυθαίρετου μπιτ σε μία γενετική ακολουθία. Μια κοινή μέθοδος εφαρμογής μετάλλαξης περιλαμβάνει τη «γέννηση» μιας τυχαίας μεταβλητής για κάθε μπιτ σε μία

γενετική ακολουθία, η οποία αυτή μεταβλητή υποδεικνύει αν θα μεταλλαχθεί ή όχι ένα συγκεκριμένο μπιτ. Η παραπάνω διαδικάσία καλείται μετάλλαξη ενός σημείου (*single point mutation*).

Σκοπός της μετάλλαξης σε έναν γενετικό αλγόριθμο, όπως αναφέρθηκε και παραπάνω είναι η διατήρηση και η εισαγωγή ποικιλότητας. Η μετάλλαξη θα πρέπει να επιτρέπει στον αλγόριθμο να αποφεύγει τοπικά ελάχιστα, πράγμα που επιτυγχάνεται με το να εμποδίζει τον πληθυσμό των ατόμων από το να γίνουν ίδια μεταξύ τους, επιβραδύνοντας ή ακόμα και σταματώντας με τον τρόπο αυτό την εξέλιξη.

Ανάλογα με τον τύπο των δεδομένων μας θα πρέπει να χρησιμοποιηθεί και ο ανάλογος τύπος μετάλλαξης. Ενδεικτικά παρατίθενται οι ακόλουθοι τύποι μεταλλάξεων:

- Μετάλλαξη ακολουθίας μπιτ (*Bit String Mutation*)

  Η μετάλλαξη προκύπτει μέσα από αλλάγη τιμής των μπιτ σε τυχαίες θέσεις. Π.χ.

$$1\ 0\ 1\ 0\ 0\ 1\ 0$$
$$\downarrow$$
$$1\ 0\ 1\ 0\ 1\ 1\ 0$$

Η πιθανότητα μετάλλαξης ενός μπιτ είναι $\frac{1}{l}$, όπου l είναι το μήκος της ακολουθίας.

- Ομοιόμορφη μετάλλαξη (*Uniform mutation*)

Στην περίπτωση αυτή μεταλλάσσεται η τιμή ενός γονιδίου-μπιτ του ατόμου, με το μπιτ αυτό να επιλέγεται από μια ομοιόμορφη τυχαία μεταβλητή της οποίας τα άνω και κάτω όρια είναι ορισμένα από τον χρήστη.

- Γκαουσιανή μετάλλαξη (Gaussian Mutation)

Η διαδικασία αυτή παρουσιάζει πολλές ομοιότητες με αυτή της ομοιόμορφης μετάλλαξης, με τη βασική διαφορά όμως ότι η τυχαία μεταβλητή ακολουθεί γκαουσιανή και όχι ομοιόμορφη κατανομή. Σε περίπτωση που η τιμή βρεθεί εκτός των ορισθέντων ορίων το νέο μπιτ «ψαλιδίζεται».

2.3.4.3 Τερματισμός (Termination)

Η γενετική διαδικασία επαναλαμβάνεται έως ότου εκπληρωθεί μια συνθήκη τερματισμού. Μερικές συνήθεις συνθήκες τερματισμού είναι [74]:

- Εύρεση λύσης που πληροί τα ελάχιστα κριτήρια
- Επίτευξη συγκεκριμένου αριθμού γενιών
- Κατακερματισμός διαθέσιμων πόρων (π.χ. χρόνος) για την επίλυση
- Η λύση με την υψηλότερη καταλληλότητα είτε είναι κοντά είτε έχει ήδη επιτευχθεί οπότε οι διαδοχικές επαναλήψεις δεν παράγουν πλέον καλύτερα αποτελέσματα
- Χειροκίνητη παρεμβολή
- Συνδυασμός των παραπάνω

Παρακάτω δίνεται ένας ψευδοκώδικας υλοποίησης ενός τυπικού γενετικού αλγορίθμου με σκοπό την καλύτερη κατανόηση της λειτουργίας των γενετικών αλγορίθμων [75].

```
function GENETIC-ALGORITHM( πληθυσμός, FITNESS-FN ) returns ένα άτομο
    inputs:     πληθυσμός, ένα σύνολο ατόμων
                FITNESS-FN, συνάρτηση που μετρά την καταλληλότητα ενός ατόμου

    repeat
        νέος_πληθυσμός ← κενό σύνολο
        loop for i from 1 to SIZE( πληθυσμός ) do
            x ← RANDOM-SELECTION( πληθυσμός, FITNESS-FN )
            y ← RANDOM-SELECTION( πληθυσμός, FITNESS-FN )
            παιδί ← REPRODUCE( x, y )
            if (μικρή τυχαία πιθανότητα) then παιδί ← MUTATE( παιδί )
            πρόσθεσε παιδί σε νέος_πληθυσμός
        πληθυσμός ← νέος_πληθυσμός
    until κάποιο άτομο είναι αρκετά κατάλληλο ή έχει περάσει αρκετός χρόνος
    return το καλύτερο άτομο από τον πληθυσμό, σύμφωνα με τη FITNESS-FN


function REPRODUCE( x, y ) returns ένα άτομο
    inputs: x, y, γονικά άτομα

    n ← LENGTH( x )
    c ← τυχαίος αριθμός από 1 μέχρι n
    return APPEND( SUBSTRING( x, 1, c ), SUBSTRING( y, c + 1, n ))
```

***Εικόνα 2.16**: Ψευδοκώδικας υλοποίησης γενετικού αλγορίθμου*

2.4 Μέθοδοι Αναγνώρισης Προτύπων

Στη μηχανική εκμάθηση, η **αναγνώριση προτύπων** είναι η ανάθεση μίας τιμής εξόδου (ή αλλιώς «ετικέτα») σε μία τιμή ή ακολουθία τιμών εισόδου ακολουθώντας έναν συγκεκριμένο αλγόριθμο. Ένα παράδειγμα της αναγνώρισης προτύπων είναι η **ταξινόμηση**, η οποία επιχειρεί να αποδώσει κάθε τιμή εισόδου σε ένα δεδομένο σύνολο από κλάσεις. Παρόλα αυτά η αναγνώριση προτύπων είναι πρόβλημα γενικότερο και δεν περιορίζεται μόνο στην ταξινόμηση. Βρίσκει εφαρμογές σε προβλήματα όπως η παλινδρόμηση (regression), και η αναγνώριση φωνής [78].

Οι αλγόριθμοι αναγνώρισης προτύπων στοχεύουν γενικά στο να παράσχουν λογικές απαντήσεις για όλες τις δυνατές εισόδους και να κάνουν ένα «ασαφές» (“fuzzy”) ταίριασμα των εισόδων αυτών. Έτσι έρχονται σε αντίθεση με τους λεγόμενους pattern matching αλγορίθμους, οι οποίοι αναζητούν ακριβείς αντιστοιχίες στην είσοδο με βάση προϋπάρχοντα μοτίβα (patterns). Τέλος η αναγνώριση προτύπων βρίσκει εφαρμογές σε πολλούς τομείς,

όπως η ψυχολογία, η ψυχιατρική, η επιστήμη των υπολογιστών και φυσικά η βιοϊατρική τεχνολογία [78].

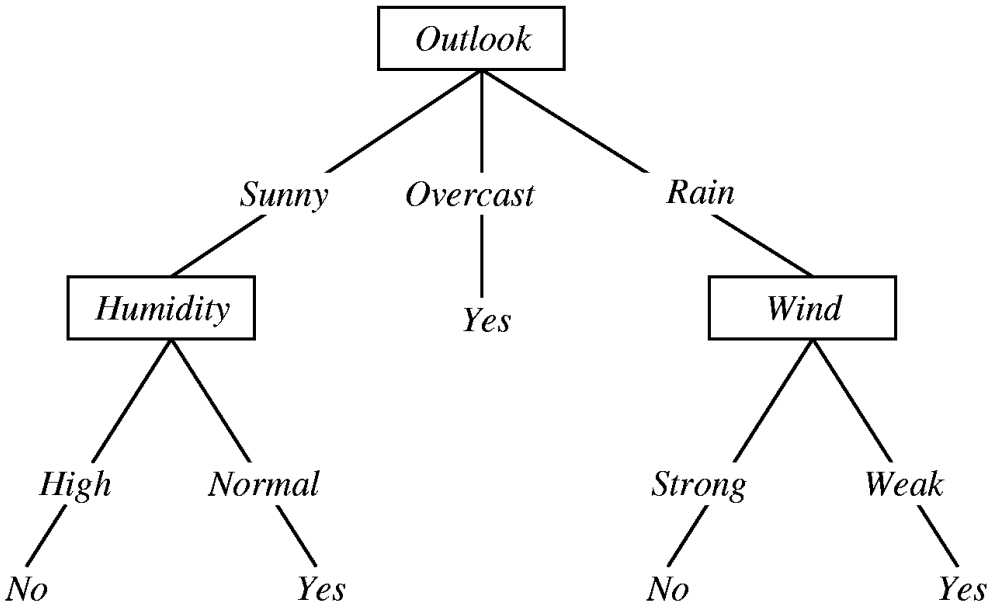
Στο πλαίσιο αυτού του κεφαλαίου θα αναφερθούν και θα αναλυθούν τεχνικές και αλγόριθμοι αναγνώρισης προτύπων, στα οποία βασιστήκαμε για την επίλυση του προβλήματος της παρούσας εργασίας. Οι τεχνικές που θα αναλυθούν είναι οι ακόλουθες

- Δένδρα Απόφασης (*Decision Trees*)
- Bayesian Ταξινομητές (*Bayesian Classifiers*)
- Νευρωνικά Δίκτυα (*Neural Networks*)

2.4.1 Δένδρα Απόφασης

Η **εκμάθηση μέσω δένδρων απόφασης** (*Decision Tree Learning*) χρησιμοποιείται στη στατιστική, στην εξόρυξη δεδομένων και στη μηχανική εκμάθηση. Χρησιμοποιεί σαν προβλεπτικό μοντέλο ένα δέντρο απόφασης, το οποίο χαρτογραφεί τις παρατηρήσεις για κάποιο γεγονός και οδηγεί σε συμπεράσματα για την τιμή της εξόδου του γεγονότος αυτού. Τα δένδρα που χρησιμοποιούνται στις τέτοιες περιπτώσεις είναι είτε **δένδρα ταξινόμησης** (*classification trees*) ή **δένδρα παλινδρόμησης** (*regression trees*). Σε αυτές τις δενδρικές δομές τα φύλλα αναπαριστούν τις κλάσεις και τα κλαδιά αναπαριστούν τους συνδυασμούς των χρησιμοποιούμενων χαρακτηριστικών, που οδηγούν στις αντίστοιχες κλάσεις. Στην ανάλυση απόφασης (decision analysis), ένα δένδρο απόφασης μπορεί να χρησιμοποιηθεί για να έχουμε μια οπτική αναπαράσταση των μοτίβων απόφασης και του τρόπου λήψης απόφασης.

Στόχος της **εκμάθησης μέσω δένδρων απόφασης** είναι η δημιουργία ενός μοντέλου, το οποίο προβλέπει την τιμή της μεταβλητής εξόδου βάσει των μεταβλητών εισόδου.



***Εικόνα 2.17**: Παράδειγμα δένδρου απόφασης*

Στην παραπάνω εικόνα απεικονίζεται ένα παράδειγμα δένδρου απόφασης. Κάθε εσωτερικός κόμβος αντιστοιχεί σε μία από τις μεταβλητές εισόδου. Κάθε φύλλο αναπαριστά την τιμή της μεταβλητής εξόδου δεδομένων των τιμών των μεταβλητών εισόδου, όπως αυτές απεικονίζονται σε μία διαδρομή στο δένδρο.

Τα δένδρα απόφασης χωρίζονται σε δύο κύριες κατηγορίες

- στα **Δένδρα Ταξινόμησης** (*Classification Trees*) και
- στα **Δένδρα Παλινδρόμησης** (*Regression Trees*)

Η διαφορά των δύο έγκειται στις περιπτώσεις που βρίσκουν εφαρμογή. Τα δένδρα ταξινόμησης χρησιμοποιούνται όταν η μεταβλητή εξόδου μας είναι κάποια κλάση και το ζητούμενο είναι με βάση τις τιμές των μεταβλητών εισόδου να προβλεφθεί σε ποια κλάση ανήκει. Αντίθετα, τα δένδρα παλινδρόμησης χρησιμοποιούνται όταν η μεταβλητή εξόδου μπορεί να λάβει πραγματικές τιμές. Στα πλαίσια αυτής της εργασίας θα μας χρειαστούν μόνο τα δένδρα ταξινόμησης και για το λόγο αυτό θα προχωρήσουμε σε ανάλυσή του στο επόμενο υποκεφάλαιο.

2.4.1.1 Δένδρα Ταξινόμησης

Τα δένδρα ταξινόμησης προσπαθούν, όπως αναφέρθηκε και προηγουμένως να αποδώσουν σε κάποιο συνδυασμό μεταβλητών εισόδου μια διακριτή τιμή-κλάση. Οι μεταβλητές στην είσοδο ενός τέτοιου δένδρου μπορούν να παίρνουν αριθμητικές ή κατηγοριοποιημένες (*categorical*) τιμές. Είναι πολύ χρήσιμα για το λόγο ότι προσφέρουν αρκετά κατανοητούς προβλέπτες σε περιπτώσεις στις οποίες υπάρχουν πολλές μεταβλητές που αλληλεπιδρούν μεταξύ τους, συνήθως με μη γραμμικό τρόπο.

Η διαδικασία κατασκευής του δένδρου έχει ως εξής:

- Αρχικά επιλέγουμε έναν αρχικό κόμβο
- Ψάχνουμε για τη δυαδική διάκριση που μας δίνει τη μεγαλύτερη πληροφορία για τις κλάσεις
- Στη συνέχεια παίρνουμε κάθε νέο κόμβο και επαναλαμβάνουμε την αναδρομική αυτή διαδικασία μέχρις ότου εκπληρωθεί κάποιο κριτήριο τερματισμού.

Για να γίνει πιο κατανοητό το πώς κατασκευάζεται ένα δένδρο απόφασης παραθέτουμε έναν ψευδοκώδικα για την κατασκευή του.

```
BuildTree(Node T, data-partition D, split attribute selection method V)
(1)   Apply V to D to find the split attribute X for node T.
(2)   Let n be the number of children of T.
(2)   if (T splits)
(3)      Partition D into D_1, ..., D_n and label note T with split attribute X
(4)      Create children nodes T_1, ..., T_n of T and label the edge (T, T_i) with predicate q_(T,T_i)
(5)      foreach i ∈ {1, .., n}
(6)         BuildTree(T_i, D_i, V)
(7)      endforeach
(8)   else
(9)      Label T with the majority class label of D
(10)  endif
```

Εικόνα 2.18: *Ψευδοκώδικας για την αναδρομική κατασκευή δένδρου ταξινόμησης*

Το προκύπτον δένδρο είναι συνήθως μεγάλο. Για το λόγο αυτό το «κλαδεύουμε» (*prune*) με χρήση διάφορων μεθόδων όπως πχ το *cross-validation*.

Ένα δένδρο ταξινόμησης μπορεί να πάρει δύο ειδών αποφάσεις, Η μία είναι πρόβλεψη σημείου, δηλαδή με την ταξινόμηση σε μία κλάση ή κατηγορία, και η άλλη είναι πρόβλεψη κατανομής ή πιθανότητας, η οποία δίνει μια πιθανότητα για κάθε μία από τις επιμέρους κλάσεις. Για τις προβλέψεις πιθανότητας κάθε τερματικός κόμβος στο δένδρο μας

δίνει μία κατανομή για τις κλάσεις. Αν ο τερματικός κόμβος αντιστοιχεί σε μια ακολουθία απαντήσεων $A = a, B = b, \dots, Q = q$, τότε ιδανικά θα προέκυπτε $\Pr(Y = y | A = a, B = b, \dots, Q = q)$ για κάθε μία από τις πιθανές τιμές y της εξόδου. Ένας απλός τρόπος για να το κατανοήσουμε αυτό είναι να χρησιμοποιήσουμε τις εμπειρικές σχετικές συχνότητες των κλάσεων στον συγκεκριμένο κόμβο. Για παράδειγμα, αν είχαμε 33 περιπτώσεις σε κάποιο συγκεκριμένο φύλλο εκ των οποίων οι 22 ανήκαν στην κατηγορία 1 και οι υπόλοιπες 11 στην κατηγορία 2, τότε το φύλλο εκείνο θα προέβλεπε: *«κατηγορία 1 με πιθανότητα* $^2/_3$ *και κατηγορία 2 με πιθανότητα* $^1/_3$*»*. Αυτή είναι η εκτίμηση **μέγιστης πιθανοφάνειας** (*maximum likelihood estimate*) της πραγματικής πιθανοτικής κατανομής και τη συμβολίζουμε με $\widehat{Pr}(.)$.

Παρότι οι εμπειρικές σχετικές συχνότητες είναι συνεπείς εκτιμήτριες των πραγματικών πιθανοτήτων, δεν υπάρχει κάτι που να μας εξαναγκάζει να τις χρησιμοποιήσουμε. Όταν ο αριθμός των κλάσεων είναι μεγάλος σε σχέση με μέγεθος των δεδομένων μας, θα μπορούσε εύκολα να υπάρχει κάποιο δείγμα που να μην ανήκει σε κάποια από τις κλάσεις. Η εμπειρική σχετική συχνότητα είναι τότε μηδενική. Αυτό είναι καλό στην περίπτωση που και η πραγματική τιμή της πιθανότητας είναι μηδενική, ενώ σε αντίθετη περίπτωση όχι και τόσο καλό.

Για τις προβλέψεις σημείου η καλύτερη στρατηγική εξαρτάται από τη συνάρτηση απωλειών. Εάν λαμβάνεται υπόψη μόνο ο ρυθμός λανθασμένης ταξινόμησης (*misclassification rate*), τότε η καλύτερη πρόβλεψη σε κάθε φύλλο είναι η κλάση με την μεγαλύτερη υπό συνθήκη πιθανότητα στο φύλλο αυτό.

2.4.1.2 Κλάδεμα δένδρων Ταξινίμησης (Classification Tree Pruning)

Το **κλάδεμα** (*prune*) είναι μια τεχνική που χρησιμοποιείται στη μηχανική εκμάθηση με σκοπό τη μείωση του μεγέθους ενός δένδρου απόφασης αφαιρώντας μέρη του δένδρου που προσφέρουν μικρή ταξινομητική ισχύ. Ο στόχος του κλαδέματος είναι διττός. Από τη μία έχει ως στόχο τη μείωση της πολυπλοκότητας του τελικού ταξινομητή και από την άλλη την καλύτερη ακρίβεια της πρόβλεψης λόγω της μείωσης του *overfitting* και της απομάκρυνσης κομματιών του ταξινομητή που βασίζονταν σε λανθασμένα δεδομένα [79].

Ένα κρίσιμο στοιχείο που ανακύπτει σε κάποιον αλγόριθμο δένδρου απόφασης είναι το βέλτιστο μέγεθος του τελικού δένδρου. Ένα τελικό δένδρο μεγάλο σε μέγεθος ενέχει τον κίνδυνο του *overfitting* στα δεδομένα της εκπαίδευσής του και την κακή γενίκευση (*generalization*) για νέα δείγματα. Αντίθετα, ένα μικρό δένδρο μπορεί να παραμερίσει σημαντικές δομικές πληροφορίες για τον δειγματικό χώρο. Μία κοινότυπη στρατηγική για την αντιμετώπιση του προβλήματος αυτού είναι να αφήνεται το δένδρο να μεγαλώνει έως ότου κάθε κόμβος αποκτήσει ένα μικρό αριθμό περιπτώσεων και εν συνεχεία να χρησιμοποιηθούν τεχνικές κλαδέματος για την αφαίρεση των κόμβων που δεν προσφέρουν

πρόσθετη πληροφορία σε αυτό. Το κλάδεμα θα πρέπει να μειώνει το μέγεθος ενός δένδρου μάθησης χωρίς να μειώνει την **ακρίβεια** (*accuracy*) του. [79]

Υπάρχουν πολλές τεχνικές για το κλάδεμα ενός δένδρου απόφασης. Οι δύο σημαντικότερες είναι το **κλάδεμα μείωσης σφάλματος** (*reduced error pruning*) και το **κλάδεμα κόστους πολυπλοκότητας** (*cost complexity pruning*). Το κλάδεμα μείωσης σφάλματος είναι ένας απλούστερος τρόπος κλαδέματος. Ξεκινάει από τα φύλλα και αντικαθιστά κάθε κόμβο με τη συχνότερα εμφανιζόμενη κλάση αυτού. Αν με την αλλαγή αυτή η προβλεπτική ακρίβεια δεν επηρεαστεί τότε η αλλαγή αυτή μένει ως έχει. Παρότι η μέθοδος αυτή είναι σχετική απλοϊκή (*naive*), έχει το πλεονέκτημα της απλότητας και της υπολογιστικής ταχύτητας. Το κλάδεμα κόστους πολυπλοκότητας λειτουργεί αρκετά διαφορετικά. Δημιουργούνται μια σειρά από δένδρα $T_0 \dots T_m$, όπου το T_0 είναι το αρχικό δένδρο και το T_m είναι μονάχα η ρίζα του δένδρου. Στο i βήμα δημιουργείται ένα νέο δένδρο αντικαθιστώντας ένα υποδένδρο από το δένδρο του βήματος $i-1$ με ένα κόμβο-φύλλο (*leaf node*) [79]. Η επιλογή του υποδένδρου που αφαιρείται γίνεται με βάση την ελαχιστοποίηση ενός κριτηρίου, ανάλογα με την περίπτωση που εξετάζεται [80].

2.4.2 Μπεϋζιανοί Ταξινομητές (Bayesian Classifiers)

Στο παρόν υποκεφάλαιο θα περιγραφούν μέθοδοι αναγνώρισης προτύπων που χρησιμοποιήθηκαν στα πλαίσια της παρούσας μελέτης. Πιο συγκεκριμένα στο σημείο αυτό θα γίνει ανασκόπηση των ταξινομητών που βασίζονται στη θεωρία απόφασης του Bayes. Η προσέγγιση που ακολουθείται βασίζεται σε πιθανολογικά επιχειρήματα τα οποία απορρέουν από τη στατιστική φύση των υπό εξέταση χαρακτηριστικών. Σκοπός της τεχνικής αυτής είναι η σχεδίαση ταξινομητών (classifiers) οι οποίοι ταξινομούν ένα άγνωστο μοτίβο (pattern) στην πιθανότερη από τις δεδομένες κλάσεις (classes). Γίνεται έτσι κατανοητό πως στόχος μας πλέον είναι να καταφέρουμε να ορίσουμε τι ακριβώς «σημαίνει πιθανότερη κλάση» [68].

Υποθέτουμε λοιπόν ότι έχουμε ένα πρόβλημα ταξινόμησης με M κλάσεις, $\omega_1,\omega_2,\dots,\omega_M$, και ένα άγνωστο μοτίβο (pattern), το οποίο αναπαριστάται από ένα διάνυσμα χαρακτηριστικών (εισόδων) $\boldsymbol{x}$. Με βάση αυτά υπολογίζουμε τις υποθετικές πιθανότητες $\boldsymbol{P(\omega_i|x),\ i=1,2,\dots,M}$, οι οποίες ονομάζονται **ύστερες (*a posteriori*)** πιθανότητες. Κάθε μία από αυτές αναπαριστά την πιθανότητα το άγνωστο μοτίβο να ανήκει στην αντίστοιχη κλάση ω_i , δεδομένου ότι το διάνυσμα των χαρακτηριστικών (εισόδων) λαμβάνει την τιμή $\boldsymbol{x}$ [68]. Αυτή είναι η βασική ιδέα στην οποία στηρίζεται η υλοποίηση ενός Bayesian ταξινομητή. Στις επόμενες ενότητες θα εμβαθύνουμε στη θεωρία απόφασης του Bayes, στην οποία και στηρίζεται, και στους τρόπους υλοποίησης ενός τέτοιου είδους ταξινομητή.

2.4.2.1 Θεωρία απόφασης Bayes (Bayes Decision Rule)

Όπως έγινε κατανοητό από αυτά που αναφέρθηκαν στην εισαγωγή του κεφαλαίου ένας Bayesian ταξινομητής βασίζεται στον κανόνα του Bayes (*Bayes' rule*). Θα εστιάσουμε λοιπόν σε αυτό το σημείο στην περίπτωση ταξινόμησης σε δύο κλάσεις, ώστε να μπορέσουμε να προχωρήσουμε αργότερα σε πιο ειδικές περιπτώσεις, δηλαδή σε ταξινόμηση σε περισσότερες των δύο κλάσεων.

Θεωρούμε τις δύο κλάσεις, ω_1,ω_2, στις οποίες ανήκουν τα πρότυπα (*patterns*) μας. Εν συνεχεία, θεωρούμε ότι οι πρότερες (a priori) πιθανότητες των κλάσεων $P(\omega_1)$, $P(\omega_2)$ είναι γνωστές. Ακόμα θεωρούμε γνωστές τις class-conditional συναρτήσεις πυκνότητας πιθανότητας (*probability density function/**pdf***) $p(\boldsymbol{x}|\omega_i),i=1,2$, με τις οποίες περιγράφεται η κατανομή των διανυσμάτων χαρακτηριστικών (feature vectors) σε κάθε μία από τις κλάσεις. Στις περιπτώσεις όπου το διάνυσμα χαρακτηριστικών παίρνει μόνο διακριτές τιμές, τότε οι συναρτήσεις πυκνότητας πιθανότητας $p(\boldsymbol{x}|\omega_i)$ γίνονται πιθανότητες $P(\boldsymbol{x}|\omega_i)$ [68].

Με βάση τις παραπάνω υποθέσεις μπορούμε τώρα να υπολογίσουμε τις υπό συνθήκη (*conditional*) πιθανότητες. Σύμφωνα με τον κανόνα του Bayes έχουμε:

$$P(\omega_i \mid x) = \frac{p(x \mid \omega_i)P(\omega_i)}{p(x)} \quad (1)$$

Όπου $p(\boldsymbol{x})$ είναι η συνάρτηση πυκνότητας πιθανότητας του $\boldsymbol{x}$ για την οποία ισχύει:

$$p(x) = \sum_{i=1}^{2} p(x \mid \omega_i)P(\omega_i) \quad (2)$$

Ο κανόνας της Bayesian ταξινόμησης, σύμφωνα με όσα αναφέρθηκαν παραπάνω, εκφράζεται [1]

$$\text{Αν} \quad P(\omega_1 \mid x) > P(\omega_2 \mid x) \quad \alpha\pi o\phi\acute{\alpha}\sigma\iota\sigma\varepsilon \quad \omega_1$$

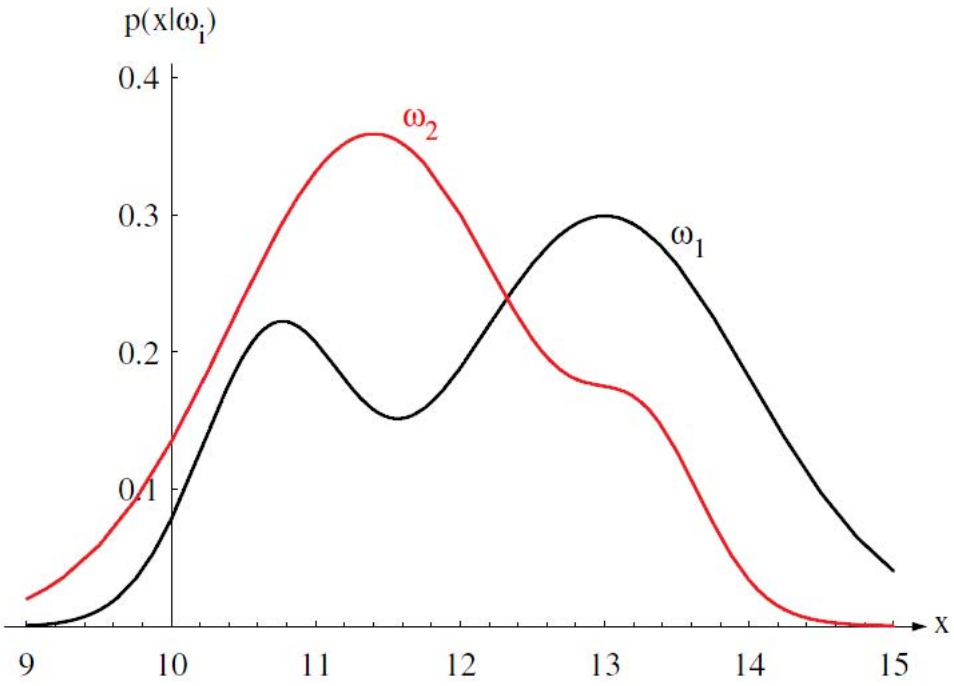
$$\text{Αν} \quad P(\omega_2 \mid x) > P(\omega_1 \mid x) \quad \alpha\pi o\phi\acute{\alpha}\sigma\iota\sigma\varepsilon \quad \omega_2$$

Η ποσότητα $p(x \mid \omega_i)$ στον κανόνα του Bayes (εξ. 1) υπολογίζεται από το σύνολο δεδομένων μας (data set) και μπορεί να θεωρηθεί μία συνάρτηση του διανύσματος-παραμέτρου ω, η οποία καλείται συνάρτησηση πιθανοφάνειας (likelihood function). Η συνάρτηση αυτή δεν είναι πιθανοτική κατανομή και το ολοκλήρωμα αυτής δεν είναι

απαραίτητα ίσο με τη μονάδα. Βάσει του παραπάνω ορισμού της πιθανοφάνειας, μπορούμε «να εκφράσουμε» τον κανόνα του Bayes σε λέξεις [81]

$$posterior \propto likelihood\ x\ prior$$

Ο κανόνας του Bayes μας δείχνει ότι έχοντας γνώση του διανύσματος χαρακτηριστικών (εισόδων), μπορούμε να υπολογίσουμε την πρότερη πιθανότητα (*prior probability*) $P(\omega_i)$ και με χρήση αυτής να καταλήξουμε στην ύστερη (*a posteriori*) πιθανότητα $P(\omega_i|\boldsymbol{x})$, την πιθανότητα δηλαδή το διάνυσμα χαρακτηριστικών $\boldsymbol{x}$ να ανήκει στην κλάση ω_i.



***Εικόνα 2.19**: Παράδειγμα συναρτήσεων πυκνότητας πιθανότητας για 2 διαφορετικές κλάσεις*

2.4.2.2 Απλοϊκός Ταξινομητής Bayes (Naive Bayes Classifier)

Ο ***απλοϊκός ταξινομητής Bayes (naive bayes classifier)*** είναι ένας πιθανοτικός ταξινομητής, οποίος βασίζεται στην εφαρμογή του κανόνα του Bayes με ισχυρές παραδοχές ανεξαρτησίας. **Σε απλά λόγια, ένας naive Bayes ταξινομητής υποθέτει ότι η παρουσία (ή απουσία) ενός συγκεκριμένου χαρακτηριστικού μίας κλάσης δεν σχετίζεται με την**

παρουσία (ή απουσία αντίστοιχα) κανενός άλλου χαρακτηριστικού. Αυτή είναι και η διαφορά του naive Bayes ταξινομητή από έναν άλλο ταξινομητή Bayes [82].

Δεδομένης της ακριβούς φύσης του πιθανοτικού μοντέλου, ένας naive Bayes ταξινομητής μπορεί να εκπαιδευτεί πολύ αποδοτικά με **επιβλεπόμενη εκμάθηση** (*supervised learning*). Σε πολλές πρακτικές εφαρμογές, η εκτίμηση των παραμέτρων ενός τέτοιου ταξινομητή γίνεται με τη μέθοδο της μέγιστης πιθανοφάνειας (*maximum likelihood*). Παρά την απλοϊκότητα της σχεδίασής τους και τις υπεραπλοποιημένες παραδοχές τους, οι naive Bayes ταξινομητές έχουν αποδειχθεί στην πράξη ότι είναι πολύ χρήσιμοι ακόμα και για πραγματικές πρακτικές εφαρμογές [82].

Το μοντέλο ενός ταξινομητή είναι ένα υπό συνθήκη μοντέλο $p(C|F_1, \dots F_n)$ μίας εξαρτημένης μεταβλητής κλάσης C με έναν μικρό αριθμό εξόδων ή κλάσεων υπό συνθήκη με τις μεταβλητές χαρακτηριστικών $F_1, \dots, F_n$. Με χρήση του κανόνα του Bayes έχουμε

$$p(C|F_1, \dots F_n) = \frac{p(C)p(F_1, \dots F_n|C)}{p(F_1, \dots F_n)}$$

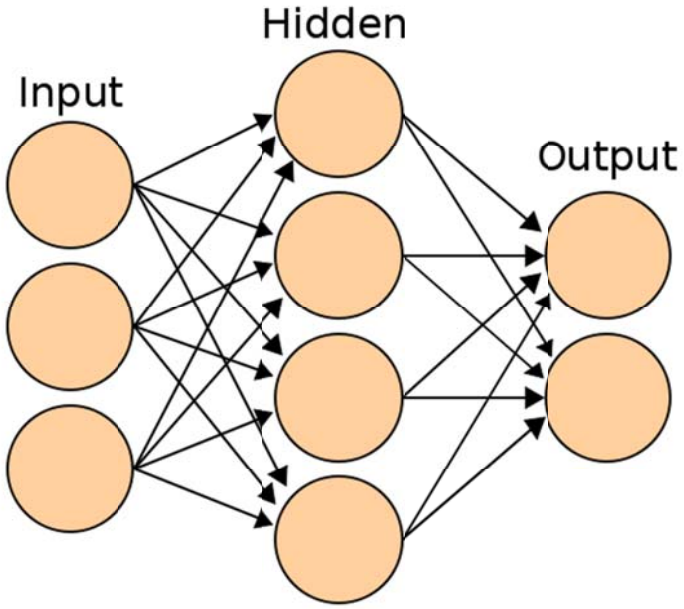
το οποίο αν θέλαμε να το εκφράσουμε με λέξεις θα καταλήγαμε σε αυτό που αναφέρθηκε και προηγουμένως

$$posterior = \frac{prior\ x\ likelihood}{evidence}$$

2.4.3 Νευρωνικά Δίκτυα (Neural Networks)

Ένα **τεχνητό νευρωνικό δικτυο** (ΤΝΔ ή απλά ΝΔ) είναι ένα υπολογιστικό μοντέλο εμπνευσμένο από τη δομή και τις λειτουργίες των βιολογικών δικτύων των νευρώνων. Ένα ΝΔ αποτελείται από ένα σύνολο διασυνδεδεμένων νευρώνων και επεξεργάζεται τις πληροφορίες χρησιμοποιώντας μια διασύνδετη προσέγγιση στον υπολογισμό. Στις περισσότερες των περιπτώσεων το ΝΔ είναι ένα προσαρμοστικό σύστημα που μεταβάλλει τη δομή του βασιζόμενο στις εξωτερικές ή εσωτερικές πληροφορίες που διαρρέουν το δίκτυο κατά τη διάρκεια της φάσης της εκπαίδευσης (*training*) του. Χρησιμοποιούνται συνήθως για

τη διαμόρφωση των σύνθετων σχέσεων μεταξύ των εισόδων και των εξαγόμενων αποτελεσμάτων ή για την ανίχνευση προτύπων (*patterns*) σε ένα σύνολο δεδομένων [84].



***Εικόνα 2.20**: Δομή ενός Νευρωνικού Δικτύου*

Η ιδέα για τη λειτουργία των ΝΔ διατυπώνεται με απλά λόγια στην παρακάτω φράση όπως την καταγράφει ο Simon Haykin στο βιβλίο του "*Neural Networks and Machine Learning*" [83]:

> *"Ένα νευρωνικό δίκτυο είναι ένας τεράστιος παράλληλος επεξεργαστής με κατανεμημένη αρχιτεκτονική, ο οποίος αποτελείται από απλές μονάδες επεξεργασίας και έχει από τη φύση του τη δυνατότητα να αποθηκεύει εμπειρική γνώση και να την καθιστά διαθέσιμη για χρήση. Μοιάζει με τον ανθρώπινο εγκέφαλο σε δύο σημεία:*
>
> 1. *Το δίκτυο προσλαμβάνει τη γνώση από το περιβάλλον του, μέσω μιας διαδικασίας μάθησης*
> 2. *Η ισχύς των συνδέσεων μεταξύ των νευρώνων, που αποκαλείται συναπτικό βάρος, χρησιμοποιείται για την αποθήκευση της γνώσης που αποκτάται"*

2.4.3.1 Λειτουργία των Νευρωνικών Δικτύων

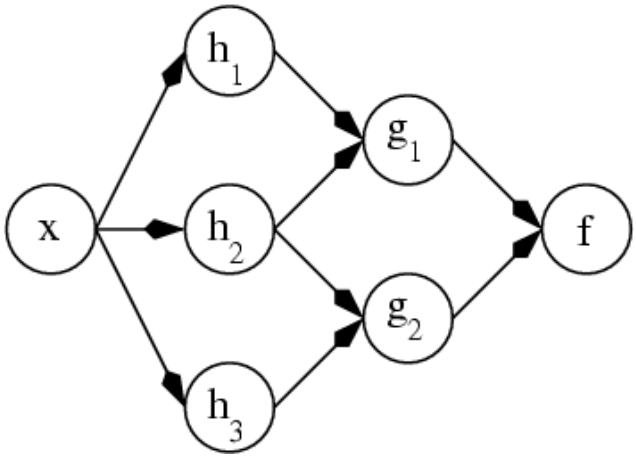
Η λέξη «δίκτυο» στον όρο τεχνητό νευρωνικό δίκτυο αναφέρεται στις διασυνδέσεις μεταξύ των νευρώνων στα διαφορετικά στρώματα κάθε συστήματος. Έστω σύστημα τριών στρωμάτων. Το πρώτο στρώμα εισάγει τους νευρώνες, οι οποίοι στέλνουν τα στοιχεία μέσω των συνάψεων στο δεύτερο στρώμα των νευρώνων, και έπειτα μέσω περισσότερων συνάψεων στο τρίτο στρώμα των νευρώνων παραγωγής. Περισσότερο σύνθετα συστήματα έχουν περισσότερα στρώματα νευρώνων με μερικούς που έχουν αυξημένα στρώματα

νευρώνων εισόδου και νευρώνων εξόδου. Οι συνάψεις αποθηκεύουν τις παραμέτρους αποκαλούμενες «βάρη» (“weights”) που χειρίζονται τα δεδομένα κατά τους υπολογισμούς [64].

Ένα ΝΔ δίκτυο ορίζεται από 3 παραμέτρους:

1. Από το πρότυπο διασύνδεσης μεταξύ των διαφορετικών στρωμάτων των νευρώνων
2. Από τη διαδικασία εκπαίδευσης για την ανανέωση των βαρών των διασυνδέσεων αυτών
3. Από τη συνάρτηση ενεργοποίησης (activation function) που μετατρέπει τη σταθμισμένη είσοδο ενός νευρώνα στην έξοδο ενεργοποίησης

Από μαθηματικής άποψης, η συνάρτηση μεταφοράς ενός νευρωνικού δικτύου $f(x)$ ορίζεται ως σύνθεση άλλων συναρτήσεων $g_i(x)$, οι οποίες με τη σειρά τους μπορούν να οριστούν και αυτές σαν σύνθεση άλλων συναρτήσεων. Αυτό μπορεί εύκολα να αναπαρασταθεί ως μία δομή δικτύου, με τα βέλη να υποδεικνύουν τις εξαρτήσεις μεταξύ των μεταβλητών. Ένας ευρύτατα χρησιμοποιούμενος τύπος σύνθεσης είναι αυτός του *μη γραμμικού σταθμισμένου αθροίσματος,* δηλαδή $f(x) = \boldsymbol{K}(\sum_i w_i g_i(x))$, όπου $\boldsymbol{K}$ είναι η συνάρτηση ενεργοποίησης, η οποία είναι προκαθορισμένη [84].



Εικόνα 2.21*: Γράφος εξαρτήσεων Νευρωνικών Δικτύων*

Κεφάλαιο 3

Μεθοδολογία και Αποτελέσματα

Στο παρόν κεφάλαιο θα αναπτύξουμε τη μεθοδολογία που ακολουθήσαμε για την επίλυση του προβλήματός μας. Σκοπός της διπλωματικής αυτής, όπως αναφέρθηκε και στο υποκεφάλαιο 1.3.2, είναι να προτείνει ένα υποσύνολο των χαρακτηριστικών (*feature subset*) που οδηγεί σε μία καλύτερη πρόβλεψη της κλινικής κατάστασης σε σχέση με το test Pap και με βάση το υποσύνολο αυτό να μπορέσουμε να παράγουμε ορισμένους κανόνες για τη βελτίωση της αξιολόγησης κάθε περιστατικού ξεχωριστά . Για το λόγο αυτό, θα αναπτυχθεί ένα σύστημα το οποίο εφαρμόζει τις τεχνικές που παρουσιάστηκαν συνοπτικά στο προηγούμενο κεφάλαιο στην κατεύθυνση της εξόρυξης δεδομενών από τη βάση μας για την βελτιστοποιημένη ταξινόμηση περιστατικών τραχηλικής ενδοεπιθηλιακής νεοπλασίας.

Αρχικά, στο υποκεφάλαιο 3.1, θα αναφερθούμε σε κάποια αποτελέσματα που προέκυψαν από την αρχική ανάλυση των δεδομένων ολόκληρης της βάσης δεδομένων των περιστατικών που είχαμε στη διάθεσή μας. Από αυτά θα φανεί ο λόγος που μας οδήγησε στο συγκεκριμένο τρόπο με τον οποίο εργαστήκαμε στα πλαίσια της συγκεκριμένης διπλωματικής, και κυρίως γιατί στραφήκαμε στις συγκεκριμένες τεχνικές που αναπτύχθηκαν στο Κεφάλαιο 2. Για την πρόταση του υποσυνόλου των χαρακτηριστικών θα παρουσιαστεί αναλυτικά η μεθοδολογία με την οποία εργαστήκαμε, λόγω του ότι επιδιώξαμε όσο το δυνατόν μεγαλύτερη επιβεβαίωση για τα αποτελέσματα που θα εξαχθούν. Πιο συγκεκριμένα, στο κομμάτι αυτό της εργασίας έγινε χρήση των γενετικών αλγορίθμων σε συνδυασμό με τη θεωρία πληροφορίας για την εξόρυξη της απαραίτητης πληροφορίας για την πλαισίωση των αποτελεσμάτων μας. Αναλυτικότερα, η μέθοδος που ακολουθήθηκε επί του θέματος αυτού θα παρουσιαστεί στο υποκεφάλαιο 3.2. Όσον αφορά τους κανόνες τους οποίους θα προτείνουμε για την αντιμετώπιση περιστατικών στο υποκεφάλαιο 3.2.1, αυτοί είναι προϊόν του υποσυνόλου που θα παράγουμε σε συνδυασμό με τεχνικές δένδρων αποφάσεων που θα χρησιμοποιήσουμε προκειμένου να οπτικοποιήσουμε τα πρότυπα (*patterns*) που εμφανίζονται στη βάση δεδομένων μας.

3.1 Εκτενής Παρουσίαση του Προβλήματος

Όπως αναφέρθηκε και στο Κεφάλαιο 1, το δείγμα πάνω στο οποίο στηρίζεται η παρούσα διπλωματική είναι μεγέθους 212 περιστατικών με την ιστολογική εξέταση αυτών να είναι είτε CIN-1 είτε CIN-2/3, δηλαδή κάποιο είδος τραχηλικής νεοπλασίας, ενώ ταυτόχρονα το αποτέλεσμα του test Pap είναι LgSIL ή HgSIL. Παρόλα αυτά, αυτό είναι μονάχα ένα μικρό μέρος της βάσης των περιστατικών τα οποία είχαμε στα χέρια μας. Μια μικρή ιδέα για τα δεδομένα της βάσης μας δίνεται από τον ακόλουθο πίνακα στον οποίο παρουσιάζονται τα περιστατικά μας σε έναν πίνακα σύγχησης (*confusion matrix*) με βάση τα αποτελέσματα της ιστολογικής και της κυτταρολογικής εξέτασης, εξαιρουμένων των περιπτώσεων ASCUS.

Πίνακας 3.1: Πίνακας σύγχυσης για το test Pap και τη βιοψία

		Pap Smear Results			
		WNL	**LGSIL**	**HGSIL**	**Ca**
Biopsy Result (Histology)	**Negative**	140 (82%)	21	5	0
	CIN 1	27	104 (69%)	16	0
	CIN 2/3	3	25	67 (73%)	0
	Ca	0	1	4	12 (100%)

Από την ανάγνωση του παραπάνω πίνακα λοιπόν οδηγούμαστε στα ακόλουθα συμπεράσματα:

*Στις 170 περιπτώσεις **WNL** αποτελέσματος του PAP, το 82% είναι ορθό και μόλις στο 18% υπερεκτιμά λανθασμένα την αλλοίωση (όχι και τόσο σημαντικό).*

*Στις 151 περιπτώσεις **LgSIL** του PAP, το 69% είναι σωστό αλλά και το 17% υποτιμά εσφαλμένα την πραγματική αλλοίωση υψηλού βαθμού.*

*Στις 92 περιπτώσεις **HgSIL** του PAP, το 73% είναι ακριβές, το 23% υπερεκτιμά την πραγματική αλλοίωση και το 4% την υποτιμά.*

*Στις 12 περιπτώσεις **καρκίνου** του PAP , παρατηρείται 100% ορθότητα.*

Παρατηρώντας αυτά βλέπουμε ότι τα περισσότερα λάθη στην αναγνώριση της πραγματικής κλινικής κατάστασης γίνονται από το test Pap σε περιπτώσεις LgSIL και HgSIL. Με βάση την παραπάνω διαπίστωση ο παραπάνω πίνακας εμφανίζει μεγάλο ενδιαφέρον στα τμήματα που φαίνονται ακολούθως.

Πίνακας 3.2: Πίνακας σύγχυσης για περιστατικά LgSIL-HgSIL και τα αντίστοιχα περιστατικά με ιστολογική CIN-1 – CIN-2+

		Pap Smear Results			
		WNL	**LGSIL**	**HGSIL**	**Ca**
Biopsy Result (Histology)	**Negative**				
	CIN 1		**104**	**16**	
	CIN 2/3		**25**	**67**	
	Ca				

Από τους παραπάνω πίνακες γίνεται κατανοητό ότι τα σημαντικότερα σφάλματα στα οποία υποπίπτει το test Pap αφορούν τις περιπτώσεις CIN-1 και CIN-2/3. Πράγμα το οποίο σημαίνει ότι εάν μπορέσουμε να βρούμε ποια είναι αυτά τα λάθη και κυρίως με ποιο συνδυασμό εξετάσεων προκύπτουν αυτά, τότε θα μπορούσαμε να αυξήσουμε σε μεγάλο βαθμό την ακρίβεια, την ευαισθησία και την ειδικότητα της πρόβλεψης της πραγματικής ιστολογικής κατάστασης των ασθενών. Αυτή η συγκεκριμένη παρατήρηση ήταν που μας έδωσε το κίνητρο να εργαστούμε στο πρόβλημα της βελτίωσης της πρόβλεψης της κλινικής κατάστασης κατά αυτόν τον τρόπο.

Καταρχήν, έγινε κατανοητό πως δε θα μπορούσαμε να εργαστούμε με όλα τα χαρακτηριστικά της βάσης δεδομένων μας, κυρίως λόγω του μεγάλου αριθμού τους. Συνολικά τα διαθέσιμα χαρακτηριστικά που είχαμε στη διάθεση μας ήταν 50. Τα 50 αυτά χαρακτηριστικά ήταν:

- **35** για τους **τύπους των ιών του HPV** που ανιχνεύθησαν
- **3** για την παρουσία ιών **High Risk, Super High Risk και Low Risk**
- **3** για τον αριθμό των υφιστάμενων **υποτύπων του ιού, των High Risk και των Low Risk**
- **5** για την παρουσία των του mRNA **των Super High Risk (*16,18,31,33,45*)**
- **1** για την παρουσία **έστω και ενός** από το **mRNA των Super High Risk**
- **1** για το **test Pap**
- **1** για την έκφραση ή όχι της πρωτεΐνης **p16**
- **1** για το αποτέλεσμα του **flow cytometry test**

Μία σημαντική ιδιομορφ**ία που παρουσίαζαν** τα δεδομένα μας ήταν το γεγονός ότι ήταν **κατηγορικά** (*categorical*). Αυτό εισήγαγε μια παραπάνω δυσκολία στη μέθοδο

εργασίας μας καθώς είναι προφανές ότι είναι σαφώς πιο εύκολος ο χειρισμός δεδομένων τα οποία είναι σε συνεχή μορφή παρά σε διακριτή. Δεδομένου του αριθμού αλλά και της φύσης των χαρακτηριστικών γίνεται κατανοητό πως δεν θα μπορούσαμε να διαχειριστούμε τα δεδομένα αυτά εάν πρώτα δε προβαίναμε σε **δραστική μείωση του αριθμού τους, χωρίς όμως ταυτόχρονα να χάνουμε σημαντική πληροφορία που κρύβεται σε αυτά.** Η δραστική αυτή μείωση θα μπορούσε να επιτευχθεί με την εξόρυξη πληροφορίας από τα δεδομένα μας και τον κατάλληλο χειρισμό της εξαχθείσας αυτής πληροφορίας.

Για αυτό το λόγο ακριβώς αποφασίσαμε ότι η **θεωρία πληροφορίας** θα μπορούσε να είναι ένα χρήσιμο εργαλείο για μας προς την κατεύθυνση αυτή. Με τον τρόπο αυτό θα μπορούσαμε να έχουμε μία εικόνα για την «ποσότητα πληροφορίας» που κρύβει κάθε ένα διαφορετικό χαρακτηριστικό για την τελική κλινική κατάσταση. Ταυτόχρονα θέλαμε όμως να υπάρχει μία κατά το δυνατόν βέλτιστη τεκμηρίωση της βαθμονόμησης αυτών των χαρακτηριστικών. Αυτό ήταν και το γεγονός που μα ώθησε στη χρήση **γενετικών αλγορίθμων** που είναι κατ' εξοχήν εργαλεία που παράγουν λύσεις σε προβλήματα βελτιστοποίησης. Με βάση αυτούς του άξονες προχωρήσαμε στην εξόρυξη πληροφορίας και με τον τρόπο αυτό παρήγαμε υποσύνολα του συνόλου χαρακτηριστικών μας. Τα υποσύνολα αυτά ελέγχθησαν και κατόπιν με αυτά προσπαθήσαμε να παράγουμε κάποιους κανόνες εκτίμησης κινδύνου γιατα επιμέρους περιστατικά. Για την παραγωγή των κανόνων αυτών χρησιμοποιήσαμε δένδρα αποφάσεων και πιο συγκεκριμένα **δένδρα ταξινόμησης**. Παρήχθησαν κάποια αρχικά δένδρα, τα οποία οπτικοποιούσαν τα πρότυπα που προέκυπταν από τη βάση μας με την εφαρμογή των συγκεκριμένων τεχνικών. Όμως σκοπός μας είναι η δημιουργία ενός όσο το δυνατόν μικρότερου και παράλληλα εύρωστου (*robust*) δένδρου, για να αποφύγουμε το *overfiiting* ενός μεγάλου δένδρου και ταυτόχρονα να υπάρχει όσο το δυνατόν καλύτερη γενίκευση (*generalization*) των εξαχθέντων αποτελεσμάτων. Έτσι οδηγηθήκαμε στη χρήση **τεχνικών κλαδέματος** (*pruning*) των δένδρων που παρήχθησαν, με σκοπό την παρουσίαση **ενός και μόνο τελικού δένδρου, στο οποίο θα απεικονίζονται κανόνες για την αντιμετώπιση των περιστατικών με τα οποία ασχολούμαστε.** Πριν προχωρήσουμε στην ανάλυση της μεθοδολογίας παραθέτουμε έναν πίνακα στον οποίο δίνονται οι τιμές που παίρνει το κάθε χαρακτηριστικό καθώς και τι σημαίνει η κάθε μία αυτή τιμή.

Πίνακας 3.3: Επεξήγηση των χαρακτηριστικών της βάσης δεδομένων μας

Χαρακτηριστικό	**Εύρος Τιμών**	**Επεξήγηση**
PAP TEST	***2/3***	***2 – LgSIL/3 - HgSIL***
HPV 6	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 11	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 16	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 18	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 26	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 31	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 33	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***

HPV 35	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 39	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 40	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 42	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 43	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 44	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 45	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 51	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 52	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 53	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 54	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 56	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 58	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 59	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 61	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 62	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 66	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 68	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 70	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 71	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 72	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 73	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 81	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 82	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 83	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 84	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 85	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HPV 89	***0/1***	***Παρουσία/Απουσία του συγκεκριμένου τύπου του HPV***
HIGH RISK/HR	***0/1***	***Παρουσία/Απουσία έστω και ενός τύπου του HPV υψηλού κινδύνου***
SUPER HIGH RISK	***0/1***	***Παρουσία/Απουσία έστω και ενός τύπου του HPV πολύ υψηλού κινδύνου (16,18,31,33,45)***
LOW RISK	***0/1***	***Παρουσία/Απουσία έστω και ενός τύπου του HPV χαμηλού κινδύνου***
No_Subtypes	***0,1,2,3,4, 5,6,7,8***	***Αριθμός όλων των τύπων του ιού που εμφανίζονται σε μία ασθενή***
No_HR	***0,1,2,3,4, 5,6,7***	***Αριθμός των τύπων υψηλού κινδύνου του ιού που εμφανίζονται σε μία ασθενή***
No_LR	***0,1,2,3***	***Αριθμός των τύπων χαμηλού κινδύνου του ιού που εμφανίζονται σε μία ασθενή***
NASBA 0/1	***0/1***	***Παρουσία/Απουσία έστω και ενός τύπου mRNA***
NASBA 16	***0/1***	***Παρουσία/Απουσία του mRNA του HPV 16***
NASBA 18	***0/1***	***Παρουσία/Απουσία του mRNA του HPV 18***
NASBA 31	***0/1***	***Παρουσία/Απουσία του mRNA του HPV 31***
NASBA 33	***0/1***	***Παρουσία/Απουσία του mRNA του HPV 33***
NASBA 45	***0/1***	***Παρουσία/Απουσία του mRNA του HPV 45***
FLOW	***0/1***	***Αρνητικό/Θετικό Flow Cytometry Test***
p16	***0/1***	***Παρουσία/Απουσία πρωτεΐνης p16***

3.2 Μεθοδολογία και Αποτελέσματα

Στα πλαίσια του υποκεφαλαίου θα γίνει λεπτομερής ανάλυση της μεθόδου εργασίας για τον προσδιορισμό υποσυνόλου των χαρακτηριστικών μας και των κανόνων που έχουμε προαναφέρει. Θα παρουσιαστεί αναλυτικά η μεθοδολογία την οποία ακολουθήσαμε για να καταλήξουμε στην παραγωγή των ζητούμενων για μας αποτελεσμάτων. Όπως θα φανεί παρακάτω εργαστήκαμε με αρκετές διαφορετικές μεθόδους και ο λόγος για τον οποίο συνέβη αυτό ήταν το γεγονός ότι θέλαμε να υπάρχει όσο το δυνατόν μεγαλύτερη επαλήθευση των αποτελεσμάτων μας, καθώς όπως εξηγήθηκε παραπάνω είναι καίρια τόσο μία πρόταση υποσυνόλου χαρακτηριστικών όσο και η παραγωγή των κανόνων που συνεπάγεται το υποσύνολο αυτό. Έτσι, με βάση τον τρόπο εργασίας ευελπιστούμε να καταλήξουμε σε συμπεράσματα τα οποία θα είναι **απολύτως τεκμηριωμένα** τόσο από τα αριθμητικά αποτελέσματα που θα προκύψουν (ποσοστά ορθής πρόβλεψης, ευαισθησία και ειδικότητα) όσο και από τις επιμέρους μεθόδους που χρησιμοποιήθηκαν.

Η μεθοδολογία που ακολουθήσαμε για να μειώσουμε τον αριθμό των χαρακτηριστικών μας, δηλαδή για την παραγωγή του ζητούμενου υποσυνόλου χαρακτηριστικών βασίζεται, όπως προαναφέρθηκε, στους **γενετικούς αλγορίθμους**, τη **θεωρία πληροφορίας** για την **εξόρυξη δεδομένων** και σε **μεθόδους ταξινόμησης**. Σε αυτό το σημείο ήρθε η ώρα να γίνει μια λεπτομερής ανάλυση του πώς εργαστήκαμε.

3.2.1 Παρουσίαση Μεθόδων και Αποτελεσμάτων

3.2.1.1 Εύρεση βέλτιστου υποσυνόλου χαρακτηριστικών με χρήση γενετικών αλγορίθμων και θεωρίας πληροφορίας με εφαρμογή τεχνικής περιτυλίγματος

Κύριο εργαλείο για να μπορέσουμε να καταλήξουμε σε υποσύνολο χαρακτηριστικών ήταν οι γενετικοί αλγόριθμοι. Μεγαλύτερο ρόλο όμως στην εξαγωγή και παράλληλα την ορθότητα των αποτελεσμάτων παίζει η παραμετροποίηση του γενετικού αλγορίθμου των οποίο θα χρησιμοποιήσουμε. Οφείλουμε να υπενθυμίσουμε ότι τα δεδομένα με τα οποία εργαζόμαστε είναι μόνο τα περιστατικά του test Pap με LgSIL και HgSIL, των οποίων η βιοψία ήταν είτε CIN-1 είτε CIN-2/3.

Όπως έγινε κατανοητό στο Κεφάλαιο 2 ένα πολύ σημαντικό ζήτημα για τους γενετικούς αλγορίθμους είναι το θέμα της επιλογής του αρχικού πληθυσμού, δηλαδή της

δεξαμενής των λύσεων μέσα στις οποίες θα αναζητήσουμε τη βέλτιστη λύση. Για το πρόβλημα αυτό, ως λύση/άτομο (*individual*) του γενετικού αλγόριθμου ορίζεται κάθε υποσύνολο χαρακτηριστικών μήκους *l*.

Βασικό ρόλο στη δημιουργία των πληθυσμών μας έπαιξε η θεωρία πληροφορίας. Συγκεκριμένα, υλοποιήσαμε δύο διαφορετικούς τρόπους για το πώς θα δημιουργείται ο αρχικός πληθυσμός στον οποίο ο γενετικός αλγόριθμος θα αναζητεί τη λύση. Ο πρώτος τρόπος περιλαμβάνει τη χρήση της ***σχετικής εντροπίας*** η οποία χρησιμοποιείται προκειμένου να ιεραρχήσουμε τα χαρακτηριστικά μας με βάση την ποσότητα της πληροφορίας που εμπεριέχεται σε κάθε ένα από αυτά (*feature ranking*). Στη συνέχεια, ανάλογα με το επιθυμητό μήκος *l* του υποσυνόλου χαρακτηριστικών, δημιουργείται ένας πληθυσμός από *l/n* υποσύνολα (όπου n το πλήθος των χαρακτηριστικών), κάθε ένα από τα οποία αποτελείται από μοναδικά χαρακτηριστικά, με βάση το μέτρο σχετικής εντροπίας του κάθε χαρακτηριστικού. Ο τρόπος αυτός δημιουργίας αρχικού πληθυσμού οδηγεί σε μικρό πληθυσμό, για το λόγο αυτό επιλέξαμε να συμπεριλαμβάνονται επιπλέον υποσύνολα λύσεων με βάση το ***κριτήριο mRMR***, που αναλύθηκε στο Κεφάλαιο 2. Σημειώνεται πως αν και ο πληθυσμός είναι αρκετά μικρότερος από τους συνήθεις πληθυσμούς γενετικών αλγόριθμων, ορισμένα άτομα αυτού είναι εξ αρχής βελτιστοποιημένα σε σχέση με τα υπόλοιπα, καθώς βάσει των μέτρων της θεωρίας πληροφορίας περιέχουν περισσότερη και πιο χρήσιμη πληροφορία. Τα άτομα αυτά είναι αυτά τα οποία εν συνεχεία θα δημιουργήσουν τους απογόνους, οι οποίοι μέσω των τεχνικών διασταύρωσης και μετάλλαξης θα οδηγήσουν στη βέλτιστη λύση. Ο τρόπος αυτός εφαρμογής του γενετικού αλγόριθμου επιλέχθηκε προκειμένου να μειωθεί το υπολογιστικό κόστος του χωρίς όμως να χάνεται χρήσιμη πληροφορία.

Ο δεύτερος τρόπος σχηματισμού του πληθυσμού είναι ακριβώς ο ίδιος με τον πρώτο, προσθέτοντας ακόμα ένα κριτήριο. Το κριτήριο αυτό έχει να κάνει με την ιεράρχηση των χαρακτηριστικών με βάση την ***καμπύλη ROC*** για κάθε επιμέρους χαρακτηριστικό και την αντίστοιχη κλάση, κατά τρόπο ανάλογο όπως συνέβη και στην πρώτη περίπτωση με κριτήριο τη σχετική εντροπία. Το επιπλέον αυτό κριτήριο χρησιμοποιήθηκε προκειμένου να δημιουργήσουμε μεγαλύτερο πληθυσμό. Με αυτό τον τρόπο γίνεται η επιλογή των αρχικών ατόμων (*individuals*), τα οποία στη συνέχεια θα υποστούν μεταλλάξεις (*mutation*) και διασταυρώσεις (*crossover*) και θα παράξουν ένα τελικό υποσύνολο χαρακτηριστικών.

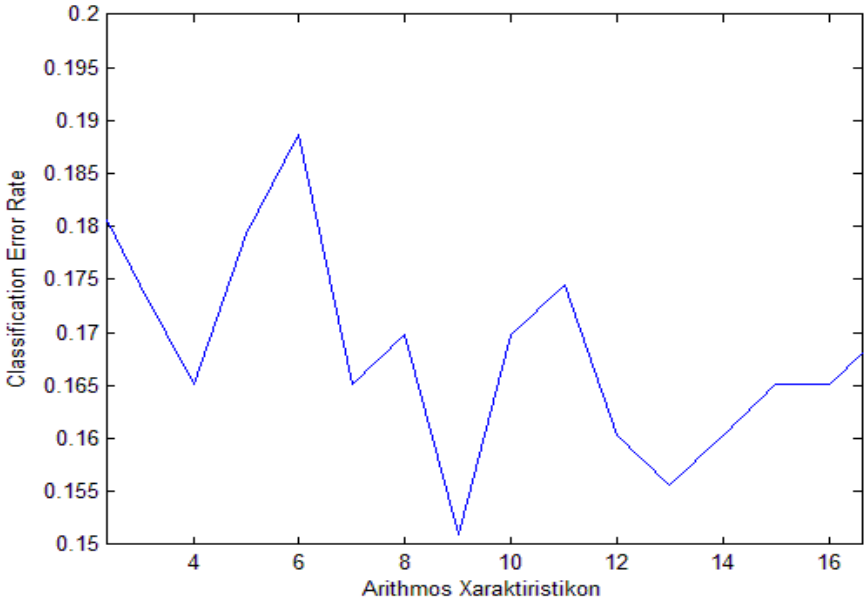
Εξίσου σημαντική παράμετρος για έναν γενετικό αλγόριθμο είναι και η επιλογή της συνάρτησης καταλληλότητας (*fitness function*). Η συνάρτηση καταλληλότητας, την οποία υλοποιήσαμε, περιλαμβάνει έναν **απλοϊκό ταξινομητή Bayes (naive Bayes Classifier)**, αλλά και ένα κριτήριο που θεσπίσαμε εμείς. Η λειτουργία της συνάρτησης καταλληλότητας έχει ως εξής:

- Αρχικά υλοποιούμε έναν ταξινομητή naive Bayes με βάση τον πληθυσμό που έχει υπολογιστεί με έναν από τους δύο τρόπους που προαναφέρθηκαν
- Στη συνέχεια, προχωρούμε στη δοκιμή του ταξινομητή μας δίνοντας όλα τα δεδομένα μας τόσο στην εκπαίδευσή του (*training*) όσο και στη δοκιμασία του *testing (resubstitution)*
- Υπολογίζουμε το Σφάλμα Ταξινόμησης (*Classification Error Rate*) του ταξινομητή
- Το τελικό αποτέλεσμα το οποίο μας επιστρέφει η fitness function μας προκύπτει από την παρακάτω εξίσωση

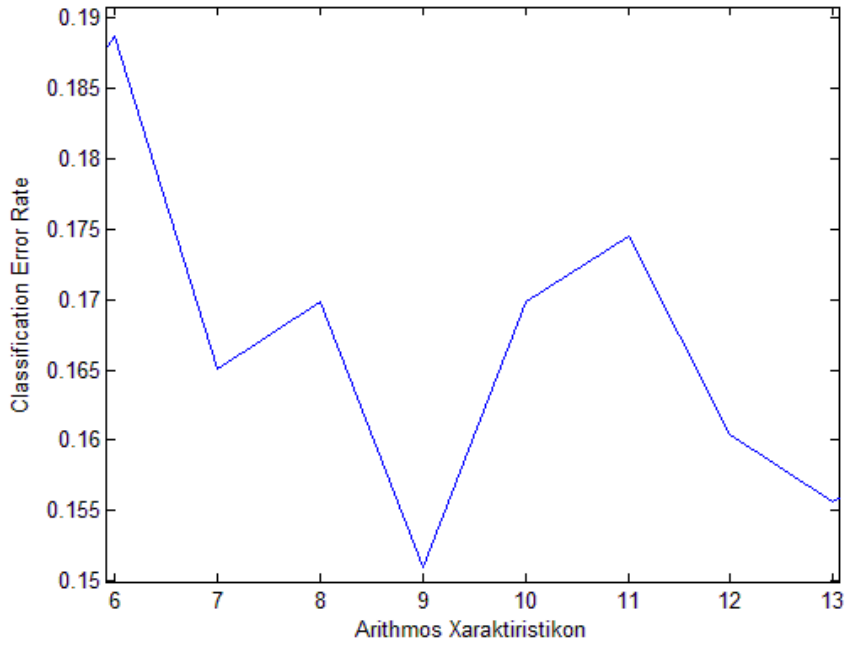
$$fitness = 100*ClassificationErrorRate + 20*mean(min(posterior\_prob))$$

Είναι πολύ σημαντικό να γίνει κατανοητό το τι ακριβώς αντιπροσωπεύει η συγκεκριμένη *fitness function* και τι ακριβώς αυτή κάνει. Ο όρος ClassificationErrorRate αντιπροσωπεύει, όπως δηλώνεται και από το όνομά του, το σφάλμα ταξινόμησης. Ο δεύτερος όμως όρος είναι αυτός που παίζει το σημαντικότερο ρόλο. Στον όρο αυτό υπολογίζεται ο μέσος όρος των μικρότερων ύστερων (*posterior*) πιθανοτήτων κατά την ταξινόμηση, δηλαδή το κατά πόσο πιθανό είναι **να μην ανήκει** το κάθε ξεχωριστό δείγμα στην κλάση που ταξινομήθηκε. Με τον τρόπο αυτό αποφεύγουμε την πιθανότητα να επιλεχθεί ως βέλτιστο ένα υποσύνολο από τον πληθυσμό μας το οποίο **ναι μεν έχει το μικρότερο σφάλμα αλλά υπάρχει μεγάλη αμφιβολία κατά τη διαδικασία της ταξινόμησης**. Όπως γίνεται σαφές από τη μεθοδολογία ακολουθούμε σε αυτή την περίπτωση μια *wrapper approach* για την εξαγωγή του υποσυνόλου χαρακτηριστικών, ενώ αντίθετα για τη δημιουργία του αρχικού πληθυσμού η προσέγγισή μας έμοιαζε περισσότερο με μία *filter approach*.

Αφού, λοιπόν, εξηγήσαμε με ποιους τρόπους παραμετροποιούμε τον γενετικό αλγόριθμο που χρησιμοποιούμε, θα προχωρήσουμε τώρα σε αυτή καθεαυτή την ανάλυσή μας. Χρησιμοποιώντας τον πρώτο τρόπο δημιουργίας του πληθυσμού που αναφέραμε παραπάνω επιχειρούμε να βρούμε ένα υποσύνολο χαρακτηριστικών για το οποίο ελαχιστοποιείται η *fitness function*. Αφού προσδιοριστεί το υποσύνολο αυτό προωθείται εν συνεχεία σε έναν naive Bayes ταξινομητή (διαφορετικό από αυτόν της συνάρτησης καταλληλότητας) και με τις διαδικασίες *resubstitution* και *cross validation* υπολογίζουμε την απόδοση αυτού, προκειμένου να ελεγχθεί περαιτέρω η λύση. Ο δευτερογενής αυτός έλεγχος πραγματοποιείται προκειμένου να ανιχνεύσουμε επιπλέον και το βέλτιστο πλήθος *l*. Για το σκοπό αυτό η ανωτέρω διαδικασία επαναλαμβάνεται για αριθμό χαρακτηριστικών *l* από 1 έως και 20. Υποθέτουμε ότι το υποσύνολο στο οποίο θέλουμε να καταλήξουμε δεν μπορεί να έχει μέγεθος πάνω από 20 χαρακτηριστικά για λόγους που έχουν εξηγηθεί προηγουμένως. Ο κώδικάς μας υλοποιήθηκε σε περιβάλλον Matlab©**.** Στις παρακάτω εικόνες απεικονίζεται το γράφημα του πλήθους των χαρακτηριστικών σε σχέση με το σφάλμα του δευτερογενούς ταξινομητή. Στην πρώτη φαίνονται τα αποτελέσματα για όλο το εύρος του αριθμού των χαρακτηριστικών, ενώ στη δεύτερη έχουμε εστιάσει στο ελάχιστο που εμφανίζει η γραφική αυτή παράσταση.



***Εικόνα 3.1**: Σφάλμα Ταξινόμησης συναρτήσει του αριθμού των χαρακτηριστικών που χρησιμοποιούμε στο υποσύνολο μας*



***Εικόνα 3.2**: Εστίαση στο σημείο που εμφανίζεται ελάχιστο σφάλμα ταξινόμησης*

Στα παραπάνω γραφήματα απεικονίζεται το σφάλμα της ταξινόμησης συναρτήσει της αύξησης του αριθμού των χαρακτηριστικών που χρησιμοποιούνται. Όπως φαίνεται από τα παραπάνω διαγράμματα παρατηρούμε ότι το σφάλμα της τελικής ταξινόμησης γίνεται ελάχιστο για 9 χαρακτηριστικά και είναι περίπου 0,15 πράγμα που σημαίνει ότι τα

αποτελέσματα της ταξινόμησης είναι ορθά κατά 85%. Τα 9 χαρακτηριστικά τα οποία επέλεξε ο γενετικός αλγόριθμος κατά τη διαδικασία αυτή είναι τα ακόλουθα:

- *HPV 66*
- *PAP TEST*
- *HPV 84*
- *HPV 11*
- *No_Subtypes*
- *HPV 40*
- *HPV 31*
- *HIGH RISK*
- *HPV 58*

Τα στατιστικά μέτρα ταξινόμησης του δευτερογενούς (τελικού σταδίου) μπεϋζιανού ταξινομητή που προέκυψαν με βάση το υποσύνολο για την τελική ταξινόμηση των δειγμάτων είναι τα ακόλουθα:

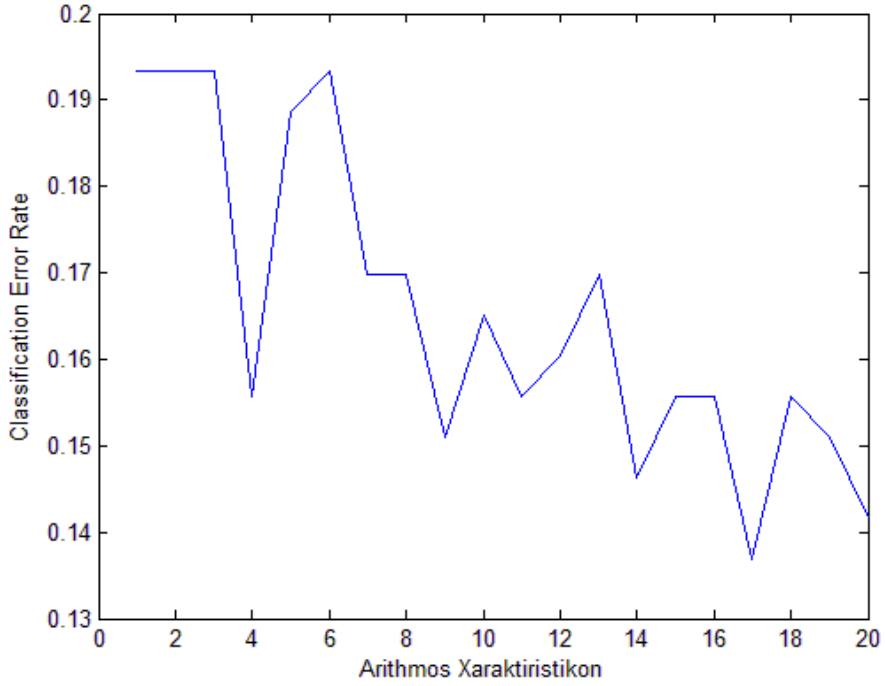
Ακρίβεια: ***84,9 %***

Ευαισθησία: ***90,8 %***

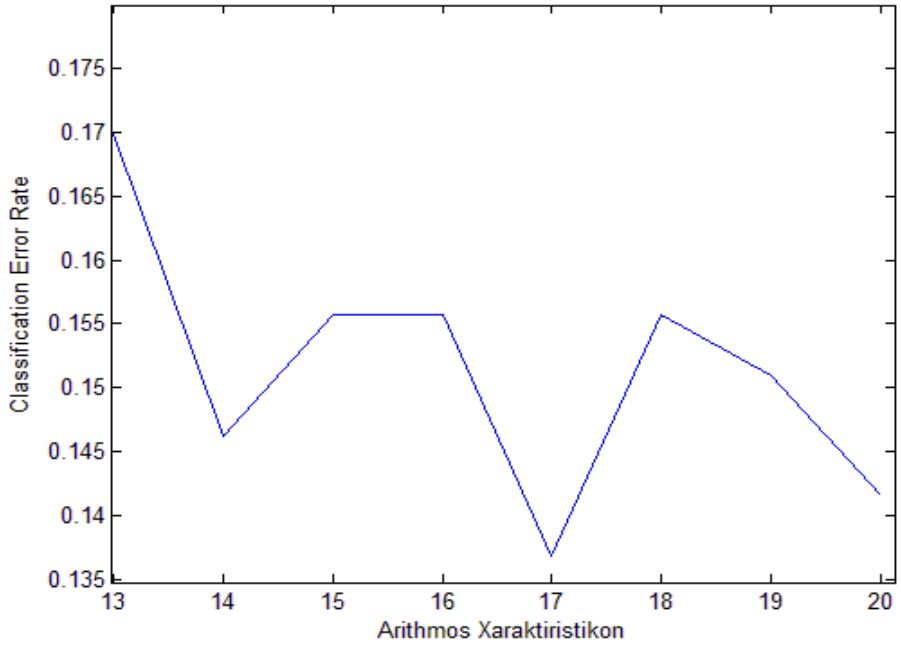
Ειδικότητα: ***77,2 %***

Σημειώνεται πως τα αποτελέσματα αυτά αφορούν τον έλεγχο του ταξινομητή με τη μέθοδο resubstitution, καθώς το πλήθος των δειγμάτων μας δεν ήταν αρκετά μεγάλο προκειμένου να παράξουμε αξιόπιστα αποτελέσματα με άλλες μεθόδους δοκιμής. Με τα αποτελέσματα αυτά αποκτήσαμε μία πρώτη εικόνα για το ποια χαρακτηριστικά θα μπορούσαν να επηρεάζουν την τελική διάγνωση. Έτσι συνεχίζουμε στο ίδιο πλαίσιο εργασίας, αυτή τη φορά όμως χρησιμοποιούμε διαφορετικό τρόπο για τη δημιουργία του αρχικού πληθυσμού. Συγκεκριμένα χρησιμοποιούμε τη δεύτερη μέθοδο την οποία αναλύσαμε παραπάνω, δηλαδή ο πληθυσμός σχηματίζεται με κριτήρια ***σχετικής εντροπίας***, κριτήρια για τη ***ROC*** και κριτήρια ***mRMR***. Είναι σαφές πως στην περίπτωση αυτή ο αρχικός πληθυσμός είναι μεγαλύτερος, οπότε και ο γενετικός αλγόριθμός μας αναζητεί λύσεις σε ένα μεγαλύτερο εύρος αρχικών ατόμων-υποσυνόλων χαρακτηριστικών. Σαν *fitness function* χρησιμοποιείται η ίδια ακριβώς με προηγουμένως για τους ίδιους ακριβώς λόγους που αναλύθηκαν προηγουμένως. Επαναλαμβάνουμε τη διαδικασία και τα βήματα που ακολουθήσαμε και προηγουμένως. Εκτελούμε έναν βρόχο από 1 έως 20, ο οποίος υποδεικνύει το μέγεθος που έχει το υποσύνολο σε κάθε μία επανάληψη. Θέλουμε να δούμε σε ποιον αριθμό χαρακτηριστικών θα εμφανίσει το σφάλμα ταξινόμησης ελάχιστο αλλά επίσης και ποια είναι

τα χαρακτηριστικά για τα οποία συμβαίνει αυτό. Εκτελώντας τα παραπάνω καταλήξαμε στις παρακάτω εικόνες που είναι αντίστοιχες με αυτές που παρουσιάστηκαν προηγουμένως.



***Εικόνα 3.3**: Σφάλμα Ταξινόμησης συναρτήσει του αριθμού των χαρακτηριστικών που χρησιμοποιούμε στο υποσύνολό μας*



***Εικόνα 3.4**: Εστίαση στο σημείο που εμφανίζεται ελάχιστο σφάλμα ταξινόμησης*

Από τα γραφήματα αυτά παρατηρούμε ότι η ελαχιστοποίηση του σφάλματος ταξινόμησης πραγματοποιείται για υποσύνολο 17 χαρακτηριστικών και το σφάλμα αυτό είναι περίπου 0,135, που αντιστοιχεί σε ακρίβεια της τάξης του 86,5%. Το υποσύνολο που επιλέχθηκε αυτή τη φορά περιείχε τα ακόλουθα χαρακτηριστικά:

- *NASBA 16*
- *NASBA 45*
- *PAP TEST*
- *FLOW*
- *NASBA 31*
- *HPV 52*
- *HPV 44*
- *HPV 56*
- *No_LR*
- *HPV 73*
- *HPV 39*
- *HPV 18*
- *No_HR*
- *HPV 85*
- *HPV 56*
- *HPV 66*

Τα στατιστικά που προέκυψαν με βάση το υποσύνολο αυτό για την τελική ταξινόμηση των δειγμάτων (με χρήση του μπεϋζιανού ταξινομητή) είναι τα ακόλουθα: (δοκιμή με resubstitution):

Ακρίβεια: ***86,3 %***

Ευαισθησία: ***90,8 %***

Ειδικότητα: ***80,4 %***

Παρατηρούμε, λοιπόν, πως με περισσότερα χαρακτηριστικά αυξάνουμε την ακρίβεια του συστήματός μας. Πιο συγκεκριμένα, αυξάνοντας τον αρχικό πληθυσμό βρίσκουμε ελάχιστο σφάλμα σε υποσύνολο με περισσότερα χαρακτηριστικά. Παρόλα αυτά είναι προτιμότερο να έχουμε λίγο μεγαλύτερο σφάλμα παρά ένα σχετικά μεγάλο υποσύνολο χαρακτηριστικών. Έτσι, με βάση την παραπάνω εργασία μας αποκτήσαμε μία εκτίμηση για το μέγεθος του υποσυνόλου των χαρακτηριστικών, και το τοποθετούμε από 9 ως 17 χαρακτηριστικά.

Πριν προχωρήσουμε, όμως, θα πρέπει να μπορέσουμε να αξιολογήσουμε αν τα στατιστικά αυτά που προκύπτουν είναι ικανοποιητικά. Το κατά πόσο είναι ικανοποιητικά ή όχι κρίνεται μέσω της σύγκρισής τους με τα αντίστοιχα στατιστικά που προκύπτουν για το test Pap στα ίδια περιστατικά στη βάση μας. Για το test Pap έχουμε

Ακρίβεια: ***80,66 %***

Ευαισθησία: ***86,67 %***

Ειδικότητα: ***72,83 %***

Είναι προφανές λοιπόν πως η μελέτη μας μέχρι στιγμής έχει νόημα αφού σε όλους ακριβώς τους στατιστικούς δείκτες παρατηρούμε άνοδο, και μάλιστα σημαντική, πράγμα που σημαίνει πως με βάση την εργασία μας θα μπορούσαμε να προβλέψουμε καλύτερα την κλινική κατάσταση των ασθενών που παρουσιάζουν test Pap LgSIL ή HgSIL σε σχέση με το test Pap μόνο του. Και ασχολούμαστε συγκεκριμένα με το test Pap γιατί από όλα τα διαθέσιμα tests είναι το μόνο το οποίο μας δίνει μια απευθείας συσχέτιση με την ιστολογία βασιζόμενο στην κυτταρολογία.

3.2.1.2 Εύρεση βέλτιστου υποσυνόλου χαρακτηριστικών με χρήση γενετικών αλγορίθμων και θεωρίας πληροφορίας με εφαρμογή τεχνικής φίλτρου

Αφού μπορέσαμε να βρούμε περίπου πόσο μεγάλο πρέπει να είναι το ζητούμενο υποσύνολό μας, δοκιμάσαμε και μία ελαφρώς διαφορετική προσέγγιση για τον προσδιορισμό του υποσυνόλου μας, βασιζόμενοι σε μια αντίστοιχη μελέτη [54]. Εργαζόμαστε και πάλι με τους γενετικούς αλγορίθμους αλλά με διαφορετικές παραμέτρους. ΟΙ διαφορές αυτές έγκεινται κυρίως στην επιλογή του αρχικού πληθυσμού και στη *fitness function*. Στην περίπτωση αυτή η επιλογή του πληθυσμού γίνεται με τρόπο τυχαίο (*random population*), ενώ σαν *fitness function* χρησιμοποιείται το κριτήριο ***mRMR***, για το οποίο επιθυμούμε τη μεγιστοποίησή του (filter approach). Αρχικά επιλέγουμε τον αριθμό των χαρακτηριστικών που επιθυμούμε να έχει το υποσύνολό μας. Εν συνεχεία, ο γενετικός αλγόριθμος με βάση αυτό τον αριθμό δημιουργεί τον αρχικό πληθυσμό με άτομα μεγέθους του αριθμού που έχουμε ορίσει πριν. Κατόπιν ο γενετικός προβαίνει στις μεταλλάξεις (*mutation*) και τις διασταυρώσεις (*crossover*) έως ότου εκπληρωθεί το κριτήριο της μεγιστοποίησης του mRMR. Τέλος, τα χαρακτηριστικά που επιλέγονται δοκιμάζονται σε ένα δέντρο ταξινόμησης για να μπορέσουμε να δούμε κατά πόσο έχουν βάση τα αποτελέσματα που εξάγονται. Ένα πρόβλημα που προκύπτει εδώ είναι το γεγονός της τυχαίας επιλογής του αρχικού πληθυσμού. Και αυτό έχει σαν αποτέλεσμα την εισαγωγή στοχαστικότητας στην εξαγωγή του τελικού

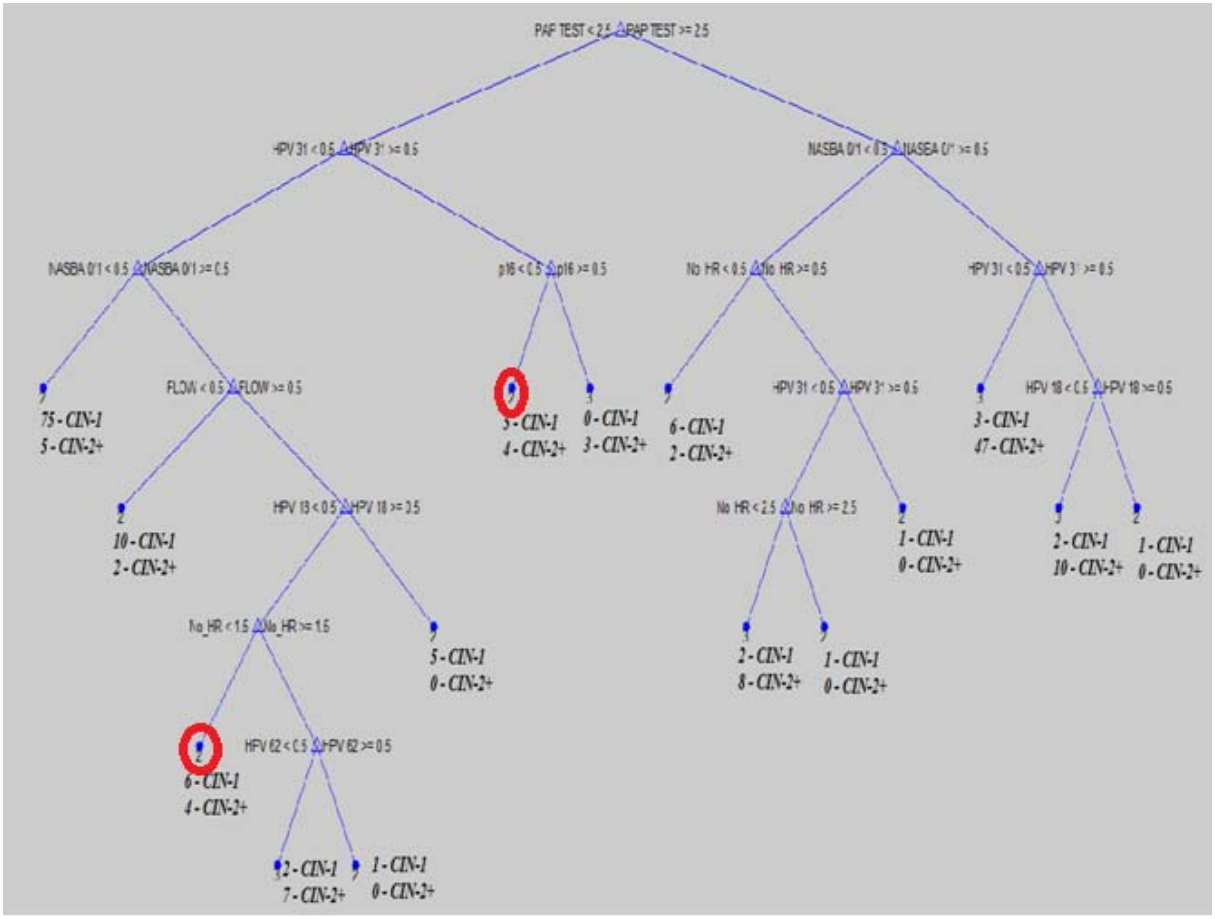
υποσυνόλου. Αυτή ακριβώς τη στοχαστικότητα θέλαμε να εξαλείψουμε και για το λόγο αυτό εφαρμόσαμε μία στρατηγική στατιστικού τύπου. Εκτελέσαμε τον αλγόριθμό μας σε ένα βρόχο 100 επαναλήψεων και σημειώναμε πόσες φορές χρησιμοποιούταν το κάθε χαρακτηριστικό στο τελικό υποσύνολο που παρήγαγε ο συγκεκριμένος γενετικός αλγόριθμος. Μετά από δοκιμές καταλήξαμε να χρησιμοποιήσουμε τα 12 συχνότερα χρησιμοποιούμενα χαρακτηριστικά τα οποία είναι τα εξής:

- *PAP TEST*
- *HPV 18*
- *HPV 31*
- *HPV 33*
- *HPV 35*
- *HPV 62*
- *No_HR*
- *NASBA 0/1*
- *NASBA 16*
- *NASBA 31*
- *FLOW*
- *p16*

Με το υποσύνολο αυτό προβαίνουμε στην κατασκευή ενός δένδρου ταξινόμησης. Με το δένδρο αυτό προσπαθούμε να ανιχνεύσουμε τα *patterns* όπως αυτά προκύπτουν με βάση τα δείγματα που έχουμε στη διάθεσή μας. Τελικά, με βάση αυτά τα χαρακτηριστικά κατασκευάστηκε ένα **δένδρο ταξινόμησης** το οποίο χρησιμοποιούσε μόλις 8 χαρακτηριστικά. Τα χαρακτηριστικά που χρησιμοποιήθηκαν στο δένδρο, και ουσιαστικά είναι και αυτά που ορίζουν τα *patterns* μας, είναι τα παρακάτω

- PAP TEST
- HPV 18
- HPV 31
- NASBA 0/1
- FLOW
- p16
- No_HR
- HPV 62

Το δένδρο το οποίο παράχθηκε τελικά φαίνεται στην παρακάτω εικόνα. Στα φύλλα του δένδρου με 2 υποδηλώνεται η κατάσταση CIN-1, ενώ με 3 η κατάσταση CIN-2+.



***Εικόνα 3.5**: Δένδρο Ταξινόμησης με βάση το σύνολο χαρακτηριστικών που έχουμε παράξει*

Αφού λοιπόν προχωρήσαμε στην εξόρυξη δεδομένων όπως προαναφέρθηκε, τα αντίστοιχα στατιστικά μέτρα που προέκυψαν για το δένδρο αυτό είναι τα εξής (δοκιμή με resubstitution):

Ακρίβεια: ***87,7 %***

Ευαισθησία: ***92,5 %***

Ειδικότητα: ***81,5 %***

Παρατηρούμε λοιπόν πως και σε αυτή την περίπτωση βρισκόμαστε αρκετά ψηλότερα σε σχέση με το test Pap, πράγμα που δείχνει την ορθότητα των διαδικασιών μας. Ένα εξίσου σημαντικό θέμα που παρατηρήσαμε από την έρευνά μας με τα δένδρα ταξινόμησης είναι ότι μπορούμε να εντοπίσουμε καταστάσεις στις οποίες παρατηρούμε πως ακόμα και η βιοψία εμφανίζει αβεβαιότητα. Πρόκειται για περιπτώσεις όπου έχουμε ταξινόμηση ενός δείγματος σε μία από τις 2 κλάσεις, αλλά οι πιθανότητες είναι πολύ κοντά στο 50-50, δηλαδή ουσιαστικά η ταξινόμηση στην κλάση είναι σχεδόν τυχαία. Δύο τέτοιες περιπτώσεις είναι κυκλωμένες στην παραπάνω εικόνα. Οι περιπτώσεις αυτές είναι πολύ σημαντικές, καθώς αν

μπορέσουμε να εντοπίσουμε ποιοι συνδυασμοί χαρακτηριστικών οδηγούν σε αυτές θα μπορούσαν να βοηθηθούν σημαντικά οι γιατροί στην επίτευξη ενός σωστού *triage*. Πολύ σημαντικό ήταν ακόμα το αποτέλεσμα του δένδρου, καθώς μας μείωσε ακόμα περισσότερο το μέγεθος του υποσυνόλου των χαρακτηριστικών που χρησιμοποιούσαμε. Αυτά ήταν που μας έκαναν να καταλάβουμε ότι περεταίρω ενασχόληση με τα δένδρα ταξινόμησης θα οδηγούσε σε ακόμα καλύτερα αποτελέσματα. Το παραπάνω δένδρο, όμως, έχει ένα μειονέκτημα το οποίο αξιολογήσαμε και κρίναμε ότι έπρεπε να προχωρήσουμε και ότι η εργασία μας δεν είχε τελειώσει εδώ. Το μειονέκτημά του είναι ότι το δένδρο είναι σχετικά μεγάλο και επίσης ότι υπάρχουν αρκετά φύλλα τα οποία περιγράφουν μικρό αριθμό περιστατικών. Αυτό σημαίνει πρακτικά ότι είναι πιθανόν να ενέχει ο κίνδυνος υπερεξειδίκευσης του δένδρου στα δεδομένα με αποτέλεσμα να μην έχουμε την ζητούμενη γενίκευση. Προς το παρόν όμως βάζουμε μία άνω τελεία στη χρήση των δένδρων και θα επανέλθουμε ξανά σε αυτά παρακάτω.

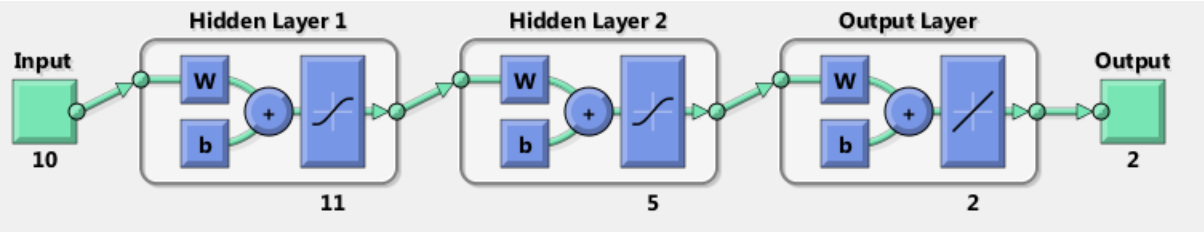
3.2.1.3 Εκτίμηση χρησιμότητας χαρακτηριστικών με βάση τη χρησιμοποίηση στα επιμέρους παραγόμενα υποσύνολα χαρακτηριστικών

Από όσα περιγράφησαν προηγουμένως, η στατιστικού τύπου στρατηγική που ακολουθήσαμε αποδείχθηκε εκ των πραγμάτων ότι απέφερε σχετικά καλά αποτελέσματα. Αποφασίσαμε έτσι να την εφαρμόσουμε και στην αρχική περίπτωση του γενετικού αλγορίθμου, αυτή τη φορά με λίγο διαφορετικό τρόπο, αλλα με τον ίδιο τελικό στόχο, δηλαδή την εξόρυξη πληροφορίας. Υλοποιήσαμε 4 διαφορετικές συναρτήσεις για τη σύσταση του αρχικού πληθυσμού του γενετικού αλγορίθμου, χρησιμοποιώντας συνδυασμούς των μέτρων θεωρίας πληροφορίας και του κριτηρίου ROC, με παρόμοια διαδικασία με αυτή που περιγράφηκε στο κεφ. 3.2.1.1. Παρατηρήσαμε προηγουμένως ότι για ένα υποσύνολο 12 χαρακτηριστικών πετύχαμε ένα αρκετά ικανοποιητικό αποτέλεσμα. Άρα θα πρέπει να δοκιμάσουμε εάν μπορούμε να βρούμε ένα ακόμα μικρότερο υποσύνολο. Έτσι, «τρέχοντας» τον γενετικό αλγόριθμο με τις διαφορετικές συναρτήσεις δημιουργίας αρχικού πληθυσμού αναζητήσαμε τα βέλτιστα υποσύνολα έως και 11 χαρακτηριστικών. Εν συνεχεία καταγράφηκε πόσες φορές εμφανιζόταν κάθε ξεχωριστό χαρακτηριστικό σε κάθε βέλτιστο υποσύνολο. Καταλήξαμε λοιπόν ότι 13 είναι τα χαρακτηριστικά που χρησιμοποιούνται συχνότερα στη σύσταση των επιμέρους υποσυνόλων μια και από το $14^{ο}$ και μετά ο αριθμός εμφάνισης έπεφτε αισθητά. Τα χαρακτηριστικά στα οποία καταλήξαμε στο σημείο αυτό είναι τα ακόλουθα:

- *PAP TEST*
- *HIGH RISK*
- *HPV 56*
- *NASBA 0/1*
- *NASBA 16*
- *HPV 62*
- *FLOW*
- *HPV31*
- *p16*
- *NASBA 31*
- *No_HR*
- *HPV 66*
- *No_Subtypes*

3.2.1.4 Αποτελέσματα απόδοσης του υποσυνόλου χαρακτηριστικών με χρήση νευρωνικών δικτύων

Στη συνέχεια, εργαστήκαμε με τα παραπάνω χαρακτηριστικά. Επιλέξαμε να εργαστούμε με νευρωνικά δίκτυα με σκοπό να ταξινομήσουμε τα δεδομένα μας στις αντίστοιχες κλάσεις. Αρχικά, λοιπόν, επιλέξαμε ένα νευρωνικό 2 στρωμάτων (*2-layer*). Το επόμενο κρίσιμο ζήτημα σε ένα νευρωνικό δίκτυο είναι η επιλογή των νευρώνων. Για την επιλογή των νευρώνων δημιουργήσαμε ένα βρόχο με σκοπό να βρει, για τους διάφορους συνδυασμούς του αριθμού των νευρώνων, το αντίστοιχο σφάλμα ταξινόμησης. Από τα παραπάνω προέκυψε ότι τα καλύτερα αποτελέσματα στην κατεύθυνση αυτή τα πετύχαμε με 11 νευρώνες στο πρώτο στρώμα και 5 στο δεύτερο. Το νευρωνικό το οποίο χρησιμοποιήσαμε τελικά είχε τη μορφή που φαίνεται στην παρακάτω εικόνα.



Εικόνα 3.6*: Δομή του νευρωνικού δικτύου που χρησιμοποιήσαμε*

Καταλήγοντας λοιπόν στο βέλτιστο νευρωνικό δίκτυο προχωρήσαμε στο επόμενο βήμα. Έχει ιδιαίτερο ενδιαφέρον να ελεγχθεί το κατά πόσο επηρεάζονται τα αποτελέσματά

μας από την αφαίρεση κάποιων εκ των χαρακτηριστικών. Αυτό πραγματοποιείται προκειμένου να μειώσουμε στο ελάχιστο δυνατό το πλήθος των χαρακτηριστικών του βέλτιστου υποσυνόλου. Ήδη από πριν είχαμε παρατηρήσει ότι τα 3 τελευταία χαρακτηριστικά δεν έπαιζαν καίριο ρόλο στη διαμόρφωση της τελικής ταξινόμησης. Εν συνεχεία λοιπόν αφαιρέσαμε ένα-ένα χαρακτηριστικά και ελέγξαμε τα στατιστικά μέτρα με σκοπό να διαπιστώσουμε μέχρι πόσα χαρακτηριστικά μπορούμε να αφαιρέσουμε χωρίς να έχουμε σημαντική άνοδο του σφάλματος. Αφαιρώντας τα τρία τελευταία χαρακτηριστικά (*No_HR, HPV 66, No_Subtypes*) παρατηρήσαμε ότι τα στατιστικά μέτρα δε **μεταβάλλονταν καθόλου και μάλιστα ήταν και αρκετά υψηλά.**

Ακρίβεια: ***90,09 %***

Ευαισθησία: ***89,17 %***

Ειδικότητα: ***91,3 %***

Πηγαίνοντας όμως στα 9 χαρακτηριστικά η ακρίβειά μας έπεφτε κάτι παραπάνω από 1 ποσοστιαία μονάδα και κρίναμε ότι δεν υπήρχε λόγος για ένα τέτοιο *tradeoff*. Έτσι, στο σημείο αυτό καταλήξαμε στο παρακάτω υποσύνολο χαρακτηριστικών, το οποίο βλέπουμε ότι παρουσιάζει αρκετά κοινά χαρακτηριστικά με όσα είχαν εξαχθεί προηγουμένως. Αυτό μόνο ως θετικό θα μπορούσε να χαρακτηριστεί καθώς παρατηρούμε πως με διάφορες μεθόδους εργασίας καταλήγουμε σε αποτελέσματα που εν ολίγοις συναληθεύονται.

- *PAP TEST*
- *HIGH RISK*
- *HPV 56*
- *NASBA 0/1*
- *NASBA 16*
- *HPV 62*
- *FLOW*
- *HPV31*
- *p16*
- *NASBA 31*

Παρατηρούμε λοιπόν ότι τα τελικά χαρακτηριστικά είναι 10 στον αριθμό, πράγμα που έρχεται σε συμφωνία με την πρώτη μας εικασία ότι το υποσύνολο μας θα πρέπει να έχει από 9 έως 17 χαρακτηριστικά.

3.2.1.5 Δένδρα Ταξινόμησης και Risk Model για την παραγωγή κανόνων

Εφόσον μπορέσαμε να καταλήξουμε σε ένα υποσύνολο για τα χαρακτηριστικά μας, τώρα θα εστιαστούμε στην προσέγγισή μας για την παραγωγή κανόνων. Θα επανέλθουμε έτσι στο σημείο αυτό στα δένδρα ταξινόμησης. Στον ακόλουθο πίνακα παρουσιάζονται τα χαρακτηριστικά στα οποία είχαμε καταλήξει στην προγενέστερη δουλειά μας με τα δένδρα ταξινόμησης και σε αυτά από την εργασία με τους γενετικούς αλγορίθμους.

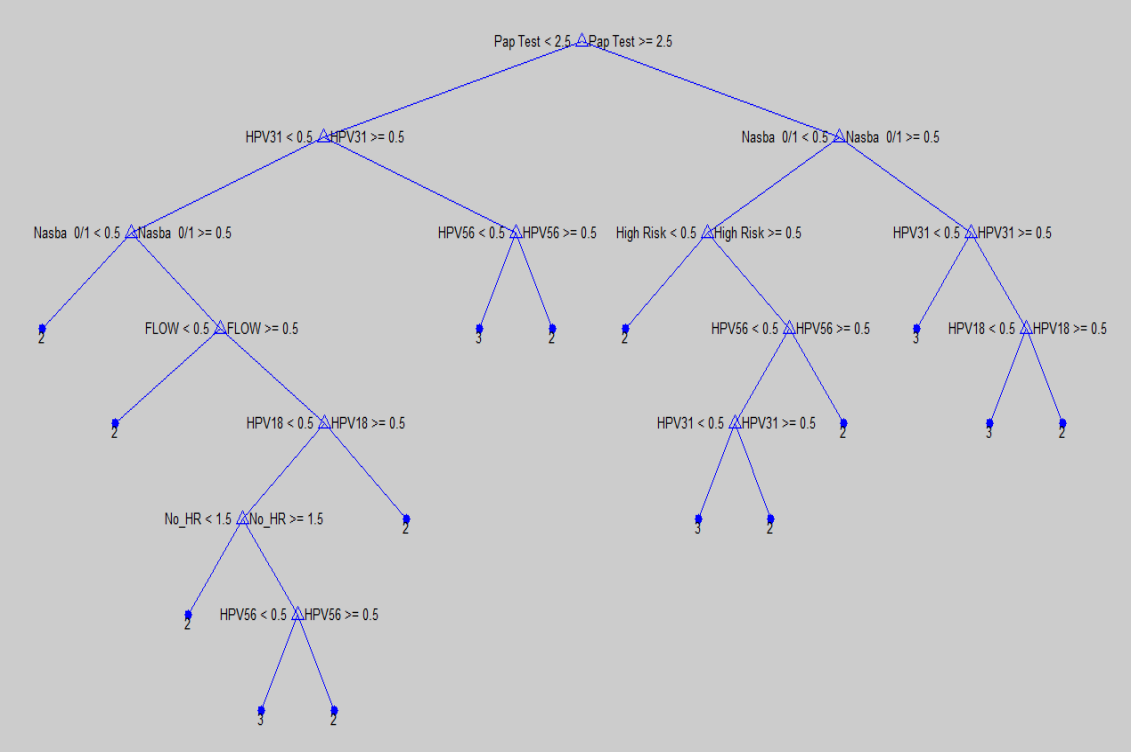
Πίνακας 3.4: Σύγκριση υποσυνόλων χαρακτηριστικών που προέκυψαν για 2 μεθόδους

Από Γενετικό Αλγόριθμο	***Από Δένδρο Ταξινόμησης***
PAP TEST	*PAP TEST*
HIGH RISK	*HPV 18*
HPV 56	*HPV 33*
NASBA 0/1	*NASBA 0/1*
HPV 62	*HPV 62*
NASBA 16	*NASBA 16*
FLOW	*FLOW*
HPV 31	*HPV 31*
p16	*p16*
NASBA 31	*NASBA 31*
No_HR	*No_HR*
HPV 66	*HPV 35*
No_Subtypes	

Στον παραπάνω πίνακα είναι σκιασμένα τα χαρακτηριστικά που η παρουσία και η σημασία τους επαληθεύεται και από τις δύο μεθόδους. Αυτά αποτελούν μία βάση για το ποια πρέπει να είναι τα χαρακτηριστικά με τα οποία θα προσπαθήσουμε να κατασκευάσουμε ένα δένδρο απόφασης, και κατ' επέκταση του κανόνες αντιμετώπισης των διαφόρων περιστατικών. Με βάση αυτά ξεκινήσαμε να δοκιμάζουμε διάφορους συνδυασμούς των χαρακτηριστικών με σκοπό να παράξουμε ένα δένδρο ταξινόμησης. Με αρκετές δοκιμές καταλήξαμε ότι το βέλτιστο δένδρο εξαγόταν με τα παρακάτω χαρακτηριστικά, τα οποία είχαν εν πολλοίς προβλεφθεί στα προηγούμενα βήματα της εργασίας μας:

- *PAP TEST*
- *HIGH RISK*
- *HPV56*
- *NASBA 0/1*
- *NASBA 16*
- *FLOW*
- *HPV31*
- *p16*
- *No_HR*
- *HPV18*

Με αυτά τα χαρακτηριστικά προχωρήσαμε στην εξαγωγή ενός δένδρου ταξινόμησης στηριζόμενοι στα δεδομένα μας. Το δένδρο αυτό φαίνεται στην ακόλουθη εικόνα.



Εικόνα 3.7*: Δένδρο Ταξινόμησης για το υποσύνολο χαρακτηριστικών*

Στην παραπάνω εικόνα με 2 υποδηλώνεται η κλινική κατάσταση CIN-1 και με 3 η CIN-2+. Το δένδρο αυτό είχε σαν αποτέλεσμα τα ακόλουθα στατιστικά μέτρα (σε resubstitution):

Ακρίβεια: ***88,21 %***

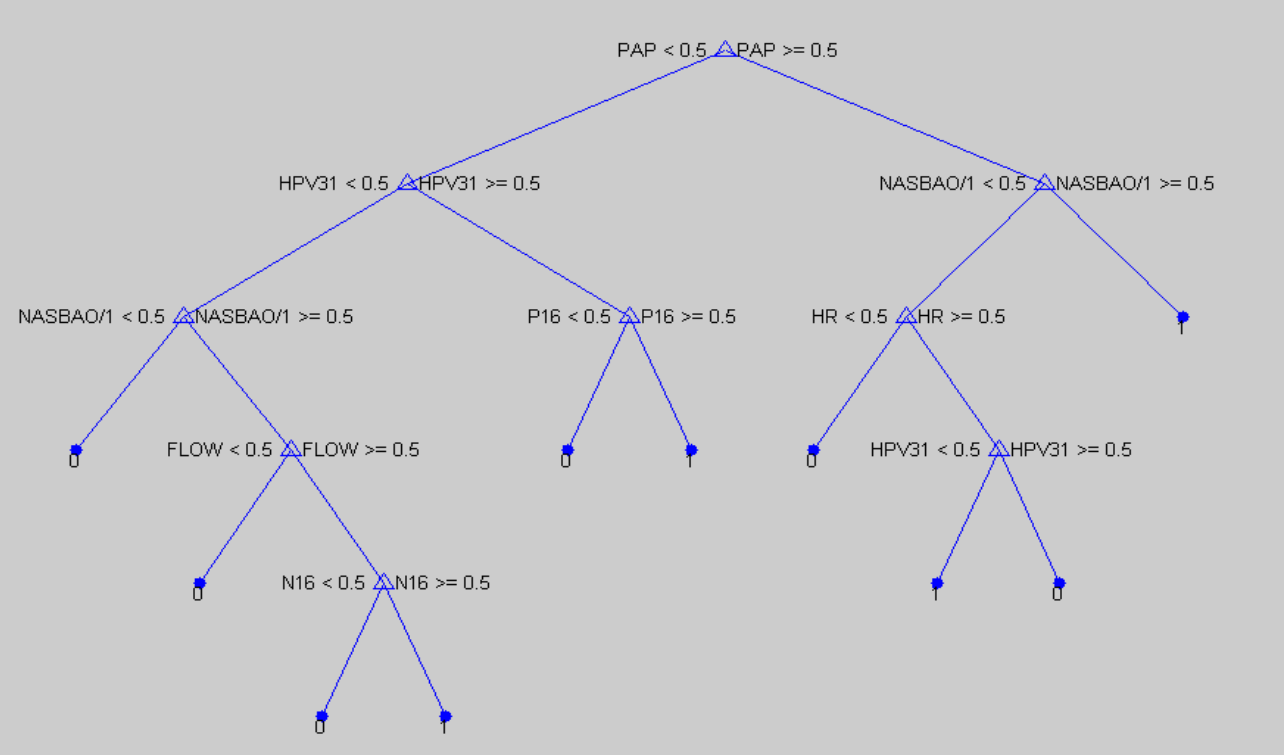
Ευαισθησία: ***90,83 %***

Ειδικότητα: ***82,50 %***

Εφόσον μπορέσαμε να κατασκευάσουμε και ένα δεύτερο δένδρο με εξίσου πολύ καλά στατιστικά μέτρα, προχωρήσαμε σε κλάδεμα (*pruning*) των δένδρων με σκοπό την παραγωγή ενός τελικού δένδρου με λιγότερα κλαδιά και φύλλα. Αυτό είναι ζητούμενο καθώς θέλουμε το τελικό δένδρο, από το οποίο θα προκύψουν και οι τελικοί κανόνες, να εμφανίζει καλή γενίκευση και να μην υπερεξειδικεύεται στα δεδομένα, πράγμα το οποίο θα το καταστήσει και εύρωστο (*robust*). **Δε θα πρέπει όμως το τελικό αυτό δένδρο να ληφθεί ως ένα κλασικό δένδρο απόφασης με την κανονική έννοια του όρου, αλλά περισσότερο σαν ένα *risk model*.** Το δένδρο αυτό δεν θα περιλαμβάνει όλα τα χαρακτηριστικά στα οποία καταλήξαμε προηγουμένως. Αυτό το κάναμε για να μπορέσουμε να καταλήξουμε σε ένα απλό και σαφές *risk model* και επίσης για να βρούμε τα βέλτιστα ελάχιστα δένδρα και να κατασκευάσουμε ένα ***random forest***. Θα ήταν εντελώς λανθασμένη η προσέγγιση του δένδρου αυτού σαν ένα κλασικό δένδρο ταξινόμησης, καθώς σκοπός του δένδρου αυτού δεν είναι να δώσει μια εικόνα για το ποιοι συνδυασμοί μας οδηγούν στις αντίστοιχες κλάσεις, αλλά να μας δώσει μία **εκτίμηση για την πιθανότητα** να ανήκει κάποιος συγκεκριμένος συνδυασμός χαρακτηριστικών σε μία κλάση. Προκειμένου να καταλήξουμε στις πιο έγκυρες αποφάσεις θέσαμε τον εξής εμπειρικό κανόνα:

«Μια απόφαση λαμβάνεται υπόψη όταν το class membership του κόμβου είναι πάνω από το 10% της κλάσης»

Με βάση όσα έχουν αναφερθεί προηγουμένως, προχωράμε με τεχνικές κλαδέματος στα δένδρα που έχουμε κατασκευάσει, εφαρμόζοντας τον παραπάνω εμπειρικό κανόνα. Έτσι καταλήγουμε τελικά στο ακόλουθο risk model όπως αυτό οπτικοποιείται στο δένδρο της ακόλουθης εικόνας.



***Εικόνα 3.8**: Το Risk Model στο οποίο κατέληξε η εργασία μας*

Παρατηρούμε, λοιπόν, ότι το παραπάνω δένδρο είναι σαφώς μικρότερο από τα προηγούμενα δένδρα, οπότε μπορούμε να πούμε ότι οι συνδυασμοί των χαρακτηριστικών που οδηγούν στα φύλλα είναι οι κανόνες τους οποίους θέλουμε να εξάγουμε. Οι κανόνες αυτοί είναι:

Πίνακας 3.5: Κανόνες που προκύπτουν από το Risk Model

PAP TEST=LgSIL + HPV 31=0 + NASBA 0/1=0 => CIN-1
PAP TEST =LgSIL + HPV 31=0 + NASBA 0/1=1 + FLOW=0 => CIN-1
PAP TEST =LgSIL + HPV 31=0 + NASBA 0/1=1 + FLOW=1 + NASBA 16=0 => CIN-1
PAP TEST =LgSIL + HPV 31=0 + NASBA 0/1=1 + FLOW=1 + NASBA 16=1 => CIN-2+
PAP TEST =LgSIL + HPV 31=1 + p16=0 => CIN-1
PAP TEST =LgSIL + HPV 31=1 + p16=1 => CIN-2+
PAP TEST =HgSIL + NASBA 0/1=1 => CIN-2+
PAP TEST =HgSIL + NASBA 0/1=0 + HR=0 => CIN-1
PAP TEST =HgSIL + NASBA 0/1=0 + HR=1 + HPV 31=0 => CIN-2+
PAP TEST =HgSIL + NASBA 0/1=0 + HR=1 + HPV 31=1 => CIN-1

Πέρα από αυτή την απλή ανάγνωση, το ***risk model*** που περιγράφεται στο δένδρο μπορεί να έχει και δεύτερη ανάγνωση η οποία παρουσιάζει ακόμα μεγαλύτερο ενδιαφέρον. Από την περαιτέρω μελέτη του ανωτέρω δέντρου εξάγονται συμπεράσματα σχετικά με τους συνδυασμούς των χαρακτηριστικών που μας οδηγούν σε καλύτερη ανίχνευση της κλινικής κατάστασης. Τα συμπεράσματα αυτά παρουσιάζονται συνοπτικά στη συνέχεια.

1. Οταν το PAP test είναι LgSIL **ΤΟΤΕ** η πιθανότητα να έχω CIN-2+ είναι 19,4%.
    a. Οταν το PAP test είναι LgSIL **ΚΑΙ** ο HPV31 είναι αρνητικός **ΤΟΤΕ** η πιθανότητα να έχω CIN-2+ είναι 15% (η πιθανότητα μειώνεται πράγμα που σημαίνει ότι αρνητικός HPV31 επιβεβαιώνει το PAP test).
        i. Όταν το PAP test είναι LgSIL **ΚΑΙ** ο HPV31 είναι αρνητικός **ΚΑΙ** NASBA=0 TOTE η πιθανότητα να έχω CIN-2+ είναι 6%. (Όπως φαίνεται ο NASBA συμπληρώνει πάρα πολύ καλά το PAP και το HPV και αρνητικός NASBA επιβεβαιώνει κατά πολύ το PAP=LgSIL).
        ii. Όταν το PAP είναι LgSIL **ΚΑΙ** ο HPV31 είναι αρνητικός **ΑΛΛΑ** ο NASBA=1 TOTE η πιθανότητα να έχω CIN-2/3 ανεβαίνει στο 35% (η εμφάνιση **έστω και ενός τύπου NASBA** ανεβάζει κατακόρυφα το ποσοστό να έχω CIN-2+).
        iii. Όταν το PAP είναι LgSIL **ΚΑΙ** ο HPV31 είναι αρνητικός **ΑΛΛΑ** NASBA=1 **ΚΑΙ** το FLOW είναι θετικό, **ΤΟΤΕ** η πιθανότητα να έχω CIN-2+ ανεβαίνει στο 45%, **ΕΠΟΜΕΝΩΣ** στην περίπτωση αυτή απαιτείται ιδιαίτερη προσοχή (ίσως επιβεβαίωση με βιοψία).
    b. Όταν το PAP είναι LgSIL **ΚΑΙ** ο HPV31 είναι θετικός **ΤΟΤΕ** η πιθανότητα να έχω CIN-2+ ανεβαίνει κατακόρυφα και συγκεκριμένα είναι 58%.

ΕΠΟΜΕΝΩΣ απαιτείται επιβεβαίωση με βιοψία. Ωστόσο, ο κόμβος αυτός έχει *class membership* ακριβώς 10%, επομένως θεωρούμε πως το ποσοστό αυτό μπορεί να μειωθεί εάν υπήρχαν στη διάθεση μας παραπάνω δεδομένα.

2. Όταν το PAP είναι HgSIL **ΤΟΤΕ** η πιθανότητα να έχω CIN-1 είναι 19,3%.
    a. Όταν το PAP είναι HgSIL **ΚΑΙ** Nasba=0 **ΤΟΤΕ** η πιθανότητα να έχω CIN-1 είναι 50%.
        i. Όταν το PAP είναι HgSIL **ΚΑΙ** Nasba=0 **ΚΑΙ** HR=0 **ΤΟΤΕ** η πιθανότητα να έχω CIN1 είναι 75%. **ΕΠΟΜΕΝΩΣ** στην περίπτωση αυτή απαιτείται ιδιαίτερη προσοχή (π.χ. επανάληψη του PAP).
        ii. Όταν το PAP είναι HgSIL **ΚΑΙ** Nasba=0 **ΚΑΙ** HR=1 TOTE η πιθανότητα να έχω CIN-1 είναι 33%.

    b. Όταν το PAP είναι HgSIL **ΚΑΙ** NASBA είναι 1 **ΤΟΤΕ** η πιθανότητα να έχω CIN1 είναι 9%. Το NASBA σε αυτή την περίπτωση επιβεβαιώνει σε πολύ μεγάλο βαθμό το Pap test).

3.3 Συμπεράσματα και σχολιασμός

Στο σημείο αυτό θα συγκεντρώσουμε όλα τα αποτελέσματα τα οποία προέκυψαν από την εργασία μας με την εφαρμογή της μεθοδολογίας που παρουσιάστηκε αναλυτικά στο προηγούμενο υποκεφάλαιο. Συνολικά τα αποτελέσματα της εργασίας μας συμπυκνώνονται στον ακόλουθο πίνακα.

Πίνακας 3.6: Συγκεντρωτικά Αποτελέσματα

Μέθοδος Εργασίας	**Αριθμός Χαρακτη-ριστικών**	**Υποσύνολο Χαρακτηριστικών**	**Στατιστικά Μέτρα Απόδοσης**
Μόνο Pap test	***1***	Pap test	*Ακρίβεια:* ***80,66 %*** *Ευαισθησία:* ***86,67 %*** *Ειδικότητα:* ***72,83 %***

Γενετικός Αλγόριθμος με fitness function το σφάλμα ταξινόμησης 2	***17***	*NASBA 16* *NASBA 45* *PAP TEST* *FLOW* *NASBA 31* *HPV 52* *HPV 44* *HPV 56* *No_LR* *HPV 73* *HPV 39* *HPV 18* *No_HR* *HPV 85* *HPV 56* *HPV 66*	*Ακρίβεια:* ***86,3 %*** *Ευαισθησία:* ***90,8 %*** *Ειδικότητα:* ***80,4 %***
Γενετικός Αλγόριθμος με fitness function το κριτήριο mRMR	***12***	*PAP TEST* *HPV 18* *HPV 31* *HPV 33* *HPV 35* *HPV 62* *No_HR* *NASBA 0/1* *NASBA 16* *NASBA 31* *FLOW* *p16*	*Ακρίβεια:* ***87,7 %*** *Ευαισθησία:* ***92,5 %*** *Ειδικότητα:* ***81,5 %***
Νευρωνικά Δίκτυα	***10***	*PAP TEST* *HIGH RISK* *HPV 56* *NASBA 0/1* *NASBA 16* *HPV 62* *FLOW* *HPV31* *p16* *NASBA 31*	*Ακρίβεια:* ***90,09 %*** *Ευαισθησία:* ***89,17 %*** *Ειδικότητα:* ***91,3 %***

Από τον παραπάνω πίνακα γίνεται παραπάνω από εμφανές το πλεονέκτημα της χρήσης τεχνικών αναγνώρισης προτύπων σε συνδυασμό με την θεωρία πληροφορίας για την εξόρυξη πληροφορίας και κατ' επέκταση την επιλογή υποσυνόλου χαρακτηριστικών για τη διάγνωση περιστατικών με τραχηλική ενδοεπιθηλιακή νεοπλασία. Τα στατιστικά μέτρα απόδοσης είναι σε κάθε περίπτωση καλύτερα, και μάλιστα σε αρκετά μεγάλο βαθμό, σε σύγκριση με τη χρησιμοποίηση μόνο της κυτταρολογίας για την εκτίμηση της κλινικής κατάστασης. Τα αποτελέσματά μας υποδεικνύουν επίσης σαφώς ποια είναι τα επιμέρους χαρακτηριστικά που εμπεριέχουν μεγάλη διαγνωστική ισχύ στις περιπτώσεις CIN. Παρατηρούμε, λοιπόν, ότι το test Pap αποτελεί βασικό χαρακτηριστικό καθως δεν λείπει από κανένα υποσύνολο χαρακτηριστικών. Μεγάλη ποσότητα πληροφορίας όμως υπάρχει και σε άλλα χαρακτηριστικά όπως το NASBA 0/1, το NASBA 16, το NASBA 31, το HPV 31, το FLOW και το p16.

Εξαιρετικά μεγάλο ενδιαφέρον παρουσιάζει η απουσία του τύπου 16 του HPV, ο οποίος θεωρείται από τους γιατρούς ως ο πλέον επικίνδυνος τύπος για την ανάπτυξη καρκίνου του τραχήλου της μήτρας. Ο HPV 16 δεν ανιχνεύτηκε σε κανένα από τα παραχθέντα υποσύνολα χαρακτηριστικών, όποιον τρόπο εργασίας και αν ακολουθήσαμε. Αν αυτό το γεγονός επαληθευτεί από τη μελέτη μεγαλύτερου αριθμού περιστατικών όταν αυτά συγκεντρωθούν στη βάση μας, θα αποτελέσει μία σημαντική ανατροπή στον τρόπο που αντιμετωπίζονται ως σήμερα τα περιστατικά με τραχηλική ενδοεπιθηλιακή νεοπλασία. Σημειώνεται, ότι η μελέτη αυτή αφορά μονάχα τα περιστατικά στα οποία το Pap test είναι LgSIL ή HgSIL και η βιοψία είναι CIN-1 ή CIN-2/3, δηλαδή η μελέτη έχει εστιαστεί στη βελτίωση της ταξινόμησης των περιστατικών με CIN. Επομένως, μπορούμε να συμπεράνουμε ότι αν και ο HPV 16 όσο και οι υπόλοιποι τύποι υψηλού κινδύνου συνδέονται άμεσα με την ανάπτυξη τραχηλικής ενδοεπιθηλιακής νεοπλασίας, δεν έχουν όλοι εξίσου σημαντικό ρόλο στη σταδιοποίηση των ενδοεπιθηλιακών αλλοιώσεων. Χαρακτηριστικά, σημαντικό ρόλο στη διάγνωση των CIN καταστάσεων φαίνεται να διαδραματίζει η παρουσία του HPV 31, καθώς όπως παρατηρούμε είναι πολλές οι περιπτώσεις στις οποίες εμφανίζεται ο HPV 31 στο υποσύνολο χαρακτηριστικών. Επιπλέον, η παρουσία του επιβεβαίωνεται από την ύπαρξη του mRNA του, αφού η παρουσία του HPV 31 στο υποσύνολο χαρακτηριστικών συνοδεύται και από το NASBA 31.

Κεφάλαιο 4

Μελλοντική Έρευνα

Στην παρούσα διπλωματική εργασία έγινε η παρουσίαση μίας μελέτης που στόχο είχε τη βελτιστοποίηση της ταξινόμησης περιστατικών τραχηλικής ενδοεπιθηλιακής νεοπλασίας και την ανάπτυξη ενός πρώτου μοντέλου αθροιστικού κινδύνου για την αντιμετώπιση των περιστατικών αυτών. Είδαμε πως η χρήση τεχνικών υπολογιστικής νοημοσύνης μπορεί να αποφέρει σημαντικά καλύτερα αποτελέσματα σε σχέση με αυτά του test Pap ως μοναδικού διαγνωστικού εργαλείου. Η μελέτη αυτή εκπονήθηκε στο πλαίσιο μιας ευρύτερης έρευνας για το θέμα του καρκίνου του τραχήλου της μήτρας. Και ακριβώς επειδή η έρευνα περιλαμβάνει εξ ορισμού τη δυναμικότητα, δεν μπορεί να είναι να στατική.

Όπως αναφέρθηκε, τα δεδομένα με τα οποία εργαστήκαμε και στα οποία εφαρμόσαμε τη μεθοδολογία μας, ήταν στο σύνολό τους 212. Είναι προφανές ότι ο αριθμός τους δεν είναι αρκετός ώστε να μας οδηγήσει σε ασφαλή αποτελέσματα γιατί δεν είναι αρκετά μεγάλος ώστε να μπορούν να περιέχονται σε αυτά αρκετές διαφορετικές περιπτώσεις επιτρέποντάς μας να εντοπίσουμε όλα τα διαφορετικά *patterns*. Εντούτοις η παρούσα εργασία μπορεί να θεωρηθεί σαν ένα αρχικό στάδιο μίας μελέτης (*preliminary work*) επί του θέματος. Με την προσθήκη περισσότερων περιστατικών θα μπορέσουμε να δοκιμάσουμε τη μεθοδολογία που περιγράφηκε και να εξάγουμε τα αντίστοιχα αποτελέσματα. Εάν αυτά αποκλίνουν σε μεγάλο βαθμό τότε ίσως θα πρέπει να γίνει επανεκτίμηση της στρατηγικής αντιμετώπισης εξ αρχής των δεδομένων. Σε περίπτωση όμως που αυτά δεν αποκλίνουν, και κυρίως δεν μειώνουν κατά πολύ τα στατιστικά μέτρα για την ανίχνευση της πραγματικής κλινικής κατάστασης, τότε θα μπορούσαμε να οδηγηθούμε σε απολύτως τεκμηριωμένα αποτελέσματα και επομένως σε ένα ακόμα πιο εύρωστο σύστημα.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] http://en.wikipedia.org/wiki/Cervical_cancer

[2] http://www.who.int/mediacentre/factsheets/fs297/en/

[3] http://globocan.iarc.fr/factsheet.asp

[4]] T.P. Canavan, N.R. Doshi, "Cervical Cancer", American Academy of Family Physicians, March 2010, http://www.aafp.org/afp/20000301/1369.html

[5]] N. MacDonald, M.B. Stanbrook, P.C. Hebert, "Human papillomavirus vaccine risk and reality", CMAJ, vol. 179 (6), September 2008

[6] PAPScreen Victoria, "Latest statistics", http://www.papscreen.org.au/browse.asp?ContainerID=c15

[7] Centers for disease control and prevention, "Sexually transmitted diseases - Genital HPV infection", http://www.cdc.gov/std/HPV/STDFact-HPV.htm

[8] The new England Journal of Medicine, "Case 10-2009 - A 23 year old woman with an abnormal Papanicolaou smear", http://www.nejm.org/doi/full/10.1056/NEJMcpc0810837

[9] http://en.wikipedia.org/wiki/Human_papillomavirus

[10] American Cancer Society, "What causes cancer of the cervix", http://www.cancer.org/docroot/CRI/content/CRI_2_2_2X_What_causes_cancer_of_the_cervix_Can_it_be_prevented_8.asp?sitearea=

[11] J.M. Marrazzo, L.A. Koutsky, N.B. Kiviat, J.M. Kuypers, K. Stine, "Papanicolaou test screening and prevalence of genital human Papillomavirus among women who have sex with women", American Journal of Public Health, vol. 91 (6), pp. 947-952, June 2001

[12] Medical Diagnostic Laboratories, http://www.mdlab.com/html/testing/hpv_typedetect.html

[13] National Cancer Institute, Benchmarks, http://benchmarks.cancer.gov/

[14] N. Munoz, F.X. Bosch, S. de Sanjose, R. Herrero, X. Castellsague, K.V. Shah, P.J. Snijders, C.J. Meijer, "Epidemiologic classification of human Papillomavirus types associated with cervical cancer", N. Engl. J. Med., vol. 348 (6), pp. 518-527, February 2003

[15] Pathology & Laboratory Medicine, http://www.archivesofpathology.org/arpaonline/default.asp?request=get-abstract&issn=1543-2165&volume=127&page=930&&

[16] J.M.M. Walboomers, M.V. Jacobs, M.M. Manos, F.X. Bosch, J.A. Kummer, K.V. Shah, P.J.F. Snijders, J. Peto, C.J.L.M. Meijer, N. Munoz, "Human Papillomavirus is a necessary cause of invasive cervical cancer worldwide", Journal of pathology, vol. 189, pp. 12-19, 1999.

[17] P.J.F. Snijders, R.D.M. Steenbergen, D.A.M. Heideman, C.J.L.M. Meijer, "HPV-mediated cervical carcinogenesis: concepts and clinical implications", Journal of Pathology, vol. 208, pp. 152-164, 2006

[18] The Age, "Expert says circumcision makes sex safer", http://www.theage.com.au/articles/2005/02/15/1108230001471.html

[19] X. Castellsague, X. Bosch, N. Munoz, C. Meijer, K. Shah, S. de Sanjose et al., "Circumcision and cervical cancer", N. Engl. J. Med., vol. 346, pp. 1105-1112, 2002

[20] J.T. Schiller, P.M. Day, R.C. Kines, "Current understanding of the mechanism of HPV infection", Gynecologic Oncology, vol. 118, pp. S12-S17, April 2010

[21] A. Chaturvedi, M.L. Gillison, "Human Papillomavirus and head and neck cancer", Epidemiology, pathogenesis and prevention of head and neck cancer, pp. 87-116, 2010

[22] http://en.wikipedia.org/wiki/Retinoblastoma_protein

[23] N. Ganguly, S.P. Parihar, "Human Papillomavirus E6 and E7 oncoproteins as risk factors for tumorigenesis", J. Biosci., vol. 34, pp. 113-123, 2009

[24] A. Chaturvedi, M.L. Gillison, "Human Papillomavirus and head and neck cancer", Epidemiology, pathogenesis and prevention of head and neck cancer, pp. 87-116, 2010

[25] http://en.wikipedia.org/wiki/Cell_cycle

[26] M.J. Conway, S. Alam, E.J. Ryndock et al., "Tissue-spanning redox gradient-dependent assembly of native human Papillomavirus type 16 virions", Journal of Virology, vol. 83 (20), pp. 10515-10526, 2009

[27]Εκπαιδευτικό διαδικτυακό πρόγραμμα Eurocytology, "Κυτταρολογία τραχήλου της μήτρας", http://www.eurocytology.eu/Static/EUROCYTOLOGY/GRE/TP1CONTENT.html

[28] American Cancer Society, "What causes cancer of the cervix", http://www.cancer.org/docroot/CRI/content/CRI_2_2_2X_What_causes_cancer_of_the_cervix_Can_it_be_prevented_8.asp?sitearea=

[29] Cancer Research UK, "Cervical Cancer – UK incidence statistics", http://info.cancerresearchuk.org/cancerstats/types/cervix/incidence/

[30] M. Demay, Practical Principles of Cytopathology, American Society for Clinical Pathology Press, 2007

[31] Medline Plus, "Cervical Cancer", http://www.nlm.nih.gov/medlineplus/ency/article/000893.htm

[32] http://en.wikipedia.org/wiki/Screening_%28medicine%29

[33] http://en.wikipedia.org/wiki/Pap_test

[34] http://www.eurocytology.eu/static/eurocytology/gre/cervical/LP1ContentDcont.html

[35] http://en.wikipedia.org/wiki/Candidiasis

[36] http://www.eurocytology.eu/static/eurocytology/gre/cervical/LP1ContentMcontA1.html

[37] http://www.eurocytology.eu/static/eurocytology/gre/cervical/LP1ContentMcontB.html

[38] http://www.eurocytology.eu/static/eurocytology/gre/cervical/LP1ContentMcont.html

[39] Denny LA and Wright TC in Best Practice and Research in Clinical Obstetrics and Gynaecology2005 vol 19,no4,pp501-505]

[40] http://en.wikipedia.org/wiki/NASBA_(molecular_biology)

[41] http://www.medscape.com/viewarticle/585223_2

[42] http://www.eurogenetica.gr/mrnatest/

[43] F.M. Carozzi, "Combined analysis of HPV DNA and p16INK4a expression to predict prognosis in ASCUS and LSIL PAP smears", Coll. Antropol., vol. 31 (2), pp. 103-106, 2007

[44] http://en.wikipedia.org/wiki/Cell_cycle

[45] http://en.wikipedia.org/wiki/Cyclin-dependent_kinase

[46] http://en.wikipedia.org/wiki/P16_%28gene%29

[47] PAPScreen Victoria, "Latest statistics", http://www.papscreen.org.au/browse.asp?ContainerID=c15

[48] R. Narimatsu, B.K. Patterson, "High-throughput cervical cancer screening using intracellular human papillomavirus E6 and E7 mRNA quantification by flow cytometry", American Society for Clinical Pathology, vol.123, pp. 716-723, 2005

[49] Cancer Research UK, "Cervical cancer – UK mortality statistics", http://info.cancerresearchuk.org/cancerstats/types/cervix/mortality/

[50] A. Trope, K. Sjoborg, A. Eskild, K. Guschieri, T. Eriksen, S. Thoresen, M. Steinbakk, V. Laurak, C.M. Jonassen, U. Westerhagen, M.B. Jacobsen, A.K. Lie, "Performance of human Papillomavirus DNA and mRNA testing strategies for women with and without cervical neoplasia", Journal of Clinical Microbiology, vol. 47 (8), pp. 2458-2464, August 2009

[51] T. Molden, I. Kraus, F. Karlsen, H. Skomedal, J.F. Nygard, B. Hagmar, "Comparison of human Papillomavirus messenger RNA and DNA detection: a cross-sectional study of 4136 women>30 years of age with a 2-year follow-up of High grade Squamous intraepithelial

lesion", Cancer Epidemiology, Biomarkers and Prevention, vol. 14 (2), pp. 367-372, February 2005

[52] H. De Vuyst και P. Claeys, "Comparison of pap smear, visual inspection with acetic acid, human papillomavirus DNA-PCR testing and cervicography".

[53] Tipaya Ekalaksananan, Chamsai Pientong, Supannee Sriamporn, Bunkerd Kongyingyoes, Prasit Pengsa, Pilaiwan Kleebkaow, Onanong Kritpetcharat, D. Max Parkin, Usefulness of combining testing for p16 protein and human papillomavirus (HPV) in cervical carcinoma screening, Gynecologic Oncology, Volume 103, Issue 1, October 2006, Pages 62-66, ISSN 0090-8258, 10.1016/j.ygyno.2006.01.033

[54] O. Ludwig and U. Nunes; " Novel Maximum-Margin Training Algorithms for Supervised Neural Networks;" IEEE Transactions on Neural Networks, vol.21, issue 6, pp. 972-984, Jun. 2010

[55] K. Guschieri, N. Wentzensen, "Human Papillomavirus mRNA and p16 detection as biomarkers for the improved diagnosis of cervical neoplasia", Cancer Epidemiology, Biomarkers and Prevention, vol. 17 (10), pp. 2536-2545, October 2008

[56] http://cervicalcancer.about.com/od/screening/a/ASCUS_pap.htm

[57] R. Narimatsu, "Cervical cancer screening using flow cytometry", American Society for Clinical Pathology, vol.123, pp. 716-723, 2005

[58] K. Denton,C. Bergeron , The Sensitivity and Specificity of p16INK4a Cytology vs HPV Testing for Detecting High-Grade Cervical Disease in the Triage of ASC-US and LSIL Pap Cytology Results

[59] J. Coste, B. Cochand-Priollet, P. de Cremoux, C. Le Galès, I. Cartier, V. Molinié, S. Labbé, M.-C. Vacher-Lavenu, P. Vielh, "Cross sectional study of conventional cervical smear, monolayer cytology, and human papillomavirus DNA testing for cervical cancer screening", British Medical Journal, vol.326, p.733, April 2003

[60] S.L. Kulasingam, J.P. Hughes, N.B. Kiviat, C.Mao, N.S. Weiss, J.M. Kuypers, L.A. Koutsky, "Evaluation of human Papillomavirus testing in primary screening for cervical abnormalities", Journal of the American Medical Association, vol.288, pp. 1749-1757, 2002

[61] J. Cuzick, A. Szarewski, H. Cubie, G. Hulman, H. Kitchener, D. Luesley, E. McGoogan, U. Menon, G. Terry, R. Edwards, C. Brooks, M. Desai, C. Gie, L. Ho, I. Jacobs, C. Pickles, P. Sasieni, "Management of women who test positive for high-risk types of human papillomavirus: the HART study", The Lancet, vol. 362, pp. 1871-1876, December 2003

[62] L.O. Sarian, S.F.M. Derchain, L.A.A. Andrade, J. Tambascia, S.S. Morais, K.J. Syrjanen, "HPV DNA test and PAP smear in detection of residual and recurrent disease following loop electrosurgical excision procedure of high-grade cervical intraepithelial neoplasia", Gynecologic Oncology, vol. 94, pp. 181-186, 2004

[63] I. Tsoumpou, M. Arbyn, M. Kyrgiou, N. Wentzensen, G. Koliopoulos, P. Martin-Hirsch, V. Malamou-Mitsi, E. Paraskevaidis, "p16INK4a immunostaining in cytological and histological specimens from the uterine cervix: a systematic review and meta-analysis", Cancer Treatment Reviews, vol. 35, pp. 210-220, 2009

[64] http://en.wikipedia.org/wiki/Artificial_neural_network

[65] Silvano Costa , Giovanni Negri, Human papillomavirus (HPV) test and PAP smear as predictors of outcome in conservatively treated adenocarcinoma in situ (AIS) of the uterine cervix

[66] http://en.wikipedia.org/wiki/Bethesda_system

[67] Κοτταρίδη Χ. , Βιολόγος, Εργαστήριο Διαγνωστικής Κυτταρολογίας, Π.Γ.Ν. Αττικόν, «Ανίχνευση υπερέκφρασης mRNA των ογκογονιδίων Ε6/Ε7 του ιού HPV με Κυτταρομετρία Ροής.»

[68] Pattern Recognition—S. Theodoridis and K. Koutroumbas,New York, NY: Academic, 2006, 3rd ed., pp. 837, ISBN: 0-12-369531-7

[69] Peng, H.C., Long, F., and Ding, C., Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, pp. 1226–1238, 2005

[70] Gray, R. M. (1990). Entropy and information theory, Springer-Verlag.

[71] http://en.wikipedia.org/wiki/Entropy_(information_theory)

[72] Thomas M. Cover, Joy A. Thomas. Elements of information theory, 2nd Edition. New York: Wiley-Interscience, 2006. ISBN 0-471-24195-4.

[73] Haykin, S. and M. Moher (2009). Communication systems, Wiley.

[74] http://en.wikipedia.org/wiki/Genetic_algorithm

[75] Russell, S. and P. Norvig (2010). The Artificial Intelligence, 3e Preview Edition, Pearson Education, Limited.

[76] http://www.obitko.com/tutorials/genetic-algorithms/selection.php

[77] http://en.wikipedia.org/wiki/Crossover_(genetic_algorithm)

[78] http://en.wikipedia.org/wiki/Pattern_recognition

[79] http://en.wikipedia.org/wiki/Pruning_(decision_trees)

[80] Mansour, Y. (1997), "Pessimistic decision tree pruning based on tree size", Proc. 14th International Conference on Machine Learning: 195–201

[81] C. M. Bishop, Pattern Recognition and Machine Learning. NewYork: Springer-Verlag, 2006.

[82] http://en.wikipedia.org/wiki/Naive_Bayes_classifier

[83] Haykin, S. S. (2009). Neural networks and learning machines, Prentice Hall.

[84] http://en.wikipedia.org/wiki/Artificial_neural_network