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

Εφαρμογή γενετικών αλγορίθμων και άλλων μεθόδων επιλογής χαρακτηριστικών για την υποστήριξη λήψης κλινικής απόφασης στη διάγνωση του καρκίνου του τραχήλου της μήτρας

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Έλενα Τόπακα

Επιβλέπων : Διονύσιος-Δημήτριος Κουτσούρης
Καθηγητής Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2016

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

Εφαρμογή γενετικών αλγορίθμων και άλλων μεθόδων επιλογής χαρακτηριστικών για την υποστήριξη λήψης κλινικής απόφασης στη διάγνωση του καρκίνου του τραχήλου της μήτρας

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Έλενα Τόπακα

Επιβλέπων : Διονύσιος-Δημήτριος Κουτσούρης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 22η Φεβρουαρίου 2016.

.....	.....	.....
Δ.-Δ. Κουτσούρης	Μ. Χαρίτου	Γ. Ματσόπουλος
Καθηγητής Ε.Μ.Π.	Ερευνήτρια Α' ΕΠΙΣΕΥ-Ε.Μ.Π.	Αν. Καθηγητής Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2016

.....

Έλενα Τόπακα

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

Περίληψη

Ο καρκίνος του τραχήλου της μήτρας αποτελεί έναν από τους πιο κοινούς τύπους καρκίνου και παρουσιάζει ένα από τα πιο υψηλά ποσοστά θνησιμότητας από καρκίνο στις γυναίκες. Παρά την ύπαρξη της κυτταρολογικής εξέτασης (εξέταση Παπανικολάου [τεστ ΠΑΠ]), η οποία είναι διαθέσιμη τα τελευταία 50+ χρόνια, καθώς και των προγραμμάτων προληπτικού πληθυσμιακού ελέγχου, ο καρκίνος του τραχήλου της μήτρας παραμένει ένα σοβαρό πρόβλημα υγείας λόγω του σχετικά υψηλού ποσοστού μη ανίχνευσης της νόσου. Στις περισσότερες των περιπτώσεων, ο καρκίνος του τραχήλου της μήτρας αναπτύσσεται ως αποτέλεσμα της υποεκτίμησης των ανωμαλιών που παρατηρούνται στην κυτταρολογική εξέταση.

Οι εξελίξεις στην κατανόηση του ρόλου της λοίμωξης από τον ιό των ανθρώπινων θηλωμάτων (HPV) και των διαφόρων τύπων του ιού στη φυσική εξέλιξη των νεοπλασιών του τραχήλου της μήτρας, είχαν ως αποτέλεσμα την παράλληλη διενέργεια της εξέτασης HPV DNA μαζί με την εξέταση Παπανικολάου. Σήμερα, η εξέταση HPV DNA τυγχάνει ευρείας αποδοχής ως βοηθητική εξέταση για τη διαλογή γυναικών που παρουσιάζουν μη φυσιολογικά ευρήματα στην κυτταρολογική εξέταση. Επιπλέον, σε πολλές ανεπτυγμένες χώρες, η εξέταση HPV DNA συμπεριλαμβάνεται και στις επίσημες κατευθυντήριες οδηγίες των προγραμμάτων πληθυσμιακού ελέγχου για τον καρκίνο του τραχήλου της μήτρας και χρησιμοποιείται είτε μόνη της είτε σε συνδυασμό με το τεστ Παπανικολάου. Ωστόσο, και οι δύο διαγνωστικές εξετάσεις παρουσιάζουν είτε υψηλή ευαισθησία είτε υψηλή ειδικότητα, αλλά όχι και τα δύο ταυτοχρόνως.

Στην παρούσα διπλωματική εργασία, παρουσιάζεται ένα ευφυές υπολογιστικό σύστημα το οποίο συνδυάζει τα αποτελέσματα της εξέτασης Παπανικολάου και του HPV DNA test, αποσκοπώντας σε πιο ισορροπημένα αποτελέσματα ως προς την ειδικότητα και την ευαισθησία για την ανίχνευση ενδοεπιθηλιακής νεοπλασίας του τραχήλου της μήτρας 2ου βαθμού ή άνω (CIN2+). Για την ανάπτυξη του συστήματος αυτού, υιοθετήθηκε ένα πλαίσιο επιλογής χαρακτηριστικών που ακολουθεί την προσέγγιση περιτυλίγματος και βασίζεται στο συνδυασμό Γενετικών Αλγορίθμων και Μπεϋζιανών Ταξινομητών. Στόχος του πλαισίου αυτού είναι η εύρεση του κατάλληλου υποσυνόλου χαρακτηριστικών, το οποίο βελτιστοποιεί την ταξινόμηση οδηγώντας σε πιο ισορροπημένα αποτελέσματα ευαισθησίας και ειδικότητας. Το παρουσιαζόμενο σύστημα μπορεί να υποστηρίξει τη λήψη κλινικών αποφάσεων για τη βελτίωση της διαχείρισης γυναικών που παραπέμπονται για κολποσκόπηση εξαιτίας θετικού αποτελέσματος σε μια διαγνωστική εξέταση.

Επιπρόσθετα, σε δεύτερο επίπεδο εξετάζεται και ένα άλλο πρόβλημα, αυτό της κατάταξης των πιο συχνών τύπων του ιού HPV ως προς την επικινδυνότητά τους για ανάπτυξη ενδοεπιθηλιακής νεοπλασίας του τραχήλου της μήτρας 2ου βαθμού ή άνω (CIN2+). Για την

επίτευξη του στόχου αυτού, χρησιμοποιήθηκαν τεχνικές επιλογής χαρακτηριστικών που ακολουθούν την προσέγγιση φιλτραρίσματος.

Λέξεις κλειδιά: Γενετικοί αλγόριθμοι, Διαγνωστικές εξετάσεις, Καρκίνος του τραχήλου της μήτρας, Ιός των ανθρωπίνων θηλωμάτων (HPV), Επιλογή χαρακτηριστικών, τύποι HPV, Ευαισθησία, Ειδικότητα

Abstract

Cervical Cancer is one of the most common types of cancer and one of the leading causes of death in women. Even though screening with cervical cytological testing (the Papanicolaou test [Pap test]) has been available for over 50 years, cervical cancer still remains a major health problem due to the high rate of non-detection of the disease. In most cases, cervical cancer develops as a result of underestimated abnormalities in the Pap test.

Advances in the understanding of the role of Human Papillomavirus (HPV) infection to the natural development of cervical neoplasia, resulted in the co-testing of Pap test with the HPV DNA test. Nowadays, HPV DNA testing is well accepted as an ancillary test and it is used for the triage of women with abnormal findings in cytology. Furthermore, in many developed countries, the official cervical cancer screening guidelines recommend HPV DNA testing to be used alone or in combination with cytology (co-testing). However, these tests are either highly sensitive or highly specific, but not both at the same time.

In this diploma thesis, an intelligent computing system, that effectively combines the results of the Pap test and the HPV DNA test, is presented. However, our focus in not on the classification's accuracy per se, but rather the creation of a system, that yields the most balanced results in terms of sensitivity and specificity for the detection of high-grade cervical intraepithelial neoplasia and cervical cancer (CIN2+). For the development of the proposed system, a wrapper feature selection framework has been adopted, which is based on the combination of Genetic Algorithms and Bayesian classifiers. The scope of this framework is the detection of the feature subset which optimizes the classification so as to achieve a balanced outcome between sensitivity and specificity. The presented system may support decision-making for the improved management of women who attend a colposcopy room following a positive test result.

Moreover, this thesis investigated an additional problem: the ranking of the most common HPV genotypes, according to the associated risk of developing high grade cervical intraepithelial neoplasia or cervical cancer (CIN2+). This was accomplished by analyzing the data with the use of filtering feature selection techniques.

Key-words: Genetic Algorithms, Diagnostic tests, Cervical cancer, Human Papillomavirus (HPV), Feature Selection, HPV genotypes, Sensitivity, Specificity

Πρόλογος

Η παρούσα διπλωματική εργασία εκπονήθηκε στο Εργαστήριο Βιοϊατρικής Τεχνολογίας της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ του Εθνικού Μετσόβιου Πολυτεχνείου σε συνεργασία με το Εργαστήριο Διαγνωστικής Κυτταρολογίας της Ιατρικής Σχολής του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών στο Πανεπιστημιακό Γενικό Νοσοκομείο «Αττικόν».

Παρατηρείται ότι οι υπάρχουσες διαγνωστικές εξετάσεις για καρκίνο του τραχήλου της μήτρας παρουσιάζουν είτε υψηλή ευαισθησία και χαμηλή ειδικότητα, είτε υψηλή ειδικότητα και χαμηλή ευαισθησία. Για τον λόγο αυτό, κρίνεται απαραίτητη η δημιουργία μιας μεθόδου, με κατώφλι τον εντοπισμό ενδοεπιθηλιακής νεοπλασίας του τραχήλου της μήτρας 2ου βαθμού ή άνω (CIN2+), η οποία να παρουσιάζει ισορροπημένη ευαισθησία και ειδικότητα. Κύριο αντικείμενο της παρούσας διπλωματικής εργασίας είναι η ανάπτυξη ενός συστήματος ταξινόμησης, το οποίο συνδυάζοντας τα αποτελέσματα της εξέτασης Παπανικολάου και του HPV DNA test, θα οδηγεί σε πιο ισορροπημένα αποτελέσματα όσον αφορά την ευαισθησία και την ειδικότητα. Για να γίνει αυτό, είναι απαραίτητη η εύρεση του βέλτιστου συνδυασμού χαρακτηριστικών, ο οποίος όταν τροφοδοτηθεί ως είσοδος στον ταξινομητή, θα ικανοποιήσει το στόχο αυτό. Η αναζήτηση του βέλτιστου υποσυνόλου χαρακτηριστικών γίνεται με τη χρήση Γενετικών Αλγορίθμων και του ταξινομητή Naïve-Bayes.

Επιπρόσθετα, η εργασία αυτή ασχολείται και με ένα άλλο πρόβλημα, το πρόβλημα της κατάταξης των πιο συχνών τύπων του ιού HPV ως προς την επικινδυνότητά τους για ανάπτυξη ενδοεπιθηλιακής νεοπλασίας του τραχήλου της μήτρας 2ου βαθμού ή άνω (CIN2+). Η κατάταξη αυτή γίνεται με τη χρήση μεθόδων φιλτραρίσματος.

Τα δεδομένα της εργασίας προέρχονται από τη βάση δεδομένων του τμήματος κυτταρολογίας της Ιατρικής Σχολής του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών και περιλαμβάνουν αποτελέσματα από τις ακόλουθες διαγνωστικές εξετάσεις: HPV DNA test, εξέταση Παπανικολάου και ιστολογική εξέταση.

Η δομή της εργασίας έχει ως εξής:

Τα Κεφάλαια 1, 2 και 3 αποτελούν το τεχνικό κομμάτι της διπλωματικής εργασίας, αφού σε αυτά περιγράφονται αναλυτικά το γενικό τεχνικό υπόβαθρο και οι χρησιμοποιούμενες μέθοδοι. Στο πρώτο Κεφάλαιο παρουσιάζονται οι Γενετικοί Αλγόριθμοι και στο δεύτερο Κεφάλαιο αναλύεται η επιλογή χαρακτηριστικών και οι διάφορες μέθοδοι υλοποίησής της. Το τρίτο Κεφάλαιο πραγματεύεται την επιλογή χαρακτηριστικών με χρήση γενετικών αλγορίθμων ακολουθώντας προσέγγιση περιτυλίγματος, με αναφορά σε προηγούμενες σχετικές μελέτες από τη βιβλιογραφία.

Ακολούθως, στο τέταρτο Κεφάλαιο παρατίθεται το βιολογικό υπόβαθρο του εξεταζόμενου προβλήματος. Αναλύονται οι έννοιες του καρκίνου του τραχήλου της μήτρας και του ιού των ανθρωπίνων θηλωμάτων (HPV), καθώς και η μεταξύ τους συσχέτιση. Επίσης αναλύεται η έννοια του προληπτικού πληθυσμιακού ελέγχου της νόσου και γίνεται επισκόπηση των ακόλουθων διαγνωστικών εξετάσεων: εξέταση Παπανικολάου, HPV DNA test και κολποσκόπηση.

Κατόπιν, στο Κεφάλαιο 5 παρουσιάζονται αναλυτικά το εξεταζόμενο πρόβλημα, η υλοποίηση του συστήματος ταξινόμησης που συνδυάζει τα αποτελέσματα του HPV DNA test και του PAP test καθώς και η μεθοδολογία που ακολουθήθηκε για την κατάταξη των HPV τύπων. Πιο συγκεκριμένα, όσον αφορά το σύστημα ταξινόμησης, αρχικά παρατίθενται οι λόγοι που οδήγησαν στην αναγκαιότητα ανάπτυξης ενός τέτοιου εργαλείου, στη συνέχεια περιγράφεται η δομή του και εν τέλει παρουσιάζεται η αρχιτεκτονική του προτεινόμενου συστήματος υποστήριξης κλινικών αποφάσεων. Στο τέλος του Κεφαλαίου παρατίθενται τα πειραματικά αποτελέσματα.

Το έκτο Κεφάλαιο αφορά στα συμπεράσματα της εργασίας που πραγματοποιήθηκε και στη συνεισφορά της στο πρόβλημα που πραγματεύεται. Γίνεται αναφορά στους μελλοντικούς ερευνητικούς στόχους και στα ανοιχτά θέματα προς μελλοντική έρευνα.

Ευχαριστίες

Στα πλαίσια της διπλωματικής μου εργασίας, θα ήθελα να εκφράσω τις ειλικρινείς μου ευχαριστίες σε όλους όσοι βοήθησαν στην περάτωσή της.

Αρχικά, θα ήθελα να εκφράσω τις θερμές ευχαριστίες μου στον κ. Δημήτριο Κουτσούρη, καθηγητή της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ του Εθνικού Μετσόβιου Πολυτεχνείου για την εμπιστοσύνη που μου επέδειξε αναθέτοντάς μου την εκπόνηση αυτής της διπλωματικής εργασίας. Επίσης, ευχαριστώ τα υπόλοιπα μέλη της τριμελούς επιτροπής, την κα. Μαρία Χαρίτου, ερευνήτρια Α' ΕΠΙΣΕΥ-ΕΜΠ, και τον κ. Γιώργο Ματσόπουλο, αναπληρωτή καθηγητή της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ του Εθνικού Μετσόβιου Πολυτεχνείου. Θα ήθελα να ευχαριστήσω και τον κ. Πέτρο Καρακίτσο, καθηγητή Κυτταρολογίας του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών και Διευθυντή του Εργαστηρίου Διαγνωστικής Κυτταρολογίας του Πανεπιστημιακού Γενικού Νοσοκομείου «Αττικόν», για τη βάση δεδομένων που μας διέθεσε, χωρίς την οποία η παρούσα εργασία δεν θα μπορούσε να είχε εκπονηθεί.

Ακόμη, θα ήθελα να ευχαριστήσω τον κ. Παναγιώτη Μπούντρη, υποψήφιο διδάκτορα της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ του Εθνικού Μετσόβιου Πολυτεχνείου, ως επιβλέποντα της διπλωματικής μου, για τη συνεργασία μας καθ' όλη τη διάρκεια εκπόνησής της. Πιο συγκεκριμένα, τον ευχαριστώ θερμά για τη βοήθεια που μου παρείχε κατά την πρώτη επαφή με το εξεταζόμενο πρόβλημα, τις υποδείξεις και την προθυμία του κατά την πορεία της έρευνας, καθώς και για τη συνεισφορά του στην παρούσα διπλωματική εργασία, όποτε αυτή κρίθηκε απαραίτητη.

Επιπλέον, ευχαριστώ όσους βρίσκονται στο οικογενειακό και φιλικό μου περιβάλλον και με στηρίζουν με την αγάπη και την εμπιστοσύνη τους όλα αυτά τα χρόνια. Ιδιαίτερα, θέλω να εκφράσω την ευγνωμοσύνη μου στη μητέρα μου, για τη διαρκή της υποστήριξη, τόσο ηθική όσο και οικονομική, καθώς και για την παροιμιώδη υπομονή που επέδειξε ουκ ολίγες φορές. Τέλος, ευχαριστώ τους φίλους μου (και ειδικά τον Αντρέα Μάντη) για την αδιάλειπτη στήριξη και βοήθειά τους, συγκαταλέγοντας σ' αυτούς και όλους όσους αγαπούν την αλήθεια και έχουν παράξει έργο για την ανθρωπότητα, δίνοντας τη δυνατότητα σε όλους εμάς και στους επόμενους από μας να συνεχίσουν το έργο τους.

Πίνακας περιεχομένων

Κατάλογος Εικόνων

Οι εικόνες έχουν γίνει με χρήση των ακόλουθων προγραμμάτων:

- Matlab
- Microsoft Visio 2013
- Microsoft Excel 2013
- Microsoft Paint 2013
- Graphmatica

Κατάλογος Πινάκων

Κεφάλαιο 1 – Γενετικοί Αλγόριθμοι

1.1 Γενετικοί Αλγόριθμοι - Εισαγωγή

Οι γενετικοί αλγόριθμοι (genetic algorithms: GA) είναι ευριστικοί αλγόριθμοι αναζήτησης (heuristic search algorithms) που προσομοιώνουν τις εξελικτικές διαδικασίες που παρατηρούνται στη φύση και οι οποίες βασίζονται στη φυσική επιλογή και στη φυσική εξέλιξη όπως αυτές περιγράφονται στη Δαρβινική εξελικτική θεωρία και στη γενετική.

Σύμφωνα με τη θεωρία της φυσικής επιλογής του Δαρβίνου [1] οι οργανισμοί που δεν είναι κατάλληλοι για το εκάστοτε περιβάλλον δεν επιβιώνουν, ενώ αυτοί που είναι, ζουν και αναπαράγουν. Συνεπώς, η ικανότητα (fitness) ενός οργανισμού μετριέται από την επιτυχία επιβίωσής του. Οι απόγονοι (offspring) μοιάζουν στους γονείς τους (parents) και συνεπώς κάθε νέα γενιά αποτελείται από οργανισμούς που μοιάζουν στα καταλληλότερα μέλη της προηγούμενης γενιάς, με αποτέλεσμα κάθε καινούρια γενιά να έχει περισσότερες πιθανότητες επιβίωσης από την προηγούμενη. Εάν το περιβάλλον αλλάζει αργά, τα διάφορα είδη προλαβαίνουν να εξελίσσονται παράλληλα με αυτό. Εάν όμως, παρουσιαστεί μια ξαφνική αλλαγή στο περιβάλλον, πιθανώς τα διάφορα είδη να αφανιστούν. Περιστασιακά μπορεί να συμβούν διάφορες μεταλλάξεις (mutations) , οι περισσότερες των οποίων οδηγούν στον γρήγορο θάνατο του μεταλλαγμένου οργανισμού, ενώ κάποιες από αυτές οδηγούν σε ένα νέο καλύτερο είδος [2].

Στις αρχές του 1950 εμπνευσμένοι από τη Δαρβινική θεωρία άρχισαν να προτείνονται διάφοροι εξελικτικοί αλγόριθμοι (evolutionary algorithms) . Αρκετοί ερευνητές, ανεξάρτητα, άρχισαν να μελετούν εξελικτικά συστήματα (evolutionary systems) βασισμένοι στην ιδέα ότι οι εξελικτικοί αλγόριθμοι θα μπορούσαν να χρησιμοποιηθούν ως εργαλείο σε προβλήματα βελτιστοποίησης στη μηχανική. Οι εξελικτικοί αλγόριθμοι αποσκοπούν στην εξέλιξη ενός πληθυσμού πιθανών λύσεων σε ένα συγκεκριμένο πρόβλημα, με τη χρήση τελεστών (operators) εμπνευσμένων από τη φυσική γενετική ποικιλότητα και τη φυσική επιλογή [3]. Στις δεκαετίες του 1950 και 1960 αναπτύχθηκαν αρκετά εξελικτικά συστήματα από πολλούς ερευνητές. Βασικές υποκατηγορίες τους είναι ο εξελικτικός προγραμματισμός (evolutionary programming), οι στρατηγικές εξέλιξης (evolution strategies), τα συστήματα ταξινόμησης (classifier systems) και ο γενετικός προγραμματισμός (genetic programming) .

Εικόνα 1: *Ανάπτυξη εξελικτικών αλγορίθμων*

Οι γενετικοί αλγόριθμοι είναι μια ειδική κλάση εξελικτικών αλγορίθμων που αρχικά είχαν επινοηθεί από τον J. Holland και τους συνεργάτες του στο University of Michigan στις αρχές της δεκαετίας του 1960 και αφού αναπτύχθηκαν είχαν προταθεί με την τελική τους μορφή το 1975 [4]. Σε αντίθεση με τους προηγούμενους ερευνητές εξελικτικών αλγορίθμων, ο αρχικός στόχος του Holland δεν ήταν ο σχεδιασμός αλγορίθμων για την επίλυση συγκεκριμένων προβλημάτων, αλλά η φορμαλιστική μελέτη του φαινομένου της προσαρμογής, όπως αυτή γίνεται στη φύση, και η ανάπτυξη μεθόδων με τις οποίες οι φυσικοί αυτοί μηχανισμοί , θα μπορούσαν να εισαχθούν σε υπολογιστικά συστήματα [3]. Αποσκοπούσε στη δημιουργία υπολογιστικών προγραμμάτων που θα «εξελίσσονταν» με τρόπο που προσομοιώνει τη φυσική επιλογή και θα ήταν ικανά να λύσουν πολύπλοκα προβλήματα τα οποία ούτε οι δημιουργοί τους δεν είναι ικανοί να κατανοήσουν πλήρως (“Computer programs that “evolve” in ways that resemble natural selection can solve complex problems even their creators do not fully understand.”) [4] . Σήμερα, ωστόσο η πιο ευρεία τους χρήση, είναι η αναζήτηση βέλτιστων λύσεων σε συστήματα που ανάγονται σε μαθηματικά προβλήματα. Ο Holland χρησιμοποίησε τεχνικές που προσομοιώνουν τη φυσική επιλογή και τελεστές από τη γενετική όπως η διασταύρωση (crossover), η μετάλλαξη (mutation) και η αντιστροφή (inversion) έτσι ώστε ένας πληθυσμός χρωματοσωμάτων (chromosomes) να μπορεί να εξελιχθεί σε έναν καινούριο πληθυσμό. Ο γενετικός αλγόριθμος του Holland με την εισαγωγή της έννοιας του πληθυσμού και των τελεστών διασταύρωσης, μετάλλαξης και αντιστροφής αποτελούσε μεγάλη καινοτομία στους εξελικτικούς αλγόριθμους.

Η ορολογία που χρησιμοποιείται στους γενετικούς αλγόριθμους είναι δανεισμένη από τη βιολογία και πιο συγκεκριμένα από τη φυσική γενετική.

Κάθε λύση σε ένα γενετικό αλγόριθμο (ΓΑ) αναπαρίσταται με ένα χρωμόσωμα (chromosome) και κάθε χρωμόσωμα αποτελείται από έναν αριθμό γονιδίων (genes). Γονίδιο (gene) είναι κάθε σύμβολο που χρησιμοποιείται για την αναπαράσταση μιας υποψήφιας λύσης. Μια ομάδα συγκεκριμένου αριθμού γονιδίων σχηματίζει ένα χρωμόσωμα/άτομο (chromosome/individual) και αντιπροσωπεύει μια πιθανή λύση. Μια ομάδα χρωματοσωμάτων συγκροτεί έναν πληθυσμό (population): μια «δεξαμενή» από πιθανές λύσεις για ένα πρόβλημα. Οι γενετικοί αλγόριθμοι χρησιμοποιούν βασικούς μηχανισμούς του εξελικτισμού: κληρονομικότητα (inheritance) , επιλογή

(selection), διασταύρωση(crossover) , μετάλλαξη (mutation), εφαρμόζοντάς τους ως τελεστές πάνω στα χρωματοσώματα. Κάθε μηχανισμός πραγματοποιείται με μια συγκεκριμένη πιθανότητα. Τα μέλη του πληθυσμού με υψηλή τιμή συνάρτησης καταλληλόλητας, η οποία θα εξηγηθεί λεπτομερειακά πιο κάτω, θεωρούνται επιλέξιμα για αναπαραγωγή και ονομάζονται γονείς(parents). Με τη χρήση τελεστών στους γονείς κατά την αναπαραγωγή, προκύπτουν νέα χρωμοσώματα , τα οποία καλούνται παιδιά/ απόγονοι (offspring). Κάθε νέα γενιά πληθυσμού αντιστοιχεί με ακόμα μια επανάληψη του γενετικού αλγόριθμου.

Η όλη διαδικασία της εξέλιξης προς τη βέλτιστη λύση υλοποιείται σε 3 βασικά στάδια: Στο πρώτο στάδιο (αρχικοποίηση) δημιουργείται τυχαία ένας αρχικός πληθυσμός απαρτιζόμενος από κάποια χρωματοσώματα/άτομα (individuals). Κάθε άτομο αντιπροσωπεύει μια λύση στο πρόβλημα και χρησιμοποιεί μια μορφή κωδικοποίησης. Ένα άτομο μπορεί να κωδικοποιηθεί ως ένα διάνυσμα αποτελούμενο από δυαδικούς αριθμούς. Κάθε δυαδικός αριθμός αντιστοιχεί σε ένα γονίδιο (gene) του ατόμου. Μια ομάδα κωδικοποιημένων γονιδίων συνθέτει ένα κωδικοποιημένο χρωμόσωμα/άτομο.

Ο αρχικός πληθυσμός (initial population) αποτελεί την πρώτη γενιά (first generation). Κάθε επανάληψη του αλγόριθμου δημιουργεί μια καινούρια γενιά. Ο γενετικός αλγόριθμος αξιολογεί κάθε μεμονωμένη λύση (χρωμόσωμα) χρησιμοποιώντας μια συνάρτηση προσαρμοστικότητας / καταλληλόλητας /ποιότητας (fitness function).

Στο δεύτερο στάδιο , το στάδιο της αναπαραγωγής (reproduction) εφαρμόζεται ο τελεστής της επιλογής (selection operator). Διάφορα άτομα από τον τρέχων πληθυσμό επιλέγονται στοχαστικά βάσει της τιμής καταλληλόλητας τους. Η τιμή καταλληλόλητας του εκάστοτε ατόμου είναι το αποτέλεσμα που προκύπτει μετά την εφαρμογή της συνάρτησης καταλληλόλητας στο κάθε άτομο. Ακολούθως, εφαρμόζεται ο τελεστής της διασταύρωσης(crossover) στα επιλεγμένα άτομα-χρωμοσώματα: τα επιλεγμένα χρωμοσώματα ανασυνδυάζονται μεταξύ τους παράγοντας καινούρια χρωμοσώματα, καθώς και ο τελεστής της μετάλλαξης (mutation): διάφορα γονίδια κάποιων χρωματοσωμάτων μεταλλάσσονται τυχαία (για παράδειγμα στην περίπτωση της δυαδικής κωδικοποίησης ορισμένα bits κάποιων εκ των επιλεγμένων χρωμοσωμάτων αναστρέφονται τυχαία). Με την εφαρμογή των τελεστών αυτών σχηματίζεται ένας νέος πληθυσμός: η επόμενη γενιά χρωματοσωμάτων.

Στο τρίτο στάδιο ο νέος πληθυσμός χρησιμοποιείται στην επόμενη επανάληψη του αλγορίθμου και η διαδικασία επαναλαμβάνεται για ένα αριθμό επαναλήψεων. Ο αλγόριθμος τερματίζεται σύμφωνα με ένα κριτήριο τερματισμού (για παράδειγμα όταν ο μέγιστος αριθμός γενεών έχει επιτευχθεί). Εν τέλει ο γενετικός αλγόριθμος επιστρέφει τον καλύτερο πληθυσμό χρωματοσωμάτων , τα οποία έχουν τις καλύτερες τιμές καταλληλόλητας. Αυτός ο πληθυσμός θα είναι η καλύτερη «δεξαμενή» λύσεων του προβλήματος.

Ψευδοκώδικας

BEGIN

INITIALISE population with random candidate solutions;

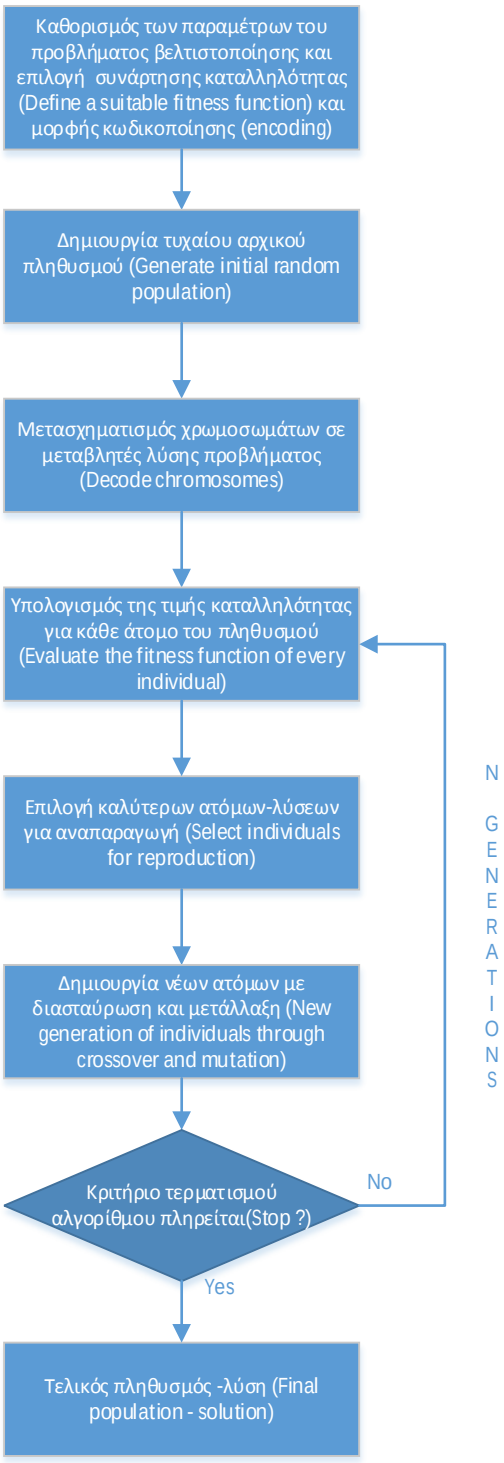
EVALUATE each candidate;

REPEAT UNTIL (termination condition) is satisfied DO

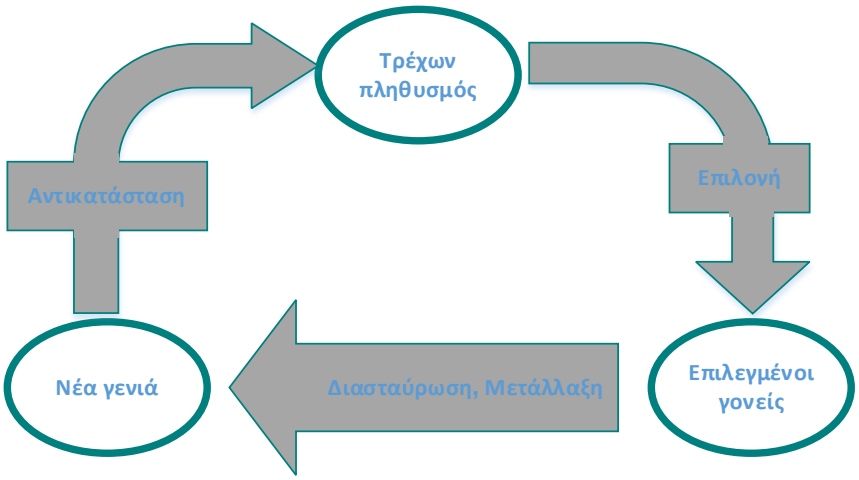
1. SELECT parents;
2. RECOMBINE pairs of parents;
3. MUTATE the resulting offspring;
4. SELECT individuals of the next generation;

END.

Η πιο πάνω διαδικασία θα περιγραφεί αναλυτικότερα στα επόμενα υποκεφάλαια.



Εικόνα 2 *: Διάγραμμα ροής τυπικού γενετικού αλγορίθμου*



Εικόνα 3: *Ένας κύκλος εξέλιξης ενός τυπικού* γενετικού *αλγορίθμου*

1.2 Χώρος/Διάστημα αναζήτησης (search space)

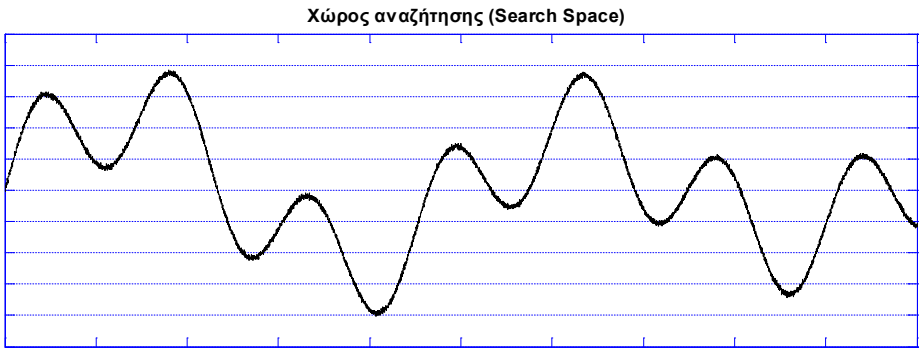
Όταν λύνουμε ένα πρόβλημα συνήθως ψάχνουμε για κάποια λύση , η οποία θα είναι η βέλτιστη των διαφόρων πιθανών άλλων λύσεων. Ο χώρος όλων των εφικτών πιθανών λύσεων αποτελεί τον χώρο αναζήτησης λύσεων του προβλήματος (search space), ή αλλιώς αναφερόμενο ως χώρο καταστάσεων (state space). Ο χώρος αναζήτησης υπονοεί κάποιου είδους απόσταση μεταξύ των πιθανών λύσεων [3].

Ορισμός 1. ***Χώρος αναζήτησης ενός προβλήματος*** ορίζεται ως το σύνολο όλων των δυνατών και έγκυρων λύσεων, μέσα στο οποίο ανήκει και η επιθυμητή λύση(ή λύσεις) του προβλήματος.

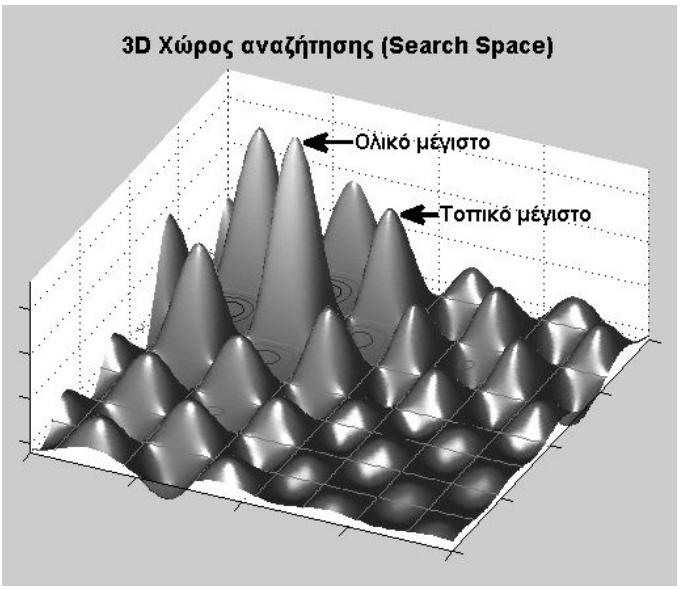
Κάθε σημείο στο χώρο αναζήτησης αναπαριστά μια πιθανή λύση. Η αποδοτικότητα κάθε πιθανής λύσης, ή αλλιώς, η αντικειμενική της αξία (objective value) σε σχέση με τον ορισμό του προβλήματος, μπορεί να καθοριστεί από την τιμή καταλληλόλητάς της στο πρόβλημα. Στόχος είναι η εύρεση της βέλτιστης λύσης (ή των βέλτιστων λύσεων), η οποία αντιστοιχεί σε ένα σημείο (ή περισσότερα) στο χώρο αναζήτησης. Η αναζήτηση λύσης, έτσι, ανάγεται στην αναζήτηση κάποιου ολικού μέγιστου ή ελάχιστου -ανάλογα με το πρόβλημα- στο χώρο αναζήτησης. Ο χώρος αναζήτησης είναι δυνατό να είναι όλος γνωστός την ώρα επίλυσης του προβλήματος, αλλά συνήθως, εκ των προτέρων, γνωρίζουμε μόνο λίγα σημεία του και παράγουμε τα υπόλοιπα καθώς η διαδικασία αναζήτησης λύσεων εξελίσσεται.

Το πρόβλημα είναι ότι η αναζήτηση μπορεί να είναι περίπλοκη. Οι γενετικοί αλγόριθμοι χρησιμοποιούνται ως μια μέθοδος αναζήτησης μιας κατάλληλης λύσης (όχι απαραίτητα της βέλτιστης). Η λύση που ευρίσκεται θεωρείται καλή λύση, κι όχι βέλτιστη, επειδή τις πλείστες φορές δεν είναι εφικτό να αποδείξεις την πραγματικά βέλτιστη λύση.

Ανάλογο του χώρου αναζήτησης στη γενετική είναι το τοπίο καταλληλότητας (fitness landscape). Ορίστηκε από το βιολόγο Sewell Wright το 1931 ως η αναπαράσταση του χώρου όλων των πιθανών γονοτύπων ενός πληθυσμού ανάλογα με την καταλληλότητά τους [3].



Εικόνα 4 : *Παράδειγμα δισδιάστατου χώρου αναζήτησης (search space)*

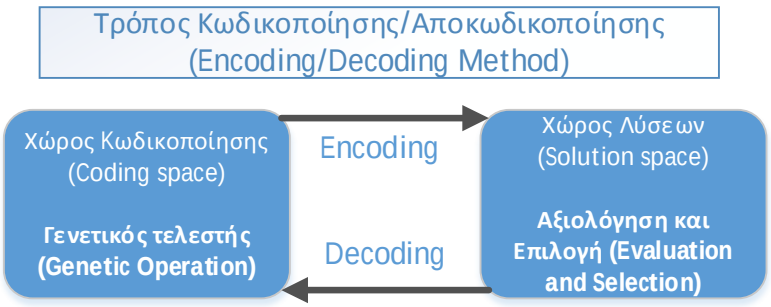


Εικόνα 5*: Παράδειγμα τρισδιάστατου χώρου αναζήτησης (search space)*

1.3 Αναπαράσταση πιθανών λύσεων – Κωδικοποίηση χρωμοσωμάτων (Encoding)

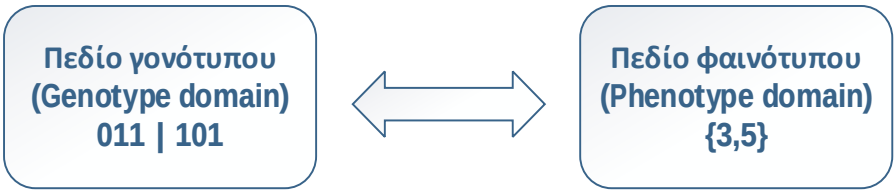
Πριν το στάδιο της αρχικοποίησης απαιτείται οι μεταβλητές απόφασης (decision variables) που αποτελούν τα χαρακτηριστικά κάθε πιθανής λύσης να κωδικοποιηθούν στα χρωμοσώματα. Κάθε παράμετρος του προβλήματος αντιστοιχεί σε ένα γονίδιο του χρωμοσώματος και πρέπει να κωδικοποιηθεί. Συνεπώς, πριν από οποιαδήποτε διαδικασία, πρέπει πρωταρχικά, να επιλεγεί μια μέθοδος αναπαράστασης των πιθανών λύσεων και κωδικοποίησής τους στο χρωμόσωμα σε

μορφή επεξεργάσιμη από τον υπολογιστή, δηλαδή με μαθηματικό τρόπο. Η κωδικοποίηση των χρωμοσωμάτων αποτελεί το πεδίο τιμών το οποίο ερευνά ο αλγόριθμος.



Εικόνα 6: *Τρόπος κωδικοποίησης και αποκωδικοποίησης*

Ο γενετικός αλγόριθμος δουλεύει παράλληλα σε δύο είδη χώρων, τον χώρο κωδικοποίησης (coding space ή search space) - αντίστοιχο του βιολογικού γονότυπου (genotype)- και στο χώρο λύσεων (solution space) – αντίστοιχο του βιολογικού φαινότυπου (phenotype). Ο φαινότυπος στη βιολογία, περιγράφει την εμφάνιση ενός ατόμου ενώ ο γονότυπος την κωδικοποιημένη πληροφορία στα χρωμοσώματα του ατόμου. Στους γενετικούς αλγόριθμους γονότυπος είναι το σύνολο των χρωμοσωμάτων που κωδικοποιούν πιθανές λύσεις και φαινότυπος είναι η αποκωδικοποίηση των χρωμοσωμάτων ώστε να προκριθούν οι καλύτεροι εκπρόσωποι του πληθυσμού. Υπάρχει ένας μετασχηματισμός μεταξύ γονότυπου και φαινότυπου, που καλείται mapping και χρησιμοποιεί την πληροφορία από το γονότυπο για να κατασκευάσει τον φαινότυπο.



Εικόνα 7: *Παράδειγμα mapping χρωμοσώματος με μεταβλητές απόφασης*

Με την έννοια χρωμόσωμα, εννοούμε μια συμβολοσειρά σταθερού, συνήθως, μήκους όπου αποθηκεύεται όλη η γενετική πληροφορία ενός ατόμου (για παράδειγμα S=1001110) . Τα γονίδια είναι τα δομικά στοιχεία του χρωμοσώματος, τα σύμβολα, δηλαδή, που συγκροτούν το χρωμόσωμα (π.χ. στην S τα bits). Locus καλείται η θέση ενός ψηφίου στην ακολουθία και alleles οι τιμές ενός ψηφίου στην ακολουθία (π.χ. στην S η τιμή ενός ψηφίου μπορεί να είναι 0 ή 1). Στον

πίνακα 1 αναφέρονται οι βασικές έννοιες του βιολογικού υπόβαθρου που είναι σχετικές με την κωδικοποίηση των χρωμοσωμάτων, καθώς και η εννοιολογική σημασία που αποκτούν στους γενετικούς αλγόριθμους.

Πίνακας 1: Βιολογικές ορολογίες – Βασική Γενετική

Βιολογική ορολογία	Βιολογική Εννοιολογική Σημασία	Εννοιολογική Σημασία στους ΓΑ
Οργανισμός (organism)	Σύνολο κανόνων: πώς ένας οργανισμός είναι κατασκευασμένος, αποτελείται από χρωμοσώματα	Αποκωδικοποίηση χρωμοσωμάτων
Χρωμόσωμα ή αλλιώς άτομο (chromosome / individual)	Συμβολοσειρά DNA, σαν μοντέλο όλου του οργανισμού, περιλαμβάνει πολλά γονίδια	Σύνολο γονιδίων, περιέχει τη λύση σε μορφή γονιδίων, Συμβολοσειρά όλων των μεταβλητών απόφασης μιας πιθανής λύσης
Γονίδιο (gene)	Συγκεκριμένη αλληλουχία/τμήμα του DNA. Αποθηκεύει μια συγκεκριμένη γενετική πληροφορία	Σύμβολο μιας μεταβλητής απόφασης , το γονίδιο περιέχει ένα κομμάτι της λύσης π.χ. εάν 23916 ένα χρωμόσωμα 2,3,8,1,6 είναι τα γονίδιά του
Χαρακτηριστικό γνώρισμα (trait)	Κάθε γονίδιο κωδικοποιεί μια συγκεκριμένη πρωτεΐνη που αντιπροσωπεύει ένα συγκεκριμένο χαρακτηριστικό γνώρισμα του ατόμου	Η αποκωδικοποίηση του γονιδίου αναπαριστά ένα συγκεκριμένο χαρακτηριστικό της λύσης
Αλληλόμορφο/αλλήλιο (allele)	Πιθανές τιμές για ένα χαρακτηριστικό (μια από τις εναλλακτικές μορφές ενός γονιδίου)	Πιθανές τιμές για ένα χαρακτηριστικό
Θέση (locus)	Θέση γονιδίου στο χρωμόσωμα	Θέση συμβόλου στο χρωμόσωμα
Γονιδίωμα (genome)	Ολικό σύνολο γενετικού υλικού που φέρεται σε ένα άτομο (Πλήρης αλληλουχία DNA που περιέχει σύνολο γενετικής πληροφορίας που είναι κωδικοποιημένο στα γονίδια)	Ολικό σύνολο κωδικοποιημένης πληροφορίας σε όλα τα χρωμοσώματα του ατόμου
Γονότυπος (genotype)	Το σύνολο γονιδίων σε ένα γονιδίωμα. Για πρακτικούς λόγους ο γονότυπος μπορεί να αναφέρεται σε ένα γονίδιο ή μια ομάδα γονιδίων που φέρουν την γενετική πληροφορία για μια συγκεκριμένη ιδιότητα του	Συγκεκριμένο σύνολο **κωδικοποιημένης** πληροφορίας σε χρωμοσώματα του ατόμου

	ατόμου	
Φαινότυπος (phenotype)	Η φυσική έκφραση του γονότυπου: η εμφάνιση, τα «ορατά» χαρακτηριστικά του ατόμου που καθορίζονται από τις πληροφορίες των γονιδίων	Τα χαρακτηριστικά της υποψήφιας λύσης: το **αποκωδικοποιημένο** περιεχόμενο ενός συγκεκριμένου χρωμοσώματος
Διασταύρωση (crossover/recombination)	Όταν ο απόγονος δύο οργανισμών έχει μισά γονίδια από τον ένα γονέα και μισά από τον άλλο	Συνδυασμός χρωματοσωμάτων για την παραγωγή απογόνων
Μετάλλαξη (mutation)	Αλλάζουν κάποια στοιχεία του DNA , συνήθως, λόγω λάθους στην αντιγραφή των γονιδίων από τους γονείς	Αλλαγή κάποιων τυχαίων γονιδίων των χρωμοσωμάτων
Καταλληλότητα (fitness)	Επιτυχία επιβίωσης του οργανισμού	Η τιμή που δίνεται σε ένα άτομο βάσει του πόσο κοντά στη βέλτιστη λύση βρίσκεται το χρωμόσωμα

Ανάλογα με το πρόβλημα, τους περιορισμούς και τις απαιτήσεις του επιλέγεται η αντίστοιχη μορφή κωδικοποίησης. Είναι πιθανό ένα πρόβλημα να επιδέχεται περισσότερες από μια κωδικοποιήσεις. Κάποιες βασικές μορφές κωδικοποίησης που έχουν ήδη χρησιμοποιηθεί είναι οι ακόλουθες: Δυαδική κωδικοποίηση (Binary encoding), Οκταδική κωδικοποίηση (Octal encoding), δεκαεξαδική κωδικοποίηση (Hexadecimal encoding), κωδικοποίηση με δομές ακέραιων τιμών στο δεκαδικό σύστημα (decimal integer encoding), κωδικοποίηση με δομές δεκαδικών τιμών στο δεκαδικό σύστημα (decimal float encoding) , κωδικοποίηση μεταλλαγής (Permutation encoding), κωδικοποίηση τιμής (Value encoding), υβριδική κωδικοποίηση, κωδικοποίηση σε μορφή αλφαριθμητικής ακολουθίας και κωδικοποίηση δέντρου.

Ο κάθε τύπος κωδικοποίησης ανάλογα με τη δομή του, μπορεί να ταξινομηθεί ως μονοδιάστατος ή δισδιάστατος. Η δυαδική, οκταδική, δεκαεξαδική κωδικοποίηση, η κωδικοποίηση μεταλλαγής και η κωδικοποίηση τιμής θεωρούνται μονοδιάστατες κωδικοποιήσεις ενώ η κωδικοποίηση δέντρου θεωρείται δισδιάστατη.

Ο τύπος κωδικοποίησης μπορεί να ταξινομηθεί και ως προς την τιμή αξιολόγησης από τη συνάρτηση καταλληλότητας. Σύμφωνα με τον Goldberg [5], η συνάρτηση καταλληλότητας για ένα συγκεκριμένο σχήμα κωδικοποίησης εξαρτάται από δύο παράγοντες: την τιμή και τη διάταξη. Παράδειγμα κωδικοποίησης που βασίζεται μόνο στη διάταξη είναι η κωδικοποίηση μεταλλαγής ενώ βασισμένο μόνο στην τιμή είναι η κωδικοποίηση τιμής. Παράδειγμα κωδικοποίησης βασισμένο και στην τιμή και στη διάταξη είναι η δυαδική κωδικοποίηση.

Είναι πιθανό ένα πρόβλημα να επιδέχεται περισσότερες από μία κωδικοποιήσεις και η επιλογή της καταλληλότερης είναι κρίσιμο βήμα σε ένα ΓΑ . Κάθε χρωμόσωμα αναπαρίσταται ως ένα διάνυσμα και υλοποιείται ως ένας πίνακας πολλών μεταβλητών.

Το χρωμόσωμα κατασκευάζεται με την τοποθέτηση όλων των κωδικοποιημένων χαρακτηριστικών το ένα πλησίον του άλλου. Έτσι, και κατά την αποκωδικοποίηση(decoding) του χρωμοσώματος γίνεται εύκολα η εύρεση των τιμών των χαρακτηριστικών.

1.3.1 Δυαδική κωδικοποίηση (Binary Encoding)

Η πιο τυπική προσέγγιση και πιο ευρέως διαδεδομένη είναι η δυαδική κωδικοποίηση (binary encoding), δηλαδή η κωδικοποίηση με χρήση των δυαδικών ψηφίων 0 και 1. Η δυαδική αναπαράσταση με το αλφάβητο {0,1} προτάθηκε από το Holland το 1975. Οι πρώτες μελέτες σχετικά με γενετικούς αλγόριθμους χρησιμοποίησαν αυτόν τον τύπο κωδικοποίησης εξαιτίας της απλότητάς της καθώς και του γεγονότος ότι είναι πιο κοντά στη γλώσσα μηχανής.

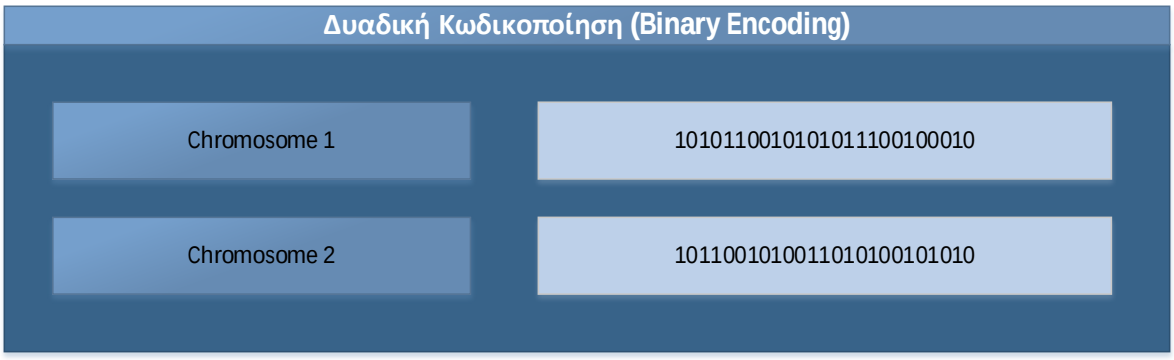
Κάθε πιθανή λύση αναπαρίσταται στο χρωμόσωμα ως μια ακολουθία 0 και 1 (binary string {0,1}) και κάθε bit δείχνει ένα χαρακτηριστικό της λύσης. Ας σημειωθεί, ότι όσο πιο μεγάλη είναι η ακολουθία των 0 και 1, (δηλαδή όσες περισσότερες είναι οι μεταβλητές απόφασης και αντίστοιχα τα γονίδια του χρωμοσώματος) τόσο περισσότερο μεγαλώνει ο χώρος αναζήτησης με εκθετικούς ρυθμούς [6].

Με τη δυαδική κωδικοποίηση προκύπτει μεγάλο πλήθος χρωμοσωμάτων ακόμα και αν ο αριθμός γονιδίων είναι μικρός [7]. Παρ' όλα αυτά η συγκεκριμένη κωδικοποίηση είναι αναποτελεσματική σε μεγάλο αριθμό προβλημάτων.

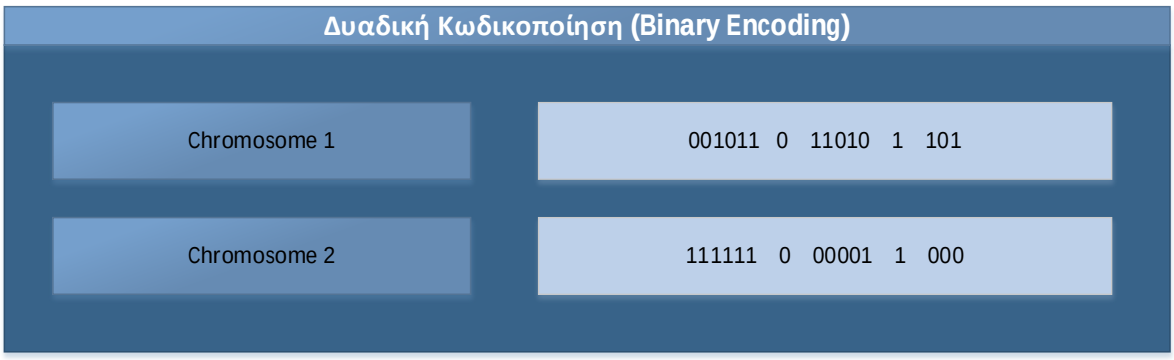
Με κάθε είδος κωδικοποίησης μπορούν να χρησιμοποιηθούν οι αντίστοιχοι μέθοδοι τελεστών αναπαραγωγής. Με τη δυαδική κωδικοποίηση οι πιο συνήθεις μέθοδοι διασταύρωσης (crossover) – που θα εξηγηθούν πιο αναλυτικά σε επόμενο υποκεφάλαιο – είναι οι 1-point crossover, N-point crossover, Uniform crossover και Arithmetic crossover. Η πιο συχνά χρησιμοποιούμενη υλοποίηση τελεστή μετάλλαξης (mutation) είναι η flip. Στη flip μετάλλαξη κάποια bits αλλάζουν από 0 σε 1 και αντιστρόφως από 1 σε 0 ανάλογα με τις τιμές των γονιδίων του χρωμοσώματος προς μετάλλαξη. Ο αριθμός των γονιδίων που θα μεταλλαχτούν καθορίζεται από την πιθανότητα μετάλλαξης (mutation rate) [8].

Η δυαδική κωδικοποίηση και η flip μετάλλαξη χρησιμοποιούνται στο πρόβλημα του σακιδίου (knapsack problem), όπου με χρήση της δυαδικής κωδικοποίησης, με την τιμή 1 υποδηλώνεται ότι ένα άτομο υπάρχει στο σακίδιο, ενώ με 0 ότι δεν υπάρχει. Το πρόβλημα του σακιδίου είναι το εξής: Δίνονται n είδη και ένα σακίδιο μεγέθους B. Κάθε είδος i είναι διαθέσιμο σε ποσότητα(size)

s_i με αξία (value) p_i . Κάθε είδος i μπορεί να συμπεριληφθεί στο σακίδιο σε οποιοδήποτε ποσοστό. Ζητείται συλλογή μέγιστης αξίας που χωράει στο σακίδιο. Στην κωδικοποίηση κάθε bit του χρωμοσώματος εκφράζει αν ένα είδος συμπεριλαμβάνεται στο σακίδιο ή όχι.



(α)



(β)

Εικόνα 8: *(α) Παράδειγμα κλασικής δυαδικής κωδικοποίησης σε χρωμοσώματα 25 γονιδίων (κάθε bit 0 ή 1 αντιστοιχεί σε διαφορετική μεταβλητή) (β) Παράδειγμα δυαδικής κωδικοποίησης σε χρωμοσώματα 5 γονιδίων (στο χρωμόσωμα 1: το 1ο γονίδιο έχει τιμή 11, το 2ο έχει τιμή 0, το 3ο έχει τιμή 26, το 4ο έχει τιμή 1 και το 5ο έχει τιμή 5)*

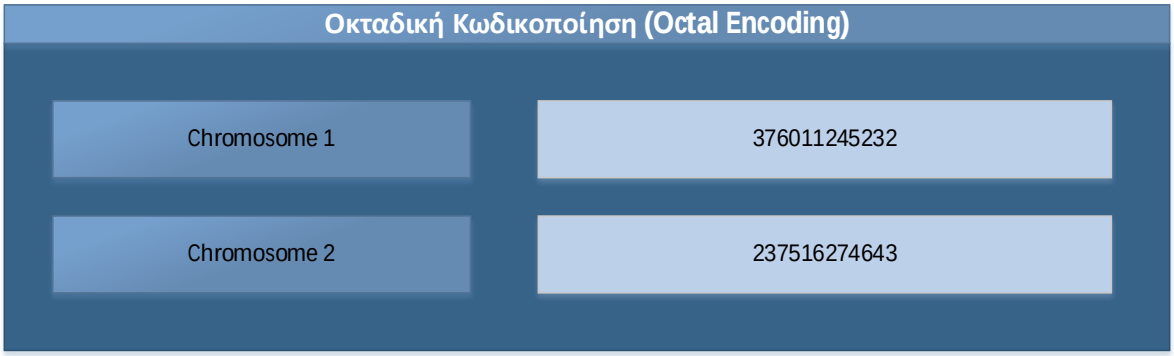
Η δυαδική κωδικοποίηση εκτός από την κλασική της μορφή, όπου κάθε μεταβλητή/γονίδιο της λύσης αναπαρίσταται με 1 bit, μπορεί να χρησιμοποιηθεί και με διαφορετικό τρόπο: κάθε γονίδιο της λύσης μπορεί να πάρει ως τιμή μια δυαδική συμβολοσειρά από 0 και 1. Αυτού του είδους η κωδικοποίηση ονομάζεται και κωδικοποίηση τιμής (περιγράφεται παρακάτω).

Στην *εικόνα 8(β)* παρουσιάζεται ένα τέτοιο παράδειγμα. Έστω ένα χρωμόσωμα που αποτελείται από 5 γονίδια όπου κάθε γονίδιο περιέχει πληροφορία σχετική με την αντίστοιχη μεταβλητή: το πρώτο γονίδιο αποτελείται από 6 bits, το δεύτερο από 1 bit, το τρίτο από 5 bits, το τέταρτο από 1 bit και το πέμπτο από 3 bits.

Συνήθως όταν γίνεται αναφορά σε δυαδική κωδικοποίηση εννοείται με την κλασική μορφή της, όπως αυτή παρουσιάζεται στην εικόνα 8(α): κάθε δυαδικό bit αναπαριστά διαφορετικό χαρακτηριστικό της λύσης.

1.3.2 Οκταδική κωδικοποίηση (Octal encoding)

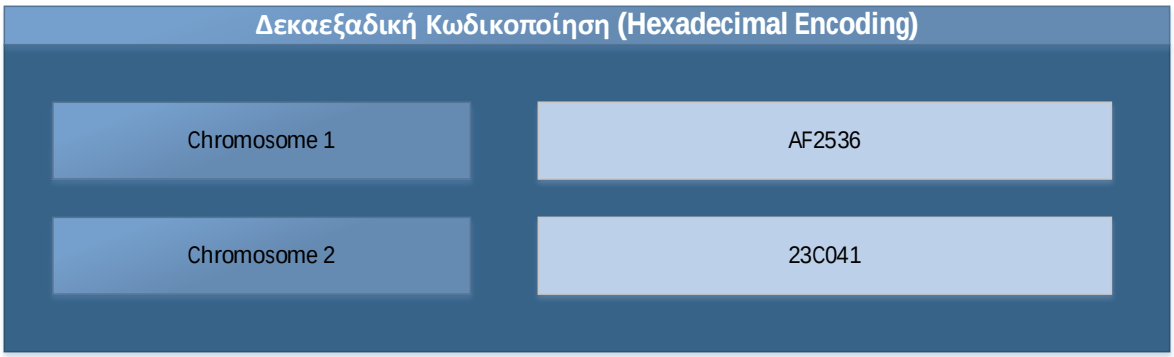
Σε αυτή την τεχνική κωδικοποίησης το χρωμόσωμα αναπαρίσταται ως μια συμβολοσειρά που χρησιμοποιεί χαρακτήρες του οκταδικού αριθμητικού συστήματος {0,1,2,3,4,5,6,7}.



Εικόνα 9: *Παράδειγμα οκταδικής κωδικοποίησης σε χρωμοσώματα 12 γονιδίων*

1.3.3 Δεκαεξαδική κωδικοποίηση (Hexadecimal encoding)

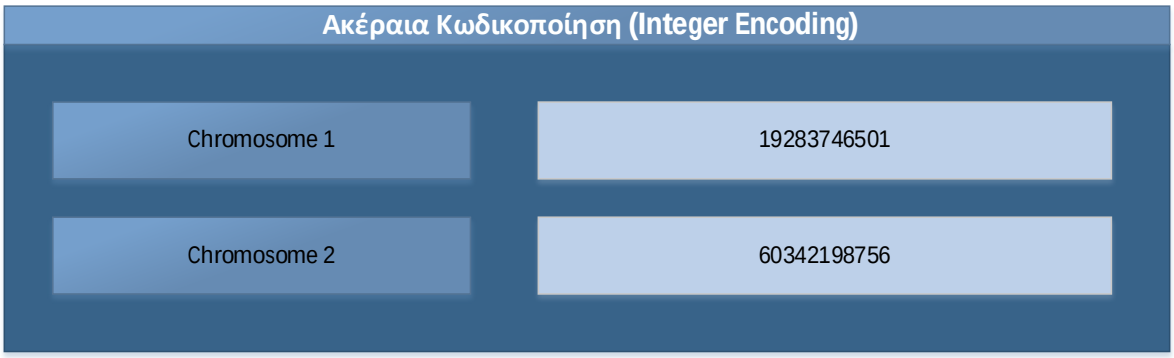
Σ' αυτή την κωδικοποίηση το χρωμόσωμα αναπαρίσταται χρησιμοποιώντας τους δεκαεξαδικούς αριθμούς {0,1,2,3,4,5,6,7,8,9,A,B,C,D,E,F} .



Εικόνα 10: *Παράδειγμα δεκαεξαδικής κωδικοποίησης σε χρωμοσώματα 6 γονιδίων*

1.3.4 Κωδικοποίηση με δομές ακεραίων (integer encoding)

Η κωδικοποίηση των λύσεων σαν δομές ακέραιων τιμών (integer encoding) χρησιμοποιεί μόνο ακέραιες τιμές από το δεκαδικό αριθμητικό σύστημα {0,1,2,3,4,5,6,7,8,9}.

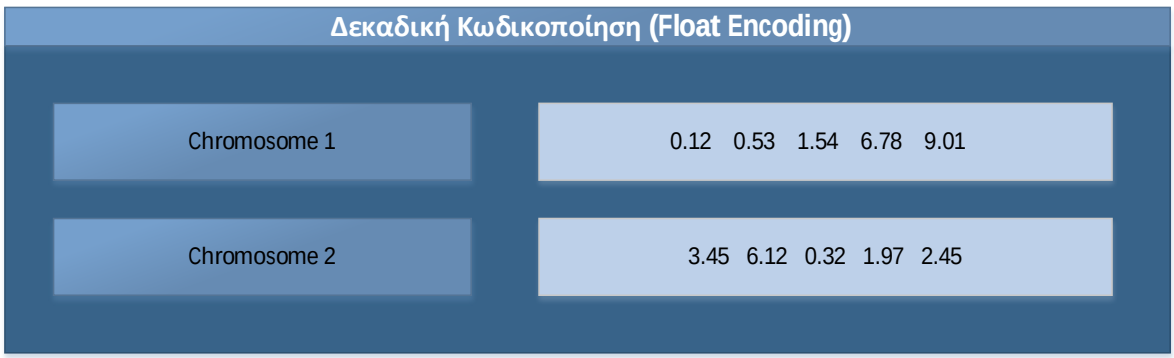


Εικόνα 11: *Παράδειγμα ακέραιας κωδικοποίησης (δεκαδικό σύστημα)*

1.3.5 Κωδικοποίηση με δομές δεκαδικών τιμών (float encoding)

Η κωδικοποίηση των λύσεων σαν δομές δεκαδικών τιμών (float encoding) χρησιμοποιεί τα ψηφία του δεκαδικού αριθμητικού συστήματος {0,1,2,3,4,5,6,7,8,9} και παράγει τιμές σε μορφή υπό διαστολή αριθμών. Αυτή η προσέγγιση παρέχει μεγαλύτερη σχεδιαστική ευχέρεια στην αναζήτηση λύσεων σε προβλήματα πραγματικών μεταβλητών [6].

Πιο πρόσφατα, η κωδικοποίηση με δομές δεκαδικών αριθμών έχει τύχει ευρείας χρήσης, κυρίως επειδή προτιμάται η αντιστοίχιση κάθε παραμέτρου με γονίδιο ενός συμβόλου. Εκτός αυτού, προσφέρει αρκετά οφέλη: Είναι πιο γρήγορος ο χειρισμός της, έχει δειχθεί εμπειρικά ότι είναι πιο συνεπής με κάθε νέο τρέξιμό της, επιτρέπει υψηλότερη ακρίβεια και είναι διαισθητικά πιο κοντά στο χώρο λύσεων του προβλήματος. Το τελευταίο σημείο, είναι ιδιαίτερα χρήσιμο επειδή επιτρέπει ευκολότερη ενσωμάτωση σε συγκεκριμένους τομείς γνώσης [6].



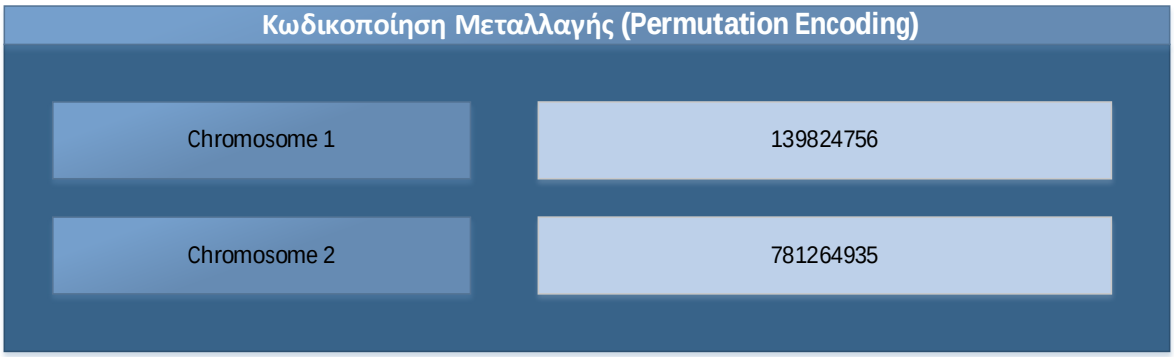
Εικόνα 12: *Παράδειγμα δεκαδικής κωδικοποίησης*

1.3.6 Κωδικοποίηση μεταλλαγής (Permutation encoding)

Η κωδικοποίηση μεταλλαγής χρησιμοποιείται σε προβλήματα διάταξης, όπως για παράδειγμα το πρόβλημα του πλανόδιου πωλητή ή προβλήματα καθορισμού σειράς στόχων. Στην κωδικοποίηση αυτή, κάθε χρωμόσωμα είναι μια ακολουθία αριθμών, η οποία αντιπροσωπεύει μια σειρά. Κάθε αριθμός αντιπροσωπεύει μια θέση σε μια σειρά.

Το πρόβλημα του πλανόδιου πωλητή είναι το εξής: Δίνονται n πόλεις, καθώς και οι μεταξύ τους αποστάσεις d_{ij}. Ένας πλανόδιος πωλητής πρέπει να τις επισκεφτεί όλες -την κάθε μια, μια φορά ακριβώς- διανύοντας τη μικρότερη απόσταση. Ζητείται η ακολουθία των πόλεων για την οποία ελαχιστοποιείται η απόσταση. Στο πρόβλημα του πλανόδιου πωλητή, η συμβολοσειρά αριθμών αντιπροσωπεύει την ακολουθία των πόλεων με τη σειρά που τις επισκέπτεται ο πλανόδιος πωλητής.

Αυτή η κωδικοποίηση είναι χρήσιμη μόνο σε προβλήματα που έχουν συγκεκριμένη διάταξη. Σε κάποιες περιπτώσεις για κάποιους τύπους μετάλλαξης και διασταύρωσης, αφού ολοκληρωθούν, θα πρέπει ακολούθως να γίνουν κάποιες διορθώσεις. Οι τελεστές διασταύρωσης που χρησιμοποιούνται στην κωδικοποίηση μεταλλαγής είναι οι Partially mapped crossover (PMX), Cycle crossover (OCX) και η Order crossover (OX). Η πιο συνήθης μέθοδος μετάλλαξης που εφαρμόζεται στα διατεταγμένα χρωμοσώματα είναι η αντιστροφή (inversion). Η αντιστροφή επιλέγει τυχαία δύο θέσεις στο χρωμόσωμα και ανταλλάζει τη διάταξη των τιμών τους.



Εικόνα 13: *Παράδειγμα κωδικοποίησης μεταλλαγής*

1.3.7 Κωδικοποίηση τιμής (Value encoding)

Εκτός από τη δυαδική αναπαράσταση που χρησιμοποιείται στην πλειονότητα των περιπτώσεων, υπάρχουν και προβλήματα που έχουν ως παραμέτρους πραγματικούς αριθμούς π.χ. προβλήματα επεξεργασίας εικόνας.

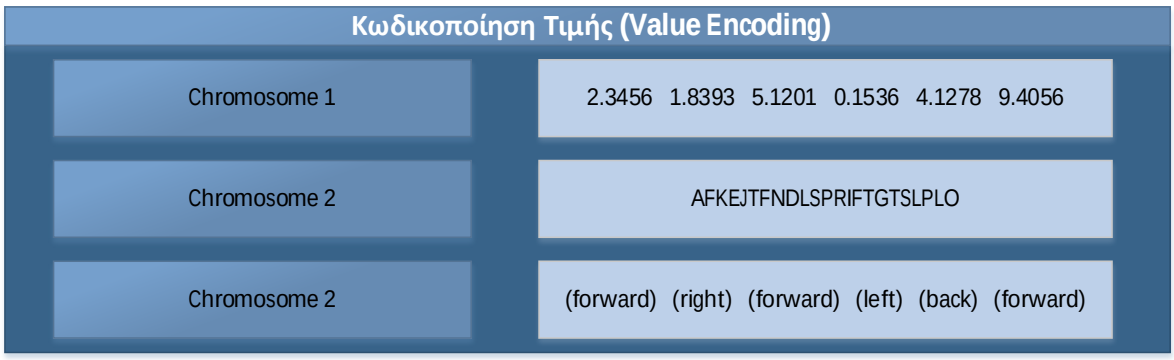
Η κωδικοποίηση τιμής (value encoding) χρησιμοποιείται σε προβλήματα που κάνουν υπολογισμούς με πιο πολύπλοκες τιμές όπως οι πραγματικοί αριθμοί. Σε αυτή την κωδικοποίηση κάθε χρωμόσωμα είναι μια ακολουθία από τιμές. Οι τιμές αυτές είναι τιμές που συνδέονται με το πρόβλημα και μπορεί να είναι ακέραιοι αριθμοί (integers), πραγματικοί αριθμοί (real numbers), αλφαριθμητικοί χαρακτήρες (chars) , τυποποιημένοι αριθμοί (form numbers) ή κάποιου είδους αντικείμενα (objects).

Στην περίπτωση που οι τιμές των χρωμοσωμάτων είναι ακέραιες οι τελεστές διασταύρωσης που χρησιμοποιούνται είναι ίδιοι με αυτούς που χρησιμοποιούνται στη δυαδική κωδικοποίηση.

Αυτή η μορφή κωδικοποίησης επιλέγεται σε πιο πολύπλοκα προβλήματα, αλλά συνήθως απαιτείται παράλληλα και η δημιουργία νέων τελεστών διασταύρωσης και μετάλλαξης προσαρμοσμένων στο υπό εξέταση πρόβλημα.

Συνηθίζεται να χρησιμοποιείται στα νευρωνικά δίκτυα (neural networks) για την εύρεση των βαρών του νευρωνικού δικτύου. Δίνεται νευρωνικό δίκτυο συγκεκριμένης αρχιτεκτονικής και ζητούνται τα βάρη για τις εισόδους των νευρώνων με τα οποία πρέπει να εκπαιδευτεί το δίκτυο έτσι ώστε να προκύψει η επιθυμητή έξοδος. Κάθε τιμή του χρωμοσώματος - που σε αυτή την περίπτωση είναι πραγματικός αριθμός- αντιπροσωπεύει το αντίστοιχο βάρος (weight) για κάθε είσοδο (input).

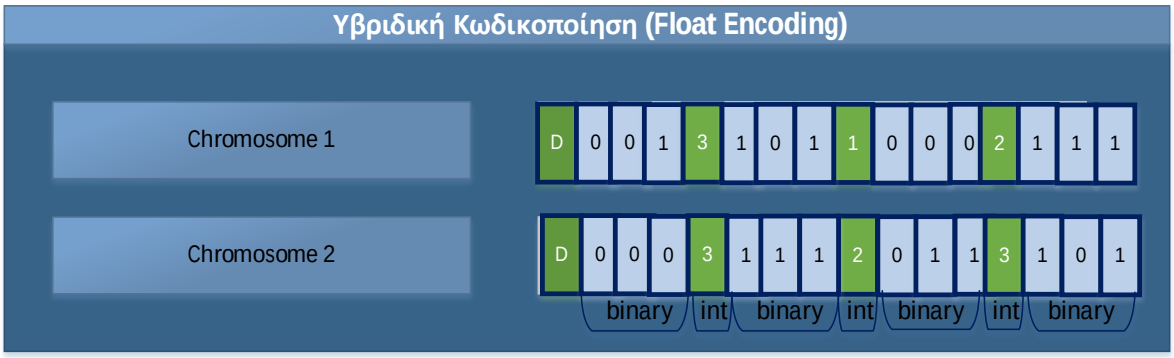
Η κωδικοποίηση με δομές ακέραιων ή δεκαδικών τιμών που αναφέρθηκαν παραπάνω είναι υποπεριπτώσεις της κωδικοποίησης με τιμή.



Εικόνα 14: *Παραδείγματα κωδικοποίησης τιμής*

1.3.8 Υβριδική κωδικοποίηση

Η υβριδική κωδικοποίηση, είναι ένας συνδυασμός της δυαδικής κωδικοποίησης και της κωδικοποίησης με δομές ακέραιων και δεκαδικών αριθμών, καθώς και χρήσης του γράμματος του αλφαβήτου D [9]. Κάθε μεταβλητή απόφασης αναπαρίσταται είτε σε δυαδική μορφή είτε σε μορφή ακεραίου ή δεκαδικού.



Εικόνα 15: *Παράδειγμα υβριδικής κωδικοποίησης*

1.3.9 Κωδικοποίηση σε μορφή αλφαριθμητικής ακολουθίας

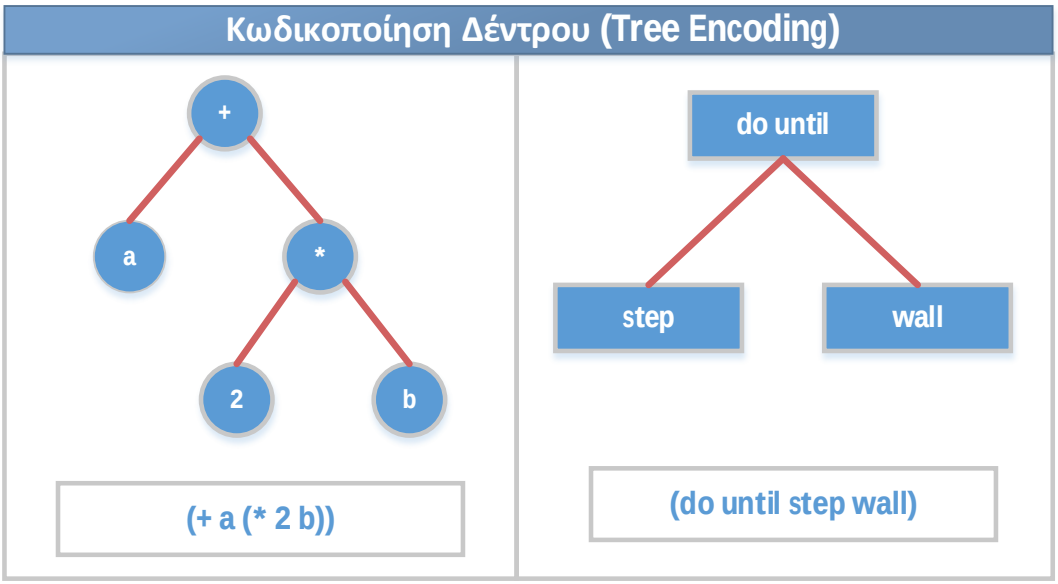
Παράδειγμα της τελευταίας μορφής κωδικοποίησης, της αλφαριθμητικής, είναι η "γραμματική" κωδικοποίηση (grammar encoding) της οποίας γίνεται χρήση στο [10].

Το μήκος της αναπαράστασης ενός χρωμοσώματος συνήθως έχει σταθερό μήκος, ωστόσο, αυτό δεν είναι απαραίτητο, μπορεί να έχει και μεταβλητό μήκος. Έχουν υλοποιηθεί γενετικοί αλγόριθμοι με μεταβλητό μήκος χρωμοσώματος [11]. Είναι εφικτό, αλλά παρουσιάζονται διάφορες δυσκολίες, τόσο στην αποκωδικοποίηση των χρωμοσωμάτων μεταβλητού μήκους για την αξιολόγησή τους από τη συνάρτηση κόστους, καθώς και στην υλοποίηση κατάλληλων τελεστών αναπαραγωγής ανάμεσα σε χρωμοσώματα διαφορετικού μήκους.

1.3.10 Κωδικοποίηση δέντρου (Tree encoding)

Η κωδικοποίηση δέντρου χρησιμοποιείται κυρίως σε εξελικτικά προγράμματα, εκφράσεις του γενετικού προγραμματισμού ή οποιαδήποτε δομή που μπορεί να κωδικοποιηθεί με τη μορφή δέντρου. Στην κωδικοποίηση δέντρου κάθε χρωμόσωμα είναι ένα δέντρο αντικειμένων, όπως συναρτήσεις ή εντολές σε γλώσσα προγραμματισμού. Συνήθως συνδυάζεται με τη γλώσσα προγραμματισμού LISP επειδή τα προγράμματα σε LISP αναπαρίστανται κατευθείαν σε μορφή

δέντρου και έτσι μπορούν και να αναλυθούν σχετικά εύκολα με τους τελεστές διασταύρωσης και μετάλλαξης.



Εικόνα 16: *Παραδείγματα κωδικοποίησης δέντρου*

1.4 Αρχικοποίηση (Initialization)

Στο στάδιο της αρχικοποίησης, το πρώτο στάδιο ενός γενετικού αλγορίθμου, παράγεται ένας αρχικός πληθυσμός από τυχαία δημιουργούντα μεμονωμένα χρωματοσώματα-λύσεις. Προτού γίνει αυτό, οι μεταβλητές απόφασης του προβλήματος έχουν ήδη κωδικοποιηθεί σε χρωμοσώματα. Αυτός ο πληθυσμός περιέχει πιθανές λύσεις του προβλήματος, λύσεις δηλαδή, που βρίσκονται στο χώρο αναζήτησης λύσεων του προβλήματος (search space). Το μέγεθος του πληθυσμού αυτού, εξαρτάται από τη φύση του εκάστοτε προβλήματος, αλλά τυπικά περιέχει έναν μεγάλο αριθμό πιθανών λύσεων (συνήθως μεταξύ 50 με 500 λύσεις). Περιστασιακά, οι λύσεις μπορούν να αναζητηθούν στις περιοχές, στις οποίες οι βέλτιστες λύσεις είναι πιθανό να βρεθούν.

Με τη δημιουργία της πρώτης γενιάς ξεκινά η διαδικασία. Σε κάθε χρωμόσωμα εφαρμόζεται ένα κριτήριο κόστους με τη χρήση της συνάρτησης καταλληλόλητας (fitness function) -αλλού αναφέρεται συνάρτηση προσαρμοστικότητας ή ποιότητας. Η συνάρτηση καταλληλόλητας υποδεικνύει πόσο κατάλληλο είναι το χρωμόσωμα υπό αξιολόγηση, ως λύση για το εξεταζόμενο πρόβλημα.

1.4.1 Αξιολόγηση χρωμοσωμάτων - Συνάρτηση καταλληλότητας (Fitness function)

Απαραίτητο συστατικό της εξελικτικής διαδικασίας είναι τα χρωμοσώματα του πληθυσμού να εμφανίζουν διαφορές στην ικανότητά τους για επιβίωση και ο βαθμός στον οποίο κάθε χρωμόσωμα διαθέτει την ικανότητα αυτή να είναι ανάλογος της δυνατότητάς του για αναπαραγωγή. Το μέτρο αυτό της ικανότητας αναφέρεται ως καταλληλότητα (fitness) και θεωρείται εγγενές χαρακτηριστικό των βιολογικών συστημάτων: η καταλληλότητα ενός γονιδίου ή χρωμοσώματος αντανακλάται στη δυνατότητά του να αναπαράγεται. Υπό αυτό το πρίσμα, οδηγούμαστε σε ταυτολογία: η καταλληλότητα αποτελεί μέτρο της δυνατότητας του ατόμου να αναπαράγεται ενώ ο ρυθμός αναπαραγωγής του είναι ανάλογος της καταλληλότητάς του [12].

Η καταλληλόλητα ενός οργανισμού ορίζεται τυπικά ως:

1) η πιθανότητα ενός οργανισμού να επιβιώσει ώστε να αναπαράγει (βιωσιμότητα)
2) ή ως η συνάρτηση του αριθμού των απογόνων που έχει ο οργανισμός (γονιμότητα) [3]

Στους ΓΑ η επιλογή των χρωμοσωμάτων -πιθανών λύσεων- προς αναπαραγωγή γίνεται με καθορισμένο τρόπο και γι’ αυτό είναι αναγκαία η χρήση μιας μεθόδου η οποία θα είναι υπεύθυνη για την συγκριτική αξιολόγηση των χρωμοσωμάτων. Ο ΓΑ αφού αποκωδικοποιήσει κάθε χρωμόσωμα του τρέχοντος πληθυσμού κάνει χρήση της λεγόμενης συνάρτησης αξιολόγησης ή καταλληλότητας (fitness function) που θα αναθέσει μια τιμή καταλληλότητας/ απόδοσης/ αξίας ικανότητας (fitness/score) σε κάθε χρωμόσωμα του τρέχοντος πληθυσμού [3]. Επομένως, παίρνει ως είσοδο την αποκωδικοποιημένη συμβολοσειρά (τα γονίδια κάθε ατόμου) και επιστρέφει την τιμή καταλληλότητάς της. Η τιμή αυτή αποτελεί και τον καθοριστικό παράγοντα επιβίωσης και πολλαπλασιασμού ή όχι του χρωμοσώματος. Η καταλληλόλητα ενός χρωμοσώματος εξαρτάται από το πόσο καλά ένα χρωμόσωμα επιλύει το υπό εξέταση πρόβλημα. Η συνάρτηση καταλληλότητας δίνει ένα ποσοτικό μέτρο αυτής της ικανότητας, της εγγύτητας δηλαδή στη βέλτιστη λύση (απόσταση ή ομοιότητα από την ιδανική λύση), αντιστοιχίζοντας τα σημεία του χώρου των υποψήφιων λύσεων και των σημείων του τοπίου καταλληλότητας (fitness landscape).

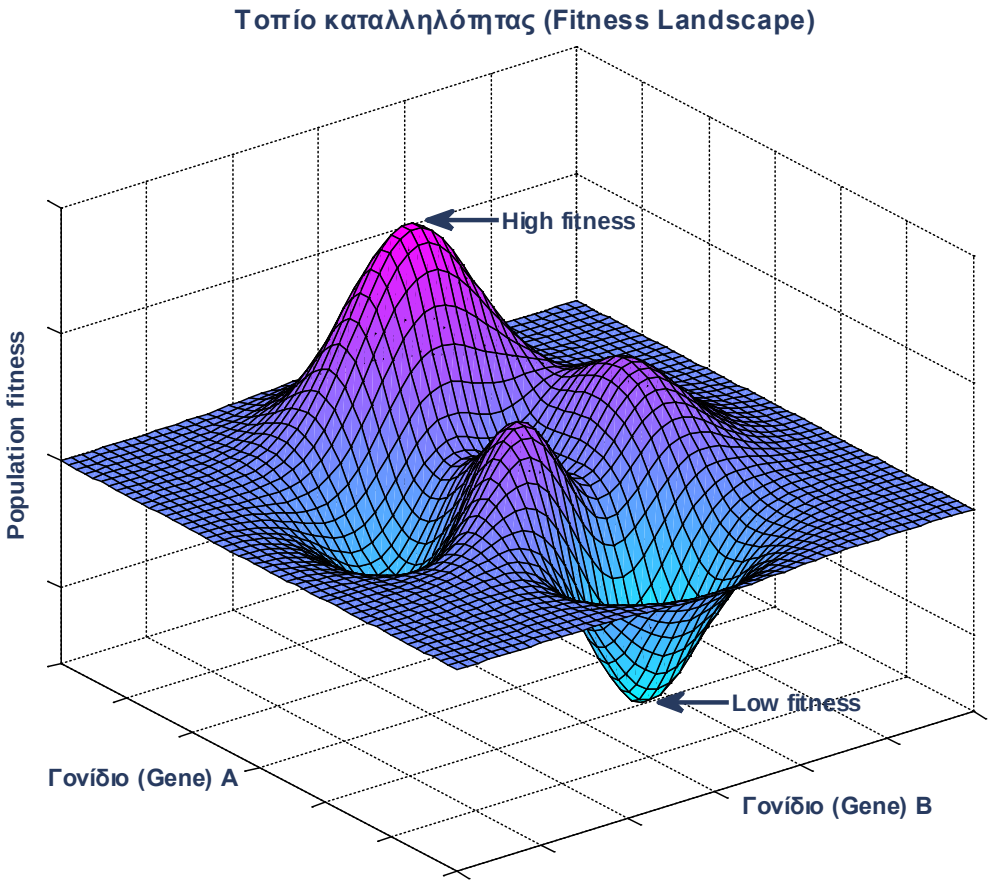
Είναι σημαντικό η συνάρτηση καταλληλότητας να είναι εύκολα υπολογίσιμη έτσι ώστε να μην επιβραδύνεται η εκτέλεση του ΓΑ.

Μια συνήθης εφαρμογή των ΓΑ είναι η βελτιστοποίηση συνάρτησης, όπου στόχος είναι η εύρεση ενός συνόλου τιμών παραμέτρων που ελαχιστοποιούν ή μεγιστοποιούν μια πολύπλοκη πολυπαραμετρική συνάρτηση.

Βελτιστοποίηση είναι η διαδικασία ρύθμισης των εισόδων ενός προβλήματος έτσι ώστε να βρεθεί η ελάχιστη ή μέγιστη έξοδος. Η είσοδος αποτελείται από μεταβλητές. Η διαδικασία ή συνάρτηση είναι η συνάρτηση καταλληλότητας (fitness function) και η έξοδος/αποτέλεσμα είναι η καταλληλόλητα - ή το κόστος. Εάν το κόστος πρέπει να ελαχιστοποιηθεί η βελτιστοποίηση γίνεται πρόβλημα ελαχιστοποίησης, ενώ αν το κόστος πρέπει να μεγιστοποιηθεί γίνεται πρόβλημα μεγιστοποίησης. Με αλλαγή πρόσημου μπορεί μια μεγιστοποίηση να μετατραπεί σε ελαχιστοποίηση και το αντίθετο. Η συνάρτηση κόστους παράγει έξοδο από ένα σύνολο μεταβλητών εισόδου. Μπορεί να είναι μια μαθηματική συνάρτηση, ένα πείραμα κοκ. Ο στόχος είναι να τροποποιηθεί η έξοδος στο επιθυμητό αποτέλεσμα βρίσκοντας τις κατάλληλες τιμές των μεταβλητών εισόδου. Το κόστος εκφράζει το πόσο διαφέρει μια πιθανή λύση από την επιθυμητή λύση [13].

Εικόνα 17: *Διάγραμμα μιας συνάρτησης ή διαδικασίας που βελτιστοποιείται: η βελτιστοποίηση αλλάζει την είσοδο έως ότου να επιτύχει την επιθυμητή έξοδο*

Στην εξελικτική υπολογιστική το τοπίο καταλληλότητας (fitness landscape) ή αλλιώς το τοπίο αξιών χρησιμοποιείται για την οπτικοποίηση της σχέσης του γενότυπου και του φαινότυπου των χρωμοσωμάτων στο εκάστοτε πρόβλημα. Ουσιαστικά, παρουσιάζεται η καταλληλότητα κάθε σημείου του χώρου αναζήτησης, δηλαδή κάθε πιθανής λύσης, σε μια γραφική παράσταση. Οι πιθανές λύσεις που παρουσιάζουν μεγάλη ομοιότητα βρίσκονται κοντά μεταξύ τους, ενώ αντίθετα αυτές που έχουν μικρή έως καθόλου ομοιότητα βρίσκονται μακρυά. Το υψόμετρο στο τοπίο καταλληλότητας καθορίζεται από την τιμή καταλληλότητας της αντίστοιχης υποψήφιας λύσης και για να ορίσουμε τη βελτιστότητα αναζητούμε το καθολικό ακρότατο (το μέγιστο ή ελάχιστο ανάλογα με την περίπτωση).



Εικόνα 18: *Παράδειγμα τοπίου καταλληλότητας (fitness landscape)*

Σε μαθηματικά προβλήματα βελτιστοποίησης συνάρτησης –έστω f(x)- είναι προφανές ότι η συνάρτηση καταλληλότητας θα πρέπει να είναι η ίδια η συνάρτηση f. Συνεπώς σε κάθε υποψήφια λύση, δηλαδή σε κάθε πιθανή τιμή της μεταβλητής x, θα αντιστοιχεί μια τιμή καταλληλότητας που θα αξιολογεί την εκάστοτε πιθανή λύση και που στην περίπτωση αυτή θα ταυτίζεται με την ίδια την εικόνα της από τη συνάρτηση f.

Η συνάρτηση καταλληλότητας (fitness function) αναφέρεται στη διεθνή βιβλιογραφία και ως συνάρτηση προσαρμοστικότητας, συνάρτηση ικανότητας (ΣΙ), συνάρτηση αξιολόγησης (evaluation function), συνάρτηση ποιότητας, αντικειμενική συνάρτηση (objective function) και συνάρτηση κόστους (cost function). Παρά την αναφορά στη βιβλιογραφία της συνάρτησης καταλληλότητας και ως συνάρτησης κόστους και ως αντικειμενικής συνάρτησης, στη θεωρία οι τρεις έννοιες μπορούν να διακριθούν. Σε πολλά προβλήματα πρακτικά ταυτίζονται αλλά αυτό δεν είναι το σενάριο σε όλες τις περιπτώσεις.

Οι αντικειμενικές συναρτήσεις είναι οι συναρτήσεις εκείνες που διέπουν το πρόβλημα. Η συνάρτηση καταλληλότητας (fitness function) βασίζεται στις τιμές των αντικειμενικών συναρτήσεων (objective function) αποσκοπώντας στη δημιουργία ενός μοναδικού κριτηρίου που θα συνοψίζει το πόσο κοντά είναι μια υποψήφια λύση στην επίτευξη όλων των επιμέρους στόχων.

Όταν το πρόβλημα έχει μόνο ένα κριτήριο διέπεται από μια μόνο αντικειμενική συνάρτηση και η επιλογή συνάρτησης καταλληλότητας είναι εύκολη αφού θα είναι ίδια με την αντικειμενική συνάρτηση ή ένα μετασχηματισμό της. Στην περίπτωση όμως προβλημάτων βελτιστοποίησης με πολλαπλά κριτήρια απαιτείται η ταυτόχρονη βελτιστοποίηση δύο ή περισσότερων αντικρουόμενων ζητημάτων με διάφορους περιορισμούς. Στο πρόβλημα βελτιστοποίησης με πολλαπλά κριτήρια κάθε κριτήριο έχει τη δική του αντικειμενική συνάρτηση. Έτσι το ολικό πρόβλημα ανάγεται σε ένα συνδυασμό όλων των αντικειμενικών συναρτήσεων σε μια μόνο συναρτησιακή μορφή που θα αποτελεί και τη συνάρτηση καταλληλότητας. Εάν το πρόβλημα βελτιστοποίησης με πολλαπλά κριτήρια είναι καλά ορισμένο δεν θα υπάρχει μια μοναδική λύση η οποία θα βελτιστοποιεί κάθε υποστόχο. Σε αυτές τις περιπτώσεις υπάρχει δυσκολία στον καθορισμό μιας ενδεχόμενης λύσης ως ιδανικότερης από κάποιες άλλες, αφού μπορεί να είναι καλύτερη ως προς ένα κριτήριο και χειρότερη ως προς κάποιο άλλο. Πρέπει σε κάθε περίπτωση ένα κριτήριο να έχει φτάσει ένα σημείο τέτοιο ώστε κάθε επιπλέον προσπάθεια βελτιστοποίησής του να συνεπάγεται την υποβάθμιση άλλων κριτηρίων. Ο στόχος είναι η εύρεση μιας τέτοιας λύσης –ανάμεσα σε πολλές που μπορεί να υπάρχουν- η οποία να είναι καλύτερη με αυτή την έννοια σε σχέση με τις υπόλοιπες. Τέτοιου είδους προβλήματα είναι γνωστά ως προβλήματα πολλαπλών στόχων (multi-objective problems). Οι ΓΑ που επιλύουν προβλήματα αυτής της κατηγορίας καλούνται ΓΑ πολλαπλών στόχων (multi-objective GAs - MOGAs) και συνήθως εφαρμόζουν μαθηματικές θεωρίες βελτιστοποίησης πολλών κριτηρίων (π.χ. βελτιστοποίηση Pareto) για την αξιολόγηση των ατόμων.

Η συνάρτηση κόστους (cost/loss function) επιστρέφει μια τιμή που αναπαριστά κάποιο κόστος σε σχέση με το πρόβλημα. Σε ένα πρόβλημα βελτιστοποίησης το ζητούμενο είναι η ελαχιστοποίηση της συνάρτησης κόστους. Μια αντικειμενική συνάρτηση -ή ο συνδυασμός των αντικειμενικών συναρτήσεων - είναι είτε η συνάρτηση κόστους, είτε η αντίθετή της (με αρνητικό πρόσημο). Στην τελευταία περίπτωση η αντικειμενική συνάρτηση πρέπει να μεγιστοποιηθεί και αναφέρεται ως συνάρτηση οφέλους (utility function). Συχνά η συνάρτηση καταλληλότητας επιλέγεται με τέτοιο τρόπο ώστε να έχει μόνο θετικές τιμές (με τον κατάλληλο μετασχηματισμό). Η επιλογή αυτή γίνεται για διευκόλυνση της εφαρμογής των τελεστών επιλογής και αναπαραγωγής οι οποίοι στηρίζονται στη συνάρτηση καταλληλότητας. Ωστόσο, η συμπεριφορά του ΓΑ είναι ανεξάρτητη από το αν η συνάρτηση καταλληλότητας έχει μόνο θετικές τιμές.

Η μέτρηση καταλληλότητας είναι η κινητήρια δύναμη του ΓΑ. Ανάλογα με το πρόβλημα πρέπει να γίνεται και η σχεδίαση της πιο ικανοποιητικής συνάρτησης καταλληλότητας. Ένα φαινομενικά δυσεπίλυτο πρόβλημα καθίσταται προσπελάσιμο από ένα ΓΑ με τη δημιουργία της κατάλληλης συνάρτησης καταλληλότητας. Ο σχεδιασμός κατάλληλων συναρτήσεων καταλληλότητας, ακόμα και για κλάσεις όμοιων προβλημάτων, μπορεί να είναι δύσβατος και για το λόγο αυτό συχνά απαιτείται εξειδικευμένη γνώση για το πρόβλημα.

Η αξιολόγηση καθώς και τα υπόλοιπα στάδια του ΓΑ αποκαλύπτουν τη μεγάλη χρησιμότητα των ΓΑ στην επίλυση δυσνόητων προβλημάτων. Ο προγραμματιστής αρκεί να σχεδιάσει την αναπαράσταση των ενδεχόμενων λύσεων και τη συνάρτηση αξιολόγησής τους χωρίς να χρειάζεται να έχει περαιτέρω γνώση των εσωτερικών διεργασιών του προβλήματος. Τα υπόλοιπα τα αναλαμβάνει ο ΓΑ.

1.5 Επιλογή (Selection)

Με την επιλογή επιλέγονται τα πιο κατάλληλα άτομα από τον τρέχων πληθυσμό για το επόμενο στάδιο, το στάδιο της αναπαραγωγής (reproduction). Τα επιλεγμένα άτομα αναφέρονται ως «γονείς» και θα χρησιμοποιηθούν στην αναπαραγωγή για τον σχηματισμό απογόνων (offspring).

Η επιλογή είναι κομβικής σημασίας στην επιτυχή περάτωση του γενετικού αλγορίθμου, αποσκοπώντας παράλληλα και στην επιλογή καλύτερων χρωμοσωμάτων, αλλά και στην επιλογή χρωμοσωμάτων που βρίσκονται πλησιέστερα στη βέλτιστη λύση.

Η πιο κοινή μέθοδος επιλογής σε ένα ΓΑ είναι η αναλογική επιλογή (fitness-proportionate selection), στην οποία ο αριθμός των φορών που ένα χρωμόσωμα αναμένεται να αναπαράγει είναι ίσος με την τιμή καταλληλότητάς του προς τον μέσο όρο των τιμών καταλληλότητας του πληθυσμού. Αυτή η πρακτική είναι ισοδύναμη σε αυτή που αναφέρουν οι βιολόγοι ως «επιλογή βιωσιμότητας» (“viability selection”) [3].

Υπάρχουν πολλές διαθέσιμες μέθοδοι για την επιλογή ατόμων/χρωματοσωμάτων. Ανάλογα με τις παραμέτρους και τις απαιτήσεις του προβλήματος μπορεί να χρησιμοποιηθεί η καταλληλότερη εκ των διαφόρων αυτών τεχνικών ή κάποια παραλλαγή τους. Παρακάτω θα αναφερθούν οι πιο συνήθεις μέθοδοι επιλογής.

1.5.1 Αναλογική επιλογή / Επιλογή ρουλέτας (Roulette wheel selection)

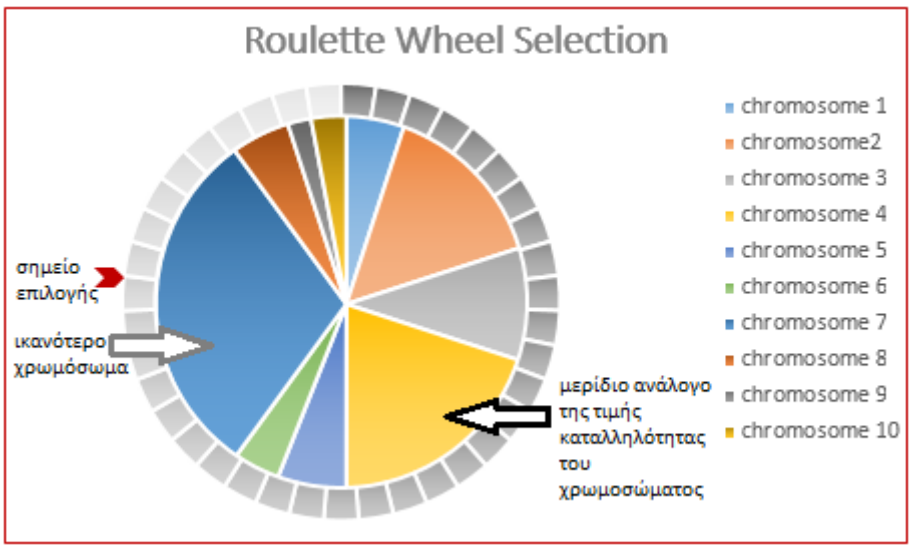
Η επιλογή ρουλέτα, αποτελεί την πιο ευρέως χρησιμοποιούμενη στοχαστική μέθοδο επιλογής και είναι μια απλή μέθοδος υλοποίησης της αναλογικής επιλογής (fitness-proportionate selection) [5]. Τα άτομα επιλέγονται με πιθανότητα επιλογής σύμφωνα με την παρακάτω εξίσωση.

$$p_{selection} = \frac{f(parent_i)}{\sum_i f(parent_i)}$$

όπου f_i είναι η τιμή καταλληλότητας (fitness value) του ατόμου i του πληθυσμού.

Τα χρωμοσώματα επιλέγονται για γονείς ανάλογα με την ικανότητά τους: όσο πιο κατάλληλο είναι ένα χρωμόσωμα τόσες περισσότερες πιθανότητες έχει να επιλεγεί. Η ικανότητα ενός χρωμοσώματος καθορίζεται από την συνάρτηση καταλληλότητας.

Ας υποθέσουμε μια ρουλέτα όπου τοποθετούνται όλα τα χρωμοσώματα του πληθυσμού, κάθε ένα εκ των οποίων παίρνει ένα κομμάτι της ρουλέτας μεγέθους ανάλογου με την τιμή καταλληλότητάς του. Ακολούθως, ρίχνεται μια μπίλια και επιλέγεται ένα χρωμόσωμα το οποίο εντάσσεται στο σύνολο των υποψήφιων γονέων. Η ρουλέτα γυρίζει τόσες φορές όσες ο αριθμός των χρωμοσωμάτων του πληθυσμού. Το χρωμόσωμα με την μεγαλύτερη ικανότητα θα επιλεχθεί περισσότερες φορές. Το νοητικό αυτό σχήμα παρουσιάζεται στην (*εικόνα 6)*.



Εικόνα 19: *Νοητικό σχήμα της επιλογής με ρουλέτα*

Η επιλογή με ρουλέτα μπορεί να υλοποιηθεί με τον ακόλουθο αλγόριθμο:

1. Υπολογισμός του αθροίσματος των τιμών καταλληλότητας όλων των χρωμοσωμάτων του πληθυσμού (έστω άθροισμα S).
2. Επιλογή τυχαίου αριθμού στο διάστημα [0,S], έστω r.
3. Αρχικοποίηση μεταβλητής d = 0 και σάρωση του πληθυσμού προσθέτοντας κάθε φορά την τιμή καταλληλότητας του τρέχοντος χρωμοσώματος στη μεταβλητή d. Όταν το άθροισμα d γίνει μεγαλύτερο του r, σταμάτα και επέστρεψε το τρέχον χρωμόσωμα.
4. Επανάληψη της διαδικασίας από το 2^{ο} έως το 4^{ο} βήμα N φορές (όπου N ο πληθάριθμος του συνόλου του πληθυσμού) .

1.5.2 Επιλογή βαθμονόμησης/ταξινόμησης/κατάταξης (Rank selection)

Τα χρωμοσώματα του πληθυσμού ταξινομούνται βάσει της τιμής της συνάρτησης κόστους και ακολούθως η επιλογή πραγματοποιείται ανάλογα με την μέγιστη αναμενόμενη τιμή του καλύτερου χρωμοσώματος κάθε γενεάς. Η αναμενόμενη τιμή (expected value) προκύπτει από τον τύπο:

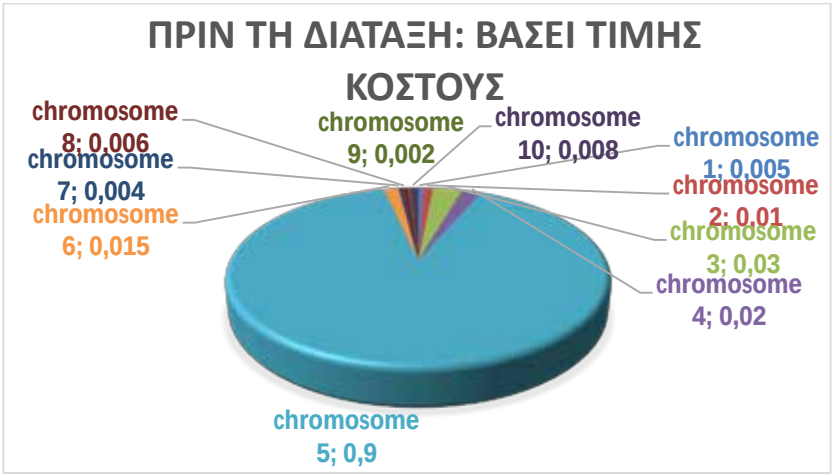
$$\mathrm{ExpVal(i,t)} = Min + (Max - Min)\frac{rank(i,t) - 1}{N - 1}$$

όπου rank(i,t) είναι η συνάρτηση που υλοποιεί την βαθμονόμηση (επιστρέφει την τιμή του χρωμοσώματος i στο χρόνο t) και πιθανόν να ακολουθεί γραμμική ή εκθετική κατανομή.

Η επιλογή βαθμονόμησης, ουσιαστικά χωρίζεται σε δύο βήματα. Στο πρώτο βήμα πραγματοποιείται η ταξινόμηση των χρωμοσωμάτων βάσει της τιμής της συνάρτησης κόστους κατά φθίνουσα ή αύξουσα σειρά ανάλογα με το αν πρόκειται για πρόβλημα μεγιστοποίησης ή ελαχιστοποίησης κι ακολούθως σε κάθε χρωμόσωμα αντιστοιχείται η τιμή της θέσης του (π.χ. το λιγότερο κατάλληλο θα έχει ικανότητα 1 , το επόμενο λιγότερο κατάλληλο τιμή 2 κοκ έως το πιο κατάλληλο που θα έχει ικανότητα N ίση με το πλήθος των χρωματοσωμάτων του πληθυσμού). Στο δεύτερο βήμα γίνεται η επιλογή του χρωμοσώματος με πιθανότητα ανάλογη της θέσης του στην ταξινομημένη λίστα χρωμοσωμάτων.

Η διαφορά αυτής της τεχνικής με την επιλογή με ρουλέτα είναι ότι στη βαθμονόμηση η επιλογή γίνεται με βάση τη θέση των χρωμοσωμάτων στην ταξινομημένη λίστα κι όχι με βάση την τιμή καταλληλότητάς τους: όσο πιο ψηλή η θέση ενός χρωμοσώματος τόσο πιο μεγάλη η πιθανότητα να επιλεγεί. Η επιλογή στη βαθμονόμηση γίνεται ανεξάρτητα από την τάξη και το μέγεθος των τιμών καταλληλότητας των χρωματοσωμάτων.

Με την επιλογή με ρουλέτα αν οι τιμές της ικανότητας διαφέρουν πολύ προκαλούνται προβλήματα. Αν για παράδειγμα η καλύτερη τιμή καταλληλότητας των χρωμοσωμάτων της ρουλέτας είναι 0.90, τότε τα άλλα χρωμοσώματα θα έχουν πολύ μικρές πιθανότητες να επιλεγούν αφού θα καταλαμβάνουν όλα μαζί μόνο το 0.10. Στην *(εικόνα 7)* παρουσιάζεται πώς επιλύνεται το πρόβλημα αυτό όταν διατάσσονται οι τιμές καταλληλότητας. Πριν τη διάταξη τα χρωμοσώματα είναι κατανεμημένα βάσει των τιμών καταλληλότητάς τους, ενώ μετά τη διάταξη είναι κατανεμημένα σύμφωνα με τη θέση κατάταξης των τιμών καταλληλότητάς τους. Με αυτή την ανακατανομή όλα τα χρωματοσώματα έχουν πιθανότητα να επιλεγούν. Η μέθοδος, όμως, της βαθμονόμησης, μπορεί να οδηγήσει σε πιο αργή σύγκλιση, επειδή με αυτή την κατανομή τα καλύτερα χρωμοσώματα δεν έχουν αρκετά μεγάλη διαφορά από τα υπόλοιπα.



(α)

(β)

Εικόνα 20: *Περίπτωση μεγάλης διαφοράς της τιμής καταλληλότητας (α) Πριν τη διάταξη (ranking) των χρωματοσωμάτων: κατανεμημένα με βάση την τιμή καταλληλότητάς τους (β) Μετά τη διάταξη των χρωματοσωμάτων: κατανομή βάσει της θέσης κατάταξης των τιμών καταλληλότητάς τους σε αύξουσα σειρά*

1.5.3 Επιλογή τουρνουά / πρωταθλημάτων / διαγωνισμών (Tournament selection)

Πραγματοποιούνται διάφορα «τουρνουά» ανάμεσα σε διαφορετικές ομάδες k τυχαία επιλεγμένων χρωμοσωμάτων του πληθυσμού, όπου k το μέγεθος του τουρνουά (tournament size). Τα k επιλεγμένα χρωμοσώματα κάθε ομάδας αναμετρώνται ανά δύο και το ικανότερο χρωμόσωμα συνεχίζει να συμμετέχει στο τουρνουά της ομάδας ενώ το λιγότερο ικανό απορρίπτεται. Οι αναμετρήσεις μεταξύ των χρωμοσωμάτων της ομάδας συνεχίζονται έως ότου να αναδειχθεί ένα χρωμόσωμα νικητής. Κατά την πραγματοποίηση ενός τουρνουά ανάμεσα στα k τυχαία επιλεγμένα χρωμοσώματα μιας ομάδας επιλέγεται ο νικητής-χρωμόσωμα ως υποψήφιος γονέας, ενώ τα υπόλοιπα χρωμοσώματα της ομάδας, καθώς και τα μη επιλεγμένα στην ομάδα χρωμοσώματα συμμετέχουν εκ νέου στη διαδικασία.

Στην επιλογή τουρνουά υπάρχουν δύο προσεγγίσεις, η στοχαστική και η ντετερμινιστική. Η διαφορά τους έγκειται στο κριτήριο που τίθεται στη φάση της ανά δύο αναμέτρησης των χρωμοσωμάτων ως προς το πιο χρωμόσωμα είναι ικανότερο. Στη στοχαστική προσέγγιση, τα k επιλεγμένα χρωμοσώματα κάθε ομάδας τοποθετούνται σε k/2 ζεύγη και τα μέλη κάθε ζεύγους αναμετρώνται με πιθανότητα να κερδίσει το ικανότερο ίση με p_i . Αντίθετα, στην ντετερμινιστική προσέγγιση ως πιο ικανό χρωμόσωμα εκ των δύο που αναμετρούνται θεωρείται αυτό με την μεγαλύτερη τιμή καταλληλότητας. Έτσι, στην μια περίπτωση η έκβαση της αναμέτρησης καθορίζεται πιθανοτικά ενώ στην άλλη ντετερμινιστικά. Στην ντετερμινιστική παραλλαγή ως νικητής της ομάδας κάθε «τουρνουά» ορίζεται αυτός με την καλύτερη τιμή καταλληλότητας οπότε εναλλακτικά αντί ανά ζεύγος αναμέτρηση εκ των k χρωμοσωμάτων της ομάδας μπορεί να χρησιμοποιηθεί ο κλασσικός γραμμικός αλγόριθμος εύρεσης μεγίστου για την ανάδειξη του νικητή ή να γίνει ταξινόμηση όλων των χρωμοσωμάτων της ομάδας ως προς την τιμή καταλληλότητας και το πιο κατάλληλο χρωμόσωμα να επιλεχθεί για αναπαραγωγή [14].

Γνωστές μέθοδοι υλοποίησης της επιλογής με τουρνουά αποτελούν οι:

- Boltzmann tournament (στοχαστική εκδοχή)
- Marriage tournament (ντετερμινιστική εκδοχή)

Στην επιλογή Boltzmann tournament αρχικοποιείται τυχαία μια υποψήφια λύση στη γειτονιά της τρέχουσας λύσης και επιλέγεται η νέα λύση με λογιστική πιθανότητα. Η διαδικασία επαναλαμβάνεται για κάποιο αριθμό δοκιμών.

Στην επιλογή Marriage tournament επιλέγεται τυχαία ένα χρωμόσωμα και ακολουθούν το μέγιστο k (ο αριθμός χρωμοσωμάτων μιας ομάδας τουρνουά) προσπάθειες έως ότου να βρεθεί κάποιο πιο κατάλληλο χρωμόσωμα. Στο πρώτο καταλληλότερο χρωμόσωμα που εντοπίζεται το «τουρνουά» της ομάδας αυτής σταματά. Νικητής κάθε τουρνουά είναι το πρώτο χρωμόσωμα που

είναι καταλληλότερο από το τυχαία επιλεγμένο χρωμόσωμα. Αν δεν εντοπιστεί χρωμόσωμα πιο κατάλληλο από το αρχικά τυχαία επιλεγμένο χρωμόσωμα νικητής θεωρείται η αρχική επιλογή.

Η επιλογή τουρνουά έχει κοινά χαρακτηριστικά με την επιλογή κατάταξης και μπορούν να θεωρηθούν ισοδύναμες λαμβάνοντας, όμως, υπόψη τους περιορισμούς και παραμέτρους του προβλήματος.

Εικόνα 21: *Παράδειγμα της επιλογής τουρνουά για μέγεθος τουρνουά k=14*

1.5.4 Αποδεκατισμός πληθυσμού (Population decimation)

Με την τεχνική του αποδεκατισμού πληθυσμού, τα χρωμοσώματα κατατάσσονται κατά φθίνουσα σειρά ως προς την τιμή της συνάρτησης καταλληλότητάς τους. Ακολούθως, επιλέγεται αυθαίρετα μια τιμή κατωφλίου. Τα χρωμοσώματα των οποίων η τιμή καταλληλότητας είναι μικρότερη από το συγκεκριμένο κατώτατο όριο απορρίπτονται. Τα υπόλοιπα χρωμοσώματα επιλέγονται για γονείς της επόμενης γενιάς.

Ο αποδεκατισμός πληθυσμού θεωρείται ντετερμινιστική τεχνική επειδή το κριτήριο επιλογής ή αποκλεισμού των χρωμοσωμάτων καθορίζεται από την ντετερμινιστική σύγκριση ανάμεσα στις τιμές της συνάρτησης κόστους και της αυθαίρετης τιμής κατωφλίου.

Ο αποδεκατισμός πληθυσμού πλεονεκτεί ως προς την απλότητα, αλλά εμπεριέχει τον κίνδυνο τα χαρακτηριστικά ενός χρωμοσώματος που έχει απομακρυνθεί από τον πληθυσμό να χαθούν εντελώς. Η απώλεια ποικιλότητας στον πληθυσμό παρατηρείται σε όλες τις τεχνικές επιλογής, αυτό, όμως, που το κάνει πιο έντονο στην προκειμένη περίπτωση, είναι ότι ξεκινά να συμβαίνει προτού ακόμα αναγνωριστεί η χρησιμότητα κάποιου χαρακτηριστικού που χάνεται. Ενώ η συνάρτηση κόστους μπορεί να είναι ενδεικτική του πόσο κατάλληλο είναι ένα χρωμόσωμα και τα χαρακτηριστικά του, δεν είναι απαραίτητα όλα τα καλά χαρακτηριστικά άμεσα συνδεδεμένα με την τιμή συνάρτησης κόστους στα πρώιμα στάδια του γενετικού.

Ο προαναφερθέντας κίνδυνος, δηλαδή η πρώιμη απώλεια κάποιων καλών χαρακτηριστικών, που συμβαίνει στις ντετερμινιστικές τεχνικές οδήγησε στη δημιουργία των στοχαστικών τεχνικών επιλογής.

1.5.5 Διαβάθμιση σίγμα (Sigmoid selection)

Στη διαβάθμιση σίγμα, η αναμενόμενη τιμή του κάθε χρωμοσώματος - πόσες δηλαδή φορές θα επιλεχθεί - εξαρτάται από τη στιγμιαία και τη μέση τιμή της συνάρτησης καταλληλότητας (f(t)), καθώς και την τυπική απόκλιση του πληθυσμού σε χρόνο t (σ(t)), σύμφωνα με τη σχέση:

$$ExpVal(i,t) = \begin{cases} 1 + \dfrac{f(i) - \overline{f(t)}}{2\sigma(t)} & \sigma(t) \neq 0 \\ 1 & \sigma(t) = 0 \end{cases}$$

Η διαβάθμιση σίγμα, όπως και αρκετές άλλες τεχνικές επιλογής, δημιουργήθηκαν έτσι ώστε να αποφευχθεί το φαινόμενο της πρώιμης σύγκλισης, ούτως ώστε δηλαδή, τα χρωμοσώματα με τις πιο υψηλές τιμές καταλληλότητας να μην μονοπωλούν την αναπαραγωγική διαδικασία. Με τη διαβάθμιση σίγμα διατηρείται σε σταθερά επίπεδα ο βαθμός συμμετοχής κατάλληλων χρωμοσωμάτων στην αναπαραγωγή καθ' όλη τη διάρκεια του αλγόριθμου, ανεξάρτητα από τη διασπορά των τιμών της συνάρτησης καταλληλότητας. Έτσι και τα λιγότερο κατάλληλα χρωμοσώματα έχουν πιθανότητα να επιλεγούν ως γονείς.

1.5.6 Τυχαία επιλογή (Random selection)

Ο τελεστής της τυχαίας επιλογής, επιλέγει εντελώς τυχαία τα χρωμοσώματα που θα χρησιμοποιηθούν για τη δημιουργία των απογόνων.

1.5.7 Επιλογή σταθερής κατάστασης (Steady-state selection)

Ο βασικός άξονας στον οποίο περιστρέφεται η επιλογή σταθερής κατάστασης είναι ότι μεγάλο μέρος των χρωμοσωμάτων πρέπει να επιζήσει στην επόμενη γενεά. Για να είναι αυτό εφικτό, αντικαθίστανται λίγα άτομα από κάθε γενιά, αλλά, όχι όλα- σε αντίθεση με τις πλείστες τεχνικές.

Η μέθοδος αυτή λειτουργεί με τον ακόλουθο τρόπο. Οι απόγονοι κάθε γενεάς ανήκουν στο σύνολο των υποψήφιων γονέων της επόμενης γενεάς. Από κάθε γενιά επιλέγονται ως γονείς τα χρωμοσώματα με τις σχετικά πιο μεγάλες τιμές καταλληλότητας και αυτά χρησιμοποιούνται στην αναπαραγωγή για την δημιουργία απογόνων. Τα χρωμοσώματα με σχετικά χαμηλή τιμή καταλληλότητας από κάθε γενιά, αντικαθίστανται από τους νέους απογόνους, τους υποψήφιους, δηλαδή, γονείς της επόμενης γενιάς. Το υπόλοιπο του πληθυσμού, δηλαδή, τα χρωμοσώματα που δεν έχουν αντικατασταθεί, επιζούν στην επόμενη γενιά και ανήκουν κι αυτά στο σύνολο των υποψήφιων γονέων της επόμενης γενιάς. Η διαδικασία επαναλαμβάνεται για τη δημιουργία των μεταγενέστερων απογόνων. Ουσιαστικά, τα λιγότερο κατάλληλα χρωμοσώματα αντικαθίστανται από τους απογόνους των ικανότερων χρωμοσωμάτων.

1.5.8 Ιεραρχική επιλογή (Hierarchical selection)

Τα χρωμοσώματα κάθε γενιάς περνάνε από πληθώρα σταδίων επιλογής. Τα αρχικά στάδια επιλογής έχουν λιγότερες απαιτήσεις και είναι ταχύτερα. Όσο πιο μεταγενέστερο είναι ένα στάδιο τόσο περισσότερες απαιτήσεις έχει.

Η τεχνική αυτή έχει το πλεονέκτημα ότι αποκλείει γρήγορα την πλειοψηφία των χρωμοσωμάτων που δεν έχουν πολλές προοπτικές. Τα εναπομείναντα χρωμοσώματα υποβάλλονται σε εκτενείς ελέγχους έως ότου βρεθεί η βέλτιστη επιλογή.

1.5.9 Ελιτισμός – Διατήρηση των ικανών (Elitism)

Ο ελιτισμός έχει προταθεί από τον Kenneth De Jong το 1975 και εξασφαλίζει ότι το καλύτερο /ένα ποσοστό από τα καλύτερα χρωμοσώματα της κάθε γενιάς θα μεταπηδήσουν στην επόμενη γενιά χωρίς να περάσουν από το στάδιο της αναπαραγωγής.

Το καλύτερο, μέχρι μια δεδομένη στιγμή του αλγορίθμου, χρωμόσωμα, (ή μερικά από τα καλύτερα μέχρι στιγμής χρωμοσώματα) μπορεί να μην έχει επιλεγεί να επιβιώσει στην επόμενη γενεά, οπότε κατά τη δημιουργία της νέας γενεάς υπάρχει μεγάλη πιθανότητα να χαθεί. Ωστόσο, το χρωμόσωμα αυτό μπορεί να είναι μια υψηλής ποιότητας λύση, ίσως ακόμα και η ολικά βέλτιστη λύση, και ως αποτέλεσμα να μην συμπεριληφθεί στον πληθυσμό. Για να αποφευχθεί η απώλεια αυτή, εκτός από μια φυσική μέθοδο επιλογής (pure selection), εφαρμόζεται και ελιτισμός. Σε κάθε γενιά τα χρωμοσώματα με την υψηλότερη τιμή καταλληλότητας αντιγράφονται κατευθείαν στην επόμενη γενιά αυτούσια, χωρίς καμιά τροποποίηση των γενετικών χαρακτηριστικών τους. Με αυτό τον τρόπο διασφαλίζεται ότι η τελική λύση θα είναι η ολικά βέλτιστη λύση, αυξάνοντας έτσι, την απόδοση του αλγορίθμου.

Αυτή η τεχνική είναι σημαντική επειδή εγγυάται ότι οι ενδεχόμενα καλές λύσεις μιας γενιάς δεν θα χαθούν ούτε θα αλλοιωθούν εξαιτίας προσμίξεων με άλλες λύσεις.

Συνήθως, η τιμή της παραμέτρου του ελιτισμού ανήκει στο διάστημα [0.05-0.20]. Για παράδειγμα, αν το μέγεθος του πληθυσμού μιας γενεάς χρωμοσωμάτων είναι N=200, τα 10-40 καλύτερα χρωμοσώματα από αυτά θα μεταφερθούν κατευθείαν στον πληθυσμό της επόμενης γενιάς.

Ο ελιτισμός παρουσιάζεται ιδιαίτερα χρήσιμος στις περιπτώσεις που η πιθανότητα των γενετικών τελεστών έχουν υψηλή τιμή. Σε αυτές τις περιπτώσεις, χρωμοσώματα που ανήκουν στο σύνολο των «υποψήφιων γονέων» , έχουν μεγάλη πιθανότητα να μην προχωρήσουν στην επόμενη γενιά και γι’ αυτό κρίνεται αναγκαίο να γίνει χρήση μιας σχετικά μεγάλης πιθανότητας ελιτισμού για τη διατήρηση των καλών λύσεων.

1.6 Αναπαραγωγή (Reproduction)

Στη φύση τα άτομα του πληθυσμού τα οποία έχουν καλύτερη ικανότητα επιβίωσης έχουν την τάση να αναπαράγονται με μεγαλύτερη συχνότητα από ότι τα υπόλοιπα άτομα του είδους. Οι οργανισμοί μέσω της αναπαραγωγής μεταβιβάζουν πολλά από τα χαρακτηριστικά τους στους απογόνους τους. Αυτά τα χαρακτηριστικά δεν μεταβιβάζονται πάντα ως πιστά αντίτυπα με αποτέλεσμα να υφίστανται παραλλαγές τους μέσα στον πληθυσμό. Οι παραλλαγές αυτές οδηγούν σε περαιτέρω διαφοροποίηση στην ικανότητα των ατόμων του πληθυσμού να ανταγωνίζονται και να επιβιώνουν.

Στους ΓΑ μετά το στάδιο της αρχικοποίησης, ακολουθεί το στάδιο της αναπαραγωγής. Στις περισσότερες υλοποιήσεις ΓΑ στο στάδιο της αναπαραγωγής χρησιμοποιούνται δύο τελεστές: ένας τελεστής ανασυνδυασμού (crossover), όπου δύο ή περισσότερα άτομα του πληθυσμού ανταλλάζουν γενετική πληροφορία και ένας τελεστής μετάλλαξης(mutation), όπου η γενετική πληροφορία ενός ατόμου μεταβάλλεται χωρίς να πραγματοποιείται ανταλλαγή γενετικής πληροφορίας. Σε κάποιες υλοποιήσεις ΓΑ ο επόμενος πληθυσμός υποψήφιων λύσεων παράγεται και με χρήση κάποιων άλλων τελεστών όπως αποίκηση εξάλειψη (colonization-extinction) και μετανάστευση (migration). Αυτοί οι τελεστές είναι σχεδιασμένοι έτσι ώστε οι ιδιότητες των γονέων να αναπαραχθούν στον απόγονό τους (offspring). Η παραγωγή καινούριου απογόνου από ένα νέο ζευγάρι επιλεγμένων γονέων κάθε φορά, συνεχίζεται έως ότου ο επιθυμητός πληθυσμός λύσεων επιτευχθεί.

1.6.1 Διασταύρωση (Crossover)

Η διασταύρωση (crossover) αναφέρεται στη βιβλιογραφία και ως ανασυνδυασμός (recombination).

Ορισμός. Διασταύρωση (crossover) είναι η διαδικασία συνδυασμού 2 ή περισσότερων λύσεων-γονέων (parent solutions) έτσι ώστε να παραχθεί μια λύση-παιδί (child solution) από αυτές.

Διαισθητικά, θα λέγαμε ότι η διασταύρωση εξυπηρετεί ανταλλαγή πληροφοριών ανάμεσα σε υποψήφιες λύσεις ή αλλιώς ανταλλαγή γενετικού υλικού ανάμεσα σε χρωμοσώματα αποσκοπώντας σε ενδεχόμενες νέες καλύτερες λύσεις.

Μια συνήθης διασταύρωση πραγματοποιείται σε 3 βασικά βήματα:

1. Ο τελεστής αναπαραγωγής επιλέγει τυχαία ένα ζεύγος ατόμων για αναπαραγωγή από το σύνολο των «υποψήφιων γονέων».

2. Ένα ή περισσότερα σημεία διασταύρωσης Σ.Δ. (crossover points) επιλέγονται τυχαία κατά μήκος των ατόμων.
3. Δημιουργούνται οι δύο απόγονοι. Ξεκινώντας από την πρώτη θέση και κινούμενοι κατά μήκος των ατόμων, ο πρώτος απόγονος αντιγράφει τα γονίδια του πρώτου γονέα, ενώ ο δεύτερος απόγονος τα γονίδια του δευτέρου γονέα. Όταν βρεθεί σημείο διασταύρωσης, η διαδικασία αντιστρέφεται: ο πρώτος απόγονος αντιγράφει τα γονίδια του δεύτερου γονέα και ο δεύτερος απόγονος τα γονίδια του πρώτου γονέα. Η ίδια διαδικασία ακολουθείται και για τα υπόλοιπα σημεία διασταύρωσης έως ότου φτάσει το τέλος των ατόμων. Έτσι προκύπτουν δύο νέα χρωμοσώματα που αποτελούν έναν ανασυνδυασμό των δύο γονέων.

Ο καθορισμός των διάφορων συνδυασμών γονέων από τα άτομα του προσωρινού πληθυσμού ίσως να επηρεάσει τη σύγκλιση του ΓΑ. Συνήθως, οι συνδυασμοί των γονέων γίνονται με τυχαίο τρόπο. Επιπλέον, κάθε άτομο του πληθυσμού επιλέγεται μια φορά ως γονέας κατά την εφαρμογή του τελεστή διασταύρωσης. Συνεπώς, στην περίπτωση της διασταύρωσης 2 γονιών με 2 απογόνους, το πλήθος των παιδιών που παράγονται είναι ίσο με το πλήθος των γονέων.

Ο τελεστής της διασταύρωσης, στον πυρήνα του μπορεί να θεωρηθεί στοχαστικός τελεστής, αφού η επιλογή των τμημάτων που θα διασταυρωθούν από κάθε γονέα, καθώς και ο τρόπος συνδυασμού των τμημάτων αυτών, είναι σε μεγάλο βαθμό στοχαστικός.

Η διασταύρωση είναι μια απαραίτητη λειτουργία που συμβάλλει καθοριστικά στην απόδοση του ΓΑ. Ανάλογα με τον τύπο του προβλήματος επιλέγεται η πιο κατάλληλη μέθοδος διασταύρωσης.

Στόχος της διασταύρωσης είναι η νέα γενιά που θα προκύψει μετά την εφαρμογή της, να αποτελείται από χρωμοσώματα που θα διαφέρουν από τους γονείς τους και θα φέρουν τον συνδυασμό των ικανότερων χαρακτηριστικών τους. Έτσι, μπορούν να προκύψουν επιτυχημένοι συνδυασμοί υψηλής ικανότητας. Ωστόσο, είναι πιθανόν, η διασταύρωση να παραγάγει απογόνους λιγότερο κατάλληλους από ότι ήταν οι γονείς τους, αλλά αυτοί δεν θα έχουν μεγάλη πιθανότητα αναπαραγωγής στον επόμενο κύκλο επειδή θα έχουν χαμηλή καταλληλότητα.

Η διασταύρωση είναι χρήσιμη και για την ανακατεύθυνση της αναζήτησης της βέλτιστης λύσης σε ανεξερεύνητες περιοχές του χώρου αναζήτησης. Με αυτό τον τρόπο διευρύνεται το εύρος του ΓΑ και αυξάνονται οι πιθανότητες του για διεκπεραίωση του στόχου του.

Η διασταύρωση λαμβάνει χώρα με μια πιθανότητα p_c , η οποία καλείται πιθανότητα διασταύρωσης (crossover probability) και είναι αναγκαίος ο προσδιορισμός της. Η πιθανότητα διασταύρωσης καθορίζει το ποσοστό των υποψήφιων γονέων στους οποίους θα εφαρμοστεί διασταύρωση και κατ' επέκταση και του ποσοστού του τελικού πληθυσμού που θα προέρχεται

από διασταύρωση. Η πιθανότητα διασταύρωσης ποικίλει ανάλογα με το πρόβλημα και είναι δυνατό να μεταβληθεί κατά τη διάρκεια εκτέλεσης του ΓΑ. Τιμές που έχουν προταθεί από ερευνητές είναι p_c=0.6 , p_c=0.95 , p_c=1 και p_c=[0.75,0.95]. Η πιθανότητα διασταύρωσης επηρεάζει το ρυθμό σύγκλισης και το χρόνο εκτέλεσης του ΓΑ. Αν είναι ίση με 1, τότε θα εφαρμόζεται συνεχώς η λειτουργία της διασταύρωσης, δηλαδή όλα τα ζεύγη του συνόλου των υποψήφιων γονέων κάθε πληθυσμού θα υπόκεινται διασταύρωση κι έτσι η αναζήτηση θα γίνει σε όλο το χώρο αναζήτησης με αποτέλεσμα τη σύγκλιση του ΓΑ στη βέλτιστη λύση αλλά με χρονοτριβή. Αν έχει μικρές τιμές, η αναζήτηση θα γίνεται με μεγάλο βήμα κι έτσι ο αλγόριθμος είναι πιθανότερο να συγκλίνει πιο σύντομα. Αν όμως το βήμα είναι πολύ μεγάλο είναι πιθανό ο ΓΑ να προσπεράσει τη βέλτιστη λύση και να ξεκινήσει να αποκλίνει. Η συνήθης τακτική που ακολουθείται είναι η επιλογή μεγάλου βήματος στην αρχή του ΓΑ και όταν ο αλγόριθμος αρχίσει να πλησιάζει τη βέλτιστη λύση τότε επανακαθορίζεται η πιθανότητα διασταύρωσης επιβάλλοντας μικρό βήμα αναζήτησης. Με αυτή την τακτική, μειώνεται ο κίνδυνος απόκλισης του ΓΑ και είναι δυνατό να αυξηθεί η ταχύτητα σύγκλισης.

Πίνακας 2: Πιθανότητα διασταύρωσης

Πιθανότητα Διασταύρωσης p_c	**Ερμηνεία**
0	Όλοι οι απόγονοι θα είναι ακριβή αντίγραφα των γονέων
1	Όλοι οι απόγονοι θα αποτελούνται από γενετικό συνδυασμό των γονέων
0.7	Οι απόγονοι θα προέρχονται από το 70% του συνόλου των «υποψήφιων γονέων»

Οι περισσότεροι τελεστές διασταύρωσης εφαρμόζονται σε δύο γονείς και παράγουν ένα, δύο ή περισσότερους απογόνους. Υπάρχουν όμως, και τελεστές με τρεις γονείς ή περισσότερους. Αν και αυτοί οι τελεστές είναι μαθηματικώς εφικτοί και εύκολα υλοποιήσιμοι, εξαιτίας μάλλον του γεγονότος ότι δεν παρατηρείται κάτι ανάλογο στη φύση, δεν είναι ιδιαίτερα διαδεδομένοι παρά την καλή τους απόδοση.

Οι κλασικοί τελεστές δύο γονέων χρησιμοποιούν μια δυαδική μάσκα διασταύρωσης της οποίας κάθε μπιτ καθορίζει ποιος γονέας θα αντιγραφεί από το παιδί. Για την παραγωγή δύο απογόνων, με χρήση της αρχικής μορφής της μάσκας παράγεται ο πρώτος απόγονος και με τη χρήση της αντίστροφης μορφής της μάσκας διασταύρωσης, όπου κάθε μπιτ της αρχικής μάσκας αντιστρέφεται (11100 έχει αντίστροφη την 00011) παράγεται κι ο δεύτερος απόγονος. Ένα σημείο διασταύρωσης στη μάσκα καθορίζει τη θέση από την οποία γενετική πληροφορία θα ξεκινήσει να συνεισφέρεται από τον άλλο γονέα. Η μάσκα διασταύρωσης για τρεις γονείς είναι μια ακολουθία

της οποίας κάθε μπιτ μπορεί να είναι ένα εκ των {0,1,2}, όπου 1 σημαίνει αντιγραφή του μπιτ στο παιδί από τον πρώτο γονέα , 2 από το δεύτερο γονέα και 0 από τον τρίτο γονέα.

Το σύνολο των απογόνων (offspring set) μετά τον τελεστή διασταύρωσης συνήθως δεν αποτελεί τον τελικό πληθυσμό της επόμενης γενεάς αφού ακόμα πάνω σε αυτόν μάλλον θα εφαρμοστούν κι άλλοι τελεστές. Όταν εφαρμοστούν όλοι οι τελεστές, το σύνολο απογόνων θα πάρει την τελική του μορφή και θα αντικαταστήσει τον πληθυσμό της τρέχουσας γενιάς.

Οι προτεινόμενες μέθοδοι διασταύρωσης αυξάνονται καθώς διάφοροι ερευνητές εισάγουν νέους τρόπους αναπαράστασης των προβλημάτων που επιλύονται με ΓΑ. Ακολούθως, παρατίθενται ενδεικτικά οι βασικές μέθοδοι διασταύρωσης (crossover).

- **Διασταύρωση ενός σημείου (One/Single Point crossover)**

Η διασταύρωση ενός σημείου θεωρείται η απλούστερη μορφή του τελεστή διασταύρωσης και έχει προταθεί από τον Holland [15]. Από κάθε ζεύγος γονέων παράγεται ένα νέο ζεύγος απογόνων.

Επιλέγεται ένα μοναδικό κοινό σημείο διασταύρωσης και στους δύο γονείς-χρωμοσώματα, με στόχο οι απόγονοι να προκύπτουν εάν σε κάθε γονέα, το τμήμα όλων των γονιδίων πέρα από αυτό το σημείο, αντιμετατεθεί με το αντίστοιχο τμήμα του άλλου. Πιο συγκεκριμένα, το αρχικό τμήμα του πρώτου απογόνου προκύπτει από την αντιγραφή της ακολουθίας του πρώτου γονέα από το σημείο έναρξης έως το σημείο διασταύρωσης, και το υπόλοιπό του τμήμα από την αντιγραφή της ακολουθίας του δεύτερου γονέα από το σημείο διασταύρωσης μέχρι το τέλος της ακολουθίας. Αντίστοιχα δημιουργείται και ο δεύτερος απόγονος παίρνοντας το αρχικό του τμήμα από το δεύτερο γονέα και το υπόλοιπο από τον πρώτο γονέα.

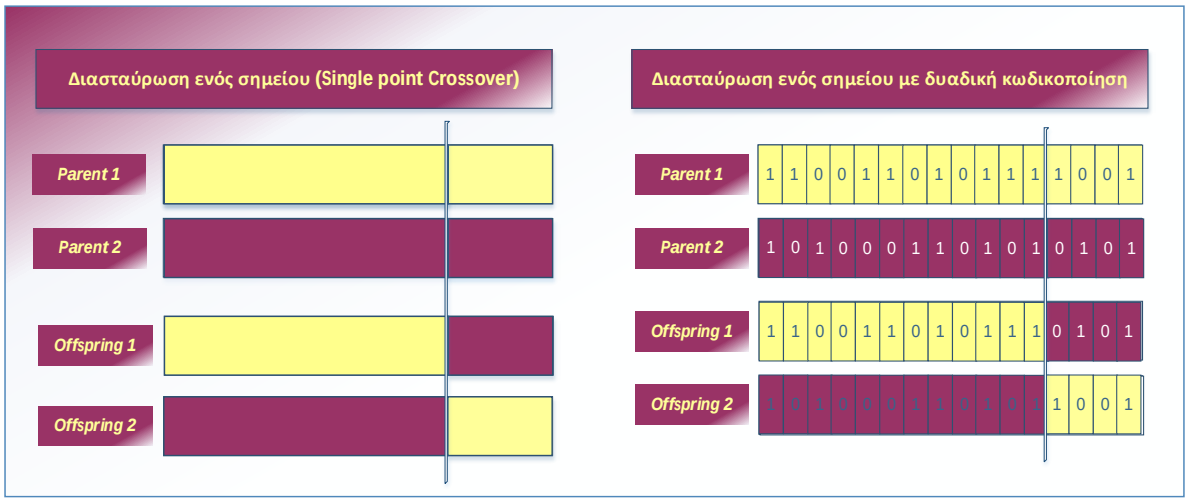
Η μάσκα διασταύρωσης ενός σημείου, θα είναι μια ακολουθία που αποτελείται από ένα πρώτο τμήμα με διαδοχικά 1 έως το σημείο διασταύρωσης και από το σημείο διασταύρωσης και μετά, από ένα δεύτερο τμήμα με διαδοχικά 0 (για παράδειγμα 111111000).

Το τυχαία επιλεγμένο, κοινό και για τους δύο γονείς, σημείο διαχωρισμού, τους χωρίζει σε δύο τμήματα. Αυτή η μέθοδος δίνει περισσότερες πιθανότητες σε γειτονικά γονίδια από κάθε άτομο-γονέα να κληρονομηθούν στον ίδιο απόγονο από ότι σε γονίδια με μια απόσταση μεταξύ τους. Συνεπώς, δημιουργείται ανεπιθύμητη συσχέτιση ανάμεσα σε γειτονικές μεταβλητές, με αποτέλεσμα να παίζει ρόλο η σειρά με την οποία οι μεταβλητές απόφασης αναπαραστάθηκαν στο χρωμόσωμα. Ειδικότερα, τα γονίδια που βρίσκονται προς το τέλος της ακολουθίας των γονέων έχουν μεγαλύτερη πιθανότητα ανταλλαγής -αγγίζει τη μονάδα- από ότι τα γονίδια που βρίσκονται στο αρχικό τμήμα των γονέων. Έτσι τα χαρακτηριστικά που βρίσκονται στην αρχή και το τέλος του ατόμου-γονέα δεν έχουν τη δυνατότητα να συνυπάρξουν στον ίδιο απόγονό του. Σε

περίπτωση που και τα δύο άκρα ενός χρωμοσώματος-γονέα περιέχουν καλό γενετικό υλικό θα παρουσιαστεί πρόβλημα, αφού κανένας απόγονός του δεν θα έχει τα καλά χαρακτηριστικά και των δύο άκρων [16]. Το φαινόμενο αυτό καλείται από τους Eshelman, Caruana και Schaffer ως προκατάληψη βάσει θέσης / πόλωση θέσεως (positional bias) και αποτελεί το κύριο μειονέκτημα της διασταύρωσης ενός σημείου, καθώς περιορίζει την αποτελεσματικότητα του ΓΑ μη μπορώντας να παραγάγει όλα τα πιθανά σχήματα (τα σχήματα περιγράφονται αναλυτικά σε επόμενο κεφάλαιο) [15].

Όσο αυξάνονται τα σημεία διασταύρωσης τόσο η επίδραση της πόλωσης θέσεως μειώνεται. Για αποφυγή των αρνητικών συνεπειών της πόλωσης θέσης, οι εξαρτώμενες μεταβλητές του προβλήματος μπορούν να τοποθετηθούν σε κοντινές θέσεις στην αναπαράσταση του χρωμοσώματος.

Παρ' όλ' αυτά, σε προβλήματα παραμετρικής βελτιστοποίησης, ο τελεστής διασταύρωσης ενός σημείου φαίνεται να είναι ανώτερος από τους υπόλοιπους τελεστές διασταύρωσης ως προς την απόδοση του ΓΑ [15].



(α) (β)

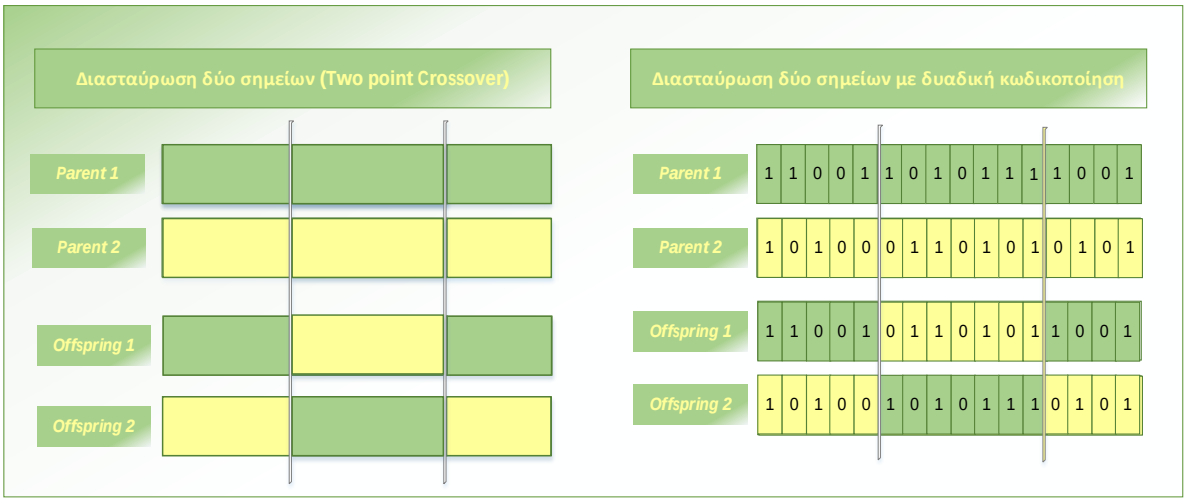
Εικόνα 22: *(α)Παράδειγμα διασταύρωσης σημείου με οποιαδήποτε κωδικοποίηση (β) Παράδειγμα διασταύρωσης σημείου με δυαδική κωδικοποίηση*

- **Διασταύρωση δύο σημείων (Two point crossover)**

Στη διασταύρωση δύο σημείων επιλέγονται δύο σημεία διασταύρωσης στους δύο γονείς-χρωμοσώματα. Τα γονίδια αριστερά και δεξιά των σημείων διασταύρωσης αντιγράφονται από τους γονείς στους απογόνους τους όπως έχουν, ενώ όλα τα γονίδια μεταξύ των δύο σημείων διασταύρωσης στα χρωμοσώματα-γονείς ανταλλάσσονται μεταξύ τους για τη δημιουργία των απογόνων. Ένα παράδειγμα μάσκας διασταύρωσης δύο σημείων είναι 111000011.

Το πρόβλημα της πόλωσης θέσης που παρατηρείται στη διασταύρωση ενός σημείου, το υπερβαίνει σε ένα βαθμό η διασταύρωση δύο σημείων αφού αντί για ένα, υπάρχουν δύο σημεία διαχωρισμού, τα οποία διαιρούν το χρωμόσωμα σε τρία τμήματα. Ο τελεστής αυτός ονομάζεται και τελεστής δακτυλίου, αφού αν θεωρήσουμε ότι το τέλος ενώνεται με την αρχή του ατόμου, τότε το κάθε άτομο διαιρείται σε δύο τμήματα, χωρίς όμως να υπάρχει κάποιο σημείο που να καθορίζει την αρχή. Με τη διασταύρωση δύο σημείων παράγονται περισσότερα σχήματα από ότι με τη διασταύρωση ενός σημείου αλλά και πάλι δεν παράγονται όλα τα πιθανά σχήματα (σχήματα αναλύονται σε επόμενο κεφάλαιο) [3]. Επειδή ακριβώς η διασταύρωση δύο σημείων είναι λιγότερο πιθανό να παρουσιάσει το πρόβλημα της πόλωσης θέσης, γενικά θεωρείται καλύτερη μέθοδος από τη διασταύρωση ενός σημείου [16].

Η προσθήκη επιπρόσθετων Σ.Δ., μπορεί να μειώσει την απόδοση του ΓΑ επειδή με την προσθήκη Σ.Δ. τα building blocks (εξηγούνται σε επόμενο κεφάλαιο) είναι πιο πιθανό να διαταραχθούν. Τα επιπρόσθετα βέβαια Σ.Δ. έχουν το πλεονέκτημα ότι καταστούν την εξερεύνηση του χώρου αναζήτησης πιο εξονυχιστική [16].



(α) (β)

Εικόνα 23: *(α) Παράδειγμα διασταύρωσης δύο σημείων με οποιαδήποτε κωδικοποίηση (β) Παράδειγμα διασταύρωσης δύο σημείων με δυαδική κωδικοποίηση*

- **Ομοιόμορφη Διασταύρωση (Uniform Crossover)**

Η γνωστή στη βιβλιογραφία ομοιόμορφη διασταύρωση, έχει προταθεί το 1989 από τον Syswerda. Η ομοιόμορφη διασταύρωση (UX) είναι από τους βασικούς τελεστές διασταύρωσης που χρησιμοποιούνται στους ΓΑ και συνδυάζεται συχνά με τη δυαδική αναπαράσταση των

χρωμοσωμάτων. Στην ομοιόμορφη διασταύρωση τα bit της ακολουθίας ενός ατόμου(παιδιού) αντιγράφονται κατά τρόπο τυχαίο είτε από τον πρώτο είτε από το δεύτερο γονιό. Η ομοιόμορφη διασταύρωση παράγει δύο απογόνους από κάθε ζεύγος γονέων με κάθε γονίδιο του καθενός παιδιού να είναι τυχαία επιλεγμένο από τα αντίστοιχα γονίδια των γονιών του.

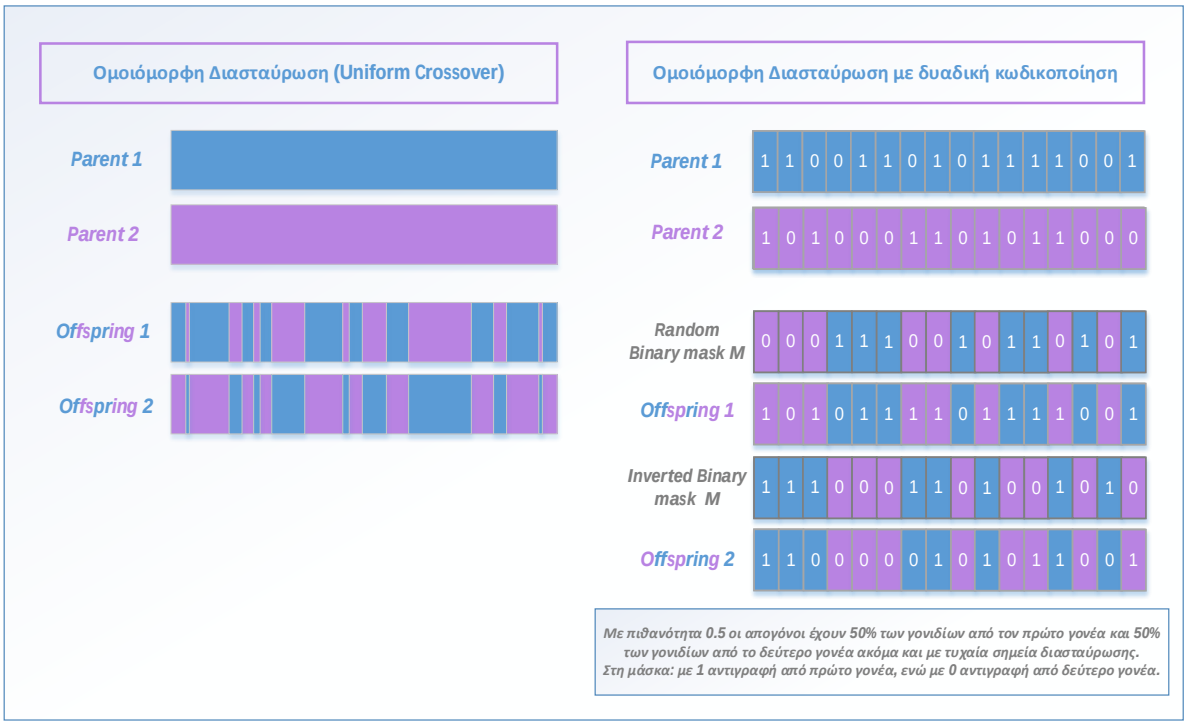
Η UX χρησιμοποιεί μια τυχαία παραχθείσα μάσκα δυαδικών bits (template/mask) μήκους ίσου με το μήκος των γονέων-χρωμοσωμάτων που προσδιορίζει ποια γονίδια του ζεύγους θα διασταυρωθούν. Για κάθε καινούριο ζεύγος γονέων του ενδεχόμενου πληθυσμού χρησιμοποιείται μια νέα τυχαία παραχθείσα μάσκα. Ανάλογα με την τιμή του δυαδικού ψηφίου της κάθε θέσης της μάσκας καθορίζεται από ποιο γονέα θα προέρχεται η γενετική πληροφορία για τη συγκεκριμένη θέση στο παιδί. Αν π.χ. το δυαδικό ψηφίο σε μια θέση της μάσκας είναι 1, τότε το παιδί παίρνει την τιμή του πρώτου γονέα για τη θέση αυτή. Αν είναι 0, παίρνει την τιμή του άλλου γονέα για το συγκεκριμένο γονίδιο. Για το άλλο παιδί είτε θα χρησιμοποιηθεί η αντίστροφη (inverted) μορφή της δυαδικής μάσκας βάσει των προαναφερθέντων, είτε θα ισχύει το αντίστροφο: όπου υπάρχει η τιμή 0 στη μάσκα, το γονίδιο αντιγράφεται από τον πρώτο γονέα, ενώ όπου υπάρχει η τιμή 1, το γονίδιο αντιγράφεται από τον δεύτερο γονέα.

Η μάσκα διασταύρωσης αποκτά τιμή 1 σε μία θέση βάσει μιας δοσμένης πιθανότητας, της λεγόμενης πιθανότητας ομοιόμορφης διασταύρωσης/πιθανότητας ανταλλαγής (uniform crossover probability/exchange probability). Η πιθανότητα ομοιόμορφης διασταύρωσης αντιστοιχεί στο ποσοστό μίξης των δύο γονέων και το παιδί προκύπτει από την τυχαία αντιγραφή των δυαδικών μπιτ από τον πρώτο ή το δεύτερο γονέα σύμφωνα με αυτό το ποσοστό. Με ποσοστό μίξης 0.5, που είναι και η πιο συνήθης επιλογή, οι απόγονοι έχουν 50% των γονιδίων από τον ένα γονέα και 50% των γονιδίων από το δεύτερο γονέα. Η πιθανότητα ανταλλαγής της ομοιόμορφης διασταύρωσης είναι μικρότερη της μονάδας επειδή διαφορετικά οι απόγονοι που θα προέκυπταν θα ήταν ίδιοι με τους γονείς τους. Η πιθανότητα ομοιόμορφης διασταύρωσης είναι συνήθως μικρότερη από αυτή που χρησιμοποιείται σε άλλες μεθόδους για αποφυγή της δραματικής μεταβολής των δύο γονέων.

Ουσιαστικά, η ομοιόμορφη διασταύρωση αποτελεί μια μέθοδο διασταύρωσης με πολλά τυχαία σημεία διασταύρωσης. Ανάλογα με την τυχαία παραχθείσα μάσκα για την διασταύρωση ενός ζεύγους γονέων μπορεί να προκύψουν από 1 έως (L – 1) σημεία διασταύρωσης σε ένα συγκεκριμένο ζεύγος, όπου L ο αριθμός των γονιδίων (ή αλλιώς το μήκος του χρωμοσώματος). Αν για παράδειγμα η μάσκα είναι 11111110000, τότε θα υπάρξει ένα σημείο διασταύρωσης, αν είναι 1111000001, 2 σημεία διασταύρωσης, 110010101001 8 σημεία διασταύρωσης ή εάν είναι 1010101010 9 σημεία διασταύρωσης. Ο αριθμός των σημείων διασταύρωσης δεν είναι σταθερός, αλλά συνήθως, υπάρχουν κατά μέσο όρο L/2 σημεία διασταύρωσης για ένα χρωμόσωμα μήκους L [17].

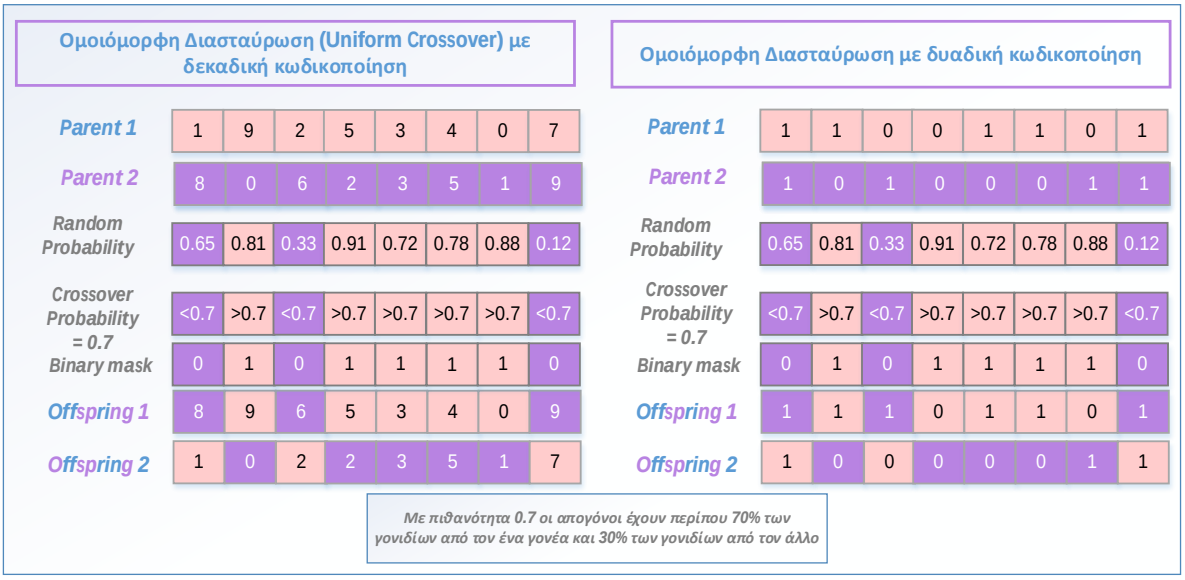
Η ομοιόμορφη διασταύρωση πάσχει από το πρόβλημα της πόλωσης κατανομής (distributional bias). Ο μέσος αριθμός των γονιδίων που διασταυρώνονται μεταξύ των γονέων συγκεντρώνεται σε μια περιοχή τιμών αντί να είναι ομοιόμορφα κατανεμημένος στο διάστημα τιμών 0 .. (L-1), όπως είναι το ζητούμενο. Πιο συγκεκριμένα ο αναμενόμενος αριθμός διασταυρώσεων είναι p*L, όπου p η πιθανότητα ανταλλαγής δύο γονιδίων σε οποιαδήποτε θέση.

Η ομοιόμορφη διασταύρωση είναι πλήρως απαλλαγμένη από οποιοδήποτε πόλωση θέσης (positional bias) επειδή κάθε γονέας χωρίζεται σε τόσα τμήματα όσα και ο αριθμός των γονιδίων που τον αποτελούν κι έτσι κάθε γονίδιο του ατόμου-γονέα μπορεί να εναλλαχθεί στους απογόνους του ανεξάρτητα από τα γειτονικά του γονίδια (η πόλωση θέσης περιγράφεται στη διασταύρωση ενός σημείου). Στην ομοιόμορφη διασταύρωση, οποιοδήποτε σχήμα συμπεριλαμβάνεται σε διαφορετικές θέσεις στους γονείς είναι πιθανό να ανασυνδυαστεί στους απογόνους. Βέβαια, αυτή η έλλειψη πόλωσης θέσης μπορεί να αποτρέψει ομάδες γονιδίων που λειτουργούν καλά μαζί από το να σχηματιστούν στον πληθυσμό, μια και η ομοιόμορφη διασταύρωση μπορεί να είναι αρκετά διασπαστική σε οποιοδήποτε σχήμα [3].



(α) (β)

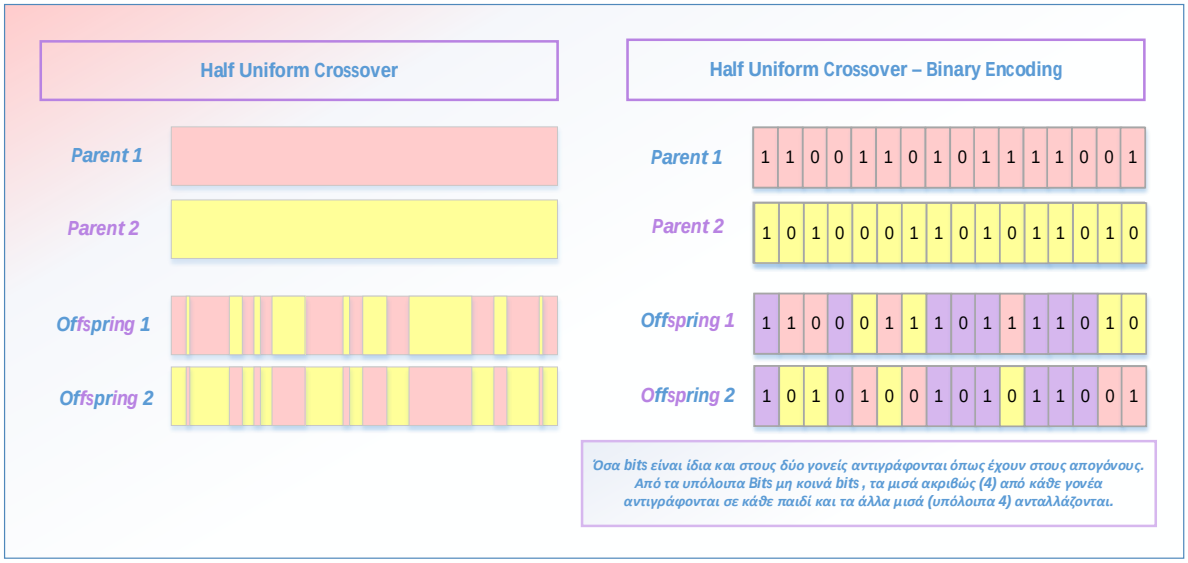
Εικόνα 24: *(α) Παράδειγμα ομοιόμορφης διασταύρωσης με 0.5 πιθανότητα διασταύρωσης (β) Παράδειγμα ομοιόμορφης διασταύρωσης με δυαδική κωδικοποίηση*



Εικόνα 25: *Παράδειγμα ομοιόμορφης διασταύρωσης με 0.7 πιθανότητα διασταύρωσης*

Μια παραλλαγή της ομοιόμορφης διασταύρωσης που χρησιμοποιήθηκε από τον Eshelman το 1990, είναι η half uniform crossover (HUX). Η μόνη διαφορά από την ομοιόμορφη διασταύρωση είναι ότι μόνο τα μισά από τα διαφορετικά bits των γονέων ανταλλάσσονται.

Στην HUX τα bits που είναι κοινά και στους δύο γονείς αντιγράφονται όπως έχουν και στους απογόνους τους, ενώ ακριβώς τα μισά από τα bits που δεν είναι κοινά ανταλλάσσονται. Επομένως, αρχικά υπολογίζεται η απόσταση Hamming (που είναι ίση με τον αριθμό των διαφορετικών bits των 2 γονιών). Το ½ της απόστασης Hamming υποδεικνύει πόσα bits από τα μη κοινά bits των δύο γονέων θα ανταλλαχθούν. Η ανταλλαγή αυτών των bits εξασφαλίζει ότι οι απόγονοι θα ισαπέχουν από τους γονείς τους, ενώ ταυτόχρονα παρέχεται κι ένας μηχανισμός ποικιλότητας.



Εικόνα 26: *Παράδειγμα HUX – binary encoding*

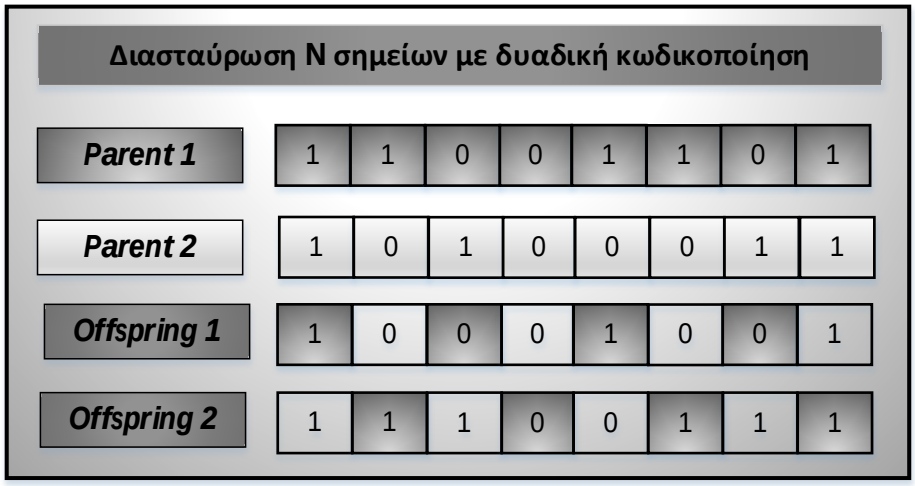
- **Διασταύρωση τριών γονέων (Three parents crossover)**

Το παιδί προέρχεται από τρεις τυχαία επιλεγμένους γονείς. Κάθε bit του πρώτου γονέα συγκρίνεται με το αντίστοιχο bit του δεύτερου γονέα. Εάν είναι τα ίδια, το bit θα συμπεριληφθεί στο παιδί. Εάν είναι διαφορετικά, τότε στο παιδί θα συμπεριληφθεί το αντίστοιχο bit από τον τρίτο γονέα.

Εικόνα 27: *Παράδειγμα διασταύρωσης με τρεις γονείς με δυαδική κωδικοποίηση*

- **Διασταύρωση Ν σημείων (N point crossover)**

Έχει προταθεί από τους Eshelman, Caruana και Schaffer το 1989. Σε αυτή την προσέγγιση επιλέγονται τόσα σημεία διασταύρωσης όσο είναι το μήκος του χρωμοσώματος. Η επιλογή πολλών σημείων διασταύρωσης μειώνει την επίδραση της πόλωσης θέσης αλλά, μειώνει και την απόδοση του ΓΑ. Το βασικό πρόβλημα που διαπιστώνεται εξαιτίας της χρήσης πολλών σημείων διασταύρωσης είναι ότι πιθανώς καλές υποψήφιες λύσεις μπορεί να χαθούν. Ωστόσο, με αυτή τη μέθοδο ο χώρος αναζήτησης εξερευνάται εξονυχιστικά και συνεπώς, απρόσιτες ενδεχομένως υποψήφιες λύσεις είναι πιο πιθανό να εντοπιστούν.



Εικόνα 28: *Παράδειγμα διασταύρωσης Ν σημείων σε χρωμοσώματα μήκους 8 με δυαδική αναπαράσταση*

- **Διασταύρωση πολλαπλών σημείων (Multi-point crossover)**

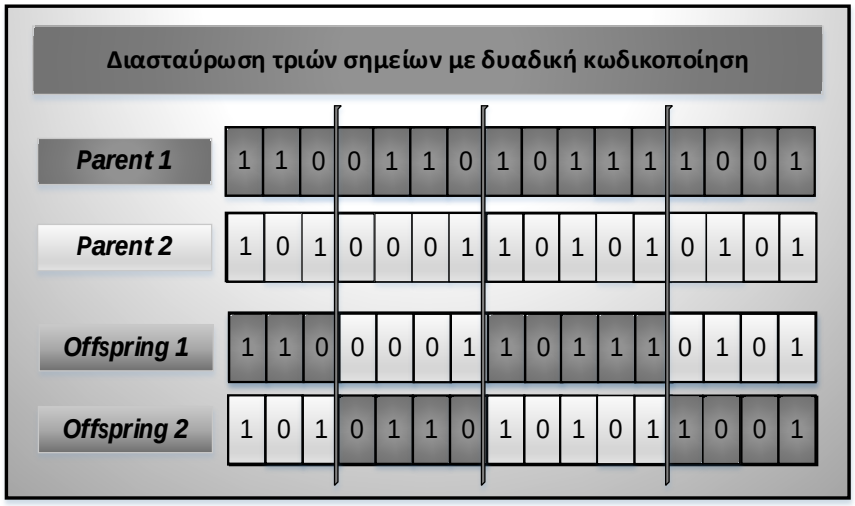
Η μέθοδος αυτή, για κάθε ζεύγος γονέων επιλέγει τυχαία έναν προκαθορισμένο αριθμό σημείων διασταύρωσης, και καθένας από τους δύο απογόνους που παράγονται παίρνει ένα τμήμα από κάθε γονέα εναλλάξ.

Μπορεί να επιλεγεί είτε μονός αριθμός σημείων διασταύρωσης είτε ζυγός αριθμός. Εάν είναι ζυγός, τα σημεία διασταύρωσης επιλέγονται τυχαία κυκλικά (τα μισά γονίδια από κάθε γονιό διατηρούνται) και η πληροφορία ανταλλάσσεται χωρίς προβλήματα. Στην περίπτωση που είναι μονός αριθμός, διαπιστώνεται πρόβλημα πόλωσης θέσης αφού η πιθανότητα ανταλλαγής ενός γονιδίου εξαρτάται από τη θέση του γονιδίου [15]. Έτσι, εάν επιλεγεί μονός αριθμός, ένα επιπλέον διαφορετικό σημείο διασταύρωσης υποθέτεται στην αρχή της ακολουθίας [16].

Η ομοιόμορφη διασταύρωση και η διασταύρωση Ν σημείων είναι υποκατηγορίες της διασταύρωσης πολλαπλών σημείων [18]. Η ομοιόμορφη διασταύρωση είναι μια γενίκευση της

διασταύρωσης Ν σημείων [19]. Με τη διασταύρωση πολλαπλών σημείων, όπως προαναφέρθηκε και στις δύο αυτές προσεγγίσεις, μειώνεται η επίδραση της πόλωσης θέσης ακόμα περισσότερο από ότι με τη διασταύρωση δύο σημείων.

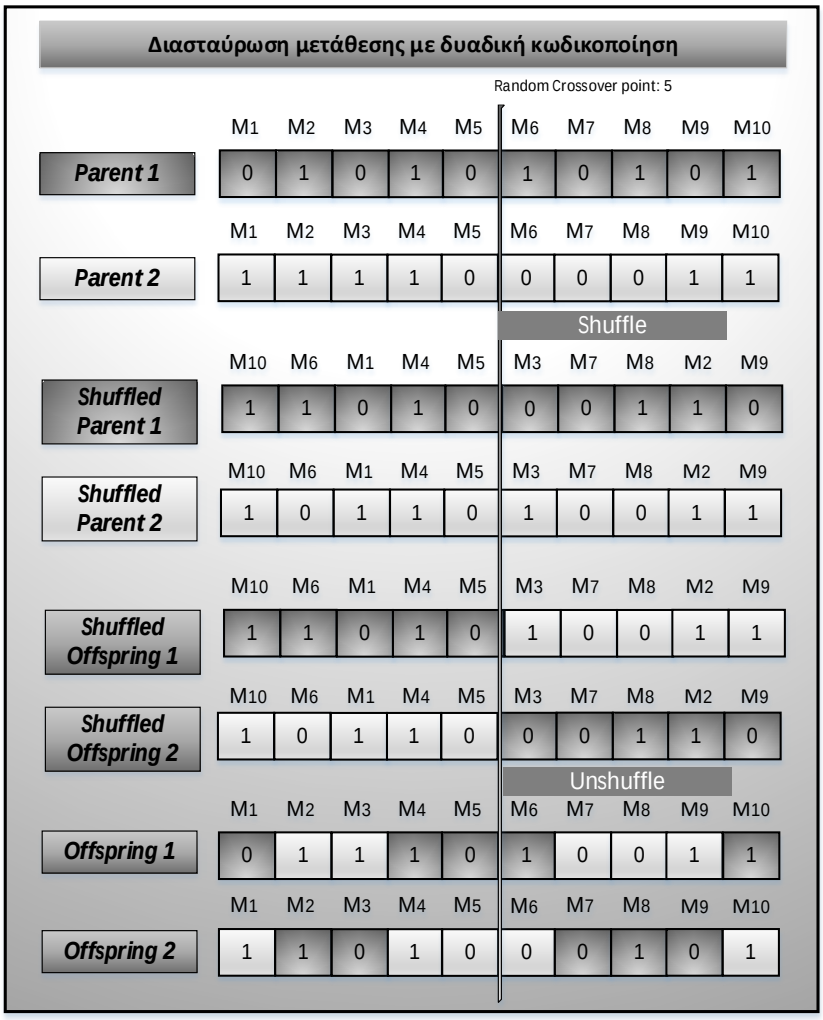
Παράδειγμα διασταύρωσης πολλαπλών σημείων είναι η διασταύρωση τριών σημείων.



Εικόνα 29: *Παράδειγμα διασταύρωσης τριών σημείων με δυαδική κωδικοποίηση και μήκος χρωμοσώματος 16*

- **Διασταύρωση μετάθεσης (Shuffle crossover)**

Η διασταύρωση μετάθεσης προτάθηκε από τους Eshelman, Caruna και Schaffer και συσχετίζεται με την ομοιόμορφη διασταύρωση. Επιλέγεται ένα Σ.Δ. όπως και στην διασταύρωση ενός σημείου. Προτού όμως γίνει η ανταλλαγή των μεταβλητών, όλες οι μεταβλητές μεταθέτονται τυχαία, με τον ίδιο όμως τρόπο, και στους δύο γονείς (αλλάζει τυχαία τις θέσεις των γονιδίων). Μετά τον ανασυνδυασμό των γονέων, αντιστρέφονται οι μεταθέσεις των μεταβλητών στους απογόνους. Έτσι αποτρέπεται οποιαδήποτε πόλωση θέσης, αφού κάθε φορά που εφαρμόζεται ο τελεστής της διασταύρωσης σε ένα ζεύγος οι μεταβλητές επανεκχωρούνται τυχαία [16]. Αντί για διασταύρωση ενός σημείου κατά τον ανασυνδυασμό των γονέων μπορεί να χρησιμοποιηθεί διασταύρωση πολλαπλών σημείων.



Εικόνα 30: *Παράδειγμα διασταύρωσης μετάθεσης με δυαδική κωδικοποίηση*

1.6.2 Μετάλλαξη (Mutation)

Η μετάλλαξη ή αλλιώς μεταλλαγή, όπως αναφέρεται στη βιβλιογραφία, συνήθως εφαρμόζεται μετά το στάδιο της διασταύρωσης. Σε αντίθεση με τη διασταύρωση εφαρμόζεται σε ένα μόνο χρωμόσωμα κάθε φορά. Στη βιολογία, η μετάλλαξη είναι μια μόνιμη αλλαγή στη σειρά των γονιδίων στο DNA ή γενικά οποιαδήποτε μεταβολή μπορεί να συμβεί στο γενετικό υλικό ενός οργανισμού. Στους ΓΑ κατ' αναλογία, ο τελεστής της μετάλλαξης αλλάζει τυχαία την τιμή ενός ή περισσοτέρων γονιδίων σε ένα χρωμόσωμα. Μπορεί να εφαρμοστεί είτε σε κάποιο αντίγραφο ενός γονέα έτσι ώστε να δημιουργήσει ένα νέο άτομο είτε σε κάποιο απόγονο ενός ζεύγους γονέων για να τον μεταλλάξει. Ο τελεστής της μετάλλαξης είναι ένας στοχαστικός τελεστής αφού το αποτέλεσμά του εξαρτάται από τυχαίες επιλογές: ποια στοιχεία του χρωμοσώματος θα μεταλλαχθούν και ποιες τιμές θα πάρουν.

Ορισμός. Μετάλλαξη (mutation) είναι η στοχαστική διαδικασία τροποποιήσεως της τιμής ενός ή περισσοτέρων χαρακτηριστικών μιας υποψήφιας λύσης.

Οι ΓΑ, εξαιτίας της στοχαστικής φύσης των τελεστών μετάλλαξης, μπορούν να διαφύγουν από τοπικά ελάχιστα / μέγιστα στο χώρο αναζήτησης, σε αντίθεση με άλλους αλγόριθμους που μένουν παγιδευμένοι εκεί. Ο στόχος της μετάλλαξης είναι η εξερεύνηση του προηγουμένως απρόσιτου χώρου αναζήτησης. Η μετάλλαξη λειτουργεί ως ασφαλιστική δικλείδα σε περίπτωση που η επιλογή και η διασταύρωση ενδεχομένως χάσουν πολύτιμες γενετικές πληροφορίες. Αντίθετα με τον τελεστή της διασταύρωσης, ο τελεστής της μετάλλαξης μπορεί να εισάγει νέες τιμές στα γονίδια των χρωμοσωμάτων (gene pool), οι οποίες δεν ήταν μέχρι τώρα παρούσες στον πληθυσμό. Έτσι, εισάγοντας καινούρια πληροφορία στον πληθυσμό, λειτουργεί ως τελεστής διατάραξης του πληθυσμού και συμβάλλει στην ποικιλομορφία των παραγόμενων γενεών. Η μετάλλαξη επιτρέπει στον ΓΑ να αποφύγει τα τοπικά ακρότατα, αποτρέποντας τα χρωμοσώματα του πληθυσμού από το να αναπτύξουν μεγάλο βαθμό ομοιότητας.

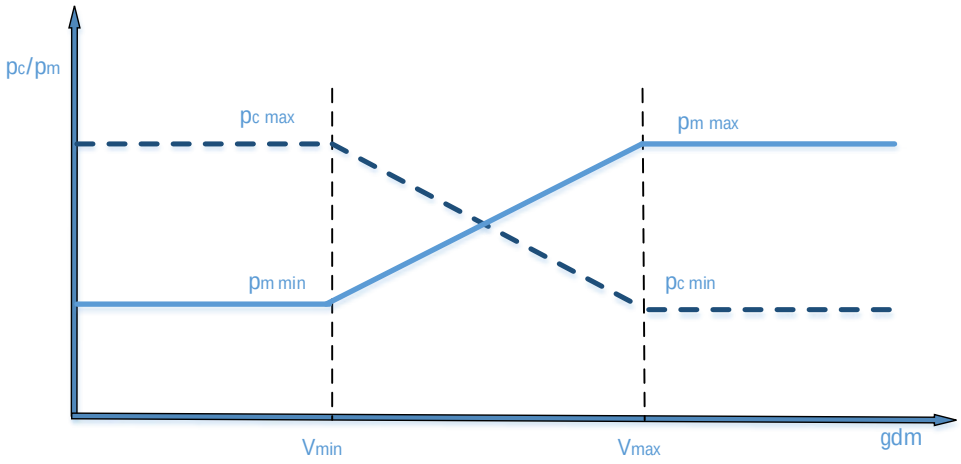
Ο τελεστής της μετάλλαξης εφαρμόζεται με μια πιθανότητα μετάλλαξης p_m. Η πιθανότητα μετάλλαξης υποδεικνύει πόσο συχνά θα μεταλλαχθούν τα γονίδια ενός χρωμοσώματος. Εάν δεν υπάρχει καθόλου μετάλλαξη, οι απόγονοι θα παραχθούν αμέσως μετά τη διασταύρωση ή θα αντιγραφούν κατευθείαν χωρίς καμία αλλαγή. Εάν εφαρμοστεί μετάλλαξη, ένα ή περισσότερα γονίδια του χρωμοσώματος θα αλλάξουν. Εάν η πιθανότητα μετάλλαξης είναι 1, τότε όλα τα γονίδια του χρωμοσώματος θα μεταλλαχθούν, ενώ εάν είναι 0, κανένα γονίδιο δεν θα μεταλλαχθεί [16].

Απαιτείται προσοχή στην επιλογή της τιμής της πιθανότητας μετάλλαξης, αφού αν είναι μεγάλη τα χρωμοσώματα δεν θα μπορούν να διατηρήσουν τα καλά δομικά στοιχεία τους, με κίνδυνο ο ΓΑ να μετατραπεί σε αλγόριθμο τυχαίας αναζήτησης και να μη συγκλίνει σε κανένα ακρότατο. Η τιμή που της ανατίθεται συνήθως είναι σχετικά χαμηλή. Για αλγόριθμους δυαδικής αναπαράστασης η τιμή της πιθανότητας μετάλλαξης επιλέγεται συνήθως στο διάστημα [0.01, 0.1]. Σε περίπτωση άλλων αναπαραστάσεων πολλές φορές είναι αρκετά μεγαλύτερη.

Οι Schaffer et al. αναφέρουν ότι η απόδοση των ΓΑ μειώνεται όταν ισχύουν ταυτόχρονα να είναι μεγάλος και ο πληθυσμός (περισσότερα από 200 άτομα) και η πιθανότητα μετάλλαξης (μεγαλύτερη από 0.05) είτε όταν συνδυάζεται μικρός πληθυσμός (λιγότερα από 20 άτομα) και μικρή πιθανότητα μετάλλαξης (μικρότερη από 0.002) [20].

Η πιθανότητα μετάλλαξης σε κάποιες παραλλαγές ΓΑ δεν είναι σταθερή αλλά μεταβάλλεται ανάλογα με την ποικιλομορφία του πληθυσμού. Υπάρχουν παραλλαγές όπου όταν η ποικιλομορφία του πληθυσμού ξεπεράσει κάποιο κατώφλι, τμήμα του πληθυσμού επαναρχικοποιείται δημιουργώντας μια saw-tooth καμπύλη ποικιλομορφίας [21]. Η επαναρχικοποίηση τμήματος του πληθυσμού μπορεί να θεωρηθεί ως στιγμιαία αύξηση της τιμής

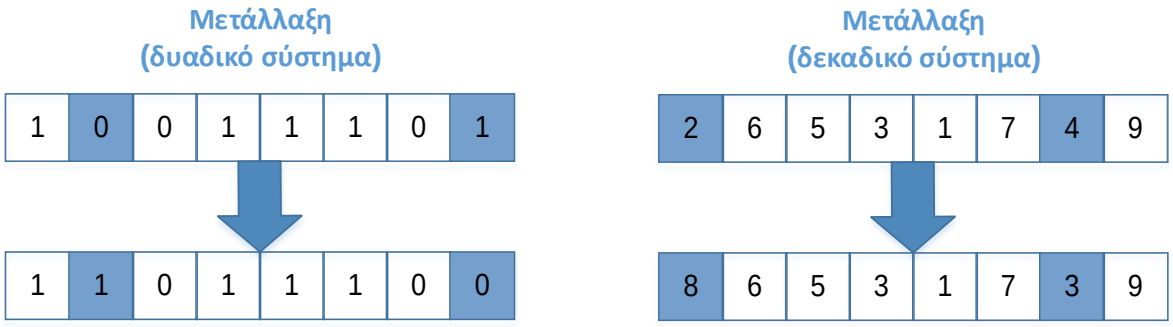
της πιθανότητας μετάλλαξης. Εναλλακτικά οι Vasconcelos et al. έχουν προτείνει μια δυναμική αλλαγή των πιθανοτήτων μετάλλαξης και διασταύρωσης: η πιθανότητα μετάλλαξης και η πιθανότητα διασταύρωσης κυμαίνονται ανάλογα με την ποικιλομορφία του πληθυσμού όπως φαίνεται στην ακόλουθη εικόνα [22]. Το μέτρο γενετικής ποικιλομορφίας gdm (genetic diversity measure) είναι ένα μέτρο ομοιότητας των στοιχείων του πληθυσμού και Vmin, Vmax οι τιμές του κάτω και άνω κατωφλίου αντίστοιχα.



Εικόνα 31: *Μεταβολή της πιθανότητας διασταύρωσης p_c και της πιθανότητας μετάλλαξης p_m συναρτήσει της ομοιότητας των στοιχείων του πληθυσμού*

Η διαδικασία ενός κλασσικού παραδείγματος απλής μορφής μετάλλαξης αναλύεται στα παρακάτω βήματα:

1. Επιλέγεται ένα χρωμόσωμα από το σύνολο των απογόνων που προέκυψαν μετά το στάδιο της διασταύρωσης (είτε κατευθείαν από το σύνολο των υποψήφιων γονέων που προέκυψαν από το στάδιο της επιλογής)
2. Με δεδομένη πιθανότητα μετάλλαξης p_m για κάθε γονίδιο του χρωμοσώματος επιλέγεται ομοιόμορφα μια τιμή στο διάστημα [0,1]. Εάν αυτή η τιμή είναι μεγαλύτερη από την πιθανότητα μετάλλαξης η διαδικασία συνεχίζεται για το επόμενο γονίδιο. Εάν είναι μικρότερη ή ίση, εφαρμόζεται μετάλλαξη στο τρέχον γονίδιο.
3. Στην περίπτωση που πραγματοποιείται μετάλλαξη θα δοθεί μια τυχαία, αλλά διαφορετική τιμή από την τρέχουσα στο γονίδιο αυτό: Αν η τιμή του γονιδίου είναι σε δυαδική μορφή τότε η τιμή του απλά αντιστρέφεται. Αν η τιμή του γονιδίου είναι σε ακέραια ή δεκαδική μορφή τότε παίρνει μια τυχαία τιμή από τις υπόλοιπες που ανήκουν στο επιτρεπτό διάστημα τιμών του γονιδίου.



Εικόνα 32: *Εφαρμογή του τελεστή της μετάλλαξης σε χρωμόσωμα δυαδικής αναπαράστασης και σε χρωμόσωμα δεκαδικής αναπαράστασης*

Ανάλογα με τον τύπο δεδομένων χρησιμοποιείται κι η ανάλογη τεχνική μετάλλαξης. Ακολούθως θα παρουσιαστούν συνοπτικά κάποιοι από αυτούς.

- **Μετάλλαξη ενός σημείου (Single Point mutation)**

Μια τυχαία μεταβλητή υποδεικνύει εάν ένα συγκεκριμένο δυαδικό bit θα μεταλλαχθεί (αντιστραφεί) ή όχι.

- **Μετάλλαξη ακολουθίας μπιτ (Bit string mutation)**

Η μετάλλαξη ακολουθίας μπιτ προκύπτει από την αλλαγή τιμής των bits σε τυχαίες θέσεις. Τα bits αντιστρέφονται σε μια τυχαία θέση με πιθανότητα *1/l* , όπου *l* το μήκος του χρωμοσώματος.

Εικόνα 33: *Παράδειγμα μετάλλαξης ακολουθίας μπιτ*

- **Ομοιόμορφη μετάλλαξη (Uniform mutation)**

Η ομοιόμορφη μετάλλαξη είναι η πιο απλή και η πιο κοινώς χρησιμοποιούμενη μέθοδος μετάλλαξης στους ΓΑ. Αντικαθιστά κάθε γονίδιο σύμφωνα με μια χαμηλή ομοιόμορφη πιθανότητα που καλείται βαθμός μετάλλαξης (mutation rate). Το δυαδικό bit που θα αντιστραφεί επιλέγεται από μια ομοιόμορφη τυχαία μεταβλητή της οποίας τα άνω και κάτω όριά της

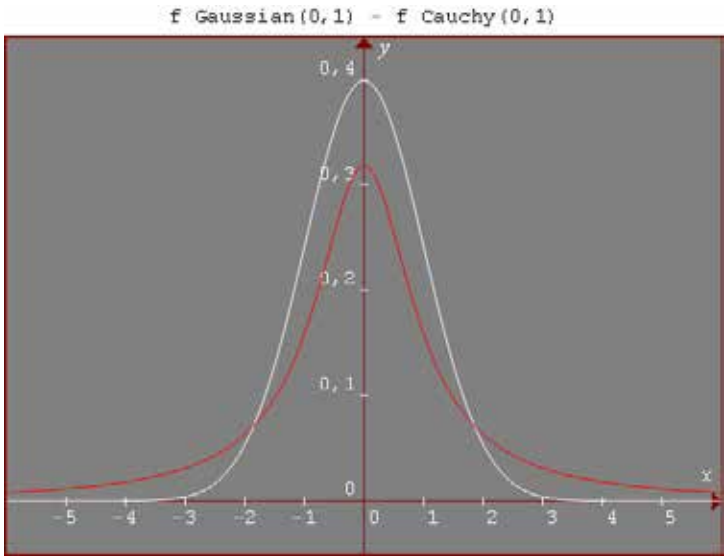
καθορίζονται από το χρήστη (Μπορεί να χρησιμοποιηθεί και με χρωμοσώματα που χρησιμοποιούν ακέραια κωδικοποίηση).

- **Γκαουσιανή μετάλλαξη (Gaussian mutation)**

Η γκαουσιανή μετάλλαξη αναπτύχθηκε από τους Rechenberg και Schwefel και καλείται και ως Classical Evolutionary Programming (CEP). Χρησιμοποιεί έναν τυχαίο αριθμό για την επιλογή του γονιδίου του οποίου θα μεταλλαχθεί η αρχική του τιμή. Ο τυχαίος αυτός αριθμός ακολουθεί μια κανονική γκαουσιανή κατανομή με μέσο μ και διασπορά σ^2 (*εικόνα 34*). Μοιάζει με την ομοιόμορφη μετάλλαξη, με μόνη διαφορά ότι η μεταβλητή που επιλέγει ποιο δυαδικό bit πρέπει να αντιστραφεί ακολουθεί κατανομή Gauss.

- **Μετάλλαξη Cauchy (Cauchy Mutation)**

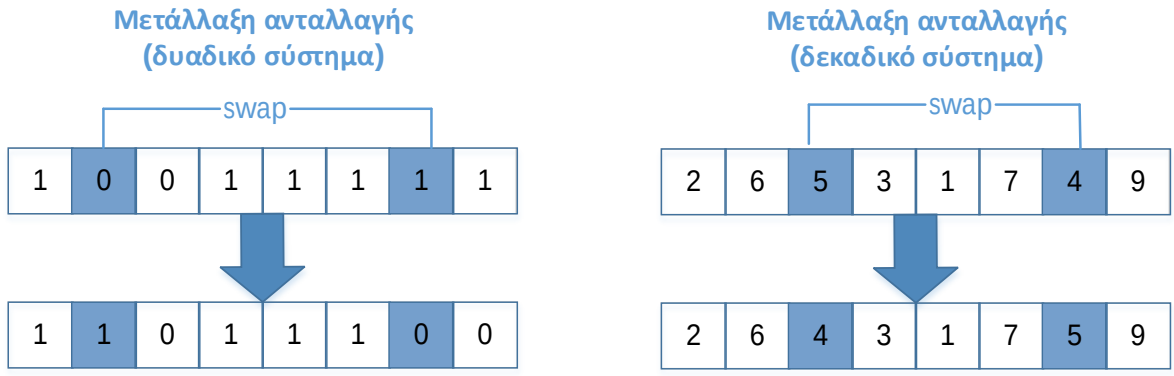
Για συναρτήσεις καταλληλότητας με ακρότατα που βρίσκονται σχετικά μακρυά μεταξύ τους, οι Yao, Liu και Lin σχεδίασαν το 1999 μια μετάλλαξη εναλλακτική της γκαουσιανής, χρησιμοποιώντας ένα τυχαίο αριθμό που ακολουθεί κατανομή Cauchy αντί Gauss (εικόνα 34). Αυτή η μέθοδος κλήθηκε Fast Evolutionary Programming (FEP).



Εικόνα 34: *Κανονική συνάρτηση πυκνότητας πιθανότητας Gauss με διασπορά σ^2=1 και συνάρτηση πυκνότητας πιθανότητας Cauchy με t=1*

- **Μετάλλαξη ανταλλαγής (Interchanging)**

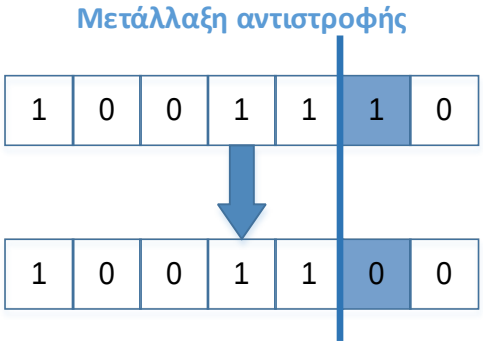
Δύο τυχαίες θέσεις της ακολουθίας επιλέγονται και τα αντίστοιχα σε αυτές τις θέσεις bits ανταλλάσσονται [16].



Εικόνα 35: *Παράδειγμα μετάλλαξης ανταλλαγής*

- **Μετάλλαξη αντιστροφής (Reversing)**

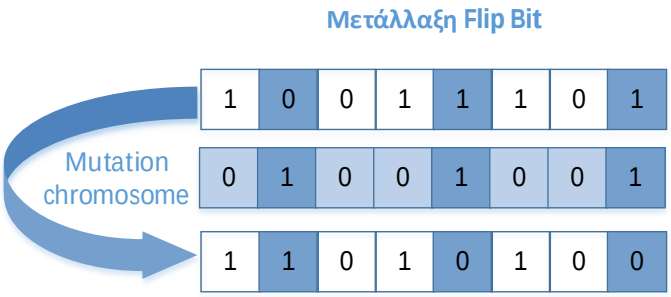
Επιλέγεται μια τυχαία θέση και τα bits δίπλα από αυτή τη θέση αντιστρέφονται [16].



Εικόνα 36: *Παράδειγμα μετάλλαξης αντιστροφής*

- **Μετάλλαξη Flip Bit (Flipping)**

Η αλλαγή ενός bit από 0 σε 1 και από 1 σε 0 είναι βασισμένη σε μια μάσκα μετάλλαξης ίδιου μήκους με τα χρωμοσώματα του πληθυσμού. Για κάθε χρωμόσωμα προς μετάλλαξη παράγεται μια τυχαία δυαδική μάσκα μετάλλαξης. Όπου υπάρχει τιμή 1 στη μάσκα το αντίστοιχο bit στο χρωμόσωμα προς μετάλλαξη αντιστρέφεται (από 0 σε 1 και από 1 σε 0).



Εικόνα 37: *Παράδειγμα μετάλλαξης Flip Bit*

- **Μετάλλαξη ορίων (Boundary)**

Αυτός ο τελεστής μετάλλαξης αντικαθιστά την τιμή του τυχαία επιλεγμένου γονιδίου (ή των τυχαία επιλεγμένων γονιδίων) είτε με το άνω είτε με το κάτω όριο των τιμών αυτού του γονιδίου. Αυτός ο τελεστής μπορεί να χρησιμοποιηθεί μόνο σε ακέραια ή δεκαδικά γονίδια.

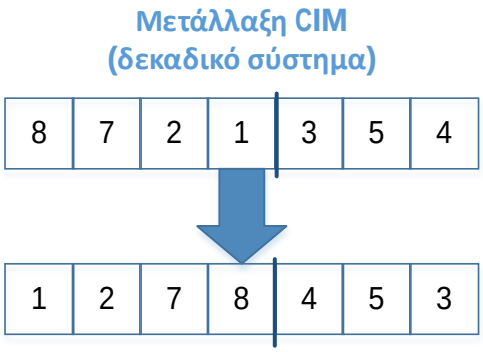
Εικόνα 38: *Παράδειγμα μετάλλαξης ορίων*

- **Μη ομοιόμορφη μετάλλαξη (Non-uniform)**

Αυτός τελεστής μετάλλαξης αυξάνει την πιθανότητα το ποσό της μετάλλαξης να γίνει 0 όσο ο αριθμός των γενεών αυξάνεται. Στις πρώτες γενιές του ΓΑ, αποτρέπει τον πληθυσμό από το να μείνει στάσιμος, ενώ ακολούθως σε αργότερα στάδια του ΓΑ, του επιτρέπει να εντοπίσει τη βέλτιστη λύση. Μπορεί να χρησιμοποιηθεί μόνο σε ακέραιους ή δεκαδικούς αριθμούς.

- **Μετάλλαξη κεντρικής αντιστροφής (Centre Inverse Mutation-CIM)**

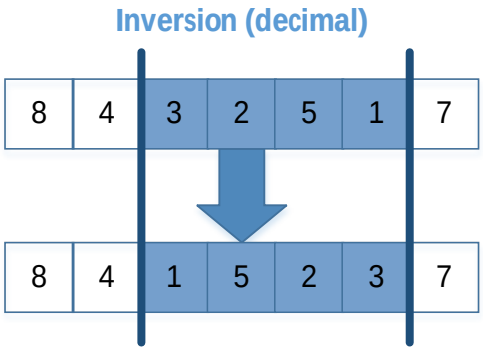
Το χρωμόσωμα χωρίζεται σε δύο τμήματα. Όλα τα γονίδια σε κάθε τμήμα αντιγράφονται και μετά αντιστρέφονται σε σειρά και τοποθετούνται στο ίδιο τμήμα.



Εικόνα 39: *Παράδειγμα CIM*

- **Αντιστροφή (Inversion)**

Επιλέγονται δύο τυχαίες θέσεις γονιδίων στο χρωμόσωμα που το χωρίζουν σε 3 τμήματα. Ακολούθως, αφήνει όπως έχουν τα τμήματα στα άκρα και αντιστρέφει το κεντρικό τους τμήμα.



Εικόνα 40: *Παράδειγμα Inversion*

1.7 Κριτήριο τερματισμού (Termination criterion)

Η γενετική διαδικασία επαναλαμβάνεται έως ότου εκπληρωθεί μια συνθήκη τερματισμού. Στην περίπτωση που είναι γνωστή η βέλτιστη τιμή του προβλήματος και ο ΓΑ απλά επιδιώκει να καθορίσει τις κατάλληλες τιμές των παραμέτρων που οδηγούν στην τιμή αυτή, τότε, όταν κάποιο χρωμόσωμα προσεγγίσει αυτή την τιμή ο αλγόριθμος θα σταματήσει. Επειδή, όμως, ο ΓΑ είναι στοχαστικής φύσεως, δεν είναι βέβαιο ότι η τιμή αυτή θα βρεθεί. Έτσι, η χρήση αυτής μόνο της συνθήκης, ως κριτηρίου τερματισμού μπορεί να αποβεί άκαρπη και να οδηγήσει τον αλγόριθμο σε άπειρο αριθμό επαναλήψεων. Συνεπώς, γίνεται απαραίτητος ο καθορισμός ενός κριτηρίου που σε συνδυασμό με την προαναφερθείσα συνθήκη να εγγυείται τον τερματισμό του ΓΑ.

Κριτήριο τερματισμού είναι η συνθήκη που πρέπει να ικανοποιηθεί, προκειμένου να σταματήσει η εκτέλεση του αλγόριθμου [23].

Κάποιες από τις συνθήκες τερματισμού είναι οι ακόλουθες:

- Εύρεση λύσης που πληροί κάποια ελάχιστα κριτήρια.
- Μέγιστος αριθμός γενεών (Maximum Generations) : ο αλγόριθμος σταματά με την επίτευξη συγκεκριμένου αριθμού γενεών. Αυτό είναι και το συνήθες κριτήριο που χρησιμοποιείται είτε μόνο του είτε σε συνδυασμό με κάποιο άλλο κριτήριο.
- Κατακερματισμός διαθέσιμων πόρων για την επίλυση, ειδικά χρόνος: ορισμός κάποιου μέγιστου επιτρεπόμενου χρονικού ορίου και τερματισμός της γενετικής διαδικασίας όταν ο συγκεκριμένος χρόνος λειτουργίας παρέλθει. Σε περίπτωση συνδυασμού με το κριτήριο μέγιστου αριθμού γενεών (όπως συνηθίζεται), όταν ο μέγιστος αριθμός γενεών έχει συμπληρωθεί πριν τον προκαθορισμένο χρόνο τερματισμού η διαδικασία τερματίζεται τότε.
- Καμία αλλαγή στην καταλληλότητα – Στασιμότητα της βελτίωσης της καταλληλότητας του πληθυσμού κάτω από ένα προκαθορισμένο κατώφλι για κάποιο δεδομένο αριθμό επαναλήψεων. Η λύση με την υψηλότερη καταλληλότητα (πρόβλημα μεγιστοποίησης) είτε προσεγγίζεται είτε έχει ήδη εντοπιστεί, οπότε, αφού άλλες επαναλήψεις δεν θα παράγουν πλέον καλύτερα αποτελέσματα, δεν υπάρχει λόγος για συνέχιση της εκτέλεσης του ΓΑ.
- Γενεές αναβολής (Stall generations): ο αλγόριθμος σταματά εάν δεν υπάρχει βελτίωση στη συνάρτηση καταλληλότητας για μια ακολουθία διαδοχικών γενεών μήκους ίσου με τον αριθμό των γενεών αναβολής.
- Όριο χρόνου αναβολής (Stall time limit): ο αλγόριθμος σταματά εάν δεν υπάρχει βελτίωση στη συνάρτηση καταλληλότητας κατά τη διάρκεια μιας περιόδου χρόνου ίσης σε δευτερόλεπτα με το όριο χρόνου αναβολής.
- Άνω όριο στον αριθμό αξιολογήσεων της συνάρτησης καταλληλότητας: όταν επιτευχθεί ένας συγκεκριμένος αριθμός αξιολογήσεων της συνάρτησης καταλληλότητας ο ΓΑ σταματά.
- Παράθυρο μεταβολής: ο ΓΑ τερματίζει αν για καθορισμένο αριθμό γενεών η μέση τιμή της καταλληλότητας του πληθυσμού δεν έχει αισθητή βελτίωση.
- Κριτήριο ποιότητας - Καλύτερο άτομο (Best individual): ο ΓΑ σταματά αν το ικανότερο χρωμόσωμα μιας γενεάς έχει καταλληλότητα καλύτερη από ένα προκαθορισμένο κατώφλι καταλληλότητας (Fitness threshold). Έτσι η αναζήτηση γίνεται πιο γρήγορα και εγγυείται τουλάχιστον μια καλή λύση.
- Μείωση της ποικιλομορφίας του πληθυσμού κάτω από ένα προκαθορισμένο κατώφλι.
- Σύγκλιση γονιδίου (Gene convergence): ο ΓΑ σταματά όταν ένα προκαθορισμένο ποσοστό των γονιδίων των χρωμοσωμάτων θεωρούνται ότι έχουν συγκλίνει. Ένα

γονίδιο θεωρείται ότι έχει συγκλίνει όταν η μέση τιμή αυτού του γονιδίου σε όλα τα χρωμοσώματα του τρέχοντος πληθυσμού είναι μικρότερη κατά ένα συγκεκριμένο ποσοστό από τη μέγιστη τιμή του γονιδίου αυτού σε όλα τα χρωμοσώματα.

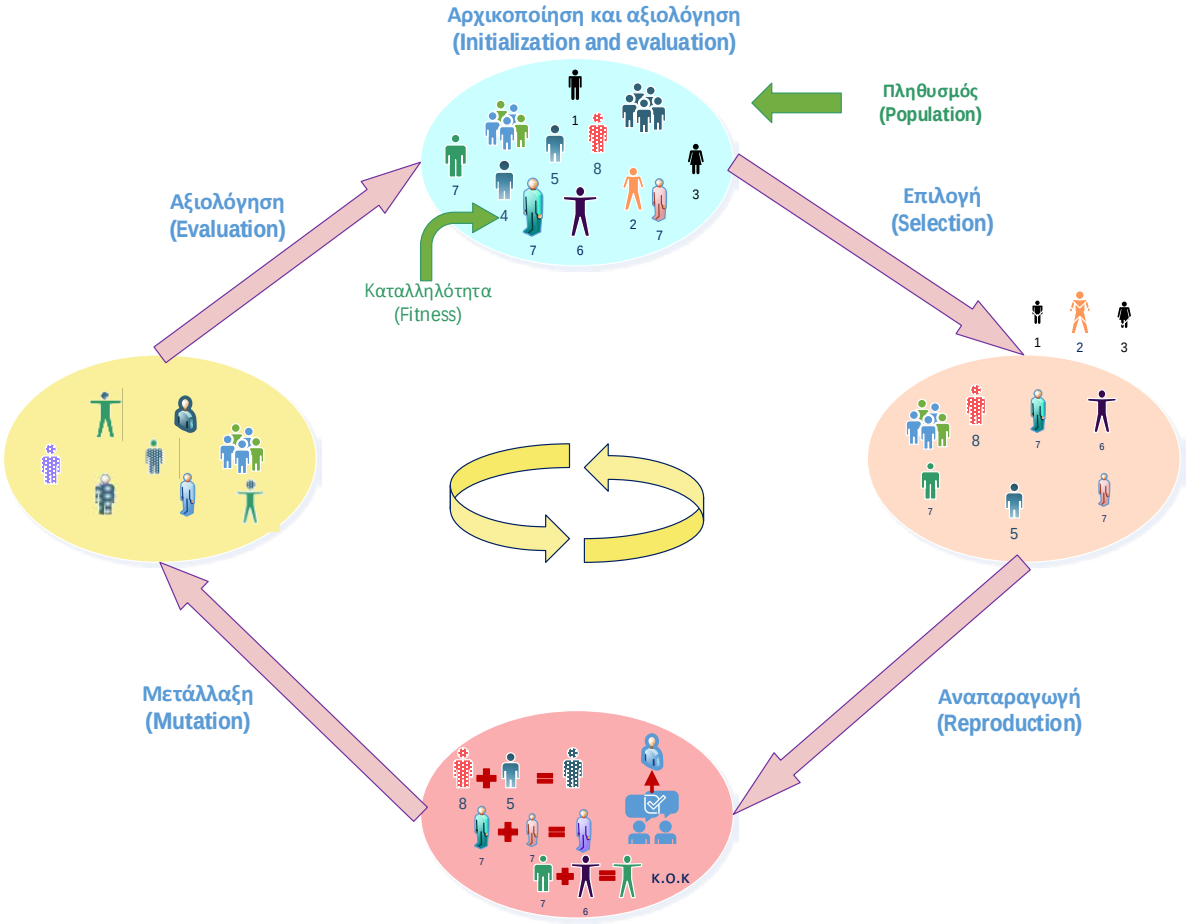
- Σύγκλιση πληθυσμού (Population Convergence) : ο ΓΑ σταματά όταν ο πληθυσμός θεωρείται ότι έχει συγκλίνει. Ο πληθυσμός θεωρείται ότι έχει συγκλίνει όταν η μέση τιμή καταλληλότητας στον τρέχων πληθυσμό είναι μικρότερη κατά ένα προκαθορισμένο ποσοστό από την καλύτερη καταλληλότητα του τρέχοντος πληθυσμού.
- Χειροκίνητη παρεμβολή.
- Κάποιος συνδυασμός των προαναφερθέντων κριτηρίων τερματισμού.

1.8 Σχεδίαση γενετικού αλγόριθμου

Για την υλοποίηση ενός γενετικού αλγόριθμου που πρόκειται να χρησιμοποιηθεί στην επίλυση ενός προβλήματος, πρέπει να παρθούν κάποιες αποφάσεις σχετικά με τα βασικά σχεδιαστικά βήματα που έχουμε προαναφέρει. Οι αποφάσεις αυτές θα είναι ανάλογες με τις παραμέτρους, τους περιορισμούς και τις απαιτήσεις του προβλήματος.

Παρακάτω αναφέρονται επιγραμματικά οι βασικές αποφάσεις σχεδίασης ενός γενετικού αλγόριθμου που πρέπει να ληφθούν [24] :

- Επιλογή του μέσου αναπαράστασης των πιθανών λύσεων (κωδικοποίησης των χρωμοσωμάτων)
- Επιλογή τρόπου δημιουργίας του αρχικού πληθυσμού των χρωμοσωμάτων
- Επιλογή της συνάρτησης καταλληλότητας για αξιολόγηση των χρωμοσωμάτων
- Επιλογή της τεχνικής φυσικής επιλογής που θα εφαρμοστεί στα χρωμοσώματα
- Επιλογή μεθόδου υλοποίησης των τελεστών αναπαραγωγής των χρωμοσωμάτων
- Καθορισμός παραμέτρων που χρησιμοποιεί ο γενετικός αλγόριθμος (μέγεθος πληθυσμού, πιθανότητες των τελεστών που θα εφαρμοστούν στον γενετικό αλγόριθμο κ.ο.κ)



Εικόνα 41: *Βρόγχος ανακύκλωσης ενός ΓΑ*

1.9 Το θεώρημα σχημάτων (The schema theory)

Το θεώρημα σχημάτων, όπως διατυπώθηκε από τον Holland το 1975 [4], επιχειρεί να εξηγήσει γιατί λειτουργούν οι γενετικές τεχνικές: «Μικρά αλλά αποτελεσματικά δομικά στοιχεία (σχήματα) λαμβάνουν εκθετικά αυξημένη πιθανότητα αναπαραγωγής από γενιά σε γενιά δίνοντας προοδευτικά καλούς απογόνους».

Η ιδέα του σχήματος είναι η βάση της θεωρίας των γενετικών αλγορίθμων. Το σχήμα επιτρέπει τον προσδιορισμό της ομοιότητας μεταξύ των χρωμοσωμάτων. Ο Holland περιέγραψε το σχήμα (schema) [25] σαν ένα πρότυπο, μια φόρμα (template) που περιγράφει ακολουθίες από χαρακτήρες που ανήκουν σε ένα υποσύνολο του πληθυσμού και παρουσιάζουν όμοιους χαρακτήρες σε συγκεκριμένες θέσεις στη συμβολοσειρά. Για παράδειγμα το 110 και το 111 είναι ταυτόσημα εάν αγνοηθεί το τελευταίο ψηφίο. Ένα σχήμα κατασκευάζεται εισάγοντας το λεγόμενο αδιάφορο σύμβολο * στο αλφάβητο Σ={0,1} των γονιδίων [23]. Έτσι το σύμβολο * παίρνει ρόλο μπαλαντέρ, δηλαδή μπορεί να αναπαριστά είτε το χαρακτήρα 0 είτε το χαρακτήρα

1. Ένα σχήμα αναπαριστά όλες τις συμβολοσειρές- ένα υποσύνολο του χώρου αναζήτησης- οι οποίες ταιριάζουν σε όλες τις θέσεις εκτός από αυτές με το αδιάφορο σύμβολο. Άρα οι δύο συμβολοσειρές του παραδείγματος μπορούν να αναπαρασταθούν με το σχήμα 11*. Ανάλογα με τις τιμές καταλληλότητας των συμβολοσειρών μπορεί να υπονοηθεί ότι οι συμβολοσειρές που ξεκινάνε με 11 είναι καλύτερες από άλλες. Αυτά τα πρότυπα ομοιότητας είναι τα σχήματα. Άλλα παραδείγματα σχημάτων και των συνεπαγόμενων συμβολοσειρών τους φαίνονται στον πίνακα 2.

Πίνακας 3: Παραδείγματα σχημάτων

Σχήμα S	Παραγόμενες συμβολοσειρές	Τάξη o(S)	Ορίζον μήκος
*10101010	**{110101010, 010101010}**	**8**	**8**
*1*010110	**{010010110, 011010110, 110010110 , 111010110 }**	**7**	**7**
*1**0	**{01000, 01010, 01100, 01110, 11000, 11010, 11100, 11110}**	**2**	**3**
**1	**{001,011,101,111}**	**1**	**0**
*********	**Όλες οι συμβολοσειρές μήκους 9**	**0**	**0**

Κάθε σχήμα αναπαριστά 2^m συμβολοσειρές όπου μ ο αριθμός των αδιάφορων συμβόλων στο σχήμα, ενώ κάθε συμβολοσειρά μήκους l, ταιριάζει σε 2^l διαφορετικά σχήματα. Υπάρχουν δύο σημαντικά μεγέθη που χαρακτηρίζουν τα σχήματα, η τάξη (order) και το ορίζον μήκος (defining length). Η τάξη ενός σχήματος S, συμβολίζεται με o(S) και ισούται με τον αριθμό των σταθερών θέσεων του σχήματος, αυτών δηλαδή που δεν περιέχουν το αδιάφορο σύμβολο *. Όσο πιο μεγάλη η τάξη ενός σχήματος τόσο πιο ειδικό είναι το συγκεκριμένο σχήμα, δηλαδή αναπαριστά λιγότερες συμβολοσειρές. Η τάξη είναι χρήσιμη στον υπολογισμό της πιθανότητας επιβίωσης του σχήματος κατά την μετάλλαξη. Το ορίζον (επίσης αναφέρεται ως οριστικό ή ορισμένο) μήκος ενός σχήματος S , συμβολίζεται με δ(S) και είναι η απόσταση της πρώτης και της τελευταίας σταθερής θέσης. Προσδιορίζει την πυκνότητα της πληροφορίας που περιέχεται στο σχήμα. Ένα σχήμα με μια σταθερή θέση έχει ορίζον μήκος μηδέν. Το ορίζον μήκος είναι χρήσιμο στον υπολογισμό της πιθανότητας επιβίωσης του σχήματος κατά τη διασταύρωση.

Το παρακάτω θεώρημα είναι εμπειρικό χωρίς να έχει γίνει κάποια φορμαλιστική μαθηματική ανάλυση. Αυτό είναι και το μεγαλύτερο μειονέκτημα των ΓΑ: δεν έχουν ακόμη αναλυθεί μαθηματικά και έτσι υπάρχει έλλειψη της πλήρους μαθηματικής επεξήγησης των λειτουργιών τους. Μη μπορώντας να επεξηγηθούν αρκετά στοιχεία της συμπεριφοράς των ΓΑ καθίσταται δύσκολη η βελτιστοποίησή τους.

Ορισμός 2. Θεώρημα σχημάτων:

Σχήματα με απόδοση άνω του μέσου όρου και με μικρό οριστικό μήκος και μικρή τάξη λαμβάνουν εκθετικά αυξανόμενες συμβολοσειρές σε διαδοχικές γενιές ενός ΓΑ.

1.10 The building block hypothesis

Κατά τα λόγια του Goldberg "… we construct better and better strings from the best partial solutions of past samplings" [5].

Σύμφωνα με την υπόθεση δομικών στοιχείων (building block hypothesis-BBH), θεωρούνται καλά σχήματα και προτιμώνται, τα σχήματα που είναι μικρής τάξης, με μικρό ορισμένο μήκος και έχουν άνω του μέσου όρου τιμή καταλληλότητας. Τα σχήματα αυτά ορίζονται ως δομικά στοιχεία (building blocks) [16]. Ένας γενετικός αλγόριθμος αναζητεί τη βέλτιστη λύση σε αυτά τα σχήματα και τους απογόνους τους. Ο Goldberg ουσιαστικά έδωσε δύο βασικές αρχές για επιλογή κωδικοποίησης σε γενετικό αλγόριθμο [5] κι αυτές οι αρχές αποτελούν και τις δύο βασικές συνιστώσες της BBH.

Η πρώτη αρχή, η αρχή του ελάχιστου αλφαβήτου, αναφέρεται στα είδη των συμβόλων που χρησιμοποιούνται στην αναπαράσταση της πληροφορίας που περιέχεται στο χρωμόσωμα. Ο Goldberg υποδεικνύει: «Επέλεξε το μικρότερο αλφάβητο που επιτρέπει μια φυσική έκφραση του προβλήματος». Η αρχή αυτή προωθεί τη χρήση δυαδικής κωδικοποίησης έναντι μιας αλφαριθμητικής κωδικοποίησης που χρησιμοποιεί πληθώρα συμβόλων. Οι δυαδικές κωδικοποιήσεις μεγιστοποιούν τον αριθμό των διαθέσιμων σχημάτων στη διαδικασία της αναζήτησης. Αν η τιμή κάθε θέσης σε ένα χρωμόσωμα μήκους l_1, με μορφή δυαδικής κωδικοποίησης, μπορεί να είναι μια εκ των {0,1,*} , τότε προκύπτουν 3^{l1} δυνατά σχήματα. Αν χρησιμοποιηθεί αλφάβητο 4 συμβόλων και του *, όπως π.χ. {a,b,c,d,*} τότε ενώ θα παράγονται 5^{l2} σχήματα, το μήκος l_2 του χρωμοσώματος που θα απαιτείται για την κωδικοποίηση ίδιου αριθμού τιμών θα είναι πολύ μικρότερο από αυτό της δυαδικής κωδικοποίησης: $l_1 < l_2$. Για παράδειγμα στην οκταδική κωδικοποίηση απαιτούνται 8 σύμβολα {0,1,2,3,4,5,6,7}. Με το αλφάβητο αυτών των 8 συμβόλων για κωδικοποίηση ακεραίων στο διάστημα [0, 63] χρειάζεται μήκος χρωμοσώματος $l_1 = 2$ *(αφού* $77 = 7*8^0 + 7*8^1 = 63$*)* και άρα προκύπτουν $9^2 = 81$ σχήματα. Η δυαδική κωδικοποίηση για να κωδικοποιήσει το ίδιο διάστημα τιμών απαιτεί μήκος χρωμοσώματος $l_2 = 6$ *(αφού* $111111 = 1*2^0 + 1*2^1 + 1*2^2 + 1*2^3 + 1*2^4 + 1*2^5 = 63$*)* και άρα προκύπτουν $3^6 = 729$ σχήματα. Αποδεικνύεται εύκολα ότι το δυαδικό αλφάβητο παράγει περισσότερα σχήματα από οποιαδήποτε άλλη κωδικοποίηση.

Η δεύτερη αρχή του Goldberg ασχολείται με την τάξη των σχημάτων και προτείνει επιλογή κωδικοποίησης τέτοια ώστε τα μικρής τάξης σχήματα (δηλαδή με μικρό αριθμό σταθερών

συμβόλων) να είναι σχετικά με το πρόβλημα και να σχετίζονται όσο το δυνατό λιγότερο έως καθόλου με τα σχήματα στις άλλες σταθερές θέσεις. Η αρχή αυτή επισημαίνει ότι όταν μια σειρά από bits πληροφορίας σε ένα χρωμόσωμα είναι στενά συνδεδεμένα μεταξύ τους, θα πρέπει να είναι λίγα στο πλήθος και οι θέσεις τους να είναι όσο το δυνατό πιο κοντά. Με αυτό τον τρόπο μειώνεται η πιθανότητα να διαχωριστούν από τους τελεστές αναπαραγωγής όταν ανταλλάσσουν γονίδια. Αν είναι διεσπαρμένα σε διάφορες θέσεις στο χρωμόσωμα τότε θα διαταράσσεται το σχήμα σε κάθε γενιά κι έτσι τα bits αυτά θα έχουν αρκετά μικρότερη επίδραση στην καταλληλόλητα του χρωμοσώματος. Παρά το ότι η τήρηση της δεύτερης αρχής είναι δύσκολη, είναι συνετό να ακολουθείται όσο πιο πολύ γίνεται. Σε περίπτωση που δεν τηρείται καθόλου, η απόδοση του ΓΑ μπορεί να πέσει σε τέτοια επίπεδα, παρόμοια με αυτά που θα έφτανε η τυχαία αναζήτηση.

Η δυνατότητα της παραγωγής όλων και πιο κατάλληλων λύσεων όταν τα δομικά στοιχεία συνδυάζονται μεταξύ τους, θεωρείται ως η κύρια αιτία της λειτουργικότητας των ΓΑ. Παρά τον υποτιθέμενο κεντρικό ρόλο των δομικών στοιχείων (building blocks) και του ανασυνδυασμού τους, η ερευνητική κοινότητα , προς το παρόν, στερείται αναλυτικών περιγραφών της ακριβής συμμετοχής που διαδραματίζει η επεξεργασία ενός σχήματος κατά την τυπική εξέλιξη μιας αναζήτησης με χρήση ΓΑ [26].

1.11 Πλεονεκτήματα Γενετικού αλγορίθμου

“Genetic algorithms are good at taking large potentially huge search spaces and navigating them, looking for optimal combinations of things, the solutions one might not otherwise find in a lifetime” – Salvatore Mangano, Computer Design, May 1995

Τα βασικά πλεονεκτήματα των γενετικών αλγορίθμων επιγραμματικά είναι τα ακόλουθα:

- Γρήγορη και αξιόπιστη επίλυση δύσκολων προβλημάτων
- Ευρεία εφαρμογή σε πληθώρα πεδίων ,περισσοτέρων από κάθε άλλη έως τώρα γνωστή μέθοδο
- Μη απαίτηση περιορισμών στις συναρτήσεις που επεξεργάζονται
- Εύκολη συνεργασία με τα υπάρχοντα μοντέλα και συστήματα
- Εφικτή συνεργασία με άλλες μεθόδους
- Δεν παίζει ρόλο η σημασία της υπό εξέταση πληροφορίας
- Εύκολα επεκτάσιμοι και εξελίξιμοι
- Έχουν εκ φύσεως το στοιχείο του παραλληλισμού

- Μόνη μέθοδος που κάνει εξερεύνηση του χώρου αναζήτησης και εκμετάλλευση της ήδη επεξεργασμένης πληροφορίας ταυτόχρονα
- Επίλυση προβλημάτων με ασυνέχειες στο διάστημα λύσης

Ένα από τα πιο σημαντικά πλεονεκτήματα των γενετικών αλγορίθμων είναι ότι μπορούν να επιλύουν γρήγορα και αξιόπιστα δύσκολα προβλήματα. Έχει αποδειχτεί θεωρητικά και εμπειρικά ότι προβλήματα με πολλές και δύσκολα προσδιορισμένες λύσεις καθώς και προβλήματα εύρεσης ακροτάτων σε συναρτήσεις που παρουσιάζουν μεγάλες διακυμάνσεις αντιμετωπίζονται πιο αποδοτικά με χρήση γενετικών αλγορίθμων [13].

Επιπλέον έχουν ευρεία εφαρμογή σε πληθώρα πεδίων, περισσότερων από κάθε άλλη έως τώρα γνωστή μέθοδο (οικονομία, βιοπληροφορική, σχεδιασμό μηχανών, επίλυση μαθηματικών εξισώσεων, πρόβλεψη σε μη γραμμικές χρονοσειρές, εκπαίδευση νευρωνικών δικτύων είναι μόνο λίγοι από τους τομείς στους οποίους εφαρμόζονται). Η επιτυχής εφαρμογή τους σε προβλήματα τόσων διαφορετικών τομέων είναι δυνατή επειδή οι διάφοροι περιορισμοί στις συναρτήσεις προς επεξεργασία (ύπαρξη παραγώγων, συνέχεια, όχι θορυβώδεις συναρτήσεις) που θεωρούνται απαραίτητοι από άλλες μεθόδους, αφήνουν αδιάφορους τους ΓΑ.

Ακόμα ένα πλεονέκτημα είναι η εύκολη συνεργασία τους με τα υπάρχοντα μοντέλα και συστήματα χωρίς να είναι αναγκαία η επανασχεδίασή τους. Αυτό είναι εφικτό επειδή οι ΓΑ χρησιμοποιούν μόνο πληροφορίες της συνάρτησης που πρόκειται να βελτιστοποιήσουν χωρίς να τους ενδιαφέρει άμεσα τι ρόλο έχει η συνάρτηση μέσα στο σύστημα.

Οι ΓΑ είναι ευέλικτοι κι έτσι είναι εφικτή η συνεργασία τους με άλλες μεθόδους: μπορούν να συμμετέχουν σε συνδυασμό με άλλες μεθόδους για τη δημιουργία μιας υβριδικής μορφής.

Εκτός των παραπάνω, οι γενετικοί αλγόριθμοι είναι εύκολα επεκτάσιμοι και εξελίξιμοι, με πολλές παραλλαγές προσαρμοσμένες στο εκάστοτε πρόβλημα ώστε να μεγιστοποιηθεί η απόδοσή τους.

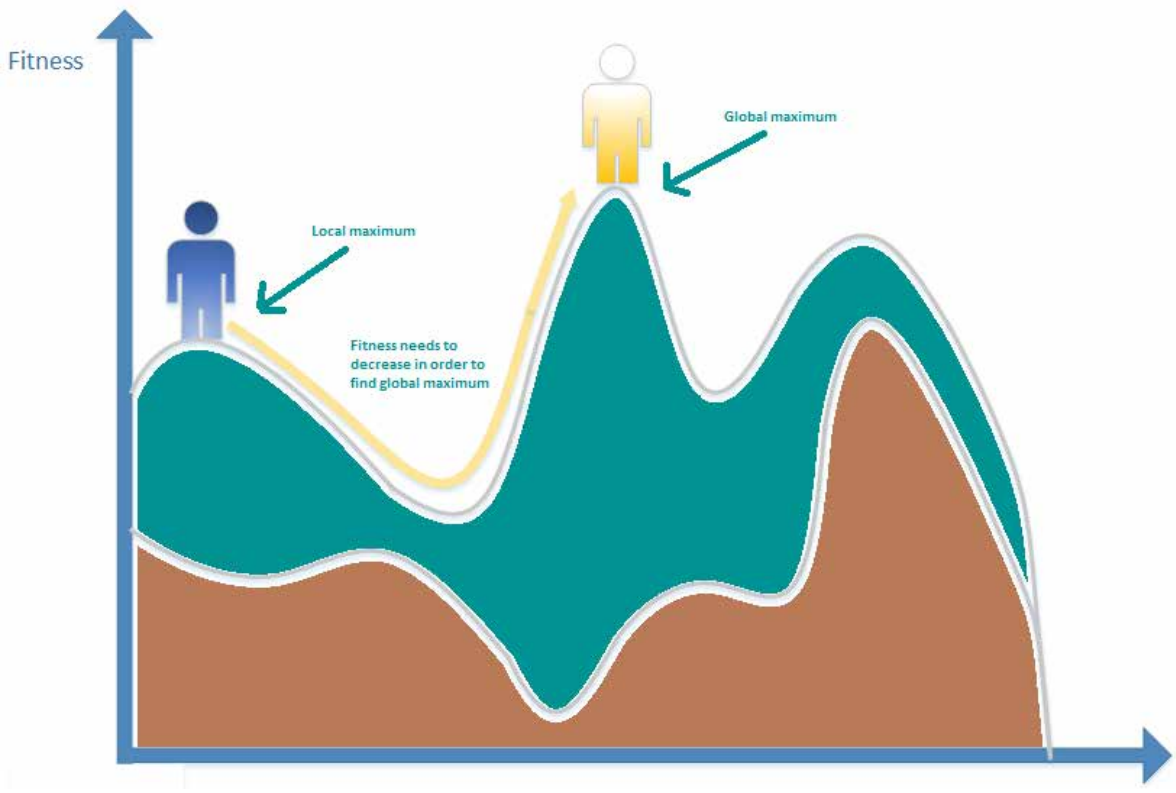
Επίσης χρήσιμο είναι το γεγονός ότι δεν παίζει ρόλο η σημασία της υπό εξέταση πληροφορίας. Η συνάρτηση καταλληλότητας εγγυάται τον επιτυχή τερματισμό του γενετικού αλγόριθμου ανεξάρτητα από τη σημασία του προβλήματος.

Οι ΓΑ είναι η μόνη μέθοδος που κάνει εξερεύνηση του χώρου αναζήτησης και εκμετάλλευση της ήδη επεξεργασμένης πληροφορίας ταυτόχρονα. Επίσης έχουν εκ φύσεως το στοιχείο του παραλληλισμού αφού σε κάθε βήμα επεξεργάζονται τεράστιο όγκο πληροφορίας, μια και κάθε χρωμόσωμα θεωρείται αντιπρόσωπος μιας ολόκληρης κλάσης άλλων χρωμοσωμάτων. Έτσι γίνεται κατορθωτή η αποδοτική σάρωση μεγάλων χώρων σε μικρούς χρόνους. Εξαιτίας της φύσης τους αυτής είναι πραγματοποιήσιμη η παράλληλη υλοποίησή τους με χρήση παράλληλων μηχανών αυξάνοντας έτσι την απόδοσή τους [23].

1.12 Περιορισμοί Γενετικού αλγόριθμου

Παρά τα πλεονεκτήματα των γενετικών αλγορίθμων παραμένουν κάποια προβλήματα. Ένα βασικό πρόβλημα είναι το πρόβλημα του χρόνου. Ο αλγόριθμος χρειάζεται αρκετό χρόνο για να εξελιχθεί και να παράγει τη βέλτιστη λύση ειδικά όταν πρέπει να καλύψει μεγάλους χώρους αναζήτησης. Οι ΓΑ έχουν υψηλή πολυπλοκότητα – συχνά εκθετική- και έτσι αδυνατούν να χειριστούν προβλήματα με τεράστιο όγκο πληροφορίας. Μπορεί να έχουν υψηλές ταχύτητες και να είναι πιο γρήγοροι από τις πλείστες μεθόδους, παραμένουν, όμως, περιθώρια βελτίωσής τους ως προς την ταχύτητα εξέλιξής τους.

Ένας άλλος περιορισμός των ΓΑ είναι ότι σε πολλά προβλήματα οι ΓΑ μπορεί να συγκλίνουν προς τοπικά βέλτιστα αντί για καθολικά βέλτιστα. Δεν μπορούν να διαπιστώσουν πότε πρέπει να θυσιάσουν βραχυπρόθεσμα μια υψηλή καταλληλότητα έτσι ώστε να αποκτήσουν μακροπρόθεσμα μεγαλύτερη (δες *εικόνα 23*). Το πρόβλημα αυτό μπορεί να μειωθεί με χρήση εναλλακτικών συναρτήσεων καταλληλότητας, αυξάνοντας το ρυθμό μετάλλαξης ή χρησιμοποιώντας άλλη μορφή αναπαράστασης των πιθανών λύσεων.



Εικόνα 42: *Το "μονοπάτι" προς το ολικό μέγιστο*

Επιπλέον η επαναλαμβανόμενη χρήση της συνάρτησης καταλληλότητας μπορεί να προκαλέσει πρόβλημα. Σε πολύπλοκα και πολυδιάστατα προβλήματα είναι πιθανόν να χρειάζεται μια

υπολογιστικά δαπανηρή συνάρτηση της οποίας μια κλίση μπορεί να πάρει ώρες να τερματίσει. Απλές βελτιστοποιήσεις δεν μπορούν να χειριστούν το πρόβλημα και τελικά χρησιμοποιείται μια απλή προσεγγιστική συνάρτηση που θυσιάζει ακρίβεια για χάρη πρακτικότητας.

Ακόμα ένας περιορισμός είναι ότι μια και η πιο κατάλληλη λύση καθορίζεται σε σχέση με τις υπόλοιπες λύσεις μπορεί να προκληθεί ασάφεια ως προς το ποιο θα πρέπει να είναι το κριτήριο τερματισμού του ΓΑ.

Επίσης, σε συγκεκριμένα προβλήματα βελτιστοποίησης άλλοι αλγόριθμοι μπορεί να είναι προτιμότεροι λόγω καλύτερης αποτελεσματικότητας. Εναλλακτικοί ή συμπληρωματικοί αλγόριθμοι μπορεί να περιλαμβάνουν εξελικτικό προγραμματισμό, αλγόριθμους σμήνους (π.χ. βελτιστοποίηση μυρμηγκοφωλιάς), προσαρμογή Gauss, στρατηγικές εξέλιξης, αναρριχητικούς αλγόριθμους. Ανάλογα με το πρόβλημα και τις υπάρχουσες πληροφορίες σχετικά με το πρόβλημα, μπορεί να καθοριστεί η καταλληλότητα των ΓΑ ή κάποιας άλλης πιο εξειδικευμένης προσέγγισης.

Κεφάλαιο 2 – Επιλογή Χαρακτηριστικών

2.1 Εισαγωγή

Ο διανυσματικός χώρος των χαρακτηριστικών (feature space) είναι ο χώρος n διαστάσεων που προκύπτει από τα n χαρακτηριστικά που χρησιμοποιούνται για την περιγραφή ενός προβλήματος. Στην *εικόνα 47* παρουσιάζονται χώροι 1, 2 και 3 διαστάσεων. Το διάνυσμα χαρακτηριστικών (feature vector) είναι το διάνυσμα που αποτελείται από τα n χαρακτηριστικά και περιγράφει ένα συγκεκριμένο δείγμα, δηλαδή εμπεριέχει τις τιμές του δείγματος για όλα τα χαρακτηριστικά όπως φαίνεται για παράδειγμα στην *εικόνα 43*. Τα χαρακτηριστικά αλλιώς αναφέρονται και ως μεταβλητές (variables). Ο αριθμός των χαρακτηριστικών μπορεί να ποικίλει.

Χαρακτηριστικά

Δείγματα

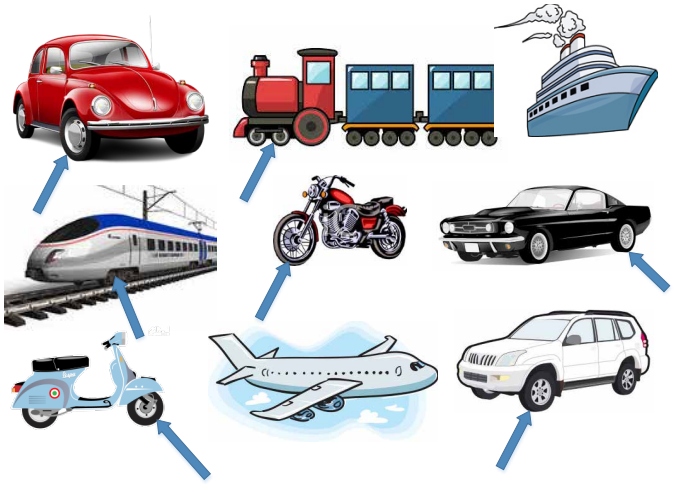
TSFD	BCS	WBC	MONO	UT	REF	QTY	COLR	MYMPH
0394	0001	CTN	0	A	00AA	10	A234	Lop
7493	0002	BPL	1	B	00AB	20	B294	Gip
0924	0001	CTN	0	A	01AA	5	K287	Mat
7734	0001	BPL	1	A	00AA	485	O485	Mat
1132	0002	BPL	1	A	00BB	50	N768	Mat
1845	0003	BPL	1	A	01AB	15	B499	Lop
0394	0001	CTN	0	B	01AA	5	L234	Gip
7734	0001	CTN	0	A	01AA	5	A234	Gip

Εικόνα 43: *Τι είναι τα χαρακτηριστικά*

Η προεπεξεργασία των δεδομένων ενός προβλήματος είναι ένα αναπόσπαστο τμήμα της αποτελεσματικής ανάλυσης δεδομένων. Η επιλογή χαρακτηριστικών (feature selection) είναι απαραίτητη σε προβλήματα όπου υπάρχει τεράστιο πλήθος δεδομένων. Αν και η ύπαρξη πολλών δεδομένων εκ πρώτης όψεως φαίνεται επιθυμητή, δημιουργεί προβλήματα επειδή εισάγει θόρυβο και μειώνει την απόδοση του συστήματος. Μερικά χαρακτηριστικά ενδεχομένως να είναι άσχετα ή περιττά και μπορεί να μπερδέψουν το σύστημά μας και συνεπώς πρέπει να αφαιρεθούν. Επιπλέον με την αύξηση του αριθμού των χαρακτηριστικών αυξάνεται και η υπολογιστική πολυπλοκότητα χωρίς να αυξάνεται η απόδοση και αφού οι υπολογιστικοί μας πόροι είναι περιορισμένοι απαιτείται μείωση του αριθμού των χαρακτηριστικών. Ως εκ τούτου, η κατάλληλη επιλογή χαρακτηριστικών έτσι ώστε να περιγράφεται επαρκώς το εκάστοτε πρόβλημα είναι ουσιώδης.

Ένα από τα μεγαλύτερα προβλήματα στην ανάλυση πολλών μεταβλητών είναι η επιλογή του συνδυασμού των μεταβλητών που παράγει το καλύτερο αποτέλεσμα. Η επιλογή χαρακτηριστικών είναι σημαντική σε:

1. μελέτες συσχέτισης: επιλογή των ανεξάρτητων μεταβλητών που επιτρέπουν την κατασκευή ενός μαθηματικού μοντέλου που να μπορεί να εξηγήσει τη συμπεριφορά μιας εξαρτημένης μεταβλητής
2. μελέτες ταξινόμησης και μοντελοποίησης: επιλογή μεταβλητών που διαχωρίζουν καλύτερα μεταξύ κατηγοριών και επιτρέπουν την κατασκευή ενός μαθηματικού μοντέλου που μπορεί να περιγράψει διαφορετικές κατηγορίες με καλή ειδικότητα και ευαισθησία [27].



Εικόνα 44: *Αναγνώριση προτύπων: ταξινόμηση βάσει των χαρακτηριστικών (τροχοί, μηχανή, τιμόνι, φτερά, χρώμα, μέγεθος), με βέλος υποδεικνύεται το χαρακτηριστικό «τροχοί»*

Σε τομείς όπως η αναγνώριση προτύπων (pattern recognition), η εκμάθηση μηχανής (machine learning) και η εξόρυξη δεδομένων (data mining), έχουν αναπτυχθεί διάφορες τεχνικές επιλογής χαρακτηριστικών.

2.1.1 Εξόρυξη δεδομένων

Ως εξόρυξη δεδομένων (data mining), θεωρείται η εξαγωγή προηγουμένως άγνωστης και δυνητικά χρήσιμης πληροφορίας από μεγάλα σύνολα δεδομένων κατόπιν ανάλυσής τους. Πρακτικά είναι μια διαδικασία ανακάλυψης των προτύπων και των σχέσεων που υπάρχουν σε αυτά τα δεδομένα. Με την ευρύτερη έννοια του όρου «εξόρυξη δεδομένων», αναφερόμαστε

στον τομέα της πληροφορικής που έχει ως σκοπό την εξαγωγή πληροφορίας από ογκώδη σύνολα δεδομένων ή από μεγάλες βάσεις δεδομένων. Στόχος της, είναι η πληροφορία που θα εξαχθεί και τα πρότυπα που θα εντοπιστούν να έχουν δομή κατανοητή προς τον άνθρωπο για να μπορέσουν να τον βοηθήσουν να πάρει τις κατάλληλες αποφάσεις. Συχνά, ο όρος «εξόρυξη δεδομένων» χρησιμοποιείται ως συνώνυμο του όρου «ανακάλυψη γνώσης» (knowledge discovery), όμως για την ακρίβεια, η εξόρυξη δεδομένων είναι κομμάτι της διαδικασίας ανακάλυψης γνώσης. Η επιλογή χαρακτηριστικών αποτελεί σημαντικό εργαλείο εξόρυξης γνώσης. Η εξόρυξη δεδομένων αντλεί υλικό από τη στατιστική, την εκμάθηση μηχανής και τα συστήματα βάσεων δεδομένων.

Εξόρυξη δεδομένων: η διαδικασία που χρησιμοποιείται για την ανακάλυψη κρυμμένης, προηγουμένως άγνωστης και δυνητικά ωφέλιμης πληροφορίας από σύνολα δεδομένων, με τέτοιο τρόπο ώστε αυτή να είναι κατανοητή και εύκολα εκμεταλλεύσιμη για τον κάτοχό της.

Η ανάγκη για εξόρυξη δεδομένων έχει αυξηθεί παράλληλα με την αύξηση του όγκου των συγκεντρωμένων δεδομένων, κάτι που έγινε εφικτό με την ευρεία διαθεσιμότητα και την εξέλιξη της τεχνολογίας των υπολογιστών. Με την ανάπτυξη της τεχνολογίας έγινε οικονομικά και χρονικά συμφέρουσα η αποθήκευση τεράστιου όγκου δεδομένων σε βάσεις δεδομένων. Η αποθήκευση όλων αυτών των δεδομένων γίνεται επειδή διαμέσου των δεδομένων που περιγράφουν ένα σύστημα, μπορείς να το κατανοήσεις καλύτερα και να εξάγεις συμπεράσματα και κανόνες που θα βοηθήσουν στην καλύτερη λειτουργία του. Η ποιότητα της περιγραφής του συστήματος πολλές φορές σχετίζεται άμεσα με το πλήθος των σχετικών με αυτό διαθέσιμων δεδομένων, αφού είναι πιθανόν περισσότερα δεδομένα να εξετάζουν περισσότερες παραμέτρους τους προβλήματος και να δίνουν πληροφορίες για περισσότερες καταστάσεις. Εφόσον σε όλα αυτά τα δεδομένα συχνά κρύβεται ωφέλιμη πληροφορία που θα μπορούσε να βελτιώσει τη λειτουργία του συστήματος οι βάσεις δεδομένων καταλήγουν να αποθηκεύουν τεράστιου μεγέθους δεδομένα.

Ο ρόλος και η χρησιμότητα της εξόρυξης δεδομένων μπορούν να προβληθούν μέσω μιας μεταφοράς: Έστω μια χώρα που έχει μια τεράστια αποκλειστική οικονομική ζώνη (AOZ), δηλαδή θαλάσσια έκταση εντός της οποίας έχει το κράτος της δικαίωμα έρευνας ή άλλης εκμετάλλευσης των θαλάσσιων πόρων. Το κράτος επενδύει στην έρευνα της AOZ της χώρας για ανίχνευση θαλάσσιων πόρων από μια εταιρία, η οποία μετά από μελέτη της εγγυάται πως κάπου μέσα στα έγκατα της θαλάσσιας περιοχής υπάρχουν κοιτάσματα πολύτιμων ορυκτών (φυσικό αέριο, πετρέλαιο κτλ), αλλά αμελεί να της πει πώς θα καταφέρουν να τα εξάγουν και να παράξουν κέρδος στην πράξη. Στην ίδια θέση βρίσκεται και ένας επιχειρηματίας, ή ένας ερευνητής που επενδύει σε μια βάση δεδομένων, η οποία αποθηκεύει όλη τη σχετική με το θέμα που τον ενδιαφέρει πληροφορία. Ενώ γνωρίζει ότι αυτή η πληροφορία είναι δυνητικά ωφέλιμη, δεν γνωρίζει πώς να εξάγει αυτή την ωφέλιμη πληροφορία από τον τεράστιο όγκο δεδομένων και

πώς να την χρησιμοποιήσει προς όφελός του. Η εξόρυξη δεδομένων έρχεται για να βοηθήσει σε ακριβώς αυτό το πρόβλημα. Δημιουργεί τις απαραίτητες «κατάλληλα διαμορφωμένες πλατφόρμες άντλησης ορυκτών κοιτασμάτων» και παρέχει τα κατάλληλα εργαλεία για τη «γεώτρηση» έτσι ώστε να εξαχθούν τα «πολύτιμα ορυκτά» και να δοθούν σε καθαρή μορφή στον ιδιοκτήτη τους.

Η πληροφορία αυτή - τα «πολύτιμα ορυκτά» - μπορεί να βρεθεί με τη μορφή προτύπων, δηλαδή μιας σειράς γεγονότων που γίνονται με συγκεκριμένη συχνότητα που υποδηλώνει μια σχέση, ένα μοτίβο ανάμεσά τους. Φορμαλιστικά ως πρότυπο (pattern) ορίζεται μια μορφή ή ένα σύνολο κανόνων που χρησιμοποιούνται για την παραγωγή γεγονότων από ένα σύνολο δεδομένων. Η εύρεση και η επεξεργασία αυτών των προτύπων είναι μια πολύπλοκη διεργασία εξαιτίας της πληθώρας των δεδομένων, πολλών εκ των οποίων δεν συνεισφέρουν στην αποκάλυψη γνώσης. Η απόκτηση γνώσης (knowledge discovery) επιτυγχάνεται με την αποτίμηση των αποτελεσμάτων και σχέσεων που παράγονται από τις μεθόδους εξόρυξης δεδομένων από ειδικούς στον εκάστοτε τομέα.

Οι διάφορες τεχνικές εξόρυξης δεδομένων εφαρμόζονται ευρέως στην εκμάθηση μηχανής (machine learning). Η εκμάθηση μηχανής αποτελεί υποκλάδο της τεχνητής νοημοσύνης (artificial intelligence- AI) και ασχολείται με την ανάπτυξη αλγορίθμων που μπορούν να «μάθουν» και να κάνουν προβλέψεις από δεδομένα. Η εκμάθηση μηχανής δημιουργεί μοντέλα/εξάγει πρότυπα από σύνολα δεδομένων μέσω ενός υπολογιστικού συστήματος. Η δημιουργία ενός μοντέλου, δηλαδή μιας απλοποιημένης, αφαιρετικής εκδοχής του περιβάλλοντος που μελετάται, ονομάζεται επαγωγική μάθηση. Η επαγωγική μάθηση επιχειρεί μια μετάβαση από το ειδικό στο γενικό: εξάγονται συμπεράσματα από ένα υπάρχον σύνολο δεδομένων με σκοπό να έχουν ισχύ και σε παρόμοια σύνολα δεδομένων που αφορούν το ίδιο είδος προβλήματος. Οι δύο βασικές κατηγορίες εκμάθησης μηχανής είναι η μάθηση με επίβλεψη (supervised learning) και η μάθηση χωρίς επίβλεψη (unsupervised learning). Στη μάθηση με επίβλεψη, το σύστημα καλείται να μάθει μια συνάρτηση από ένα σύνολο δεδομένων η οποία περιγράφει ένα μοντέλο. Διακρίνεται σε δύο είδη προβλημάτων, τα προβλήματα ταξινόμησης (classification, πρόβλεψη διακριτής κατηγορίας) και τα προβλήματα παρεμβολής/παλινδρόμησης (regression). Στη μάθηση χωρίς επίβλεψη, το σύστημα πρέπει μόνο του να ανακαλύψει συσχετίσεις ή ομάδες από ένα σύνολο δεδομένων, δημιουργώντας πρότυπα, χωρίς να είναι γνωστό αν υπάρχουν, πόσα και ποια είναι. Παραδείγματα προτύπων είναι οι κανόνες συσχέτισης (association rules) και οι ομάδες(clusters) οι οποίες προκύπτουν από τη διαδικασία της ομαδοποίησης (clustering, ανάδειξη ομάδων όμοιων δειγμάτων).

Η εξόρυξη δεδομένων και η εκμάθηση μηχανής επειδή χρησιμοποιούν πολλές κοινές μεθόδους επικαλύπτονται αρκετά. Η διάκρισή τους έγκειται στο ότι η εκμάθηση μηχανής

επικεντρώνεται στην πρόβλεψη, η οποία γίνεται με βάση γνωστές ιδιότητες που μαθαίνονται από τα δεδομένα εκπαίδευσης, ενώ η εξόρυξη δεδομένων επικεντρώνεται στην ανακάλυψη προηγουμένως άγνωστων ιδιοτήτων που κρύβονται στα δεδομένα. Η εκμάθηση μηχανής συχνά χρησιμοποιεί μεθόδους εξόρυξης δεδομένων ως ένα βήμα προεπεξεργασίας των δεδομένων για να βελτιώσει με αυτό τον τρόπο την ακρίβεια του συστήματος εκμάθησης.

Η εκμάθηση μηχανής χρησιμοποιεί και πολλές στατιστικές μεθόδους με αποτέλεσμα να παρουσιάζει επικάλυψη σε πολλά σημεία και με την στατιστική. Η στατιστική περιέχει θεωρίες που χρησιμοποιούνται για τη μελέτη δεδομένων και των σχέσεων μεταξύ αυτών, οι οποίες χρησιμοποιούνται και στην εκμάθηση μηχανής και στην εξόρυξη δεδομένων. Η εκμάθηση μηχανής θεωρείται η τομή της στατιστικής και της τεχνητής νοημοσύνης. Σε αντίθεση, όμως με τη στατιστική, στην εκμάθηση μηχανής υπάρχει ιδιαίτερο ενδιαφέρον για τις υπολογιστικές ιδιότητες των χρησιμοποιούμενων στατιστικών μεθόδων. Η υπολογιστική πολυπλοκότητα μιας μεθόδου, δηλαδή πόσο αυξάνει η απαίτηση του αλγορίθμου σε υπολογιστικούς πόρους (χώρο και χρόνο εκτέλεσης) όταν αυξάνει το μέγεθος της εισόδου, αποτελεί βασική παράμετρο στην εκμάθηση μηχανής.

Οι τεχνικές επιλογής χαρακτηριστικών αποτελούν βασικό εργαλείο στην εκμάθηση μηχανής και την εξόρυξη δεδομένων. Εκτός από τις διάφορες ήδη υπάρχουσες τεχνικές επιλογής χαρακτηριστικών που είχαν αναπτυχθεί στους τομείς της αναγνώρισης προτύπων (pattern recognition), της εκμάθησης μηχανής (machine learning) και της εξόρυξης δεδομένων (data mining), νέες τεχνικές έχουν αναπτυχθεί εξαιτίας της ανάγκης εφαρμογής τους σε προβλήματα βιοπληροφορικής (bioinformatics). Η εξόρυξη δεδομένων σε βιολογικά και ιατρικά δεδομένα είναι ιδιαίτερα ενδιαφέρουσα λόγω των σημαντικών συμπερασμάτων που μπορούν να εξαχθούν από τέτοιου είδους δεδομένα. Η επιλογή χαρακτηριστικών είναι ιδιαιτέρως χρήσιμη στον τομέα της ιατρικής, εξαιτίας του μεγάλου μεγέθους των δεδομένων που ανακτώνται από τις διάφορες διαγνωστικές εξετάσεις. Ειδικά όσον αφορά διαγνωστικές εξετάσεις, η δυσκολία που παρουσιάζεται είναι ότι η σχέση μεταξύ του αριθμού των χαρακτηριστικών κάθε δείγματος και του αριθμού των δειγμάτων είναι δυσανάλογη. Αντί να έχουμε μεγάλο αριθμό δειγμάτων και μικρό αριθμό πιθανών χαρακτηριστικών, συνήθως έχουμε μικρό αριθμό δειγμάτων και τεράστιο αριθμό πιθανών χαρακτηριστικών.

Ας πάρουμε για παράδειγμα την επιλογή γονιδίων από δεδομένα μικροσυστοιχιών έτσι ώστε να αποφασιστεί εάν ένας ασθενής είναι υγιής ή όχι. Τα χαρακτηριστικά σε αυτή την περίπτωση είναι οι συντελεστές έκφρασης γονιδίων ανάλογα με την ποσότητα mRNA στο δείγμα ασθενούς. Είναι πιθανό να υπάρχουν μόνο 100 δείγματα ασθενών ενώ τα χαρακτηριστικά μπορεί να φτάνουν σε αριθμό από 6000 έως και 60000 [28]. Δεν σχετίζονται όμως και δεν είναι αναγκαίο να

ελεγχθούν όλα τα χαρακτηριστικά για μια συγκεκριμένη ασθένεια. Γι' αυτό είναι απαραίτητη η μείωση του αριθμού γονιδίων προς εξέταση.

Εικόνα 45: *Δείγμα για εκπαίδευση και αξιολόγηση*

Επιλογή χαρακτηριστικών (feature selection/ FS) είναι η μέθοδος που χρησιμοποιείται για την επιλογή του βέλτιστου ή υποβέλτιστου υποσυνόλου χαρακτηριστικών- ενός συνόλου χαρακτηριστικών που αποδίδουν καλύτερα σε ένα πρόβλημα ταξινόμησης (classification), ομαδοποίησης (clustering) ή παλινδρόμησης (regression) - από μια μεγάλη «δεξαμενή» πιθανών χρήσιμων χαρακτηριστικών (feature pool).

Ταξινόμηση (classification) είναι μια διαδικασία κατάταξης αντικειμένων σε προκαθορισμένες κατηγορίες ανάλογα με τα χαρακτηριστικά τους και υλοποιείται με επιβλεπόμενο αλγόριθμο εκμάθησης (supervised learning).

Ομαδοποίηση (clustering) είναι η οργάνωση μιας συλλογής από αντικείμενα σε ομάδες (clusters) με βάση κάποιο μέτρο ομοιότητας και γίνεται με χρήση αλγόριθμου μη επιβλεπόμενης εκμάθησης (unsupervised learning).

Παλινδρόμηση (regression) είναι μια στατιστική τεχνική μοντελοποίησης για την έρευνα της συσχέτισης μεταξύ μιας εξαρτώμενης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών. Η ανάλυση της παλινδρόμησης μας δίνει πληροφορίες για τη μεταβολή της εξαρτώμενης μεταβλητής όταν μεταβάλλεται μία από τις ανεξάρτητες μεταβλητές, ενώ οι υπόλοιπες ανεξάρτητες διατηρούνται σταθερές, δηλαδή εξακριβώνει την αιτιώδη επίδραση μιας μεταβλητής σε μια άλλη. Όταν χρησιμοποιείται ως τεχνική εξόρυξης δεδομένων έχει ως αποτέλεσμα ένα μοντέλο που χρησιμοποιείται για πρόβλεψη και πρόγνωση των τιμών της εξαρτημένης μεταβλητής για τα νέα δεδομένα.

Στην εικόνα 44 παρουσιάζεται το πρόβλημα της ταξινόμησης διαφόρων αντικειμένων (δειγμάτων) σε 5 κατηγορίες: αυτοκίνητα, αεροπλάνα, πλεούμενα, τρένα, μοτοσυκλέτες. Κάθε αντικείμενο έχει διάφορα γνωρίσματα (χαρακτηριστικά) π.χ. την ύπαρξη ή όχι τροχών, μηχανής, τιμονιού, φτερών, βαγονιών, το χρώμα, το μέγεθος κοκ. Βάσει αυτών των χαρακτηριστικών θα καθοριστεί σε ποια κατηγορία ανήκει κάθε αντικείμενο. Ωστόσο, είναι πιθανό να υπάρχει πληθώρα χαρακτηριστικών, εκ των οποίων μόνο κάποια να είναι σημαντικά (π.χ. ύπαρξη τροχών), εξού και η ανάγκη για επιλογή χαρακτηριστικών.

Ας σημειωθεί ότι ενώ η επιλογή χαρακτηριστικών μπορεί να χρησιμοποιηθεί για την τροφοδότηση διάφορων μοντέλων, για λόγους ευκολίας, θα μιλάμε συνήθως για επιλογή χαρακτηριστικών για ταξινόμηση.

2.2 Κατάρα της διαστασιμότητας (Curse of dimensionality)

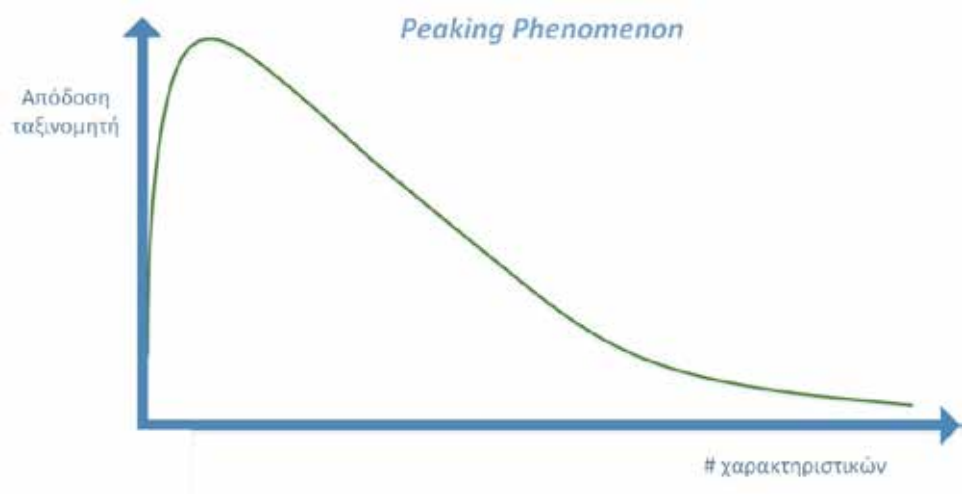
Η κατάρα της διαστασιμότητας (curse of dimensionality) είναι ένα από τα μεγαλύτερα προβλήματα, μετά την υπερπροσαρμογή (overfitting), στον τομέα της εκμάθησης μηχανής (machine learning) και ιδιαίτερα στην αναγνώριση προτύπων. Ωστόσο, επηρεάζει κι άλλα πεδία εκτός από την εκμάθηση μηχανής.

Ο όρος κατάρα της διαστασιμότητας (curse of dimensionality) έχει επινοηθεί από τον R.E. Bellman το 1961, όταν μελετούσε θέματα δυναμικής βελτιστοποίησης. Ο Bellman, όταν αρχικά χρησιμοποίησε τον όρο αυτό, αναφερόταν στην εκθετική αύξηση της υπολογιστικής πολυπλοκότητας που προκαλείται από τη γραμμική αύξηση του αριθμού των διαστάσεων ενός προβλήματος. Με την πάροδο του χρόνου άλλαξε λίγο η έννοια του όρου με αποτέλεσμα να σχετίζεται πλέον περισσότερο με την αραιότητα των δεδομένων στις υψηλές διαστάσεις.

Η κατάρα της διαστασιμότητας αναφέρεται σε φαινόμενα που προκύπτουν όταν αναλύουμε και οργανώνουμε δεδομένα σε χώρους πολλών διαστάσεων - όπως χώροι εκατοντάδων ή ακόμα και χιλιάδων διαστάσεων - φαινόμενα τα οποία, σε χώρους μικρών διαστάσεων, όπως ο τρισδιάστατος χώρος, δεν παρουσιάζονται. Ενώ η ίδια μέθοδος σε χώρο μικρών διαστάσεων εφαρμόζεται χωρίς κανένα πρόβλημα, μόλις αυξηθεί αρκετά η διαστασιμότητα τότε αρχίζουν τα προβλήματα (να σημειώσουμε ότι ο αριθμός των διαστάσεων ενός προβλήματος αντιστοιχεί στον αριθμό των χρησιμοποιούμενων χαρακτηριστικών).

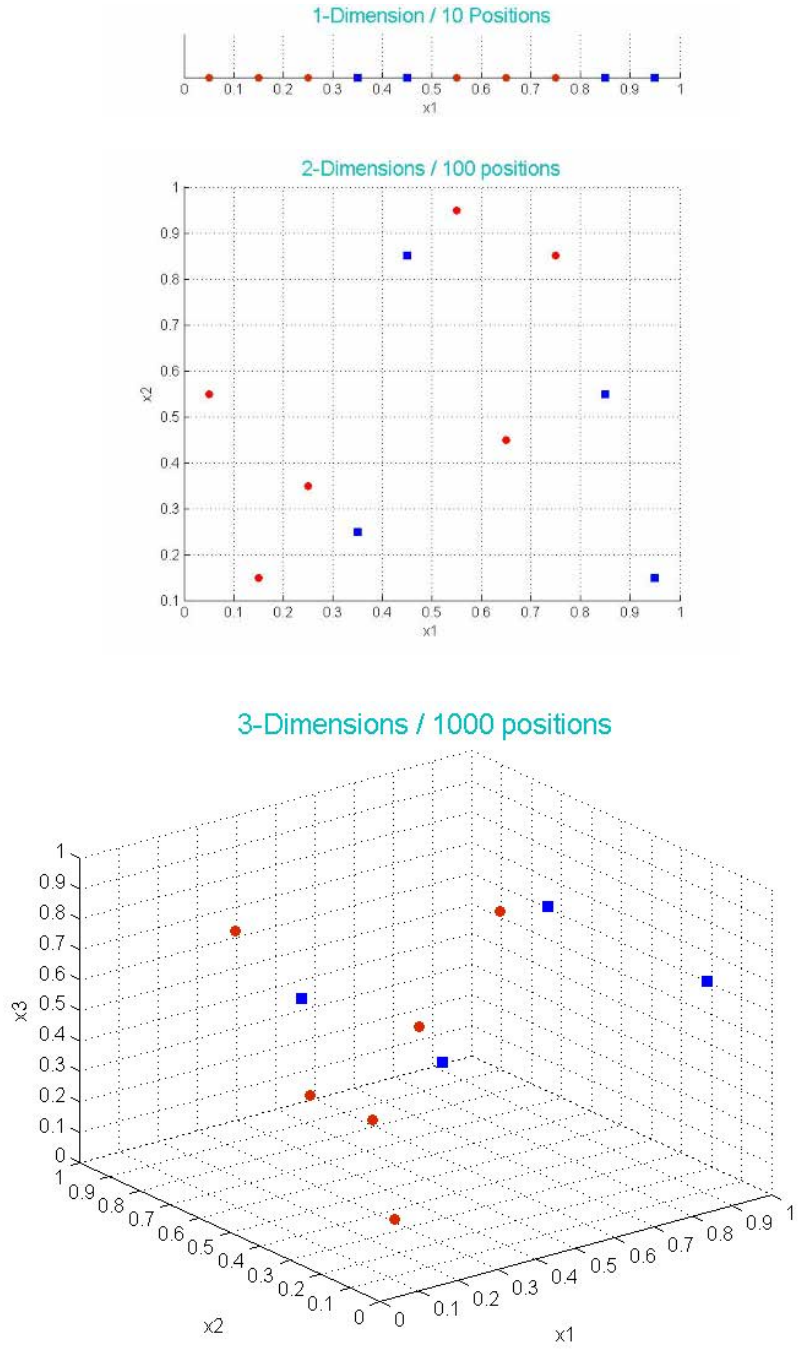
Ειδικά στην αναγνώριση προτύπων, ένα παράδειγμα θα ήταν η απόδοση ενός ταξινομητή σε σχέση με τον αριθμό των χαρακτηριστικών για σταθερό αριθμό παραδειγμάτων εκπαίδευσης. Η δυσκολία που παρουσιάζεται στην εφαρμογή μεθόδων εκμάθησης μηχανής σε δεδομένα ενός πολυδιάστατου χώρου οφείλεται στο ότι καθώς ο όγκος μεταξύ διαφορετικών παραδειγμάτων εκπαίδευσης αυξάνεται ραγδαία, τα δεδομένα γίνονται αραιώς κατανεμημένα και ταξινομούνται δύσκολα. Έτσι με σταθερό αριθμό παραδειγμάτων εκπαίδευσης, η προβλεψιμότητα ενός αλγόριθμου εκμάθησης μειώνεται καθώς η διαστασιμότητα αυξάνεται. Οπότε όσο αυξάνει ο αριθμός των διαστάσεων αυξάνεται εκθετικά και το πλήθος των απαιτούμενων παραδειγμάτων εκπαίδευσης. Το φαινόμενο αυτό είναι γνωστό και ως peaking phenomenon [29] ή ως φαινόμενο Hughes (Hughes effect και το κατέδειξε ο Hughes το 1968 καθώς μελετούσε διακριτή ταξινόμηση.

Στην παρακάτω γραφική παράσταση βλέπουμε ότι με την αύξηση του αριθμού των χαρακτηριστικών, η απόδοση του ταξινομητή αυξάνεται αρχικά ως ένα σημείο, ενώ ακολούθως από το σημείο αυτό και μετά μειώνεται εκθετικά.



Εικόνα 46: *Peaking Phenomenon/Hughes effect*

Πρόβλημα: Με την αύξηση της διαστασιμότητας d, ο όγκος του χώρου αυξάνεται τόσο γρήγορα ώστε τα διαθέσιμα δεδομένα γίνονται αραιά κατανεμημένα (sparse) όπως φαίνεται και στην *εικόνα 47*. Εξαιτίας της αραιότητας των δεδομένων οι μέθοδοι που απαιτούν στατιστική σημαντικότητα δεν επιτυγχάνουν. Για να είναι εφικτό ένα στατιστικά αξιόπιστο αποτέλεσμα με χρήση της ίδιας μεθόδου, το πλήθος δεδομένων απαιτείται να αυξάνεται εκθετικά σε σχέση με τη διαστασιμότητα. Στην *εικόνα 47* βλέπουμε ότι για να προκύψει η ίδια πυκνότητα τιμών δεδομένων στους άξονες για d=1 αρκούν 10 σημεία, για d=2 απαιτούνται 100 σημεία και για d=3 απαιτούνται 1000 σημεία.



Εικόνα 47: *Curse Dimensionality representation*

Η οργάνωση και η αναζήτηση δεδομένων συχνά στηρίζεται στον εντοπισμό περιοχών όπου σχηματίζονται ομάδες με αντικείμενα που έχουν κοινές ιδιότητες. Όταν η διαστασιμότητα είναι μεγάλη , όλα τα αντικείμενα φαίνονται να είναι αραιά κατανεμημένα και με πολλές ανομοιότητες μεταξύ τους με αποτέλεσμα οι συνήθεις στρατηγικές οργάνωσης να είναι λιγότερο αποδοτικές. Επιπλέον όσο μεγαλύτερη είναι η διαστασιμότητα τόσο πιο μεγάλο είναι το υπολογιστικό κόστος.

Η κατάρα της διαστασιμότητας αποτελείται από δύο σκέλη, το ένα είναι τα πολυδιάστατα δεδομένα και το άλλο ο αλγόριθμος που χρησιμοποιείται στο εκάστοτε πρόβλημα. Ένας

αλγόριθμος που χρειάζεται χρόνο ή μνήμη εκθετικά ανάλογο/η του αριθμού των διαστάσεων των δεδομένων προφανώς θα παρουσιάζει προβλήματα σε πολυδιάστατους χώρους. Κατά συνέπεια, ένας τρόπος επίλυσης της κατάρας της διαστασιμότητας είναι η αλλαγή του αλγόριθμου και η επιλογή μιας μεθόδου που να λειτουργεί καλύτερα στις διαστάσεις των δεδομένων του συγκεκριμένου προβλήματος. Ο άλλος τρόπος επίλυσης είναι η προεπεξεργασία των δεδομένων έτσι ώστε οι διαστάσεις των δεδομένων να είναι στα όρια της καλής λειτουργίας του αλγόριθμου που θα χρησιμοποιηθεί. Η μείωση της διαστασιμότητας των δεδομένων (dimension/feature reduction), ωστόσο, πρέπει να γίνει με τέτοιο τρόπο ώστε να μην υπάρχει σημαντική απώλεια πληροφορίας.

Υπάρχουν δύο τρόποι να μειώσουμε τις διαστάσεις ενός διανύσματος χαρακτηριστικών (feature vector):

- Feature extraction (εξαγωγή χαρακτηριστικών): Μετασχηματισμός των υπάρχοντων δεδομένων σε χώρο μικρότερων διαστάσεων – Εξαγωγή m καινούριων χαρακτηριστικών μέσω γραμμικού ή μη γραμμικού συνδυασμού των δοσμένων d χαρακτηριστικών
- Feature selection (επιλογή χαρακτηριστικών): Επιλογή ενός υποσυνόλου μεγέθους m από το σύνολο των υπάρχοντων d χαρακτηριστικών χωρίς την εφαρμογή μετασχηματισμού, όπου $m < d$.

Η επιλογή χαρακτηριστικών μπορεί να θεωρηθεί ως μια ειδική περίπτωση της εξαγωγής χαρακτηριστικών.

2.3 Επιλογή χαρακτηριστικών

2.3.1 Δημιουργία βέλτιστου υποσυνόλου

Η επιλογή χαρακτηριστικών διαφέρει από άλλες μεθόδους μείωσης της διάστασης όπως η εξαγωγή χαρακτηριστικών ως προς το ότι δεν μετασχηματίζει το χώρο των δεδομένων, δεν αλλοιώνει δηλαδή την αρχική παρουσίαση των μεταβλητών, απλά απορρίπτει τις λιγότερο σημαντικές συνιστώσες. Έτσι, διατηρείται η αρχική σημασιολογία επιτρέποντας στον ειδικό του τομέα να μπορέσει να τις ερμηνεύσει.

Σε πρακτικές εφαρμογές, η επιλογή του βέλτιστου υποσυνόλου χαρακτηριστικών δεν είναι εφικτή αφού θα ήταν υπολογιστικά πολύπλοκο να ψάξουμε όλο το χώρο των πιθανών υποσυνόλων χαρακτηριστικών. Για το λόγο αυτό αναζητούμε προσεγγίσεις του βέλτιστου

υποσυνόλου. Το ερώτημα είναι δοσμένου ενός αριθμού χαρακτηριστικών, πως μπορεί κανείς να επιλέξει τα πιο σημαντικά από αυτά, έτσι ώστε να μειώσει τον αριθμό τους και παράλληλα να διατηρήσει όσο το δυνατό περισσότερη χρήσιμη πληροφορία. Το πρόβλημα είναι διττό: από τη μια τα χαρακτηριστικά με την περισσότερη πληροφορία θα πρέπει να βρεθούν και να χρησιμοποιηθούν και από την άλλη τα άσχετα χαρακτηριστικά ή χαρακτηριστικά που μοιράζονται τις ίδιες πληροφορίες πρέπει να αποφευχθούν. Η χρήση άσχετων ή περιττών χαρακτηριστικά απλά αυξάνει τη διαστασιμότητα του χώρου χαρακτηριστικών και κατ' επέκταση και την πολυπλοκότητα του προβλήματος χωρίς να προσφέρει κάποιο κέρδος στην απόδοση. Συχνά, χαρακτηριστικά που αποδίδουν χαμηλά ξεχωριστά, συνδυασμένα με άλλα δίνουν πολύ καλά αποτελέσματα.

Συναφή χαρακτηριστικά (Relevant Features): Δίνουν πληροφορία μόνα τους ή σε συνδυασμό με άλλα.

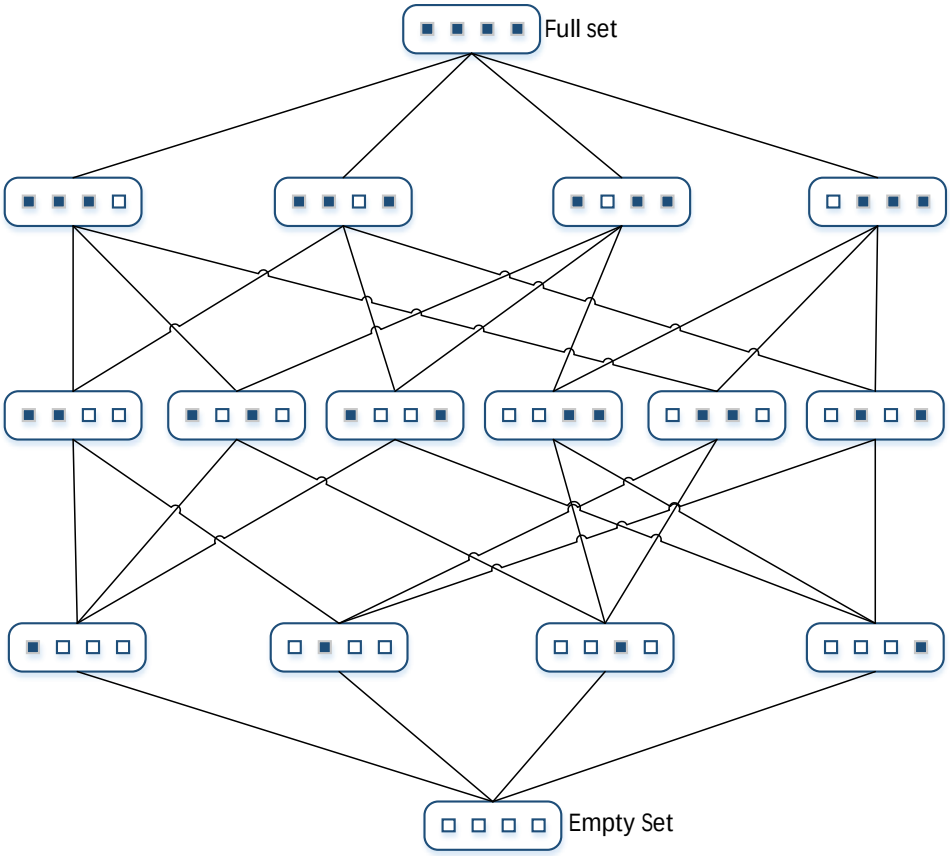
Περιττά/Πλεονάζοντα χαρακτηριστικά (Redundant features): Μπορούν να αφαιρεθούν επειδή υπάρχει κι άλλο υποσύνολο χαρακτηριστικών που δίνει την ίδια πληροφορία. Η προσθήκη τους απλά υποβαθμίζει την απόδοση. Για παράδειγμα, παρόλο που δύο χαρακτηριστικά μπορεί να φέρουν σημαντική πληροφορία όταν μελετώνται ξεχωριστά, όταν συνδυαστούν μπορεί να αποδίδουν μικρό κέρδος εξαιτίας της υψηλής μεταξύ τους συσχέτισης. Έτσι αυξάνεται η πολυπλοκότητα χωρίς την ανάλογη αύξηση του κέρδους.

Άσχετα χαρακτηριστικά (Irrelevant features): Δεν προσφέρουν χρήσιμη πληροφορία ούτε μόνα τους ούτε σε συνδυασμό με άλλα και δεν πρέπει να περιλαμβάνονται στο υποσύνολο.

Θα μπορούσαμε να ορίσουμε το πρόβλημα FS στην ταξινόμηση ως εξής: Δεδομένου ενός συνόλου N χαρακτηριστικών, επέλεξε το βέλτιστο υποσύνολο από *l* χαρακτηριστικά που βελτιστοποιεί ένα κριτήριο / μια αντικειμενική συνάρτηση αξιολόγησης: π.χ. μεγιστοποιεί την ακρίβεια ταξινόμησης (classification accuracy) ή ελαχιστοποιεί το σφάλμα ταξινόμησης (classification error). Ο στόχος της FS είναι να καθορίσει το ελάχιστο, αλλά επαρκές υποσύνολο χαρακτηριστικών που θα καταλήξει σε σχεδόν την ίδια πιθανοτική κατανομή των κλάσεων με αυτή που θα προέκυπτε αν χρησιμοποιούνταν όλα τα χαρακτηριστικά. Ωστόσο, επειδή ενδιαφερόμαστε για χαρακτηριστικά που μπορεί να έχουν υψηλή αμοιβαία συσχέτιση, ιδανικά θα έπρεπε να σχηματίσουμε όλους τους πιθανούς συνδυασμούς διανυσμάτων *l* χαρακτηριστικών από τα N χαρακτηριστικά που ήταν αρχικά διαθέσιμα και ακολούθως να αξιολογήσουμε το κάθε ένα από αυτά με στόχο να επιλέξουμε το καλύτερο διάνυσμα χαρακτηριστικών (feature vector selection, wrapper approach).

Η FS μπορεί να θεωρηθεί ως ένα πρόβλημα αναζήτησης, όπου κάθε κατάσταση στο χώρο αναζήτησης αντιπροσωπεύει ένα υποψήφιο υποσύνολο χαρακτηριστικών. Έστω ένα αρχικό σύνολο 4 χαρακτηριστικών (F1,F2,F3,F4) και το αντίστοιχο δυαδικό διάνυσμα 4 διαστάσεων

(1,1,1,1), όπου η τιμή κάθε θέσης του διανύσματος δείχνει εάν το αντίστοιχο χαρακτηριστικό θα επιλεγεί. Επομένως το διάνυσμα (1,0,0,0) υποδεικνύει ότι θα επιλεγεί μόνο το πρώτο χαρακτηριστικό F1, ενώ το (1,1,1,1) σημαίνει ότι θα επιλεγούν και τα 4 χαρακτηριστικά. Για n χαρακτηριστικά ο συνολικός αριθμός των πιθανών υποσυνόλων αγνοώντας το διάνυσμα (0,0,0,0) είναι 2^n-1 . Άρα, στο παράδειγμά μας, όπως φαίνεται και στην *εικόνα 48,* προκύπτουν 2^4-1 =15 πιθανά υποσύνολα (εκτός του κενού συνόλου). Όπως μπορούμε να παρατηρήσουμε ο αριθμός των υποσυνόλων αυξάνεται εκθετικά όσο αυξάνεται ο αριθμός των χαρακτηριστικών. Εξαιτίας αυτού, αλγόριθμοι brute force που κάνουν εξαντλητική αναζήτηση δεν μπορούν να χρησιμοποιηθούν [30].



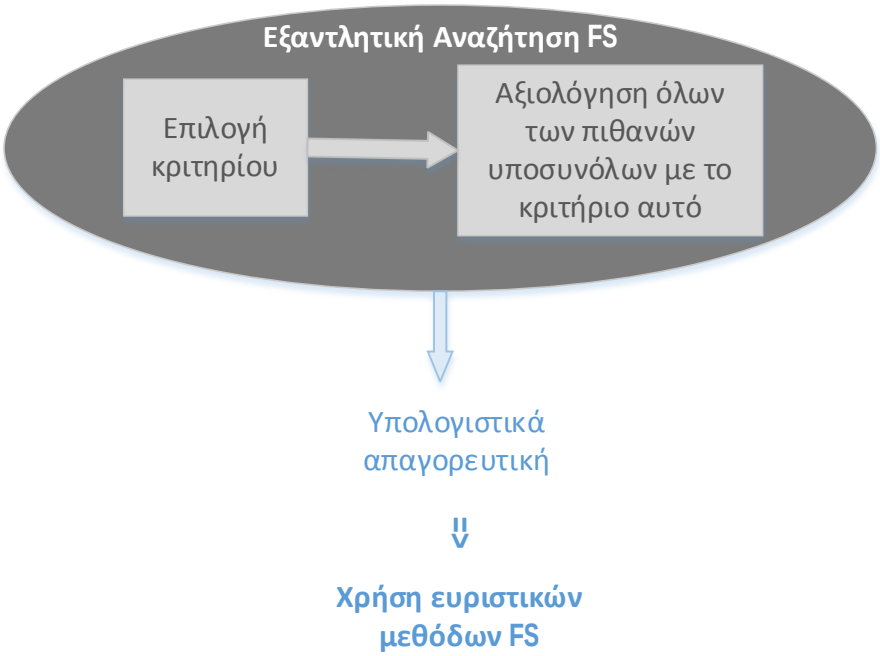
Εικόνα 48: *FS ως πρόβλημα αναζήτησης: Παράδειγμα 4 χαρακτηριστικών*

Εύρεση του αριθμού όλων των υποσυνόλων συγκεκριμένου μεγέθους: Ο συνολικός αριθμός διανυσμάτων με *l* χαρακτηριστικά, επιλεγμένα από το αρχικό σύνολο χαρακτηριστικών μεγέθους N, όπου *l* < N, υπολογίζεται από τον ακόλουθο τύπο:

$$\binom{N}{l} = \frac{N!}{l!\,(N-l)!}$$

Στο προηγούμενο παράδειγμα δηλαδή (*εικόνα 48*), το πλήθος των υποσυνόλων που περιέχουν 3 χαρακτηριστικά ισούται με $\binom{4}{3} = \frac{4!}{3!(4-3)!} = 4$, ενώ το πλήθος των υποσυνόλων που περιέχουν 2 χαρακτηριστικά ισούται με $\binom{4}{2} = \frac{4!}{2!(4-2)!} = 6$.

Επειδή η εξαντλητική αναζήτηση όλων των υποψήφιων διανυσμάτων, παρόλο που θα έδινε το βέλτιστο αποτέλεσμα (optimal FS), είναι υπολογιστικά απαγορευτική, χρησιμοποιούνται ντετερμινιστικές ή στοχαστικές μέθοδοι αναζήτησης, όπως π.χ. ευριστικές, για την αποφυγή της εξαντλητικής αναζήτησης. Οι ευριστικές μέθοδοι ανήκουν στην υποβέλτιστη επιλογή χαρακτηριστικών (suboptimal FS).



Εικόνα 49: *Curse Dimensionality representation*

Όπως έχουμε δει στην *εικόνα 48* υπάρχουν δύο ακραίες περιπτώσεις υποσυνόλων: αυτού με όλα τα χαρακτηριστικά επιλεγμένα και του κενού υποσυνόλου. Το βέλτιστο υποσύνολο κατά πάσα πιθανότητα βρίσκεται κάπου ανάμεσα σε αυτές τις δύο περιπτώσεις. Για την εύρεσή του τίθεται το ερώτημα από πού να ξεκινήσουμε την αναζήτηση: από το κενό υποσύνολο ή το πλήρες. Εάν οι διαστάσεις του προβλήματος είναι μικρές, οπουδήποτε κι αν ξεκινήσουμε θα καταλήξουμε πολύ γρήγορα στο βέλτιστο υποσύνολο. Εάν όμως, έχουμε μεγάλες διαστάσεις (N>20), όπως γίνεται συνήθως, τότε το σημείο εκκίνησης της αναζήτησης έχει ιδιαίτερη σημασία αφού επηρεάζει καθοριστικά το υπολογιστικό κόστος αναζήτησης [30]. Για το σκοπό αυτό υπάρχουν διάφορες στρατηγικές αναζήτησης οι οποίες θα αναλυθούν παρακάτω.

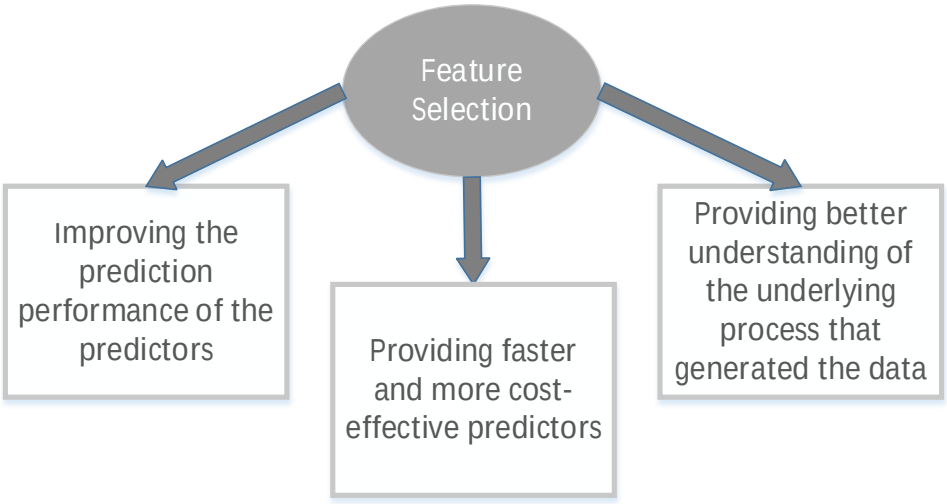
2.3.2 Πλεονεκτήματα επιλογής χαρακτηριστικών

Εκτός από το πρωταρχικό μείζονος σημασίας πλεονέκτημα που είναι η βελτίωση της απόδοσης μέσω της αποφυγής της κατάρας της διαστασιμότητας είναι πιθανό να προκύψουν κι άλλα οφέλη μέσω της επιλογής χαρακτηριστικών. Η μείωση των χαρακτηριστικών μπορεί να διευκολύνει την ερμηνεία των δεδομένων αφού μια απεικόνιση που ορίζεται βάσει λίγων χαρακτηριστικών (data visualization) είναι πιο εύκολα ερμηνεύσιμη συγκριτικά με μία απεικόνιση που ορίζεται βάσει πολλών χαρακτηριστικών.

Επιπλέον, υπάρχει κέρδος στη μείωση του υπολογιστικού κόστους με αποτέλεσμα τη δημιουργία πιο γρήγορων και πιο αποτελεσματικών μοντέλων. Συγκεκριμένα στους ταξινομητές, χρησιμοποιώντας λιγότερα χαρακτηριστικά, η εκπαίδευσή τους, καθώς και η διαδικασία κατάταξης γίνονται πολύ πιο γρήγορα και παράλληλα μειώνονται και οι απαιτήσεις σε μνήμη.

Επιπρόσθετα, με τη διάκριση των πιο σημαντικών χαρακτηριστικών ως προς το αποτέλεσμα μιας διαδικασίας, είναι πιθανό να αποκτηθεί καλύτερη διαίσθηση για το πραγματικό πρόβλημα, δίνοντας έτσι τη δυνατότητα στους ειδικούς του τομέα να το αντιμετωπίσουν πιο αποτελεσματικά. Το πλεονέκτημα αυτό είναι ιδιαίτερα χρήσιμο σε προβλήματα βιοπληροφορικής όπου για παράδειγμα δίνεται η δυνατότητα να αναγνωριστούν γονίδια που σχετίζονται με διάφορες νόσους.

Τέλος, σε προβλήματα αναγνώρισης προτύπων, εκτός από τη βελτίωση της απόδοσης βοηθά και στην αποφυγή της υπερπροσαρμογής (overfitting).



Εικόνα 50: *Feature selection objective*

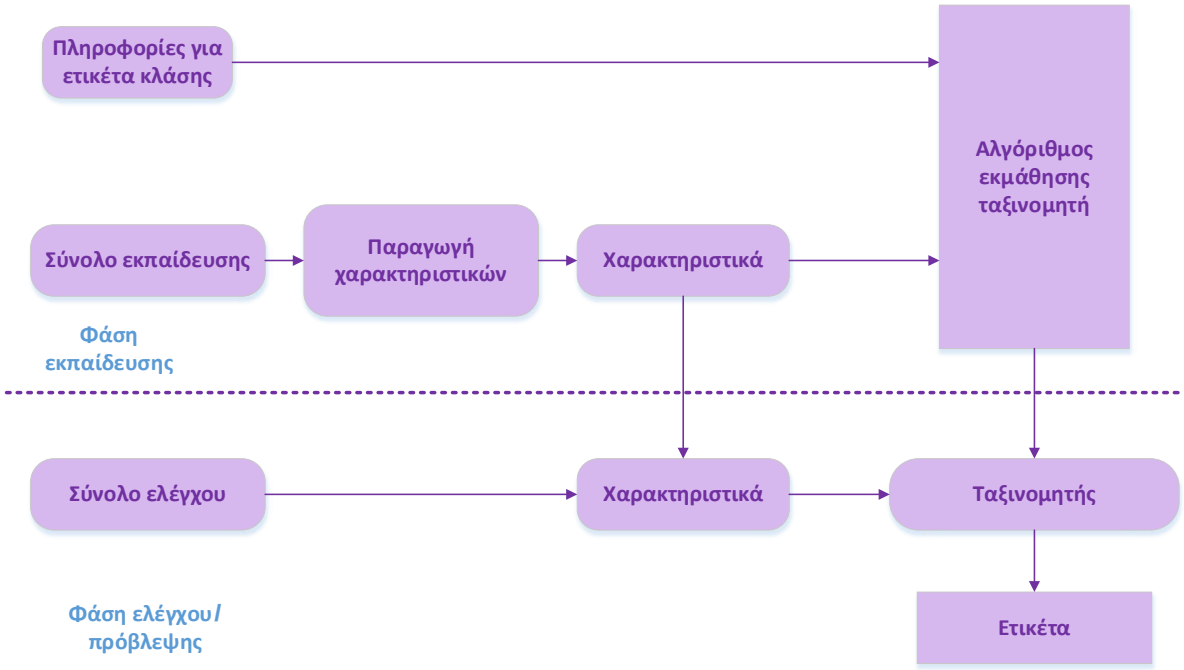
2.4 Επιλογή χαρακτηριστικών σε προβλήματα ταξινόμησης

2.4.1 Ταξινόμηση

Η ταξινόμηση (classification) είναι μια διαδικασία κατάταξης αντικειμένων σε κατηγορίες ανάλογα με τα χαρακτηριστικά τους. Για να εξεταστεί ένα αντικείμενο, αρχικά μελετώνται τα ιδιαίτερά του χαρακτηριστικά και στη συνέχεια γίνεται η κατάλληλη κατηγοριοποίησή του. Η ταξινόμηση κατατάσσει τα δεδομένα ανάλογα με τις κοινές τους ιδιότητες σύμφωνα με ένα προκαθορισμένο μοντέλο, αφού εντοπίσει τους συσχετισμούς ανάμεσά τους.

Κάθε ταξινομητής περιλαμβάνει δύο φάσεις: τη φάση εκπαίδευσης και τη φάση ελέγχου. Στη φάση εκπαίδευσης χρησιμοποιείται ένα σύνολο από δείγματα (training set) ώστε να «εκπαιδευτεί» ο ταξινομητής, δηλαδή να εντοπιστούν οι βέλτιστες παράμετροι για τη λειτουργία του. Ένα δείγμα εκπαίδευσης περιγράφεται από ένα διάνυσμα τιμών κάποιων χαρακτηριστικών και μια ετικέτα κλάσης. Για παράδειγμα, σε ένα πρόβλημα ιατρικής διάγνωσης, τα χαρακτηριστικά μπορεί να είναι η ηλικία, το βάρος, η ομάδα αίματος και η αρτηριακή πίεση ενός ασθενή, ενώ η ετικέτα της κλάσης μπορεί να υποδεικνύει αν ο ασθενής πάσχει από μια ασθένεια, π.χ. κάποια καρδιακή πάθηση, ή όχι. Επειδή ακριβώς η ταξινόμηση απαιτεί τη χρήση ενός συνόλου δεδομένων για εκπαίδευση, ανήκει στην κατηγορία αλγορίθμων επιβλεπόμενης μάθησης. Ο σκοπός του επαγωγικού αλγόριθμου εκμάθησης είναι να δημιουργήσει ένα ταξινομητή που θα είναι χρήσιμος να ταξινομήσει μελλοντικά δείγματα. Ο ταξινομητής είναι μια απεικόνιση από το χώρο των τιμών των χαρακτηριστικών στο σύνολο των τιμών των κλάσεων. Στη δεύτερη φάση, τη φάση ελέγχου, γίνεται έλεγχος της απόδοσης του ταξινομητή σε ένα άλλο σύνολο δεδομένων (test/testing set) διαφορετικό από το σύνολο εκπαίδευσης. Η ακρίβεια του ταξινομητή μπορεί να εκτιμηθεί χρησιμοποιώντας διάφορες τεχνικές εκτίμησης της ακρίβειας [31].

Στην *εικόνα 51* φαίνεται η γενική διαδικασία της ταξινόμησης δεδομένων χωριζόμενη στις δύο προαναφερθείσες φάσεις. Κατά την παραγωγή των χαρακτηριστικών, τα δεδομένα αναλύονται ώστε να προκύψει ένα σύνολο χαρακτηριστικών, βάσει κάποιων μοντέλων παραγωγής χαρακτηριστικών. Τα χαρακτηριστικά μπορεί να παίρνουν ως τιμή μια κατηγορία (π.χ. για ομάδα αίματος : “Α”,”Β”, “ΑΒ”, “Ο”), ένα μέγεθος (π.χ. «μεγάλο», «μέτριο», «μικρό»), μια τιμή ακεραίου (π.χ. για ηλικία), μια τιμή πραγματικού αριθμού (π.χ. για ύψος). Αφού αναπαρασταθούν τα δεδομένα ως αυτά τα χαρακτηριστικά, ο αλγόριθμος εκμάθησης θα χρησιμοποιήσει την ετικέτα της κλάσης που ανήκει κάθε δείγμα, καθώς και χαρακτηριστικά κάθε δείγματος ώστε να εκπαιδευτεί ο ταξινομητής.



Εικόνα 51: *Η γενική διαδικασία ταξινόμησης δεδομένων*

Το πρόβλημα της ταξινόμησης μπορεί να περιγραφεί ως εξής: Δοσμένου ενός συνόλου δεδομένων $X=\{x_1,x_2,\ldots,x_n\}$ -όπου κάθε διάνυσμα $x_i \in X$ αποτελεί την αναπαράσταση ενός αντικειμένου ως προς κάποια χαρακτηριστικά- και ενός συνόλου κλάσεων $C=\{c_1,c_2,\ldots,c_m\}$ –όπου C ένα σύνολο διακριτών τιμών- πρέπει να καθοριστεί σε ποια κλάση $c_j \in C$ (για ακρίβεια σε ποια ετικέτα κλάσης) ανήκει κάθε ένα από τα στοιχεία του συνόλου X. Η κατάταξη των δεδομένων σε κατηγορίες ορίζεται μέσω της απεικόνισης $f: X \rightarrow C$, που σημαίνει ότι κάθε χαρακτηριστικό x_i αντιστοιχεί σε μια κλάση c_j. Στόχος μας είναι να εκτιμηθεί η *f* από ένα σύνολο παραδειγμάτων. Τα παραδείγματα είναι διανύσματα $x_1,x_2,\ldots,x_k \in X$ που έχουν ήδη κατηγοριοποιηθεί στις κλάσεις $c_1,c_2,\ldots,c_l \in C$. Το σύνολο των παραδειγμάτων αυτών $D=\{(x_1,c_1),\ldots,(x_k, c_l)\}$ ονομάζεται σύνολο εκπαίδευσης (training set). Με χρήση του συνόλου D επιχειρείται η δημιουργία μιας απεικόνισης $g: X \rightarrow Y$ η οποία να είναι μια όσο το δυνατό καλύτερη προσέγγιση της *f*. Τη δημιουργία της *g* υλοποιεί ένας αλγόριθμος επαγωγής (induction algorithm). Ένα σύνολο παραμέτρων ϑ καθορίζει το τι θα κάνει η *g*. Ο αλγόριθμος επαγωγής επιλέγει τις παραμέτρους ϑ έχοντας στόχο η *g* να κατατάσσει όσο πιο σωστά γίνεται τα παραδείγματα του συνόλου D. Η απεικόνιση *g* ονομάζεται ταξινομητής (classifier) και η διαδικασία επιλογής των κατάλληλων παραμέτρων ονομάζεται εκπαίδευση του ταξινομητή. Οι αλγόριθμοι επαγωγής που χρησιμοποιούνται για την ταξινόμηση αναφέρονται επίσης ως ταξινομητές.

Ταξινομητές: μια πλειάδα μεθόδων που αποσκοπούν στην ταξινόμηση δεδομένων. Είναι αλγόριθμοι που δέχονται ως είσοδο (input) άγνωστα δεδομένα (πρότυπα) σε μορφή διανύσματος

(vector) και δίνουν ως έξοδο (output) την κλάση στην οποία ανήκει το καθένα. Υπάρχουν δύο ειδών ταξινομητές: οι δυαδικοί οι οποίοι ταξινομούν τα δεδομένα σε δύο κλάσεις και οι ταξινομητές πολλών κλάσεων που κατηγοριοποιούν τα δεδομένα σε περισσότερες από δύο κλάσεις.

Μετά την επιλογή των παραμέτρων ϑ, γίνεται έλεγχος για να εξακριβωθεί πόσο καλή προσέγγιση της f είναι η απεικόνιση g. Αυτό γίνεται με τη χρήση κάποιων παραδειγμάτων, του λεγόμενου συνόλου ελέγχου (test set), των οποίων η κλάση είναι ήδη γνωστή. Ελέγχεται αν η g κατατάσσει ορθά τα παραδείγματα αυτά σε κλάσεις μέσω ενός μέτρου αξιολόγησης, της ακρίβειας ταξινόμησης (classification accuracy), που είναι το ποσοστό των σωστών κατατάξεων προς τον αριθμό των παραδειγμάτων του συνόλου ελέγχου. Είναι μείζονος σημασίας τα παραδείγματα του συνόλου ελέγχου να μην έχουν χρησιμοποιηθεί στο σύνολο εκπαίδευσης D επειδή τα παραδείγματα του συνόλου εκπαίδευσης έχουν ήδη χρησιμοποιηθεί για τον προσδιορισμό των παραμέτρων ϑ και είναι επόμενο ότι η g θα τα κατατάξει σωστά κι έτσι η μέτρηση της ακρίβειας ταξινόμησης δεν θα είναι αντιπροσωπευτική της απόδοσης του ταξινομητή. Συνεπώς το σύνολο ελέγχου απαρτίζεται από παραδείγματα τα οποία δεν χρησιμοποιήθηκαν κατά το στάδιο της εκπαίδευσης.

Είναι πιθανό, η απεικόνιση g, που προκύπτει από το στάδιο της εκπαίδευσης, να κατατάσσει ορθά τα παραδείγματα εκπαίδευσης αλλά να μην έχει καλή ικανότητα γενίκευσης: όταν δηλαδή επιχειρεί να κατατάξει τα παραδείγματα του συνόλου ελέγχου το σφάλμα ταξινόμησης (classification error) είναι μεγάλο. Κάποια άλλη απεικόνιση g' μπορεί να μην κατατάσσει τόσο καλά τα παραδείγματα εκπαίδευσης, αλλά να έχει καλύτερη ικανότητα γενίκευσης. Το φαινόμενο αυτό καλείται υπερπροσαρμογή ή υπερεκπαίδευση (overfitting) και παρουσιάζεται όταν ένας ταξινομητής προσαρμόζεται πολύ καλά στα δείγματα του συνόλου εκπαίδευσης, σε σημείο που επηρεάζεται από τον τυχαίο θόρυβο που υπάρχει σ' αυτά.

Για την αντιμετώπιση του φαινομένου της υπερπροσαρμογής σε προβλήματα ταξινόμησης, ιδιαίτερα στις περιπτώσεις που ο αριθμός των δειγμάτων του συνόλου εκπαίδευσης είναι μικρός σε σύγκριση με τον αριθμό των χαρακτηριστικών κάθε δείγματος που πρέπει να κατηγοριοποιηθεί, είναι απαραίτητη η χρήση τεχνικών επιλογής χαρακτηριστικών. Στις περιπτώσεις αυτές, μια αποτελεσματική ταξινόμηση μπορεί να γίνει μόνο μέσω της επιλογής κατάλληλου υποσυνόλου χαρακτηριστικών.

2.4.2 Επιλογή χαρακτηριστικών για ταξινόμηση

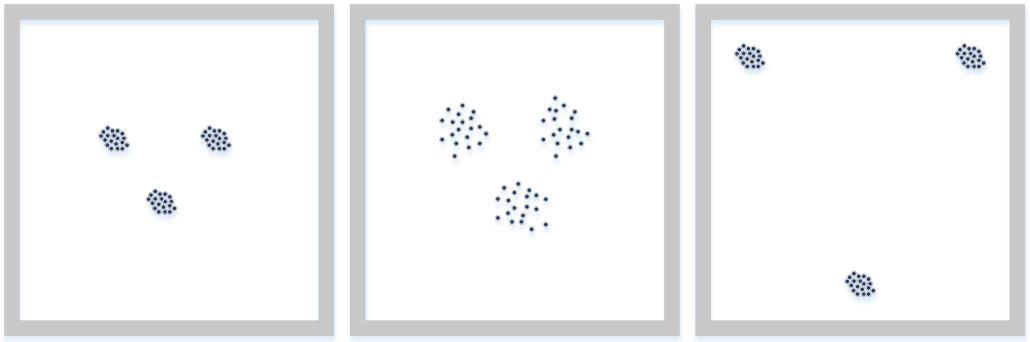
Η επιλογή χαρακτηριστικών εφαρμόζεται τόσο σε επιβλεπόμενη (supervised learning) όσο και σε μη επιβλεπόμενη εκμάθηση (unsupervised learning). Ιδιαίτερα όμως, στο πρόβλημα της επιβλεπόμενης εκμάθησης , δηλαδή σε προβλήματα ταξινόμησης, προσφέρει πολλά οφέλη και είναι απαραίτητη. Όπως έχει προαναφερθεί η κατάρα της διαστασιμότητας/φαινόμενο Hughes είναι βασικό πρόβλημα στην ταξινόμηση: για σταθερό αριθμό δειγμάτων εκπαίδευσης, ο βαθμός ταξινόμησης (classification rate) του ταξινομητή μειώνεται μετά από ένα σημείο (peak) εξαιτίας υπερπροσαρμογής (overfitting) (*εικόνα 46*). Αυτό καθιστά αναγκαία τη χρήση τεχνικών επιλογής χαρακτηριστικών σε προβλήματα ταξινόμησης.

Το κύριο όφελος που προσφέρει είναι ότι βελτιώνει τις απαραίτητες ικανότητες γενίκευσης που θα πρέπει να έχει ο ταξινομητής, αποφεύγοντας κατ' ακολουθία την υπερπροσαρμογή (overfitting) και βελτιώνοντας την απόδοσή του, δηλαδή την ικανότητα προβλεψιμότητάς του. Κατά τη σχεδίαση ενός ταξινομητή βασικό σχεδιαστικό βήμα είναι η επιλογή του μεγέθους του, δηλαδή του αριθμού των παραμέτρων εισόδου του. Το μέγεθος αυτό δεν μπορεί να ξεπερνά ένα όριο επειδή αυτό θα επιδρούσε αρνητικά στις ικανότητες γενίκευσης του ταξινομητή, στον τρόπο, δηλαδή, που χειρίζεται νέα δεδομένα εκτός αυτών που χρησιμοποιήθηκαν για την εκπαίδευσή του. Με την αύξηση του μεγέθους ο ταξινομητής τείνει να προσαρμόζεται στις τιμές που χρησιμοποιήθηκαν κατά την εκπαίδευσή του με αποτέλεσμα να αυξάνεται το σφάλμα ταξινόμησης των νέων δεδομένων. Ιδιαίτερα στην περίπτωση που τα δεδομένα εκπαίδευσης είναι λίγα, η χρήση μικρότερου αριθμού χαρακτηριστικών μπορεί να μειώσει τον κίνδυνο υπερπροσαρμογής των παραμέτρων του ταξινομητή στα δεδομένα εκπαίδευσης [32]. Έστω N ο αριθμός προτύπων εκπαίδευσης του ταξινομητή και K ο αριθμός των ελεύθερων παραμέτρων ταξινόμησης. Όσο μεγαλύτερος ο λόγος N/K τόσο καλύτερη είναι η ικανότητα γενίκευσης του ταξινομητή. Αύξηση των χαρακτηριστικών αντιστοιχεί σε αύξηση του K. Συνεπώς, για πεπερασμένο και μικρό συνήθως N, η μείωση των χαρακτηριστικών συνδράμει στη σχεδίαση ενός ταξινομητή με υψηλές ικανότητες γενίκευσης.

Επιπλέον, ένα άλλο βασικό βήμα για τη σχεδίαση του ταξινομητή είναι η αποτίμηση της λειτουργίας του, εκτιμώντας την πιθανότητα εσφαλμένης ταξινόμησης. Η πιθανότητα σφάλματος μειώνεται όσο αυξάνεται ο λόγος του αριθμού των προτύπων εκπαίδευσης ως προς τη διαστασιμότητα N/d, δηλαδή όσο μειώνεται ο αριθμός των χαρακτηριστικών.

Επιπρόσθετα, μια επιτυχής επιλογή χαρακτηριστικών που προσεγγίζει το βέλτιστο υποσύνολο χαρακτηριστικών έχει ως συνεπακόλουθο την απλοποίηση της σχεδίασης του ταξινομητή. Αν επιλεγούν χαρακτηριστικά μικρής διακριτικής ικανότητας η απόδοση του ταξινομητή θα είναι πολύ χαμηλή, ενώ αν επιλεγούν χαρακτηριστικά με μεγάλη διακριτική ικανότητα που φέρουν

μεγάλη ποσότητα πληροφορίας, το κέρδος μείωσης της πολυπλοκότητας του ταξινομητή θα είναι σημαντικό. Επομένως, είναι ζωτικής σημασίας η εφαρμογή μιας καλής επιλογής χαρακτηριστικών που θα στοχεύει στην επιλογή χαρακτηριστικών που έχουν μεγάλη απόσταση μεταξύ διαφορετικών κλάσεων (large between-class distance) αλλά ταυτόχρονα μικρή διακύμανση στο εσωτερικό μιας κλάσης (small within-class variation). Δηλαδή, επιδιώκεται τα χαρακτηριστικά που επιλέγονται να λαμβάνουν απομακρυσμένες τιμές όταν ανήκουν σε διαφορετικές κλάσεις και κοντινές τιμές όταν βρίσκονται στην ίδια κλάση. Ακολουθούνται διάφορες στρατηγικές για να επιτευχθεί αυτό. Μια στρατηγική είναι να εξεταστούν τα χαρακτηριστικά ένα προς ένα μέσω ενός μέτρου διαχωρισιμότητας των κλάσεων και να απορριφθούν αυτά που παρουσιάζουν μικρή διακριτική ικανότητα (filter approach). Μια εναλλακτική στρατηγική είναι να εξεταστούν διάφοροι συνδυασμοί χαρακτηριστικών και να ελεγχθεί ποιοι συνδυασμοί καταλήγουν σε καλύτερη απόδοση, ανεξάρτητα από τη ξεχωριστή διακριτική ικανότητα των επιμέρους χαρακτηριστικών (wrapper approach). Οι στρατηγικές αυτές αναλύονται περαιτέρω ακολούθως. Για καλύτερη διακριτική ικανότητα ενός χαρακτηριστικού κάποτε είναι χρήσιμη η εφαρμογή γραμμικού ή μη γραμμικού μετασχηματισμού σε αυτό.



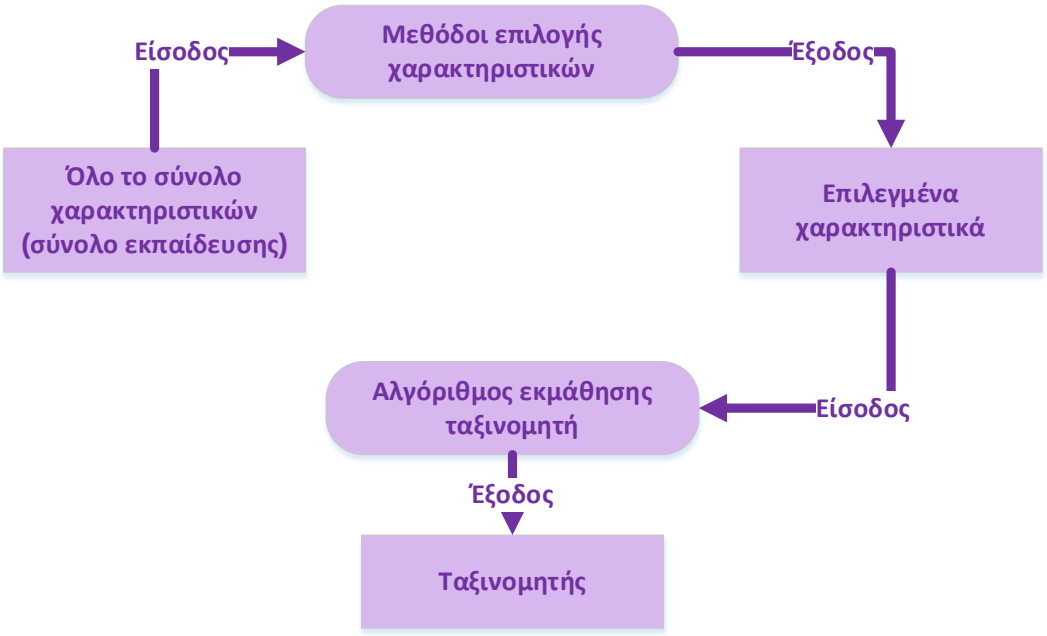
Εικόνα 52: *(α) small within-class variation and small between-class distance (β) large within-class variation and small between-class distance (γ) small within-class variation and large between-class distance*

Η χρήση μικρότερου υποσυνόλου χαρακτηριστικών βελτιώνει την ακρίβεια ταξινόμησης εξαλείφοντας τα χαρακτηριστικά που εισάγουν θόρυβο στα δεδομένα. Επιπλέον η χρήση μικρού συνόλου χαρακτηριστικών αυξάνει την αξιοπιστία της εκτιμώμενης απόδοσης του ταξινομητή [33]. Άλλο ένα πλεονέκτημα που παρέχει η επιλογή χαρακτηριστικών στην ταξινόμηση είναι ότι η εύρεση των χαρακτηριστικών με την περισσότερη πληροφορία και του τρόπου με τον οποίο αλληλεπιδρούν - μέσω του ταξινομητή – μπορεί να προσφέρει γνώσεις σχετικά με τον τομέα του προβλήματος.

Σε σχέση με το υπολογιστικό κόστος, εφόσον κάθε χαρακτηριστικό που χρησιμοποιείται στην διεργασία της ταξινόμησης μπορεί να αυξήσει το κόστος και το χρόνο εκτέλεσης του συστήματος, όταν τα καλύτερα χαρακτηριστικά για ένα συγκεκριμένο ταξινομητή έχουν αναγνωριστεί, ο χρόνος και το κόστος υπολογισμού που απαιτείται για την ταξινόμηση θα μειωθούν [34]. Αν η ανάπτυξη του ταξινομητή εμπλέκει αναζήτηση ή έναν αλγόριθμο εκμάθησης, η μείωση των χαρακτηριστικών θα μειώσει επίσης και το χώρο αναζήτησης που θα χρειάζεται να εξερευνηθεί από τον αλγόριθμο εκμάθησης και συνεπώς θα μπορεί να μειώσει και το χρόνο που απαιτείται για την εκμάθηση μιας επαρκώς ακριβούς συνάρτησης ταξινόμησης [32].

Επιγραμματικά, η μείωση της διαστασιμότητας που μπορεί να επιτευχθεί μέσω της επιλογής χαρακτηριστικών σε προβλήματα ταξινόμησης είναι αναγκαία, ώστε να μειωθεί η πολυπλοκότητα του ταξινομητή και το υπολογιστικό κόστος, να επιταχυνθεί η διαδικασία εκπαίδευσης, να βελτιωθεί η απόδοση πρόβλεψης αποφεύγοντας την υπερπροσαρμογή (overfitting) και γενικώς να κάνει πιο εύκολο το χειρισμό του συνόλου των δεδομένων.

Στην *εικόνα 53* παρουσιάζεται σχηματικά ένα γενικό πλαίσιο επιλογής χαρακτηριστικών σε προβλήματα ταξινόμησης (όπως θα δούμε παρακάτω η επιλογή χαρακτηριστικών που ακολουθεί αυτή τη σχηματική αναπαράσταση ανήκει στην κατηγορία μεθόδων φιλτραρίσματος).



Εικόνα 53: *Γενικό πλαίσιο εφαρμογής επιλογής χαρακτηριστικών σε προβλήματα ταξινόμησης*

2.5 Τεχνικές Επιλογής χαρακτηριστικών

Η επιλογή χαρακτηριστικών προϋποθέτει τον καθορισμό δύο βασικών συνιστωσών της:

- τη διαδικασία παραγωγής - τον τρόπο αναζήτησης των υποψήφιων υποσυνόλων χαρακτηριστικών και
- το κριτήριο αξιολόγησης βάσει του οποίου θα γίνει η επιλογή του βέλτιστου υποσυνόλου

Σύμφωνα με την επιλογή της κατάλληλης μεθόδου για κάθε προαναφερθείσα συνιστώσα, η FS μπορεί να διακριθεί σε κατηγορίες που διαχωρίζονται με βάση τη διαδικασία παραγωγής/ τον τρόπο αναζήτησης των υποψήφιων υποσυνόλων και σε κατηγορίες που διαχωρίζονται ανάλογα με τη μεθοδολογία αξιολόγησης για την επιλογή του βέλτιστου υποσυνόλου χαρακτηριστικών.

Συνεπώς, σύμφωνα με τη διαδικασία παραγωγής των χαρακτηριστικών προκύπτουν δύο βασικοί τύποι: η κατάταξη μεμονωμένων χαρακτηριστικών (Individual Feature Ranking, IFR) και η επιλογή υποσυνόλων χαρακτηριστικών (Feature Subset Selection, FSS). Στον *πίνακα 4* παρουσιάζεται η κατηγοριοποίηση των μεθόδων επιλογής χαρακτηριστικών βάσει της διαδικασίας παραγωγής, καθώς και κάποια παραδείγματα τέτοιων μεθόδων.

Πίνακας 4: Κατηγοριοποίηση μεθόδων επιλογής χαρακτηριστικών βάσει της διαδικασίας παραγωγής

Κατηγορία	Τρόπος Αναζήτησης	Παραδείγματα
Κατάταξη μεμονωμένων χαρακτηριστικών (Individual Feature Ranking)	Συγκρίνει τη σχέση κάθε μεμονωμένου χαρακτηριστικού με το συγκεκριμένο πρόβλημα	Οι περισσότεροι μέθοδοι φιλτραρίσματος
Επιλογή υποσυνόλου χαρακτηριστικών (Feature Subset Selection)	Εξαντλητική: εξέταση όλων των υποψήφιων λύσεων	Branch and Bound, Best-first search (BFS)
	Ευρετική Ντετερμινιστική: άπληστη προσέγγιση για την επιλογή των χαρακτηριστικών	Sequential forward selection (SFS), Sequential backward selection (SFS), Sequential floating forward selection (SFFS), Sequential floating backward selection (SFBS)
	Τυχαία (μη ντετερμινιστική ευρετική): αναζητεί τη βέλτιστη λύση με τυχαίο τρόπο	Simulated annealing (SA), Las Vegas Filter (LVF), Tabu Search (TS), Genetic Algorithms (GA)

Στις τεχνικές κατάταξης μεμονωμένων χαρακτηριστικών (IFR), κάθε χαρακτηριστικό αξιολογείται ξεχωριστά ως προς τη σχέση του με την κλάση του προβλήματος. Ακολούθως, τα χαρακτηριστικά κατατάσσονται και τα καλύτερα από αυτά επιλέγονται. Οι αλγόριθμοι κατάταξης μεμονωμένων χαρακτηριστικών χαρακτηρίζονται από απλότητα, κλιμάκωση και σχετικά ανεκτή χρονική πολυπλοκότητα αφού το υπολογιστικό κόστος αφορά κυρίως τον υπολογισμό και την κατάταξη των n χαρακτηριστικών. Είναι εμπειρικώς επιτυχείς αλγόριθμοι αφού ικανοποιούν το βασικό τους στόχο, την απομάκρυνση των άσχετων χαρακτηριστικών. Οι τεχνικές επιλογής υποσυνόλων χαρακτηριστικών (FSS), αναζητούν το βέλτιστο υποσύνολο χαρακτηριστικών ως προς την απόδοσή του στην ταξινόμηση του συνόλου δεδομένων. Αυτοί οι αλγόριθμοι πλεονεκτούν σε σχέση με τους αλγόριθμου IFR επειδή εκτός από τη σχέση των χαρακτηριστικών με τις κλάσεις του προβλήματος λαμβάνεται υπόψη και η συσχέτιση ανάμεσα στα ίδια τα χαρακτηριστικά. Η αδυναμία αυτών των τεχνικών έγκειται στο μεγάλο υπολογιστικό κόστος που απαιτείται για την αναζήτηση στους διάφορους συνδυασμούς χαρακτηριστικών που απαρτίζουν τα υποσύνολα. Για την εύρεση του βέλτιστου υποσυνόλου, από όλα τα πιθανά υποσύνολα που προκύπτουν από το σύνολο δεδομένων, υπάρχουν διάφοροι τρόποι αναζήτησης.

Ανάλογα με τον τρόπο αναζήτησης των υποψήφιων λύσεων, η FSS μπορεί να χωριστεί σε διάφορες στρατηγικές αναζήτησης (search strategies). Η στρατηγική αναζήτησης απαιτείται έτσι ώστε να επιλεγούν τα πιθανά υποσύνολα χαρακτηριστικών που ακολούθως θα αξιολογηθούν από ένα κριτήριο / αντικειμενική συνάρτηση. Κάποιες βασικές στρατηγικές αναζήτησης είναι οι εξαντλητικοί (exhaustive), οι ευρετικοί/ευριστικοί (heuristic) και οι τυχαίοι (random) αλγόριθμοι [35]. Υπάρχουν επίσης και διάφοροι υβριδικοί (hybrid) αλγόριθμοι αναζήτησης. Οι ευρετικές ή τυχαίες στρατηγικές επιχειρούν τη μείωση της πολυπλοκότητας διακυβεύοντας την βελτιστότητα.

Στην **εξαντλητική/πλήρη αναζήτηση** εξερευνάται ολόκληρος ο χώρος συμπεριλαμβανομένου και των 2^N διαφορετικών υποσυνόλων. Όπως έχουμε ήδη αναφέρει και εξηγήσει η εξαντλητική αναζήτηση είναι εκθετική και συνεπώς απαγορευτική. Υπάρχουν διάφοροι εξαντλητικοί αλγόριθμοι αλλά δεν μπορούν να χρησιμοποιηθούν σε δεδομένα υψηλών διαστάσεων.

Οι **ευρετικές ντετερμινιστικές στρατηγικές** παράγουν προσθετικά ή αφαιρετικά τα υποψήφια υποσύνολα με σειριακή πρόσθεση ή αφαίρεση χαρακτηριστικών. Μειώνουν τον αριθμό καταστάσεων από εκθετικό σε πολυωνυμικό αλλά έχουν την τάση να παγιδεύονται σε τοπικά ελάχιστα. Παραδείγματα ευρετικών στρατηγικών είναι η ακολουθιακή προς τα εμπρός επιλογή (sequential forward selection), η ακολουθιακή προς τα πίσω απαλοιφή (sequential backward elimination) και η αμφίδρομη αναζήτηση (bidirectional search) που είναι διαφορετικές εκδοχές άπληστης αναζήτησης με διαφορά το σημείο εκκίνησης της αναζήτησης. Οι στρατηγικές αυτές είναι σχετικά υπολογιστικά συμφέρουσες και ανθεκτικές αλλά εξαιτίας της άπληστης προσέγγισης που ακολουθούν συχνά βρίσκουν μόνο τοπικά βέλτιστα.

Στην **στοχαστική/τυχαία ή αλλιώς μη ντετερμινιστική ευρετική αναζήτηση (random search)** γίνεται τυχαία εξερεύνηση του χώρου αναζήτησης. Σημείο εκκίνησης είναι ένα τυχαία επιλεγμένο υποσύνολο και η επόμενη κατεύθυνση καθορίζεται από μια δοσμένη πιθανότητα. Επιλέγουν το ολικά ή τοπικά βέλτιστο υποσύνολο με τυχαίο τρόπο σε προκαθορισμένο αριθμό επαναλήψεων. Παράδειγμα στοχαστικής αναζήτησης είναι οι γενετικοί αλγόριθμοι.

Στον *πίνακα 5* παρουσιάζεται μια σύγκριση της εξαντλητικής, της ευρετικής και της τυχαίας αναζήτησης.

Πίνακας 5: Σύγκριση εξαντλητικής, ντετερμινιστικά ευρετικής και τυχαίας αναζήτησης [35]

Στρατηγική αναζήτησης	Ακρίβεια	Πολυπλοκότητα	Πλεονεκτήματα	Μειονεκτήματα
Εξαντλητική αναζήτηση	Πάντα βρίσκει το βέλτιστο υποσύνολο	Εκθετική (exponential)	Πολύ ακριβής	Πολύ μεγάλη πολυπλοκότητα
Ευρετική αναζήτηση (ντετερμινιστική)	Καλή αν δεν χρειαστούν οπισθοδρομήσεις	Τετραγωνική (quadratic)	Απλή και γρήγορη	Οπισθοδρόμηση δεν είναι εφικτή, εγκλωβισμός σε τοπικά ελάχιστα
Τυχαία αναζήτηση (μη ντετερμινιστική)	Καλή με τις κατάλληλες παραμέτρους	Γενικά χαμηλή	Αποφεύγει τα τοπικά ελάχιστα	Δυσκολία επιλογής κατάλληλων παραμέτρων

Η επιλογή υποσυνόλου χαρακτηριστικών μπορεί να διαχωριστεί σε 3 κατηγορίες ανάλογα με το αν το κριτήριο αξιολόγησης εξαρτάται ή όχι από τον αλγόριθμο εκμάθησης που χρησιμοποιήθηκε για την κατασκευή του ταξινομητή: βαθμωτές μέθοδοι ή μέθοδοι φιλτραρίσματος (filter methods), μέθοδοι «περιτυλίγματος» ή αναδίπλωσης (wrapper methods) και μέθοδοι ενσωμάτωσης (embedded methods) [32].

Τόσο οι μέθοδοι φιλτραρίσματος όσο και η μέθοδοι περιτυλίγματος μπορούν να χρησιμοποιήσουν τις διάφορες στρατηγικές αναζήτησης για εξερεύνηση του χώρου όλων των πιθανών συνδυασμών χαρακτηριστικών [36].

Στις **τεχνικές φιλτραρίσματος (filter approach)** η FS είναι ανεξάρτητη από τη μέθοδο εκμάθησης που χρησιμοποιήθηκε για την κατασκευή του επιλεγμένου ταξινομητή. Αλλιώς ονομάζεται και βαθμωτή επιλογή χαρακτηριστικών (scalar FS) επειδή τα χαρακτηριστικά τυγχάνουν αξιολόγησης το καθένα ξεχωριστά: κατατάσσονται σύμφωνα με ένα συγκεκριμένο κριτήριο κατάταξης (ranking criterion) και τα *l* χαρακτηριστικά με τις καλύτερες αντίστοιχες τιμές επιλέγονται για να συμπεριληφθούν στο βέλτιστο υποσύνολο δημιουργώντας έτσι ένα διάνυσμα χαρακτηριστικών *l* διαστάσεων το οποίο θα χρησιμοποιηθεί ως διάνυσμα εισόδου του ταξινομητή. Η κατάταξη πραγματοποιείται ανάλογα με τη συσχέτιση κάθε χαρακτηριστικού με τις κατηγορίες δηλαδή ανάλογα με την ικανότητα ταξινόμησής του. Για το λόγο αυτό οι αλγόριθμοι

φιλτραρίσματος σχετίζονται άμεσα με τους αλγόριθμους κατάταξης χαρακτηριστικών (feature ranking algorithms).

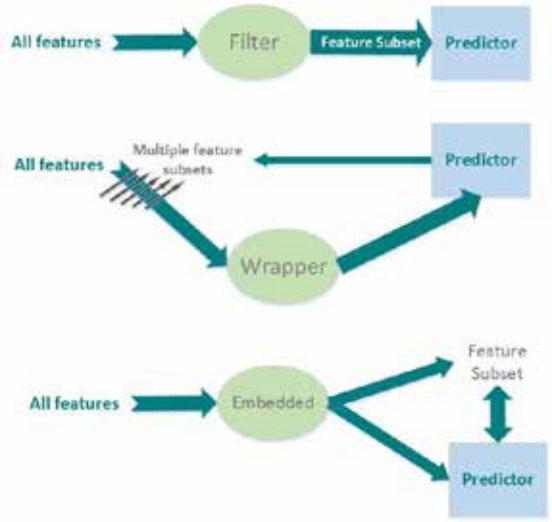
Στις **τεχνικές περιτυλίγματος (wrapper approach)** η FS «περιτυλίγεται» γύρω από τον αλγόριθμο εκμάθησης που χρησιμοποιείται, έτσι ώστε να αξιολογήσει τα υποψήφια υποσύνολα χαρακτηριστικών σύμφωνα με την απόδοση ταξινόμησής τους. Οι μέθοδοι περιτυλίγματος χρησιμοποιούν τον ίδιο τον ταξινομητή για να αξιολογήσουν το κάθε υποσύνολο χαρακτηριστικών για την επιλογή του καλύτερου διανύσματος χαρακτηριστικών: κάθε υποσύνολο χαρακτηριστικών χρησιμοποιείται ως διάνυσμα εισόδου για την εκπαίδευση και τη δοκιμή ενός συγκεκριμένου ταξινομητή και ακολούθως εκτιμάται το σφάλμα ταξινόμησής του. Το υποσύνολο με το μικρότερο σφάλμα ταξινόμησης επιλέγεται ως το καλύτερο. Επομένως, η wrapper FS είναι εξαρτώμενη από τον ταξινομητή και πραγματοποιείται για ένα συγκεκριμένο ταξινομητή μόνο. Οι αλγόριθμοι περιτυλίγματος αναφέρονται και ως αλγόριθμοι επιλογής υποσυνόλου (subset selection algorithms) επειδή εξετάζουν υποσύνολα χαρακτηριστικών και όχι κάθε χαρακτηριστικό ξεχωριστά. Στις μεθόδους περιτυλίγματος εξαιτίας των πολλών υποψήφιων υποσυνόλων χαρακτηριστικών χρησιμοποιούνται ευριστικές μέθοδοι για την εύρεση του βέλτιστου υποσυνόλου. Μια από τις καλύτερες μεθόδους περιτυλίγματος είναι οι Γενετικοί Αλγόριθμοι. Οι μέθοδοι περιτυλίγματος μπορούν να διαχωριστούν σε δύο υποκατηγορίες: τους ντετερμινιστικούς και τους τυχαίους αλγόριθμους [37].

Το βασικό πλεονέκτημα των μεθόδων φιλτραρίσματος έναντι των μεθόδων περιτυλίγματος είναι ότι έχουν χαμηλότερο υπολογιστικό κόστος και πολυπλοκότητα, δηλαδή είναι πιο γρήγορες και πιο απλές. Ιδιαίτερα στην περίπτωση δεδομένων υψηλών διαστάσεων, οι μέθοδοι περιτυλίγματος έχουν πολύ πιο υψηλό υπολογιστικό κόστος επειδή κάθε συνδυασμός χαρακτηριστικών πρέπει να αξιολογηθεί από τον ταξινομητή. Αντίθετα η FS με φιλτράρισμα εκτελείται μόνο μία φορά και με το αποτέλεσμά της μπορούν να δοκιμαστούν διάφοροι ταξινομητές. Επιπλέον, υποστηρίζεται ότι οι μέθοδοι φιλτραρίσματος έχουν καλύτερες ιδιότητες γενίκευσης επειδή δεν εξαρτώνται από οποιοδήποτε αλγόριθμο εκμάθησης σε αντίθεση με τις μεθόδους περιτυλίγματος που έχουν μεγαλύτερη πιθανότητα υπερπροσαρμογής (overfitting). Ωστόσο, για τον ίδιο λόγο το κύριο μειονέκτημα των μεθόδων φιλτραρίσματος είναι ότι αγνοούν την αλληλεπίδραση με τον ταξινομητή: ο καλύτερος συνδυασμός χαρακτηριστικών μπορεί να μην είναι ανεξάρτητος από τις επαγωγικές και αναπαραστατικές προτιμήσεις του αλγόριθμου μάθησης που θα χρησιμοποιηθεί για την κατασκευή του ταξινομητή [32]. Αντιθέτως, οι μέθοδοι περιτυλίγματος έχουν καλύτερη απόδοση επειδή η FS βελτιστοποιείται ειδικά για τον επιλεγμένο ταξινομητή και κατά συνέπεια μπορούν να βρουν ένα μικρό υποσύνολο χαρακτηριστικών με μεγάλη ακρίβεια. Επιπρόσθετα, οι μέθοδοι φιλτραρίσματος ελέγχοντας ένα-ένα τα χαρακτηριστικά αγνοούν τις εξαρτήσεις που έχουν αυτά μεταξύ τους, κάτι το οποίο μπορεί να

οδηγήσει σε χαμηλή απόδοση του ταξινομητή. Για τη διόρθωση του προβλήματος αυτού έχουν προταθεί τεχνικές φιλτραρίσματος πολυμεταβλητών (multivariate filter techniques) που στοχεύουν στον συνυπολογισμό σε ένα βαθμό αυτών των εξαρτήσεων. Ένα άλλο μειονέκτημα των μεθόδων φιλτραρίσματος είναι ότι τείνουν να καταλήγουν σε ένα υποσύνολο μεγάλου μεγέθους.

Τα πιο συνήθη στατιστικά μέτρα που χρησιμοποιούνται ως κριτήρια κατάταξης στις μεθόδους φιλτραρίσματος σε προβλήματα ταξινόμησης -όχι όλα με τον αυστηρό μαθηματικό ορισμό- είναι μέτρα απόστασης ή διαχωρισιμότητας των κλάσεων, μέτρα συσχέτισης (correlation) και μέτρα αμοιβαίας πληροφορίας (mutual information). Ένα μέτρο διαχωρισιμότητας που χρησιμοποιείται ευρέως είναι η καμπύλη ROC (ROC curve). Μια συνήθης μεθοδολογία βασισμένη στην αμοιβαία πληροφορία είναι η μέθοδος ελαχίστου πλεονασμού και μέγιστης συνάφειας (minimum redundancy maximum relevance mRMR). Άλλα στατιστικά μέτρα είναι ο στατιστικός έλεγχος υπόθεσης μέσω t-test και η απόσταση Bhattacharyya.

Η τρίτη κατηγορία τεχνικών FS, οι **τεχνικές ενσωμάτωσης (embedded techniques)** μοιάζουν με τις τεχνικές περιτυλίγματος. Υλοποιούν κι αυτές μια ευριστική μέθοδο που αποσκοπεί στην εύρεση του καλύτερου υποσυνόλου χαρακτηριστικών για ένα συγκεκριμένο ταξινομητή. Οι ενσωματωμένες τεχνικές, όμως, σε αντίθεση με τις τεχνικές περιτυλίγματος δεν αντιμετωπίζουν τον αρχικό ταξινομητή ως μαύρο κουτί αλλά παίρνουν πληροφορίες από αυτόν έτσι ώστε να βελτιώσουν τη διαδικασία επιλογής χαρακτηριστικών. Η αναζήτηση για το βέλτιστο υποσύνολο χαρακτηριστικών εμπεριέχεται στην κατασκευή του ίδιου του ταξινομητή ως μέρος της διαδικασίας εκπαίδευσης. Οι ενσωματωμένες τεχνικές έχουν το πλεονέκτημα που έχουν και οι τεχνικές περιτυλίγματος, της αλληλεπίδρασης με τον ταξινομητή, με το επιπλέον πλεονέκτημα να έχουν λιγότερες υπολογιστικές απαιτήσεις και να έχουν μικρότερο κίνδυνο για υπερπροσαρμογή. Σχηματικά μπορούμε να δούμε και τις 3 κατηγορίες FS στην *εικόνα 54*.



Εικόνα 54: *Κατηγορίες τεχνικών FS: filter, wrapper, embedded techniques*

Στον *πίνακα 6* γίνεται μια σύγκριση των filter, wrapper και embedded μεθόδων επιλογής χαρακτηριστικών. Στους *πίνακα 7* και *πίνακα 8* παρουσιάζονται οι πιο συνηθισμένες μέθοδοι FS, τα πλεονεκτήματα και τα μειονεκτήματά τους, καθώς και κάποια παραδείγματα κάθε τεχνικής.

Πίνακας 6: Σύγκριση filter, wrapper και embedded μεθόδων [35]

Μέθοδοι φιλτραρίσματος (Filter Methods)	Μέθοδοι περιτυλίγματος (Wrapper methods)	Ενσωματωμένες μέθοδοι (Embedded Methods)
Οι μέθοδοι φιλτραρίσματος φαίνεται να είναι λιγότερο βέλτιστες	Οι μέθοδοι περιτυλίγματος είναι μια καλύτερη εναλλακτική σε προβλήματα επιβλεπόμενης μάθησης	Εάν υπάρχουν πολλά άσχετα χαρακτηριστικά στο σύνολο δεδομένων τότε μειώνεται πολύ η απόδοση των ενσωματωμένων μεθόδων
Είναι πιο γρήγορες από τις μεθόδους περιτυλίγματος	Απαιτούν περισσότερο χρόνο εκτέλεσης από τις μεθόδους φιλτραρίσματος	Είναι πιο γρήγορες από τις μεθόδους περιτυλίγματος
Επιδεικνύουν καλύτερη ικανότητα γενίκευσης από τις μεθόδους περιτυλίγματος	Υπάρχει έλλειψη γενικότητας επειδή συνδέονται με ένα ταξινομητή	Έχουν έλλειψη γενικότητας επειδή εξαρτώνται από ένα ταξινομητή
Επιλέγουν μεγάλο υποσύνολο χαρακτηριστικών	Είναι πιο ακριβείς από τις μεθόδους φιλτραρίσματος: επιτυγχάνουν καλύτερα ποσοστά ταξινόμησης.	Είναι λιγότερο επιρρεπείς σε υπερπροσαρμογή
Το υπολογιστικό κόστος είναι μικρότερο για μεγάλα σύνολα δεδομένων	Το υπολογιστικό κόστος είναι μεγαλύτερο για μεγάλα σύνολα δεδομένων από ότι στις μεθόδους φιλτραρίσματος	Το υπολογιστικό κόστος είναι μικρότερο σε σύγκριση με τις μεθόδους περιτυλίγματος
Ανεξάρτητοι από τον αλγόριθμο ταξινόμησης	Εξαρτώνται από τον αλγόριθμο ταξινόμησης	Εξαρτώνται από τον αλγόριθμο ταξινόμησης

Πίνακας 7: Κατηγοριοποιήσεις τεχνικών επιλογής χαρακτηριστικών [37]

Μέθοδος FS	Πλεονεκτήματα	Μειονεκτήματα	Παραδείγματα
Βαθμωτή/ Φιλτραρίσματος (Filter)	Univariate		
	Γρήγορες Ανεξάρτητες από τον ταξινομητή Κλιμακωτές	Αγνοούν τις εξαρτήσεις των χαρακτηριστικών Αγνοούν την αλληλεπίδραση με τον ταξινομητή	Ευκλείδεια απόσταση χ^2 i-test Κέρδος πληροφορίας
	Multivariate		
	Μοντελοποιούν τις εξαρτήσεις των χαρακτηριστικών Ανεξάρτητες από τον ταξινομητή Καλύτερη υπολογιστική πολυπλοκότητα από τις μεθόδους περιτυλίγματος	Πιο αργές από τις univariate τεχνικές Λιγότερο κλιμακωτές από τις univariate τεχνικές Αγνοούν την αλληλεπίδραση με τον ταξινομητή	Correlation-based feature selection (CFS) Markov blanket filter (MBF) Fat correlation-based feature selection (FCBF)
Περιτυλίγματος (Wrapper)	Ντετερμινιστικές (Deterministic)		
	Απλές Αλληλεπιδρούν με τον ταξινομητή Λιγότερο υπολογιστικά απαιτητικές από ότι οι τυχαίες μέθοδοι	Κίνδυνος υπερπροσαρμογής Πιο επιρρεπείς από τις τυχαίες να παγιδευτούν σε τοπικό βέλτιστο (άπληστη αναζήτηση) Εξαρτώνται από τον ταξινομητή	Sequential forward selection (SFS) Sequential backward elimination (SBE) Plus q take-away t Beam search
	Τυχαίες (Stochastic)		
	Λιγότερο επιρρεπείς σε τοπικά βέλτιστα Αλληλεπιδρούν με τον ταξινομητή Μοντελοποιούν τις εξαρτήσεις των χαρακτηριστικών	Υπολογιστικά απαιτητικές Επιλογή εξαρτώμενη από τον ταξινομητή Μεγαλύτερος κίνδυνος υπερπροσαρμογής από ότι οι ντετερμινιστικοί αλγόριθμοι	Simulated annealing Randomized hill climbing Γενετικοί αλγόριθμοι Estimation of distribution algorithms
Ενσωματωμένη (Embedded)			
	Αλληλεπιδρούν με τον ταξινομητή	Η επιλογή εξαρτάται από τον ταξινομητή	Δέντρα απόφασης Weighted naïve Bayes

	Καλύτερη υπολογιστική πολυπλοκότητα από ότι οι μέθοδοι περιτυλίγματος Μοντελοποιεί τις εξαρτήσεις των χαρακτηριστικών		FS using the weight vector of SVM

Πίνακας 8: Κατηγοριοποίηση μεθόδων FS βάσει κριτηρίου αξιολόγησης - Παραδείγματα filter και wrapper μεθόδων

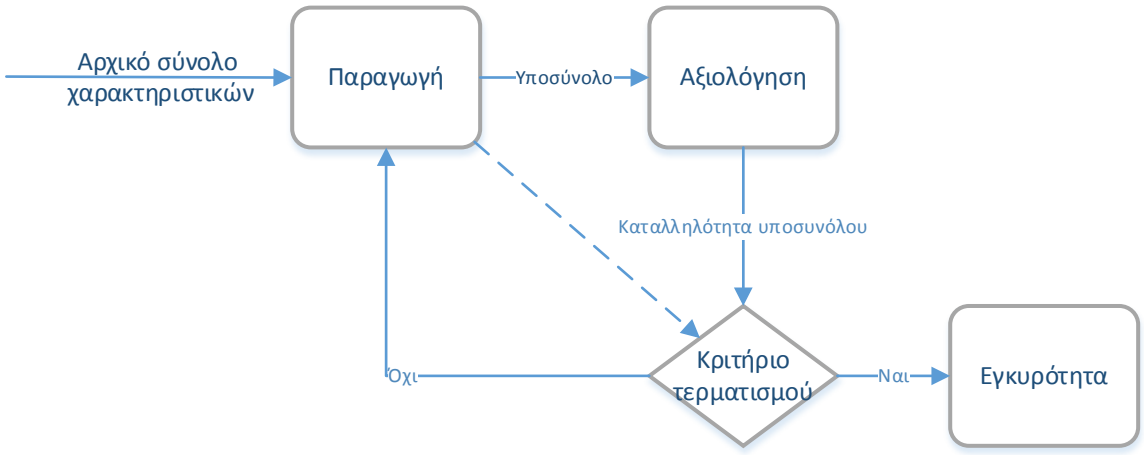
Κατηγορία FS	Κριτήριο Αξιολόγησης	Παραδείγματα
Μέθοδοι φιλτραρίσματος (filter)	Συσχέτιση (correlation): το μέτρο αυτό αξιολογεί την ικανότητα μιας μεταβλητής να προβλέψει μια άλλη	Συντελεστής συσχέτισης Pearson, Κέρδος πληροφορίας
	Απόσταση (distance): καθορίζει τον βαθμό διαχωρισμού ανάμεσα στις διάφορες κλάσεις	Κριτήριο Fischer, Απόσταση Bhattacharyya Studen's t-test, Sincich, Relief
	Συνοχή (consistency): εντοπίζει τον ελάχιστο αριθμό χαρακτηριστικών που μπορούν να διαχωρίσουν τις κλάσεις	Ρυθμός ασυνέχειας (Inconsistency rate)
Μέθοδοι περιτυλίγματος (wrapper)	Ταξινόμηση (classification): η απόδοση της ταξινόμησης του αλγόριθμου εκμάθησης	Decision trees, Naïve Bayesian Classifier, Selective Bayesian Classifier, Neural Network

2.5.1 Διαδικασία επιλογής χαρακτηριστικών

Μια τυπική διαδικασία επιλογής χαρακτηριστικών αποτελείται από 4 βασικά βήματα:

- Μια διαδικασία παραγωγής πιθανού υποσυνόλου χαρακτηριστικών μέσω μιας στρατηγικής αναζήτησης
- Ένα κριτήριο/αντικειμενική συνάρτηση αξιολόγησης του πιθανού υποσυνόλου χαρακτηριστικών: μετρά την καταλληλότητα του υποσυνόλου που παράχθηκε από την στρατηγική αναζήτησης. Αν είναι καλύτερο, αντικαθιστά το προηγούμενο καλύτερο υποσύνολο. Αυτό γίνεται είτε μέσω τεχνικών φιλτραρίσματος (μέτρα απόστασης, μέτρα πληροφορίας, μέτρα συσχέτισης) ή μέσω τεχνικών περιτυλίγματος (ακρίβεια ή σφάλμα ταξινόμησης) ή μέσω ενσωματωμένων τεχνικών.
- Ένα κριτήριο τερματισμού: ανάλογα με τη στρατηγική αναζήτησης μια διαδικασία επιλογής χαρακτηριστικών μπορεί να τρέχει πολύ χωρίς λόγο. Οι στρατηγικές αναζήτησης και το κριτήριο αξιολόγησης μπορεί να επηρεάσουν την επιλογή ενός κριτηρίου τερματισμού. Κάποια παραδείγματα κριτηρίων τερματισμού είναι ένας προκαθορισμένος αριθμός χαρακτηριστικών που πρέπει να επιλεγούν ή ένας προκαθορισμένος αριθμός επαναλήψεων. [38]
- Μια διαδικασία ελέγχου σχετικά με το αν το καλύτερο υποσύνολο που προέκυψε από τη διαδικασία FS είναι έγκυρο: αυτό το βήμα είναι προαιρετικό και δεν συμπεριλαμβάνεται στην επιλογή χαρακτηριστικών. Απλά εξετάζει και συγκρίνει τα αποτελέσματα του υποσυνόλου της τεχνικής FS που έχει χρησιμοποιηθεί με τα αποτελέσματα άλλων υποσυνόλων που επιλέχθηκαν βάσει άλλων τεχνικών επιλογής χαρακτηριστικών.

Αυτή η διαδικασία φαίνεται σχηματικά στην *εικόνα 55*.



Εικόνα 55: *Επιλογή χαρακτηριστικών (τροποποιημένο διάγραμμα από [38])*

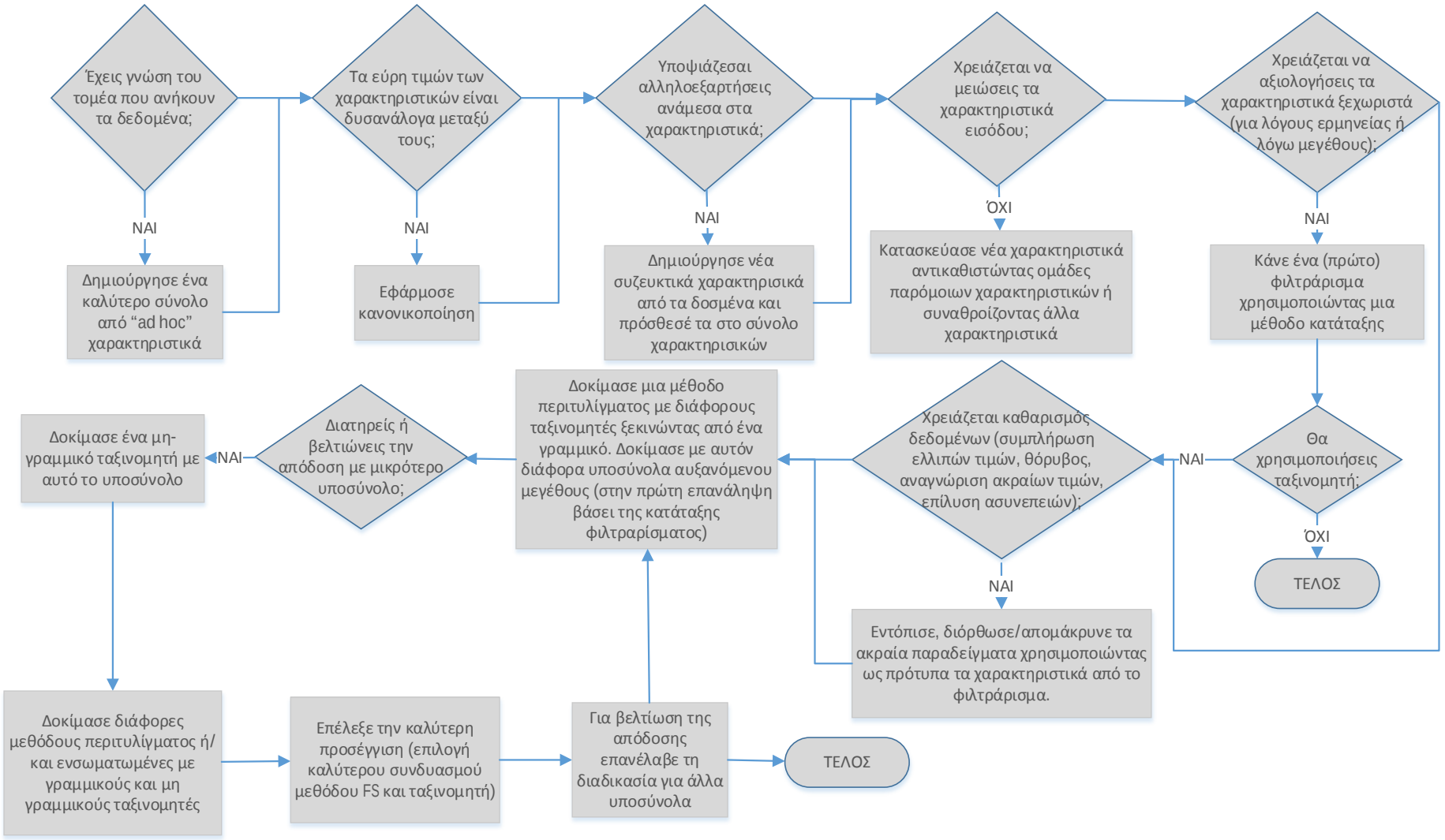
2.5.2 Υβριδική μέθοδος επιλογής χαρακτηριστικών

Ανάλογα με τη μέθοδο FS που θα χρησιμοποιηθεί δίνεται διαφορετική βαρύτητα σε κάποια πλεονεκτήματα της FS σε σχέση με άλλα. Όταν στόχος μας είναι η επιλογή και δημιουργία των υποσυνόλων που είναι χρήσιμα για την κατασκευή ενός καλού ταξινομητή, τότε η κατάταξη όλων των πιθανών συναφή χαρακτηριστικών (relevant features) και η επιλογή των πρώτων καλύτερων (μέθοδος φιλτραρίσματος), πολλές φορές δεν επαρκεί για το πρόβλημα της ταξινόμησης. Η επιλογή των πιο συναφή χαρακτηριστικών συνήθως δεν είναι βέλτιστη για την κατασκευή ενός ταξινομητή, ιδιαίτερα όταν συμπεριλαμβάνονται περιττά/πλεονάζοντα χαρακτηριστικά (redundant features). Αντίθετα, ένα υποσύνολο από χρήσιμα χαρακτηριστικά μπορεί να αποκλείει πολλά περιττά, αλλά συναφή χαρακτηριστικά (μέθοδοι περιτυλίγματος). Τα συναφή χαρακτηριστικά δεν είναι απαραίτητα και τα πιο χρήσιμα για ένα δεδομένο ταξινομητή. Έτσι μέθοδοι επιλογής υποσυνόλων χαρακτηριστικών (feature subset selection), όπως οι μέθοδοι περιτυλίγματος, αξιολογούν υποσύνολα από χαρακτηριστικά ανάλογα με τη χρησιμότητά τους στο δοσμένο ταξινομητή ενώ μέθοδοι φιλτραρίσματος αξιολογούν το κάθε χαρακτηριστικό ξεχωριστά ανάλογα με τη συνάφειά του. Ουσιαστικά, οι διάφορες μέθοδοι αξιολογούν τη στατιστική σημαντικότητα (statistical significance) της συνάφειας (relevance) των χαρακτηριστικών διαφορετικά.

Κάποιες φορές, μπορεί να γίνει χρήση μιας υβριδικής μεθόδου (hybrid method), δηλαδή ενός συνδυασμού των δύο μεθόδων: αρχικά μπορεί να χρησιμοποιηθεί μια μέθοδος φιλτραρίσματος σε ένα πρόβλημα ταξινόμησης πολύ μεγάλων διαστάσεων για σκοπούς προκαταρκτικής γρήγορης μείωσης των διαστάσεων των δεδομένων, με αφαίρεση των άσχετων χαρακτηριστικών ή των λιγότερο συναφή, και ακολούθως στους διάφορους συνδυασμούς των υπόλοιπων

χαρακτηριστικών να εφαρμοστεί μια μέθοδος περιτυλίγματος για την επιλογή του βέλτιστου υποσυνόλου από αυτά (μείωση των χαρακτηριστικών συνεπάγεται μείωση του αριθμού των πιθανών συνδυασμών χαρακτηριστικών δηλαδή των πιθανών υποσυνόλων). Οι ενσωματωμένες μέθοδοι ακολουθούν παρόμοια λογική με αυτή των μεθόδων περιτυλίγματος, αλλά το κάνουν πιο αποδοτικά, αποσκοπώντας στην παράλληλη βελτιστοποίηση τόσο της χρησιμότητας των χαρακτηριστικών, όσο και της διατήρησης ενός μικρού πλήθους χαρακτηριστικών στο επιλεγμένο υποσύνολο [28].

Στην *εικόνα 56* παρουσιάζεται μια προτεινόμενη διαδικασία για εφαρμογή μιας υβριδικής επιλογής χαρακτηριστικών. Όπως βλέπουμε, μια καλή προτεινόμενη διαδικασία για επιλογή ενός υποσυνόλου χαρακτηριστικών που να μειώνει τις διαστάσεις του χώρου δεδομένων και παράλληλα να δίνει καλά αποτελέσματα περιλαμβάνει τις 2 φάσεις, όπως τις περιγράψαμε πιο πάνω: την αρχική ελάττωση του αριθμού των χαρακτηριστικών με μια μέθοδο φιλτραρίσματος και την επιλογή του καλύτερου συνδυασμού χαρακτηριστικών από αυτά που έμειναν με μια μέθοδο περιτυλίγματος για την εύρεση του βέλτιστου δυνατού υποσυνόλου.



Εικόνα 56: *Ροή εργασιών για εφαρμογή υβριδικής επιλογής χαρακτηριστικών*

Εν συνεχεία θα δούμε αναλυτικότερα ορισμένα μέτρα διαχωρισιμότητας και κάποια άλλα στατιστικά μέτρα που χρησιμοποιούνται σε βασικές τεχνικές φιλτραρίσματος, καθώς και κάποιες στρατηγικές αναζήτησης που χρησιμοποιούνται σε συνδυασμό με τεχνικές περιτυλίγματος για την επιλογή χαρακτηριστικών.

2.6 Επιλογή χαρακτηριστικών με μεθόδους filter-Ανάλυση χρησιμοποιούμενων μεθόδων φιλτραρίσματος χαρακτηριστικών

Οι τεχνικές φιλτραρίσματος αποτιμούν ξεχωριστά το κάθε χαρακτηριστικό με κάποιο μέτρο. Η κατάταξη των χαρακτηριστικών είναι βασισμένη στη συνάφεια των μεμονωμένων χαρακτηριστικών με την ετικέτα κλάσης. Έστω ένα σύνολο από m παραδείγματα $\{x_k,y_k\}$ (k=1,2,3,…,m), με κάθε παράδειγμα να αποτελείται από n μεταβλητές εισόδου $x_{k,i}$ (i=1,2,3,…,n) και μια μεταβλητή εξόδου y_k. Η κατάταξη των μεταβλητών χρησιμοποιεί ένα κριτήριο / συνάρτηση βαθμολόγησης S(i) η οποία υπολογίζεται από τις τιμές $x_{k,i}$ και y_k , k=1,2,3,…,m. Κατά σύμβαση, θεωρούμε ότι μια υψηλή βαθμολογία αποτελεί ένδειξη μιας χρήσιμης μεταβλητής και ως εκ τούτου κατατάσσουμε τις μεταβλητές σε φθίνουσα σειρά σε σχέση με το S(i) [28].

Οι μέθοδοι φιλτραρίσματος μπορεί:

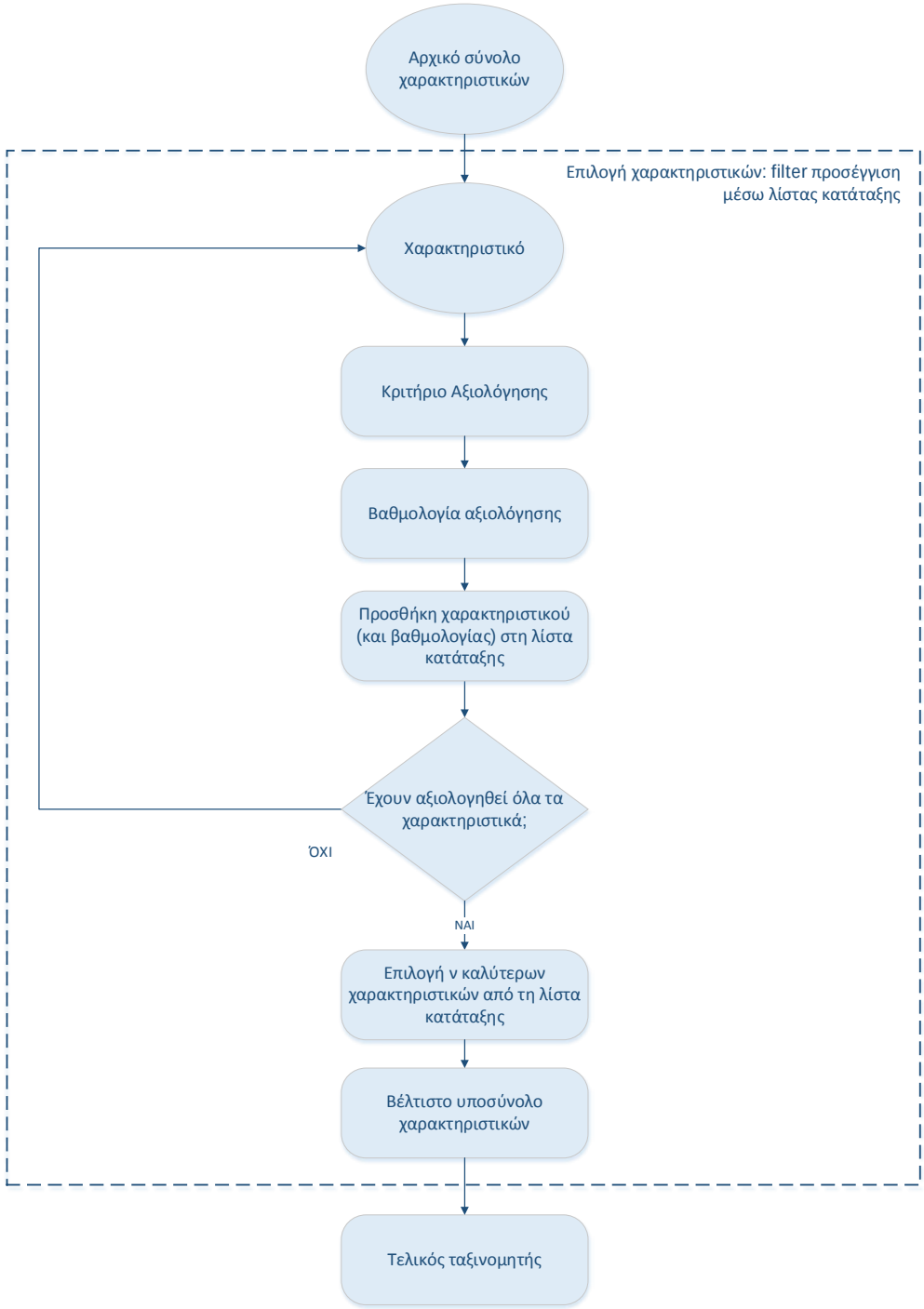
1. είτε να είναι καθαρά τεχνικές κατάταξης χαρακτηριστικών με το βέλτιστο υποσύνολο να προκύπτει στο τέλος, από τη λίστα κατάταξης
2. είτε να χρησιμοποιούν μια στρατηγική αναζήτησης για την παραγωγή του βέλτιστου υποσυνόλου χαρακτηριστικών με χρήση ενός κριτηρίου αξιολόγησης.

Για τις τεχνικές κατάταξης χαρακτηριστικών (εικόνα 57) απαιτούνται 3 βήματα:

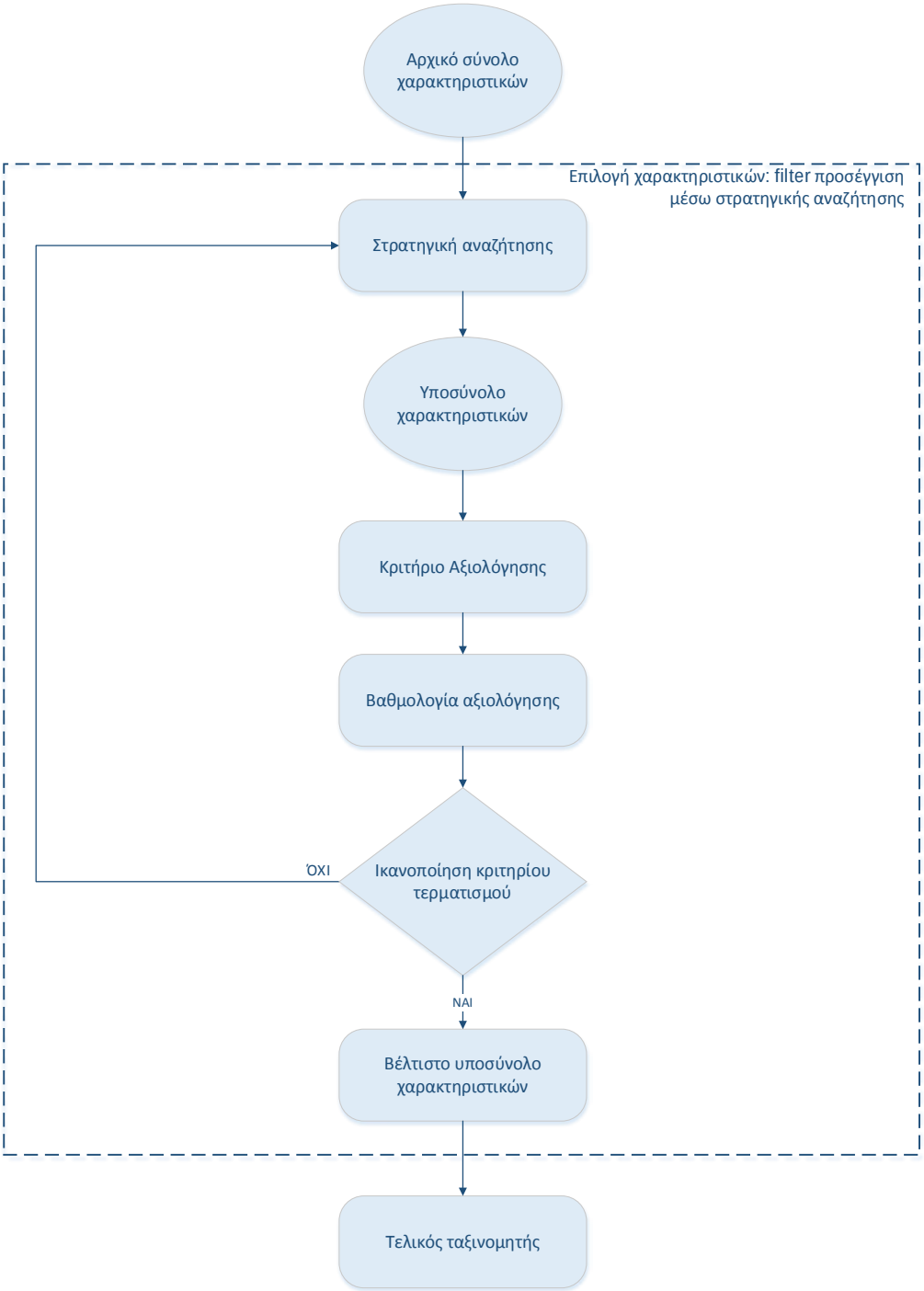
1. μεμονωμένη αξιολόγηση κάθε χαρακτηριστικού με ένα κριτήριο αξιολόγησης
2. δημιουργία μιας λίστας κατάταξης και
3. επιλογή των πρώτων χαρακτηριστικών της λίστας κατάταξης για τον σχηματισμό του βέλτιστου υποσυνόλου χαρακτηριστικών

Για την παραγωγή του βέλτιστου υποσυνόλου χαρακτηριστικών μέσω στρατηγικής αναζήτησης (*εικόνα 58*) η διαδικασία που ακολουθείται είναι:

1. παραγωγή ενός πιθανού υποσυνόλου μέσω μιας στρατηγικής αναζήτησης
2. αξιολόγηση του πιθανού υποσυνόλου με ένα κριτήριο αξιολόγησης (μάλλον στατιστικό μέτρο)
3. έλεγχος αν ικανοποιείται το κριτήριο τερματισμού: αν ναι, η διαδικασία τερματίζει και το υποσύνολο είναι το βέλτιστο, αν όχι, η διαδικασία επαναλαμβάνεται



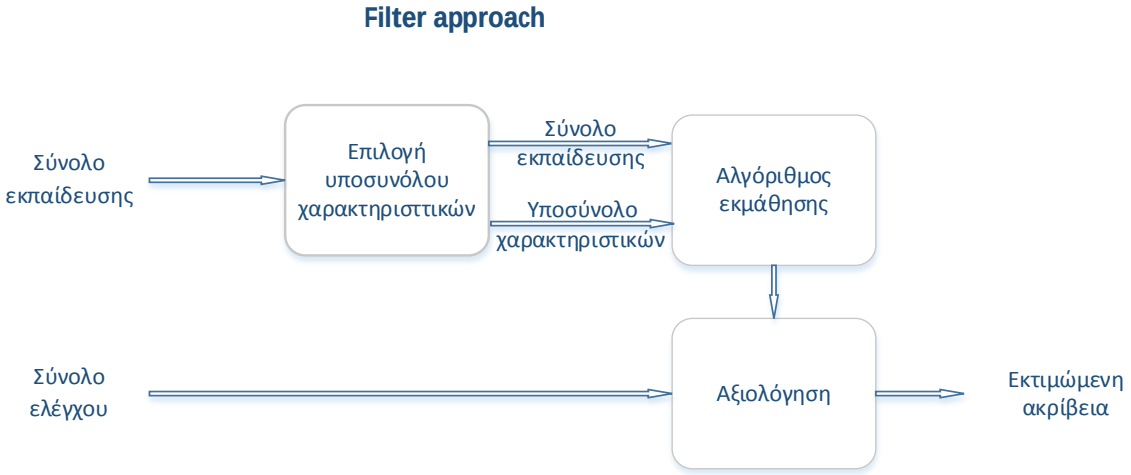
Εικόνα 57: *Παραγωγή βέλτιστου υποσυνόλου μέσω λίστας κατάταξης (filter προσέγγιση)*



Εικόνα 58: *Παραγωγή βέλτιστου υποσυνόλου μέσω στρατηγικής αναζήτησης (filter προσέγγιση)*

Στην εικόνα 59 βλέπουμε τη χρήση τεχνικής φιλτραρίσματος σε ένα πρόβλημα ταξινόμησης. Αρχικά από το σύνολο εκπαίδευσης γίνεται η επιλογή του βέλτιστου υποσυνόλου χαρακτηριστικών ανεξάρτητα από τον ταξινομητή, μέσω κάποιου κριτηρίου αξιολόγησης (μέτρα απόστασης, μέτρα πληροφορίας, μέτρα συσχέτισης) και ακολούθως το βέλτιστο υποσύνολο χαρακτηριστικών χρησιμοποιείται για την εκπαίδευση του ταξινομητή. Στη συνέχεια με το σύνολο

ελέγχου γίνεται αξιολόγηση της μεθόδου επιλογής χαρακτηριστικών έτσι ώστε να μπορούμε να την συγκρίνουμε με άλλες μεθόδους επιλογής χαρακτηριστικών.



Εικόνα 59: *Η filter προσέγγιση για επιλογή υποσυνόλου χαρακτηριστικών σε πρόβλημα ταξινόμησης (τροποποιημένο διάγραμμα από [39])*

Οι βασικές και πιο ευρέως χρησιμοποιούμενες τεχνικές φιλτραρίσματος: στατιστικός έλεγχος υπόθεσης (student's t test), καμπύλες ROC, mRMR και Relief (εκ των οποίων οι τρεις τελευταίες έχουν χρησιμοποιηθεί σε αυτή τη διπλωματική θέση), αναλύονται παρακάτω.

2.6.1 Επιλογή χαρακτηριστικών βασισμένη σε στατιστικά μέτρα

2.6.1.1 Στατιστικός έλεγχος υπόθεσης (Student's t-test)

Η πιο απλή τεχνική επιλογής χαρακτηριστικών filtering είναι η χρήση του στατιστικού ελέγχου υπόθεσης t-test για την αποτίμηση της διαχωριστικής ικανότητας κλάσεων του κάθε χαρακτηριστικού. Με χρήση του t-test καθορίζεται, εάν οι κατανομές των τιμών ενός χαρακτηριστικού για δύο διαφορετικές κλάσεις είναι διακριτές. Αυτή η τεχνική είναι ιδιαίτερα χρήσιμη όταν πρόκειται για χαρακτηριστικά που ακολουθούν κανονική κατανομή (Gaussian distribution).

Έστω x τυχαία μεταβλητή που αναπαριστά κάποιο συγκεκριμένο χαρακτηριστικό. Αν $N(=N_1+N_2)$, ο αριθμός όλων των τιμών των δειγμάτων του χαρακτηριστικού x, θεωρούμε n_i, όπου

$i=1,2,\ldots,N_1$, τα δείγματα του χαρακτηριστικού x στην κλάση ω_1 με μέση τιμή μ_1 και κατ' αντιστοιχία m_j, $j=1,2,\ldots,N_2$ τα δείγματα του χαρακτηριστικού x στην κλάση ω_2 με μέση τιμή μ_2. Θα προσπαθήσουμε να διαπιστώσουμε αν οι τιμές που παίρνει το χαρακτηριστικό x για τις διαφορετικές κλάσεις, ω_1,ω_2, διαφέρουν σημαντικά. Για να το κάνουμε αυτό, αντιμετωπίζουμε το πρόβλημα εντός του πλαισίου στατιστικού ελέγχου υπόθεσης, εξακριβώνοντας ποια εκ των δύο παρακάτω στατιστικών υποθέσεων είναι η ορθή (στατιστική υπόθεση είναι ένας ισχυρισμός που αναφέρεται στην κατανομή μιας ή περισσότερων τυχαίων μεταβλητών) [40]:

H_0: Οι τιμές του χαρακτηριστικού x δεν διαφέρουν σημαντικά (μηδενική υπόθεση ή null hypothesis)

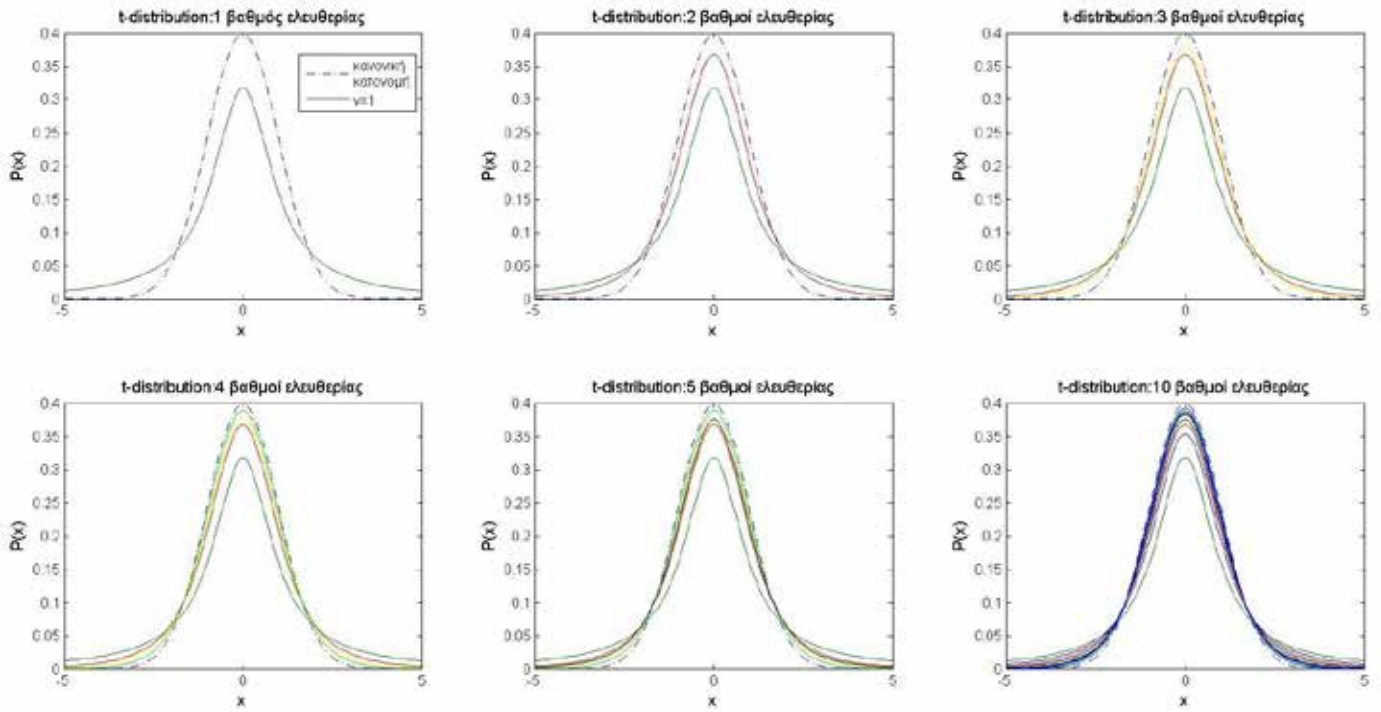
H_1: Οι τιμές του χαρακτηριστικού x διαφέρουν σημαντικά (εναλλακτική υπόθεση ή alternative hypothesis)

Στην περίπτωση που ισχύει η μηδενική υπόθεση, το χαρακτηριστικό απορρίπτεται επειδή δεν θα μπορεί να χωρίσει τα δείγματα ορθά σε κλάσεις. Σε αντίθετη περίπτωση που ισχύει η εναλλακτική υπόθεση, το χαρακτηριστικό επιλέγεται, αφού θα κάνει εύκολη τη διάκριση σε κλάσεις.

Η απόφαση αν θα απορριφθεί ή όχι η μηδενική υπόθεση λαμβάνεται εξετάζοντας πειραματικά τις στατιστικές πληροφορίες που προκύπτουν από τις παρατηρηθείσες τιμές της τυχαίας μεταβλητής x. Εξαιτίας της αξιοποίησης στατιστικής πληροφορίας είναι προφανές ότι οποιαδήποτε απόφαση ενέχει μια πιθανότητα σφάλματος, που αναφέρεται και ως significance level, την τιμή της οποίας σε πρακτικές εφαρμογές την προκαθορίζουμε. Οι τιμές της τυχαίας μεταβλητής x, χρησιμοποιούνται από μια στατιστική συνάρτηση ελέγχου για τον υπολογισμό ενός αριθμού (test statistic ή t-statistic) ο οποίος θα καθορίσει αν θα απορριφθεί ή όχι η H_0. Η στατιστική συνάρτηση ελέγχου χρησιμοποιείται για τη μέτρηση της διαφοράς των δεδομένων από αυτό που αναμένεται να συμβαίνει αν η H_0 είναι αληθής. Η διαδικασία λήψης της απόφασης ονομάζεται έλεγχος της στατιστικής υπόθεσης (test of statistical hypothesis).

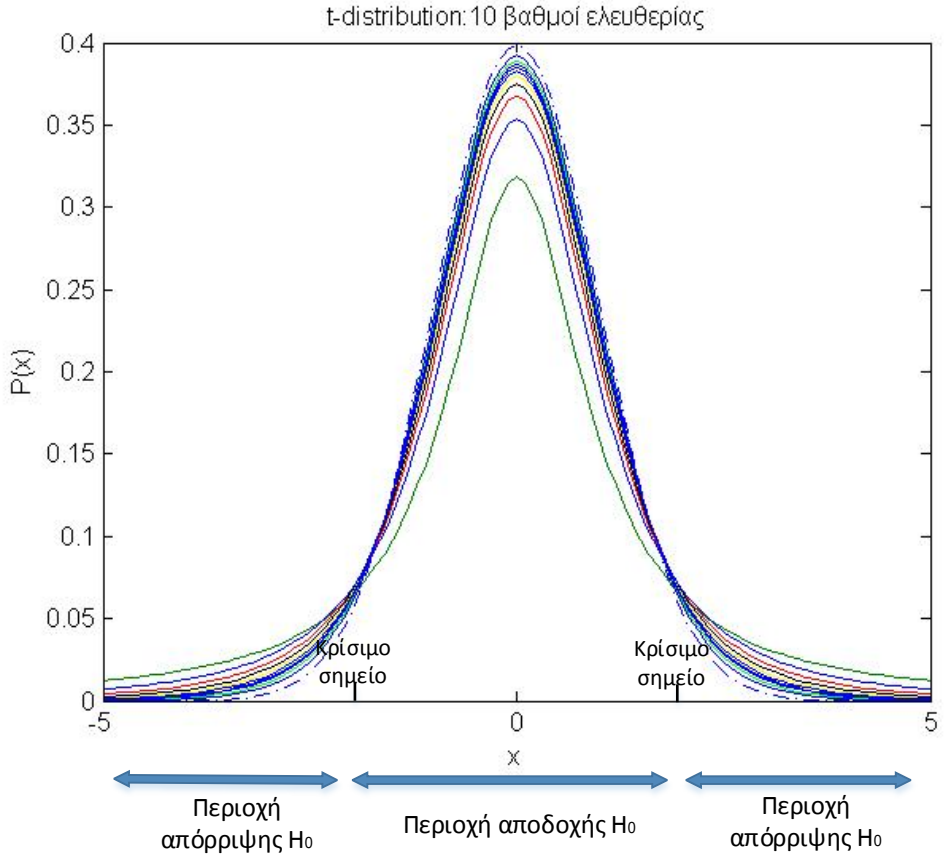
Η στατιστική συνάρτηση ελέγχου ακολουθεί την κατανομή πιθανότητας student's t-distribution. Οι t-κατανομές (t-distributions) περιγράφουν τις κατανομές των δειγμάτων των διαφόρων κλάσεων και μοιάζουν σχηματικά με τις κανονικές κατανομές με κάποιες διαφορές (αναφέρονται ως student's distributions επειδή όταν ο Gosset το 1908 τις χρησιμοποίησε σε ένα paper, το έκανε υπό το ψευδώνυμο "Student"). Ενώ μια κανονική κατανομή περιγράφει ολόκληρο τον πληθυσμό (θεωρητικά άπειρο αριθμό δειγμάτων), οι t-κατανομές περιγράφουν τα διάφορα δείγματα που έχουν παρθεί από τον πληθυσμό. Για n δείγματα από τον πληθυσμό δειγμάτων, οι t-κατανομές θα έχουν $v=n-1$ βαθμούς ελευθερίας (υπάρχουν κι άλλοι τρόποι υπολογισμού των βαθμών ελευθερίας αλλά αυτός είναι ο πιο απλός). Στο πρόβλημά μας η t-κατανομή θα έχει συνολικά $(N_1 - 1) + (N_2 - 1) = N-2$ βαθμούς ελευθερίας επειδή θα έχει (N_1-1)

βαθμούς ελευθερίας για την κλάση ω_1 και (N_2–1) βαθμούς ελευθερίας για την κλάση ω_2. Όσο πιο μεγάλος είναι ο αριθμός των δειγμάτων που ανήκουν σε μια κλάση, άρα και των βαθμών ελευθερίας, τόσο πιο πολύ οι t-κατανομές της κλάσης αυτής προσεγγίζουν την κανονική κατανομή. Με τη βοήθεια της t-κατανομής και της τιμής του t-test αξιολογείται η στατιστική σημαντικότητα της διαφοράς ανάμεσα στις δύο κλάσεις.



Εικόνα 60: *Παραδείγματα t-κατανομής για διάφορους βαθμούς ελευθερίας*

Αν η τιμή του t-test που προκύπτει από τα διαθέσιμα δείγματα, όπως έχει υπολογιστεί μέσω της στατιστικής συνάρτησης ελέγχου, απέχει πολύ από το μηδέν τότε ανήκει στην περιοχή απόρριψης της H_0 και επιλέγεται η εναλλακτική υπόθεση H_1. Αντίθετα, αν η τιμή του t-test δεν απέχει πολύ από το μηδέν τότε ανήκει στην περιοχή αποδοχής της H_0 και επιλέγεται η μηδενική υπόθεση H_0. Η τιμή της παραμέτρου η οποία διαχωρίζει την περιοχή αποδοχής από την περιοχή απόρριψης λέγεται κρίσιμο σημείο.



Εικόνα 61: *Παράδειγμα περιοχής αποδοχής H_0 και απόρριψης H_0*

Η τιμή του κρίσιμου σημείου στις t-κατανομές καθορίζεται ανάλογα με τους βαθμούς ελευθερίας και την προκαθορισμένη τιμή της πιθανότητας εσφαλμένης απόφασης (ρ) μέσω μιας στατιστικής συνάρτησης. Με βάση το κρίσιμο σημείο, μπορούμε να προσδιορίσουμε το διάστημα αποδοχής H_0. Στον π*ίνακα 9* παρουσιάζονται κάποιες επιλεγμένες τιμές του κρίσιμου σημείου.

Πίνακας 9: Τιμές κρίσιμου σημείου για τους διάφορους συνδυασμούς βαθμών ελευθερίας και πιθανοτήτων σφάλματος

Βαθμοί ελευθερίας	*1 - ρ*	*0.8*	*0.9*	*0.95*	*0.98*	*0.99*
1		3.08	6.31	12.71	31.82	63.66
2		1.89	2.92	4.43	6.97	9.93
3		1.64	2.35	3.18	4.54	5.84
4		1.53	2.13	2.78	3.75	4.60
5		1.48	2.02	2.57	3.37	4.03
6		1.44	1.94	2.45	3.14	3.71
7		1.41	1.90	2.37	3.00	3.50
8		1.40	1.86	2.31	2.90	3.36
9		1.38	1.83	2.26	2.82	3.25
10		1.37	1.81	2.23	2.76	3.17
11		1.36	1.80	2.20	2.72	3.11
12		1.36	1.78	2.18	2.68	3.06
13		1.35	1.77	2.16	2.65	3.01
14		1.35	1.76	2.15	2.62	2.98
15		1.34	1.75	2.13	2.60	2.95
16		1.34	1.75	2.12	2.58	2.92
17		1.33	1.74	2.11	2.57	2.90
18		1.33	1.73	2.10	2.55	2.88
19		1.33	1.73	2.09	2.54	2.86
20		1.33	1.73	2.09	2.53	2.85
21		1.32	1.72	2,08	2.52	2.83
22		1.32	1.72	2.07	2.51	2.82
23		1.32	1.71	2.07	2.50	2.81
∞		1.28	1.65	1.96	2.33	2.58

Υπάρχουν διάφορες στατιστικές συναρτήσεις ελέγχου που μπορούν να χρησιμοποιηθούν στο student's t-test. Μια από αυτές, είναι αυτή που χρησιμοποιείται στο welch's t-test (παραλλαγή του student's t-test με τη διαφορά ότι δεν υποθέτει ίσες αποκλίσεις), η οποία ελέγχει τη στατιστική υπόθεση αν οι μέσες τιμές του χαρακτηριστικού για τις δύο κλάσεις είναι ίσες

λαμβάνοντας υπόψη και την απόκλιση των τιμών κάθε κλάσης: υπολογίζεται η διαφορά των αντίστοιχων μέσων τιμών ενός συγκεκριμένου χαρακτηριστικού για τις δύο κλάσεις και ακολούθως εξετάζεται αν αυτή η διαφορά απέχει πολύ ή όχι από το μηδέν. Η στατιστική συνάρτηση ελέγχου που χρησιμοποιείται στο welch's t-test υπολογίζεται βάσει του ακόλουθου τύπου:

$$t = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} - D_m \text{ με υποθέσεις } \begin{matrix} H_0 : D_m = m_1' - m_2' = 0 \\ H_1 : D_m = m_1' - m_2' \neq 0 \end{matrix}$$

όπου μ_1, μ_2 οι μέσες τιμές των δύο κλάσεων , s_1, s_2 οι αντίστοιχες τυπικές αποκλίσεις και N_1 , N_2 το πλήθος των δειγμάτων που ανήκουν σε κάθε κλάση.

Υποθέτοντας ότι ισχύει η H_0 , δηλαδή $\Delta_\mu = 0$, προκύπτει:

$$t = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

Η συγκεκριμένη στατιστική συνάρτηση ελέγχου έχει καλύτερα αποτελέσματα όταν οι τιμές των δειγμάτων των δύο κλάσεων έχουν άνισες αποκλίσεις και τα πλήθη των δειγμάτων σε κάθε κλάση, N_1 και N_2 , είναι επίσης άνισα. Συνήθως χρησιμοποιείται όταν οι τιμές των δύο κλάσεων δεν επικαλύπτονται.

<u>Παράδειγμα:</u>

Οι τιμές των δειγμάτων ενός χαρακτηριστικού σε δύο κλάσεις είναι οι παρακάτω:

Τιμές των N_1 (=10) δειγμάτων που ανήκουν στην κλάση ω_1:

3.6 4.1 3.7 3.4 3.9 3.5 3.7 3.5 4.1 3.8

Τιμές των N_2 (=10) δειγμάτων που ανήκουν στην κλάση ω_2:

3.3 3.6 3.2 3.6 3.0 3.1 3.4 3.1 2.8 3.4

Για να διαπιστώσουμε αν το χαρακτηριστικό έχει αρκετή πληροφορία για να το επιλέξουμε στο βέλτιστο υποσύνολο χαρακτηριστικών χρησιμοποιούμε το t-test. Θα ελέγξουμε αν οι τιμές του χαρακτηριστικού για τις δύο κλάσεις διαφέρουν σημαντικά.

ω_1: $\mu_1 = 3.73$ και $s_1^2 = 0.0601$

ω_2: $\mu_2 = 3.25$ και $s_2^2 = 0.0672$

Θεωρούμε τις ακόλουθες στατιστικές υποθέσεις σχετικά με τη διαφορά των μέσων τιμών των δύο κλάσεων:

H_0: $\Delta_\mu = \mu_1' - \mu_2' = 0$ (οι τιμές των δύο κλάσεων δεν διαφέρουν)

H_1: $\Delta_\mu = \mu_1' - \mu_2' \neq 0$ (οι τιμές των δύο κλάσεων διαφέρουν)

Υπολογίζουμε την test-statistic θεωρώντας ότι ισχύει η H_0 με πιθανότητα σφάλματος 0.01:

$$t = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} - (m_1' - m_2') = \frac{3.73 - 3.25}{\sqrt{\frac{0.0601}{10} + \frac{0.0672}{10}}} - 0 = 4.25$$

Υπολογίζουμε τους βαθμούς ελευθερίας:

$v = (N_1 - 1) + (N_2 - 1) = 18$

Ακολούθως ελέγχουμε αν η τιμή του t-test ανήκει στο διάστημα αποδοχής της H_0 για t-distribution 18 βαθμών ελευθερίας και πιθανότητα σφάλματος 0.01 σύμφωνα με τον *πίνακα 9*. Η τιμή του κρίσιμου σημείου για πιθανότητα σφάλματος 0.01 είναι 2.88 άρα το διάστημα αποδοχής της H_0 είναι [-2.88, 2.88]. Η τιμή του t-statistic στο παράδειγμα αυτό, είναι μεγαλύτερη από την τιμή του κρίσιμου σημείου (4.25 > 2.88), οπότε ισχύει η H_1: οι μέσες τιμές διαφέρουν σημαντικά και άρα το χαρακτηριστικό έχει καλή διακριτική ικανότητα μεταξύ των κλάσεων.

Εφαρμόζουμε την παραπάνω διαδικασία, όπως έχει περιγραφεί, σε όλα τα χαρακτηριστικά: υπολογίζουμε την τιμή του t-test για κάθε χαρακτηριστικό και βάσει της τιμής αυτής επιλέγουμε τα χαρακτηριστικά που έχουν καλή διαχωριστική ικανότητα των κλάσεων. Με αυτό τον τρόπο η τιμή του t-test χρησιμοποιείται ως μέτρο της διαχωριστικής ικανότητας των κλάσεων του κάθε χαρακτηριστικού.

2.6.1.2 Άλλα μέτρα

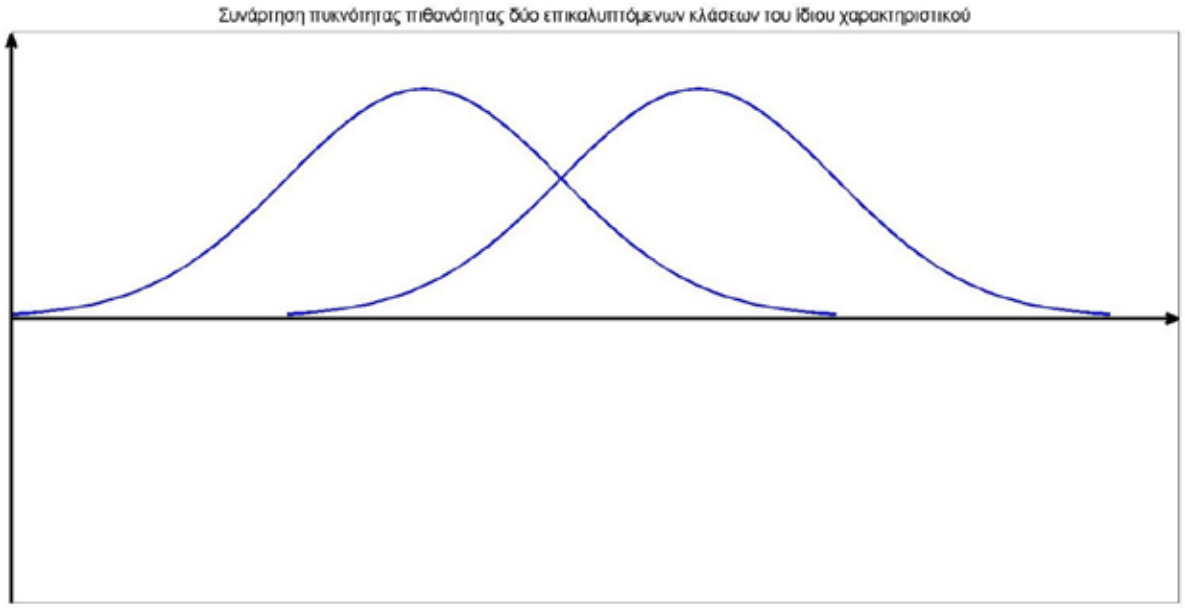
Παρόμοια με τη χρήση της τιμής του student's t-test ως μέτρο διαχωρισιμότητας των κλάσεων, έχουν επίσης προταθεί και τα ακόλουθα μέτρα βλ. [40]:

- Απόκλιση (divergence)
- Απόσταση Bhattacharyya
- Κριτήριο FDR (Fisher's Discriminant Ratio)

2.6.2 Επιλογή χαρακτηριστικών με χρήση καμπυλών ROC

Στατιστικές μέθοδοι όπως ο στατιστικός έλεγχος υποθέσεων που περιγράψαμε προηγουμένως, παρέχουν στατιστική πληροφορία σχετικά με τη διαφορά των μέσων τιμών των κλάσεων ενός χαρακτηριστικού. Ο ρόλος αυτών των μεθόδων είναι καταλυτικός, αλλά όχι καταληκτικός: ενώ η πληροφορία που παρέχουν είναι χρήσιμη κι επαρκής για την απόρριψη χαρακτηριστικών, σε περίπτωση που οι μέσες τιμές των κλάσεων ενός χαρακτηριστικού

βρίσκονται κοντά, είναι ανεπαρκής, και δεν εγγυάται ότι χαρακτηριστικά που έχουν περάσει τον στατιστικό έλεγχο υποθέσεων έχουν απαραίτητα καλή διαχωριστική ικανότητα κλάσεων. Οι μέσες τιμές μπορεί να διαφέρουν σημαντικά, εν τούτοις η έκταση γύρω από τις μέσες τιμές μπορεί να είναι αρκετά μεγάλη ώστε να μην είναι διακριτός ο διαχωρισμός των κλάσεων. Για το λόγο αυτό, είναι απαραίτητες τεχνικές οι οποίες παρέχουν πληροφορία για επικαλυπτόμενες κλάσεις (εικόνα 62).



Εικόνα 62: *Επικαλυπτόμενες συναρτήσεις πυκνότητας πιθανότητας δύο κλάσεων του ίδιου χαρακτηριστικού*

Η καμπύλη ROC (Receiver Operating Characteristics curve) είναι μια γραφική παράσταση που χρησιμοποιείται για να απεικονίσει την απόδοση ενός δυαδικού ταξινομητή. Έχει ιδιότητες που την κάνουν ιδιαίτερα χρήσιμη σε προβλήματα με επικαλυπτόμενες κατανομές κλάσεων και άνισα σφάλματα ταξινόμησης, επειδή παρέχει πληροφορίες σχετικά με την επικάλυψη μεταξύ των κλάσεων.

(Ιστορικό πλαίσιο: Η καμπύλη ROC είχε αναπτυχθεί από ηλεκτρολόγους μηχανικούς κατά τον Β' Παγκόσμιο Πόλεμο για εντοπισμό των αντίπαλων αεροσκαφών στα πεδίο της μάχης. Μετά την επίθεση στο Pearl Harbor το 1941, ο στρατός των ΗΠΑ ξεκίνησε μια έρευνα με στόχο να αυξήσει την προβλεψιμότητα στα ορθώς ανιχνευθέντα ιαπωνικά αεροσκάφη από τα σήματα του ραδιοεντοπιστή. Αργότερα, οι καμπύλες ROC ξεκίνησαν να χρησιμοποιούνται εκτενώς στην ιατρική και στον τομέα της τεχνικής νοημοσύνης.)

Η ανάλυση ROC χρησιμοποιείται για την αξιολόγηση και σύγκριση διαφόρων ταξινομητών μεταξύ τους. Για να την χρησιμοποιήσουμε στο πρόβλημα της επιλογής χαρακτηριστικών, θεωρούμε ταξινομητές με είσοδο ένα μόνο χαρακτηριστικό (single variable classifiers): αξιολογούμε κάθε χαρακτηριστικό σύμφωνα με την ατομική του ικανότητα πρόβλεψης, χρησιμοποιώντας ως κριτήριο την απόδοση ενός ταξινομητή κατασκευασμένου με μια μόνο μεταβλητή. Ένας τέτοιος ταξινομητής προκύπτει θέτοντας ένα κατώφλι θ στην τιμή του χαρακτηριστικού (όπου θ θα ανήκει στο εύρος τιμών του υπό εξέταση χαρακτηριστικού), το οποίο θα λειτουργεί ως διαχωριστικό όριο ανάμεσα στις δύο κλάσεις [28].

Μια καμπύλη ROC είναι μια δισδιάστατη γραφική παράσταση του true positive rate, γνωστού και ως ευαισθησία, συναρτήσει του false positive rate, που υπολογίζεται ως η διαφορά (1-ειδικότητα), για ένα δυαδικό ταξινομητή, του οποίου το κατώφλι διάκρισης των δύο κλάσεων ποικίλει. Κάθε σημείο της καμπύλης ROC αναπαριστά ένα ζευγάρι SN/(1-SP) που αντιστοιχεί σε ένα συγκεκριμένο κατώφλι απόφασης ενός χαρακτηριστικού [41]. Η ευαισθησία(SN) και η ειδικότητα(SP) είναι στατιστικές μετρήσεις της απόδοσης μιας δοκιμής δυαδικής ταξινόμησης. Η ευαισθησία (sensitivity), γνωστή και ως true positive rate, μετρά την αναλογία των θετικών δειγμάτων που αναγνωρίστηκαν ορθά, ενώ η ειδικότητα(specificity/true negative rate) μετρά την αναλογία των αρνητικών δειγμάτων τα οποία αναγνωρίστηκαν ορθά.

Έστω ένα πρόβλημα ταξινόμησης δύο κλάσεων του οποίου τα αποτελέσματα χαρακτηρίζονται ως θετικά (positive) ή ως αρνητικά (negative). Οι πιθανές καταστάσεις που μπορούν να προκύψουν είναι 4:

- True positive (TP): η απόφαση του ταξινομητή είναι θετική και συμφωνεί με την πραγματική τιμή της κλάσης
- True negative (TN): η απόφαση του ταξινομητή είναι αρνητική και συμφωνεί με την πραγματική τιμή της κλάσης
- False Positive (FP): η απόφαση του ταξινομητή είναι θετική αλλά η πραγματική τιμή της κλάσης είναι αρνητική
- False negative (FN): η απόφαση του ταξινομητή είναι αρνητική αλλά η πραγματική τιμή της κλάσης είναι θετική

Ένα παράδειγμα από τον πραγματικό κόσμο θα ήταν μια διαγνωστική εξέταση που προσπαθεί να αποφασίσει εάν ένα άτομο έχει μια συγκεκριμένη ασθένεια: θετική εξέταση σημαίνει το άτομο ασθενεί από τη συγκεκριμένη ασθένεια ενώ αρνητική εξέταση σημαίνει το άτομο δεν έχει την ασθένεια. Η κατάσταση FP παρουσιάζεται όταν η διαγνωστική εξέταση έχει θετικό αποτέλεσμα ενώ στην πραγματικότητα το άτομο δεν έχει την ασθένεια και η κατάσταση FN

παρουσιάζεται όταν η διαγνωστική εξέταση έχει αρνητικό αποτέλεσμα, αλλά στην πραγματικότητα το άτομο έχει την ασθένεια. Οι καταστάσεις TP και TN παρουσιάζονται όταν το αποτέλεσμα της διαγνωστικής εξέτασης συμφωνεί με την πραγματικότητα. Στον *πίνακα 10* παρουσιάζονται οι πιθανές καταστάσεις.

Πίνακας 10: Πιθανές καταστάσεις

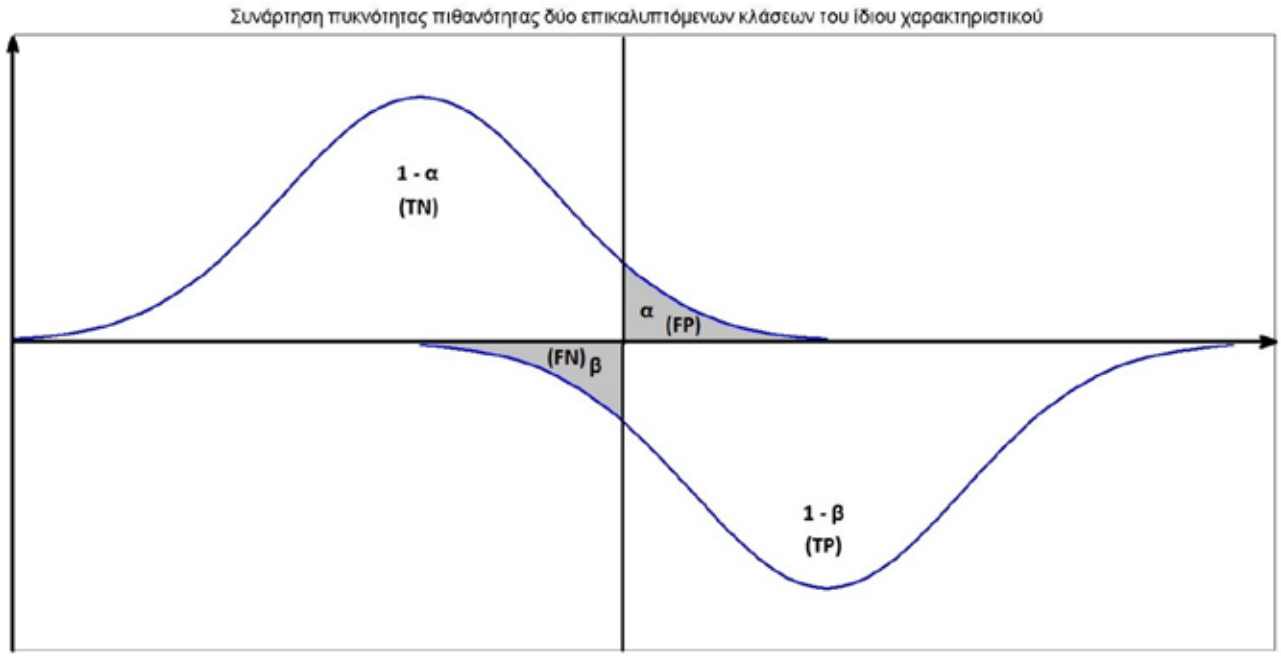
	Παρουσία νόσου / Πραγματική τιμή κλάσης (P)	Απουσία νόσου / Πραγματική τιμή κλάσης (N)
Εξέταση θετική / Θετική πρόβλεψη ταξινομητή (P)	TP	FP
Εξέταση αρνητική / Αρνητική πρόβλεψη ταξινομητή (N)	FN	TN

Ευαισθησία (SN/TPR): $SN = TPR = \frac{TP}{TP+FN}$

Ειδικότητα (SP/TNR): $SP = TNR = \frac{TN}{TN+FP}$

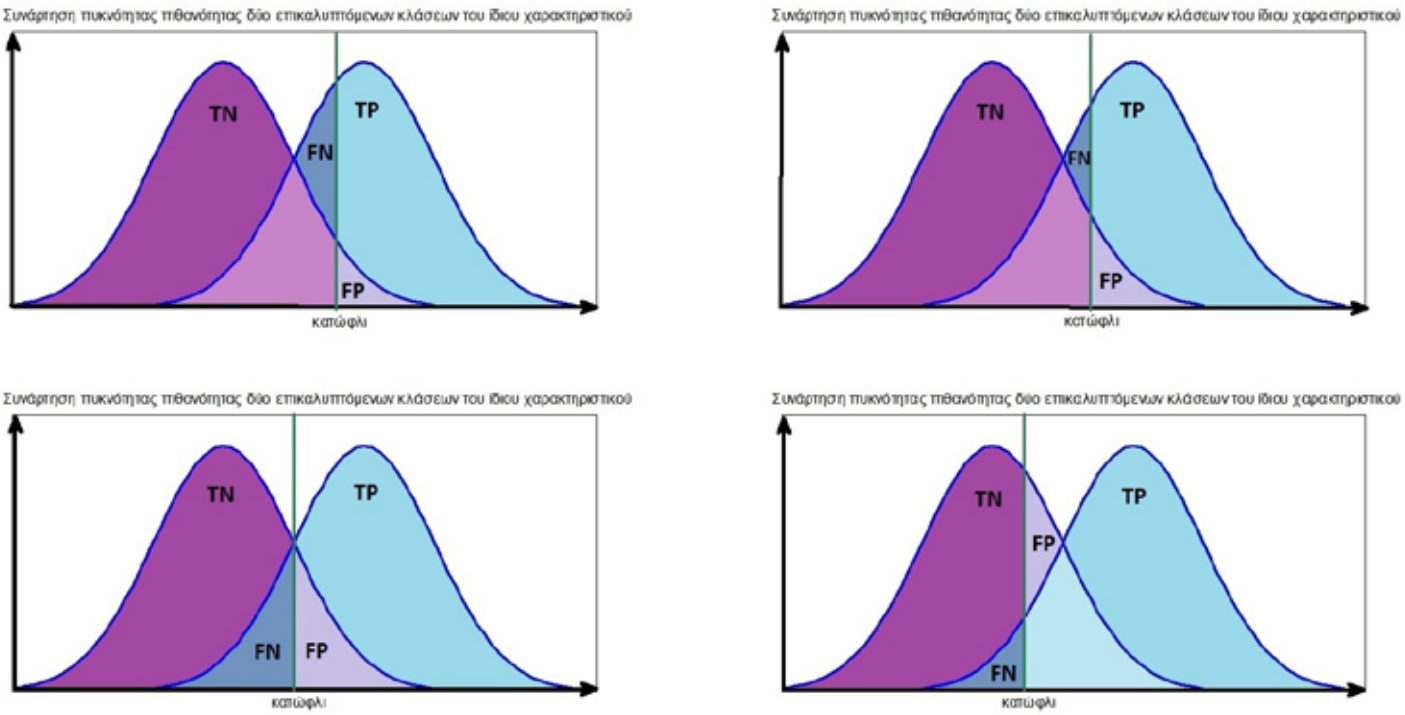
(1-Ειδικότητα) : $FPR = 1 - TNR = 1 - \frac{TN}{TN+FP} = \frac{FP}{TN+FP}$

Στην εικόνα 63 παρουσιάζεται ένα παράδειγμα δύο επικαλυπτόμενων συναρτήσεων πυκνότητας πιθανότητας που περιγράφουν την κατανομή ενός χαρακτηριστικού σε δύο κλάσεις, μαζί με ένα κατώφλι (η μια συνάρτηση πυκνότητας πιθανότητας είναι αντεστραμμένη έτσι ώστε να είναι πιο ευδιάκριτη). Οι τιμές αριστερά του κατωφλίου ανήκουν στην κλάση ω1 , ενώ οι τιμές δεξιά του κατωφλίου ανήκουν στην κλάση ω2. Βάσει του κατωφλίου, η απόφαση που θα παρθεί μπορεί να καταλήγει σε εσφαλμένο συμπέρασμα για την κλάση ω1 με πιθανότητα α, ή σε ορθό συμπέρασμα με πιθανότητα (1-α). Ομοίως, εσφαλμένη απόφαση σχετικά με την κλάση ω2 μπορεί να παρθεί με πιθανότητα β, και ορθή απόφαση με πιθανότητα (1-β). Η πιθανότητα εσφαλμένης απόφασης α ή β ισούται με την σκιασμένη περιοχή κάτω από την αντίστοιχη καμπύλη, όπως φαίνεται στην εικόνα 63. Αν θεωρήσουμε ότι η κλάση ω1 αντιστοιχεί σε αρνητική έξοδο (N) του ταξινομητή και η κλάση ω2 σε θετική έξοδο (P), τότε η πιθανότητα (1-β) αντιστοιχεί στην πιθανότητα κατάστασης TP και η πιθανότητα α στην πιθανότητα κατάστασης FP.



Εικόνα 63: *Επικαλυπτόμενες συναρτήσεις πυκνότητας πιθανότητας δύο κλάσεων του ίδιου χαρακτηριστικού (η μια είναι αντεστραμμένη για να είναι πιο ευδιάκριτη) μαζί με ένα κατώφλι*

Η ROC καμπύλη, όπως έχουμε αναφέρει, είναι μια γραφική παράσταση του true positive rate συναρτήσει του false positive rate ή αλλιώς της SN συναρτήσει του (1-SP). Στην εικόνα 64 παρουσιάζεται η εφαρμογή κατωφλίου απόφασης ενός ταξινομητή καθώς σαρώνει τις επικαλυπτόμενες συναρτήσεις πυκνότητας πιθανότητας δύο κλάσεων του ίδιου χαρακτηριστικού. Από κάθε σάρωση με διαφορετική τιμή κατωφλίου προκύπτουν διαφορετικές τιμές α και β και κάθε φορά προστίθεται ένα σημείο (α, (1-β)) , δηλαδή (FPR,TPR) στην καμπύλη ROC. Σε κάθε τιμή κατωφλίου αντιστοιχεί διαφορετικό ποσοστό TP,FP, TN και FN αποτελεσμάτων. Η καμπύλη ROC παριστά ένα σύνολο σημείων (FPR,TPR) για άπειρες δυνατές τιμές κατωφλίου.



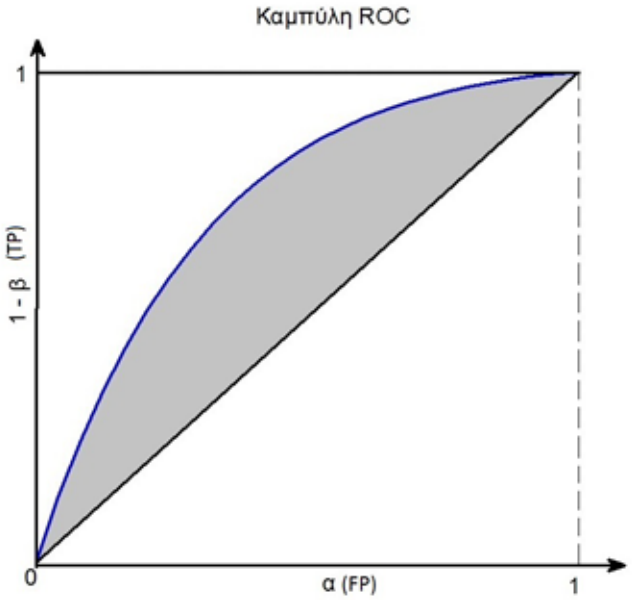
Εικόνα 64: *Εφαρμογή διαφόρων τιμών κατωφλίου σε επικαλυπτόμενες συναρτήσεις πυκνότητας πιθανότητας δύο κλάσεων του ίδιου χαρακτηριστικού*

Στην εικόνα 65 παρουσιάζεται η επακόλουθη καμπύλη ROC. Η καμπύλη ROC ενός συγκεκριμένου ταξινομητή επιδεικνύει την απόδοσή του καθώς εναλλάσσονται η ευαισθησία και η ειδικότητα: μια αύξηση στην ευαισθησία θα συνοδεύεται από μια μείωση στην ειδικότητα. Η καμπύλη ROC ενός τυχαίου ταξινομητή περνά από τα σημεία (0,0) και (1,1). Στο σημείο (0,0) όλα ταξινομούνται ως αρνητικά, δεν ταξινομείται κανένα ως θετικό (always negative): ο ταξινομητής βρίσκει όλες τις αρνητικές περιπτώσεις ορθά και όλες τις θετικές περιπτώσεις εσφαλμένα. Στο σημείο (1,1) όλα ταξινομούνται ως θετικά (always positive) κι επομένως ο ταξινομητής κατατάσσει όλες τις θετικές περιπτώσεις ορθά και όλες τις αρνητικές περιπτώσεις εσφαλμένα. Το σημείο (0,1) αναπαριστά τον ιδανικό ταξινομητή και το σημείο (1,0) αναπαριστά τον ταξινομητή που τα βρίσκει όλα εσφαλμένα.

Εάν οι δύο κατανομές των κλάσεων επικαλύπτονται πλήρως (εικόνα 67α), τότε για οποιοδήποτε τιμή του κατωφλίου θα ισχύει α = (1 – β), δηλαδή TPR = FPR. Αυτή η περίπτωση ROC καμπύλης είναι αυτή του τυχαίου ταξινομητή και αντιστοιχεί στην ευθεία διαγώνιο y=x από το σημείο (0,0) έως το σημείο (1,1) , γνωστή και ως no-discrimination line . Εάν ένας ταξινομητής κατατάσσει τυχαία τις μισές φορές στη θετική κλάση, αναμένεται να βρίσκει ορθά τα μισά θετικά και τα μισά αρνητικά: αυτό αναπαρίσταται ως το σημείο (0.5,0.5) στο χώρο ROC. Εάν κατατάσσει στη θετική κλάση 85% των φορών, αναμένεται να βρίσκει 85% των θετικών αποτελεσμάτων ορθά,

αλλά το FPR θα αυξηθεί επίσης κατά 85%: σημείο (0.85,0.85). Επομένως, ένας τυχαίος ταξινομητής θα παράγει ένα σημείο ROC που θα «ολισθαίνει» κατά μήκος της διαγωνίου y=x (εικόνα 67α) ανάλογα με τη συχνότητα με την οποία προβλέπει τη θετική κλάση. Ένας ταξινομητής με καμπύλη ROC στην άνω τριγωνική περιοχή της διαγωνίου y=x (εικόνα 65), έχει καλύτερη απόδοση από τον τυχαίο ταξινομητή και ένας ταξινομητής με καμπύλη ROC στην κάτω τριγωνική περιοχή, χειρότερη. Εάν βρίσκεται στην κάτω περιοχή, τότε ο ταξινομητής είναι πιο συχνά εσφαλμένος από ότι ορθός.

Στην εικόνα 65 βλέπουμε την επακόλουθη καμπύλη ROC από το παράδειγμα της εικόνας 62 σε σχέση με την καμπύλη ROC y=x. Καθώς οι κατανομές των κλάσεων απομακρύνονται, η αντίστοιχη καμπύλη ROC διαφοροποιείται από τη διαγώνιο y=x: όσο μικρότερη είναι η επικάλυψη των κλάσεων, τόσο μεγαλύτερο είναι το εμβαδό της σκιασμένης περιοχής μεταξύ της καμπύλης και της διαγωνίου [40].

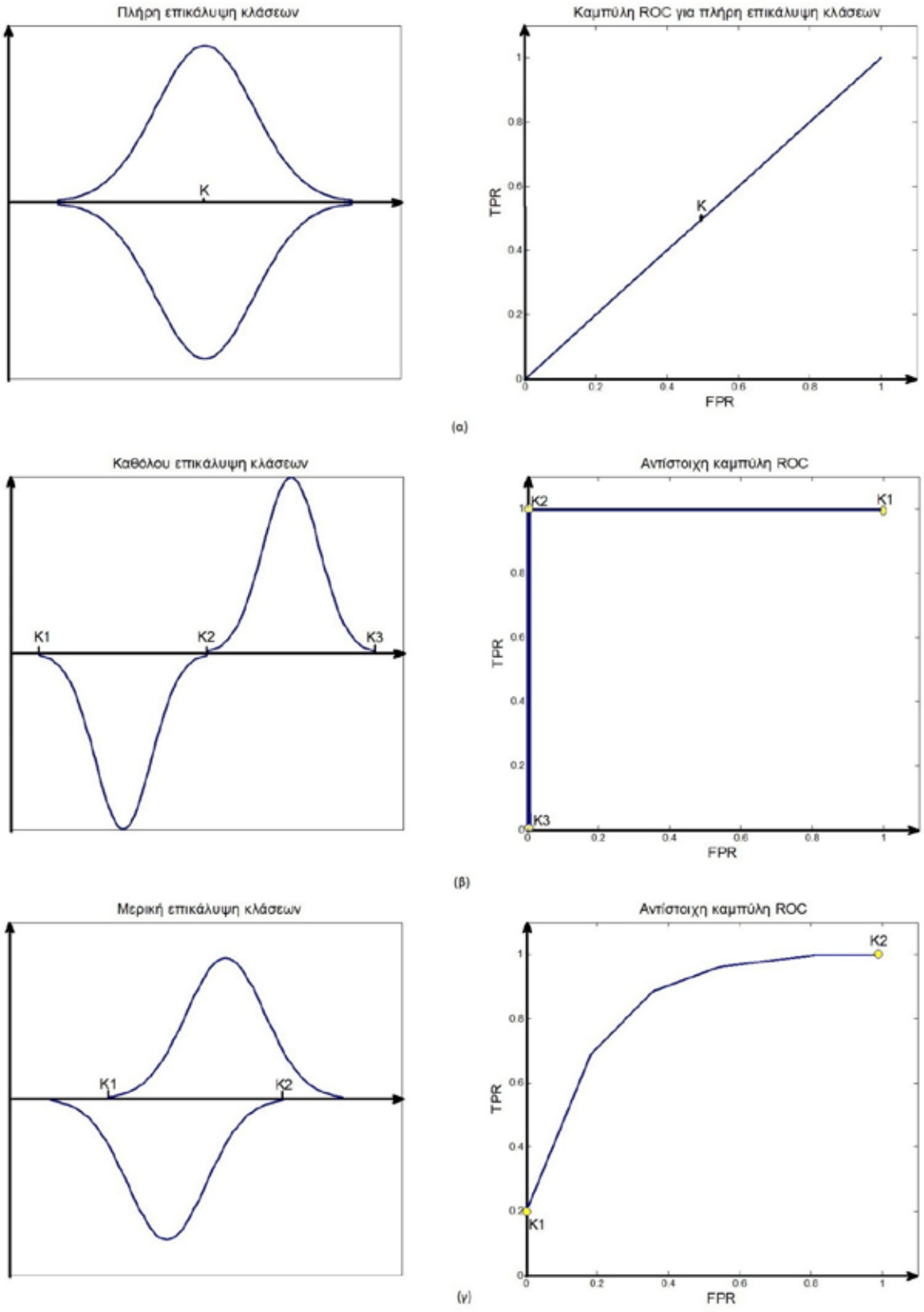


Εικόνα 65: *Επακόλουθη καμπύλη ROC από εικόνα 62 σε σχέση με την καμπύλη ROC όπου TPR=FPR (όσο μεγαλύτερη η σκιασμένη περιοχή τόσο μικρότερη είναι η επικάλυψη των κλάσεων)*

Η άλλη ακραία περίπτωση είναι όταν οι δύο κλάσεις είναι πλήρως διαχωρισμένες, δηλαδή δεν παρουσιάζουν καμία επικάλυψη (εικόνα 66β). Σε αυτή την περίπτωση, αν το κατώφλι μετακινείται έτσι ώστε να καλύψει όλο το εύρος των τιμών του α στο [0,1], το (1 – β) παραμένει ίσο με μονάδα. Επομένως, όπως φαίνεται στην εικόνα 65, το εμβαδό του τριγώνου που σχηματίζεται ανάμεσα στην καμπύλη ROC όταν οι κλάσεις είναι πλήρως διαχωρισμένες και στη διαγώνιο y=x θα είναι ½ .

Συνεπώς, βάσει των δύο ακραίων περιπτώσεων που έχουμε περιγράψει, παρατηρούμε ότι το εμβαδό της περιοχής ανάμεσα στην καμπύλη ROC και τη διαγώνιο y=x, θα κυμαίνεται από μηδέν για κατανομές που επικαλύπτονται πλήρως έως ½ για κατανομές που είναι πλήρως διαχωρισμένες. Το εμβαδό ανάμεσα στην καμπύλη ROC και τη διαγώνιο αποτελεί μέτρο της απόδοσης ενός ταξινομητή και κατ' επέκταση της διακριτικής ικανότητας ενός συγκεκριμένου χαρακτηριστικού.

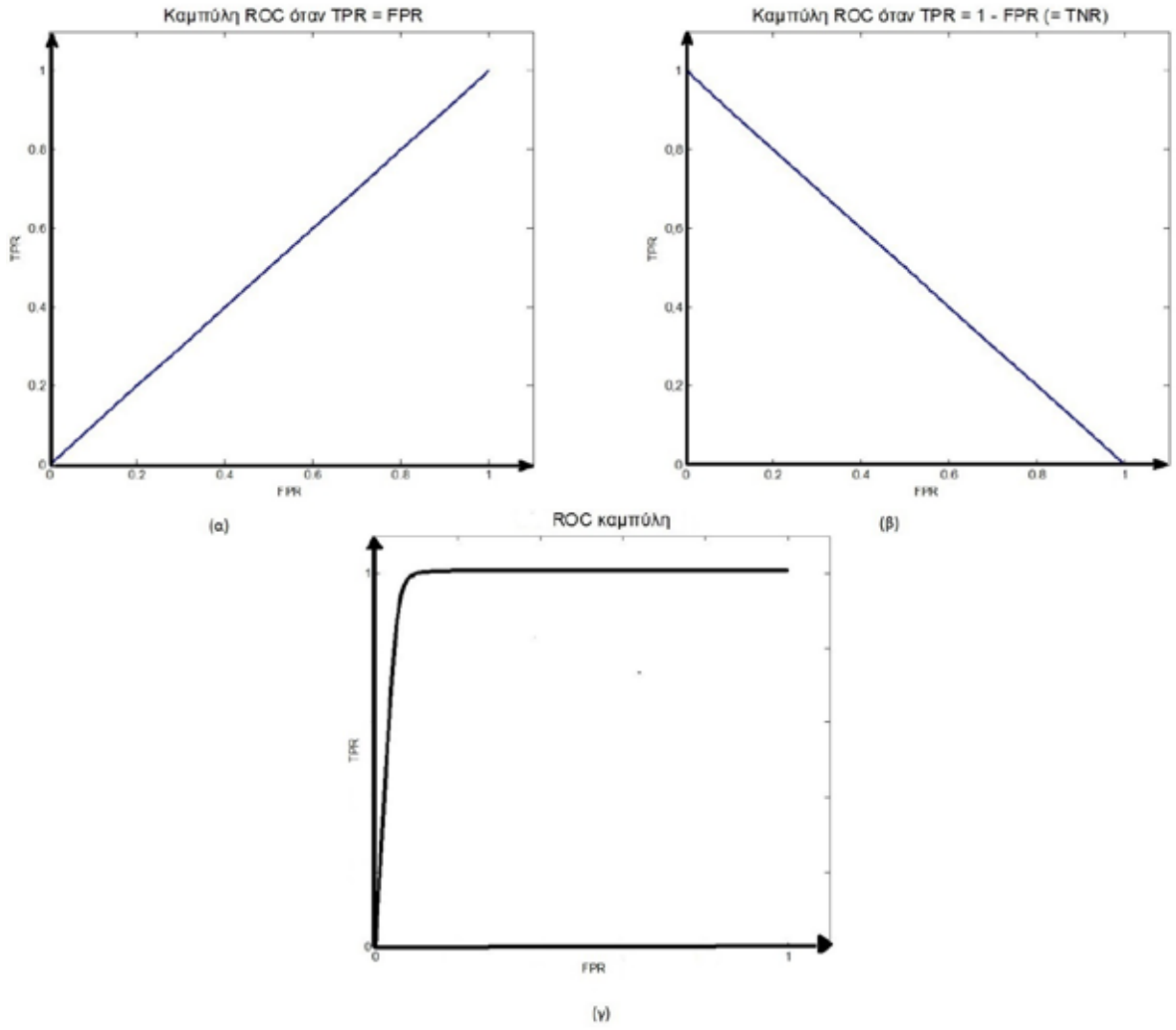
Στην εικόνα 66 παρουσιάζονται οι 3 περιπτώσεις όπου οι κατανομές των κλάσεων παρουσιάζουν πλήρη, καθόλου ή μερική επικάλυψη, μαζί με τις αντίστοιχες ROC καμπύλες.



Εικόνα 66: *(α) Πλήρη(χείριστη περίπτωση), (β) καθόλου(ιδανική περίπτωση) και (γ) μερική επικάλυψη κλάσεων (η θετική κλάση είναι αντεστραμμένη), με τις αντίστοιχες καμπύλες ROC, καθώς και με σημεία που παριστούν διάφορες τιμές κατωφλίου*

Η διαγώνιος y= 1-x (εικόνα 67β) αναπαριστά ταξινομητές οι οποίοι αποδίδουν εξίσου καλά και στις δύο κλάσεις , δηλαδή TPR=1-FPR =TNR : αριστερά της διαγωνίου y=1-x περιέχονται ταξινομητές που αποδίδουν καλύτερα στα αρνητικά παρά στα θετικά, ενώ δεξιά της διαγωνίου η απόδοση στα θετικά υπερτερεί.

Μια καμπύλη ROC με ιδανική ακρίβεια, θα ξεκινούσε από το σημείο (0,0), θα συνέχιζε κατακόρυφα έως το σημείο (0,1) και ακολούθως οριζόντια έως το σημείο (1,1) στη δεξιά άνω άκρη της γραφικής παράστασης (εικόνα 67γ). Στην πράξη, σε πραγματικά δεδομένα, οι καμπύλες ROC βρίσκονται κάπου στην άνω αριστερά περιοχή της γραφικής, ανάμεσα στην ιδανική καμπύλη ROC και την ευθεία y=x [41].



Εικόνα 67: *Καμπύλη ROC (α) όταν TPR=1-FPR=TNR, (β) όταν TPR = FPR και (γ) που προσεγγίζει την ιδανική καμπύλη ROC*

Μια παραλλαγή της χρήσης του εμβαδού ανάμεσα στην καμπύλη ROC και τη διαγώνιο, ως διαχωριστικό μέτρο, είναι η χρήση του εμβαδού της περιοχής κάτω από την καμπύλη ROC (area

under the curve - AUC) έως τους άξονες XY, το οποίο αξιολογεί πόσο καλά ένα χαρακτηριστικό μπορεί να διακρίνει μεταξύ δύο κλάσεων και παρέχει μια συνολική εκτίμηση για την απόδοσή του σε όρους ευαισθησίας και ειδικότητας. Μεγαλύτερες τιμές AUC υποδεικνύουν, κατά μέσο όρο, καλύτερη απόδοση. Επιπλέον παρουσιάζεται ενδιαφέρον στην ερμηνεία του AUC: εάν πάρουμε δύο τυχαία επιλεγμένα δείγματα, εκ των οποίων το ένα να ανήκει στη θετική κλάση και το άλλο στην αρνητική, το AUC θα ισούται με την πιθανότητα το θετικό δείγμα να έχει μεγαλύτερη πιθανότητα πρόβλεψης από το αρνητικό δείγμα. Άρα το AUC δίνει την πιθανότητα ο ταξινομητής να κατατάσσει ορθά τέτοια ζευγάρια δειγμάτων. Το εμβαδό κάτω από την καμπύλη ROC μπορεί να κυμαίνεται από ½ σε κατανομές με πλήρη επικάλυψη έως 1 σε κατανομές που είναι πλήρως διαχωρισμένες μεταξύ τους (αν και θεωρητικά 0≤AUC≤1, σε ρεαλιστικούς ταξινομητές ισχύει 0.5≤AUC≤1 , αφού η χείριστη περίπτωση, δηλαδή του τυχαίου ταξινομητή, δίνει τη διαγώνιο y=x με AUC=0.5) [42]. Το εμβαδό ανάμεσα στην καμπύλη ROC και τη διαγώνιο, που όπως προαναφέραμε χρησιμοποιείται κι αυτό ως μέτρο, ουσιαστικά ισούται με (AUC- ½). Καλύτερα χαρακτηριστικά, και στα δύο αυτά μέτρα, προφανώς θεωρούνται αυτά που έχουν υψηλότερη τιμή.

Πρακτικά, η καμπύλη ROC μπορεί εύκολα να κατασκευαστεί, μετακινώντας το κατώφλι και υπολογίζοντας τα ποσοστά εσφαλμένης και ορθής διάκρισης σε κάθε κατώφλι, για όλα τα χαρακτηριστικά του συνόλου εκπαίδευσης. Θεωρητικά το κατώφλι παίρνει τιμές από -∞ έως +∞, ωστόσο, σε πρακτικές εφαρμογές, προεπιλέγουμε ένα σύνολο, πεπερασμένο σε πλήθος, που αποτελείται από συγκεκριμένες τιμές κατωφλίου τις οποίες θεωρούμε κατάλληλες. Εμείς επιλέξαμε να χρησιμοποιήσουμε ως μέτρο αξιολόγησης της διαχωριστικής ικανότητας κάθε χαρακτηριστικού το εμβαδό του άνω τριγώνου που σχηματίζεται ανάμεσα στην καμπύλη ROC και τη διαγώνιο.

Το μειονέκτημα της χρήσης της καμπύλης ROC στο πρόβλημα της επιλογής χαρακτηριστικών, είναι ότι σε περίπτωση που ο αριθμός των χαρακτηριστικών που διαχωρίζουν επιτυχώς όλα τα δείγματα σε κλάσεις είναι μεγάλος (δηλαδή όταν δεν παρατηρείται επικάλυψη των δύο κλάσεων σε πολλά χαρακτηριστικά), τότε κριτήρια που είναι βασισμένα σε ποσοστά επιτυχούς ταξινόμησης, όπως αυτά της ανάλυσης ROC, δεν θα μπορούν να κάνουν διάκριση ανάμεσα στα υψηλότερα στην κατάταξη χαρακτηριστικά [28]. Συνεπώς, η χρήση της καμπύλης ROC έχει νόημα μόνο όταν έχουμε αρκετά χαρακτηριστικά για τα οποία παρατηρείται επικάλυψη των κατανομών των κλάσεων ή όταν χρησιμοποιείται μόνο σε ένα πρώτο στάδιο προεπεξεργασίας με σκοπό τον αποκλεισμό των άσχετων χαρακτηριστικών.

2.6.3 Τεχνική mRMR (minimum redundancy maximum relevance)

Η τεχνική φιλτραρίσματος μειωμένης περιττής πληροφορίας και μέγιστης συνάφειας (ή ελάχιστου πλεονασμού και μέγιστης σχετικότητας), γνωστή ως mRMR (minimum redundancy maximum relevance), έχει προταθεί από τους Peng et al [43] και είναι βασισμένη σε μέτρα αμοιβαίας πληροφορίας. Αμοιβαία πληροφορία δύο τυχαίων μεταβλητών είναι το μέτρο της μεταξύ τους αμοιβαίας εξάρτησης. Ο αλγόριθμος mRMR είναι ιδιαίτερα χρήσιμος στην περίπτωση που έχουμε να κάνουμε με χαρακτηριστικά μεγάλου μεγέθους, ή όταν αντιμετωπίζουμε προβλήματα επιλογής χαρακτηριστικών από χιλιάδες υποψήφια χαρακτηριστικά.

Το πλεονέκτημά της mRMR σε σχέση με τις υπόλοιπες μεθόδους φιλτραρίσματος είναι ότι λογαριάζει στην αξιολόγηση των χαρακτηριστικών τις συσχετίσεις που μπορεί να υπάρχουν ανάμεσα στα χαρακτηριστικά [43]. Σε άλλες μεθόδους, εάν ένα χαρακτηριστικό x_i καταταχθεί υψηλά, άλλα χαρακτηριστικά που έχουν υψηλή συσχέτιση με το x_i , είναι πολύ πιθανόν να καταταχθούν επίσης υψηλά και κατά συνέπεια να επιλεγούν. Αυτό όμως δεν είναι αποδοτικό, αφού, όπως έχουμε εξηγήσει στο *2.3.1*, αυτά τα χαρακτηριστικά λόγω της υψηλής συσχέτισης μεταξύ τους είναι πλεονάζοντα/περιττά και δεν αποφέρουν ουσιαστικό κέρδος, παρά μόνο επιβαρύνουν σημαντικά το υπολογιστικό κόστος. Στην mRMR, συναφή χαρακτηριστικά (relevant features) και περιττά χαρακτηριστικά (redundant features) εξετάζονται ταυτόχρονα.

Η μέθοδος mRMR μελετά την αμοιβαία πληροφορία ανάμεσα στο κάθε χαρακτηριστικό και την στοχευόμενη κλάση και παράλληλα λαμβάνει υπόψη και τις εξαρτήσεις μεταξύ των χαρακτηριστικών. Τα χαρακτηριστικά που ικανοποιούν τους δύο υποστόχους - να έχουν την υψηλότερη συνάφεια με την κλάση και την ελάχιστη περιττή πληροφορία ανάμεσά τους - επιλέγονται. Ουσιαστικά, τα επιλεγμένα χαρακτηριστικά πρέπει να είναι ως προς την αμοιβαία πληροφορία όσο το δυνατόν ανόμοια το ένα προς το άλλο, ενώ να εξακολουθούν να είναι, το καθένα ξεχωριστά, όσο το δυνατόν πιο όμοια με τη μεταβλητή της κλάσης. Η μεγιστοποίηση της ανομοιότητας των χαρακτηριστικών μπορεί να επιτευχθεί με διάφορους τρόπους, όπως π.χ. μεγιστοποιώντας τις αμοιβαίες ευκλείδειες αποστάσεις τους ή ελαχιστοποιώντας τις ανά ζεύγος συσχετίσεις τους [43]. Ως μέτρο της συνάφειας κάθε χαρακτηριστικού με την κλάση, μπορεί να χρησιμοποιηθεί η μεγιστοποίηση της αμοιβαίας πληροφορίας τους ή κάποιο άλλο από τα συνήθη κριτήρια που μεγιστοποιούν τη συνάφεια.

Η περιττή πληροφορία όλων των χαρακτηριστικών σε ένα υποσύνολο χαρακτηριστικών – με κριτήριο την αμοιβαία πληροφορία - είναι η μέση τιμή όλων των τιμών αμοιβαίας πληροφορίας

ανάμεσα σε κάθε ζεύγος χαρακτηριστικών. Η ελαχιστοποίηση της περιττής πληροφορίας (minimum redundancy) παρουσιάζεται στον παρακάτω τύπο:

$$\min W_I\,,\;\; W_I = \frac{1}{|S|^2}\sum_{i,j\in S} I(i,j)$$

, όπου S το υποσύνολο των ήδη επιλεγμένων χαρακτηριστικών και I (i,j) η αμοιβαία πληροφορία ανάμεσα στα χαρακτηριστικά i και j .

Η αμοιβαία πληροφορία I(i, j) ορίζεται ως:

$$I\,(i,j) = \iint p(i,j)\,\log\frac{p(i,j)}{p(i)p(j)}\,di\,dj\;\geq 0$$

, όπου p(i, j) η συνάρτηση της από κοινού κατανομής πιθανότητας (joint probability distribution function) των μεταβλητών i και j και p(i), p(j) οι συναρτήσεις των περιθώριων κατανομών πιθανότητας (marginal distribution probability functions) των μεταβλητών i και j αντίστοιχα.

Η αμοιβαία πληροφορία καθορίζει την εξάρτηση δύο μεταβλητών συγκρίνοντας την από κοινού τους πιθανότητα με τις περιθώριες πιθανότητες. Εάν δύο χαρακτηριστικά, έστω i και j, είναι δύο ανεξάρτητες τυχαίες μεταβλητές τότε η αμοιβαία πληροφορία τους I(i, j) θα ισούται με 0 (εάν i, j ανεξάρτητες τότε p(i, j) = p(i)p(j) και άρα $\log\frac{p(i,j)}{p(i)p(j)} = \log 1 = 0$).

Η συνάφεια ενός υποσυνόλου χαρακτηριστικών με μια συγκεκριμένη κλάση ορίζεται ως η μέση τιμή όλων των τιμών αμοιβαίας πληροφορίας ανάμεσα σε κάθε μεμονωμένο χαρακτηριστικό και τη συγκεκριμένη κλάση. Η μεγιστοποίηση της συνάφειας (maximum relevance) δίνεται από τον ακόλουθο τύπο:

$$max\,V_I\,,\;\; V_I = \frac{1}{|S|}\sum_{i\in S} I(i,h)$$

, όπου h μια από τις κλάσεις του προβλήματος.

Στόχος είναι η ταυτόχρονη βελτιστοποίηση και των δύο συνθηκών: ελαχιστοποίηση της περιττής πληροφορίας και μεγιστοποίηση της συνάφειας. Επειδή, όμως, η αύξηση της συνάφειας συνήθως συνοδεύεται με αύξηση της περιττής πληροφορίας, δεν υπάρχει μοναδικό υποσύνολο που να υπερέχει έναντι των άλλων με βάση και τα δύο κριτήρια. Για το λόγο αυτό, το κριτήριο mRMR συνδυάζει τα δύο προαναφερθέντα μέτρα σε μια αντικειμενική συνάρτηση. Τέτοια κριτήρια μπορεί να είναι η μεγιστοποίηση της διαφοράς των μέτρων max (V-W) ή του λόγου τους max ($\frac{V}{W}$).

$$\max_{i\in\Omega_S}\;[\;\; I(i,h) - \frac{1}{|S|}\sum_{j\in S} I(i,j)\;]$$

$$\max_{i \in \Omega_S} \{ \ I(i,h) / [\frac{1}{|S|} \sum_{j \in S} I(i,j) \ ] \}$$

, όπου Ω το σύνολο όλων των χαρακτηριστικών.

Η μέθοδος mRMR δημιουργήθηκε ώστε να καλύψει δυσκολίες που υπήρχαν με μια άλλη μέθοδο, η οποία χρησιμοποιούσε για τον υπολογισμό της εξάρτησης των χαρακτηριστικών με την κλάση, το κριτήριο μέγιστης εξάρτησης (max-dependency), που επίσης ορίζεται με όρους αμοιβαίας πληροφορίας. Η mRMR θεωρείται ως η πιο ευρέως χρησιμοποιούμενη προσέγγιση του κριτηρίου μέγιστης εξάρτησης. Βάσει της θεωρίας πληροφορίας, το βέλτιστο υποσύνολο χαρακτηριστικών είναι αυτό που έχει τη μέγιστη αμοιβαία πληροφορία με την κλάση. Το κριτήριο μέγιστης εξάρτησης χρησιμοποιείται για την εύρεση του βέλτιστου υποσυνόλου χαρακτηριστικών S με m χαρακτηριστικά, τα οποία θα έχουν από κοινού, τη μεγαλύτερη αμοιβαία πληροφορία και κατ' επέκταση τη μεγαλύτερη εξάρτηση, με την κλάση h. Το υποσύνολο χαρακτηριστικών που πρέπει να επιλεγεί ως βέλτιστο, είναι αυτό που μεγιστοποιεί το κριτήριο μέγιστης εξάρτησης, το οποίο παρουσιάζεται στον επόμενο τύπο.

$$max \ V_I \ , \ \ V_I = \ I(\{x_i \ , i = 1, \dots, m\}, h) = \ I \ (S_m, h) \ , \acute{o}\pi ou \ S_m = \{x_1, x_2, \dots, x_m\}$$

Για m>1 η αμοιβαία πληροφορία $I \ (S_m, h)$ ορίζεται ως:

$$I \ (S_m, h) = \ \iint p(S_m, h) \ \log \frac{p(S_m, h)}{p(S_m) p(h)} \ dS_m \ dh$$

$$= \ \iiint p(S_{m-1}, x_m, h) \ \log \frac{p(S_{m-1}, x_m, h)}{p(S_{m-1}, x_m) p(h)} \ dS_{m-1} \ dx_m \ dh$$

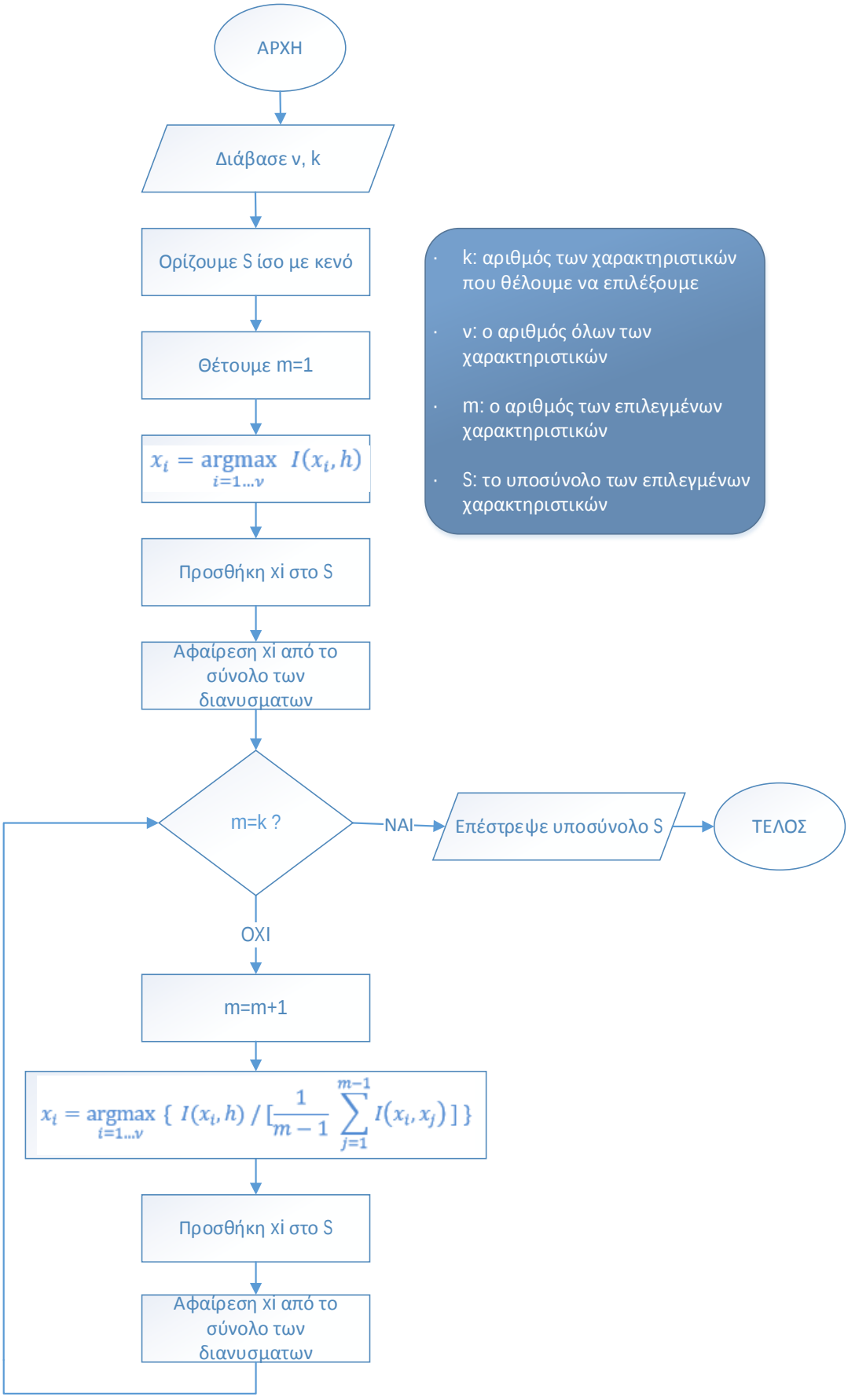
$$= \ \iint \dots \iint p(x_1, x_2, \dots, x_m, h) \ \log \frac{p(x_1, x_2, \dots, x_m, h)}{p(x_1, x_2, \dots, x_m) p(h)} \ dx_1 \ dx_2 \ \dots \ dx_m \ dh$$

, όπου $p(S_m)$ η από κοινού κατανομή των x_1, x_2, x_3, … , x_m και $p(S_m,h)$ η από κοινού κατανομή των x_1, x_2, x_3, … , x_m και h.

Το κριτήριο μέγιστης εξάρτησης αξιολογεί διάφορα υποσύνολα χαρακτηριστικών και επιλέγει το βέλτιστο από αυτά. Αντίθετα, η αντικειμενική συνάρτηση που χρησιμοποιείται στη μέθοδο mRMR, αντί να αξιολογεί υποσύνολα χαρακτηριστικών, αξιολογεί τα χαρακτηριστικά μεμονωμένα και βήμα προς βήμα δημιουργεί το σύνολο S που αποτελείται από m βέλτιστα χαρακτηριστικά. Επειδή το κριτήριο συνάφειας που χρησιμοποιείται στην mRMR, εξετάζει την αμοιβαία πληροφορία κάθε μεμονωμένου χαρακτηριστικού με την κλάση, αν χρησιμοποιηθεί μόνο του, προκύπτουν πολλά περιττά χαρακτηριστικά. Συνεπώς, στην αντικειμενική συνάρτηση mRMR είναι απολύτως απαραίτητο και το κριτήριο μειωμένης περιττής πληροφορίας μαζί με αυτό της μέγιστης συνάφειας, κάτι που είναι αχρείαστο στο κριτήριο μέγιστης εξάρτησης.

Παρόλο που το κριτήριο της μέγιστης εξάρτησης είναι αποδοτικό στις περιπτώσεις που απαιτείται η επιλογή λίγων χαρακτηριστικών από ένα μεγάλο σύνολο, δεν είναι κατάλληλο στις περιπτώσεις όπου χρειάζεται να επιτευχθεί ταξινόμηση με υψηλή ακρίβεια. Βασικό πλεονέκτημα της μεθόδου mRMR είναι ότι η εκτίμηση της συνάφειας και της περιττής πληροφορίας είναι χαμηλών διαστάσεων προβλήματα που αφορούν δύο μόνο μεταβλητές. Αυτό καθιστά την mRMR ένα πολύ πιο εύκολο πρόβλημα από αυτό του άμεσου υπολογισμού της πυκνότητας ή της αμοιβαίας πληροφορίας πολλών μεταβλητών σε ένα υψηλών διαστάσεων χώρο. Αν και το πρόβλημα της μέγιστης εξάρτησης έχει μεγαλύτερη θεωρητική αξία, είναι δύσκολη μια ακριβής εκτίμηση της εξάρτησης του υποσυνόλου χαρακτηριστικών με την κλάση αφού είναι δύσκολο να προσδιοριστούν ακριβώς οι πυκνότητες πιθανότητας p(x1, … ,xm) και p(x1, … , xm, h) και επομένως η αμοιβαία πληροφορία I({x1, …, xm}, h). Η mRMR αποφεύγει τον υπολογισμό των ποσοτήτων αυτών, υπολογίζοντας μόνο τις ποσότητες p(xi, xj) και p(xi, h) , κάτι που είναι προφανώς πιο εύκολο και ακριβές. Συνεπώς, η εκτίμηση της μέγιστης συνάφειας μέσω του κριτηρίου mRMR είναι πιο αξιόπιστη και παράλληλα προσεγγίζει πολύ καλά τη θεωρητική αξία του κριτηρίου της μέγιστης εξάρτησης. Επιπρόσθετα, το κριτήριο συνάφειας που χρησιμοποιείται στην mRMR έχει μικρότερο υπολογιστικό κόστος και υψηλότερη ταχύτητα από το κριτήριο του άμεσου υπολογισμού της αμοιβαίας πληροφορίας ανάμεσα σε όλο το υποσύνολο S και την κλάση h.

Ο προτεινόμενος αλγόριθμος από τους Peng et al [43] που χρησιμοποιείται για τη βελτιστοποίηση της αντικειμενικής συνάρτησης της mRMR είναι ένας απλός ευριστικός αλγόριθμος: προκειμένου η διαδικασία mRMR να γίνει υπολογιστικά εφικτή, το υποσύνολο σχηματίζεται χρησιμοποιώντας μια forward μέθοδο άπληστης αναζήτησης (βλ. 7.1.1 sequential forward selection). Αρχικά επιλέγεται το πιο συναφές χαρακτηριστικό, αυτό δηλαδή με τη μεγαλύτερη αμοιβαία πληροφορία με την κλάση. Ακολούθως, σε κάθε επανάληψη επιλέγεται το χαρακτηριστικό με τη μεγαλύτερη συνάφεια με την κλάση και ταυτόχρονα τη μικρότερη ομοιότητα με τα ήδη επιλεγμένα χαρακτηριστικά. Βασικό μειονέκτημα του forward αλγόριθμου άπληστης αναζήτησης είναι ότι δεν έχει τη δυνατότητα να αφαιρεί περιττά χαρακτηριστικά από αυτά που έχει ήδη επιλέξει. Στην *εικόνα 68* παρουσιάζεται το διάγραμμα ροής του ψευδοκώδικα ενός ενδεικτικού αλγόριθμου mRMR.



Εικόνα 68: *Διάγραμμα ροής ενδεικτικού αλγόριθμου mRMR*

2.6.4 Αλγόριθμος Relief (RELevance In Estimated Features)

Ο αλγόριθμος Relief (RELevance In Estimated Features), ο οποίος έχει προταθεί από τους Kira και Rendell [44] για προβλήματα δυαδικής ταξινόμησης, επιλέγει τα συναφή χαρακτηριστικά με χρήση μιας στατιστικής μεθόδου: τον κανόνα του πλησιέστερου γείτονα (nearest-neighbor rule). Η μέθοδος αυτή, μπορεί να μην εντοπίζει το μικρότερο δυνατό υποσύνολο χαρακτηριστικών, αλλά επειδή επιλέγει μόνο τα στατιστικώς σχετικά χαρακτηριστικά, καταλήγει σε ένα αρκετά μικρό υποσύνολο το οποίο αποτελείται από αυτά. Ο αλγόριθμος RELIEF είναι αρκετά γρήγορος, αφού απαιτεί γραμμικό χρόνο ως προς τον αριθμό των δοσμένων χαρακτηριστικών και των δειγμάτων εκπαίδευσης [44]. Στα βασικά του πλεονεκτήματα συγκαταλέγονται: η καλή του ακρίβεια ακόμα κι όταν τα χαρακτηριστικά αλληλεπιδρούν, η ανεξαρτησία του από ευριστικούς μηχανισμούς, η χαμηλή του πολυπλοκότητα και η χρήση του τόσο σε διακριτά όσο και σε συνεχή χαρακτηριστικά. Τα μειονεκτήματά του περιλαμβάνουν: ανεπάρκεια να διακρίνει τα περιττά χαρακτηριστικά, περιορισμένη αποδοτικότητα όταν ο αριθμός των στιγμιότυπων εκπαίδευσης είναι μικρός, μικρή ανεκτικότητα στο θόρυβο και μη εφαρμογή σε ελλιπή δεδομένα.

Ο αλγόριθμος είναι βασισμένος σε μια απλή αρχή. Έχει στόχο να κατατάξει αντικείμενα με παρόμοιες ιδιότητες σε μια κλάση. Κάποιες από αυτές τις ιδιότητες (χαρακτηριστικά) είναι πολύ σημαντικές για την ταξινόμηση, ενώ άλλες είναι λιγότερο σημαντικές. Για παράδειγμα ας θεωρήσουμε το πρόβλημα του διαχωρισμού αυτοκινήτων από αεροπλάνα (*εικόνα 69*). Έστω ότι έχουμε πληθώρα αυτοκινήτων και αεροπλάνων. Οι δύο κλάσεις έχουν αρκετές γνωστές κοινές ιδιότητες: αποτελούν και οι δύο μέσα μεταφοράς, χρησιμοποιούν και οι δύο μηχανή, τιμόνι κτλ. Ωστόσο μπορείς να ισχυριστείς ότι δεν έχει κάθε αεροπλάνο μηχανή: κάποια αεροπλάνα είναι ανεμόπτερα που χρησιμοποιούν σταθερές πτέρυγες που με την προς τα εμπρός κίνησή τους παράγουν ικανή άνωση για την ανύψωσή τους χωρίς τη βοήθεια μηχανής. Εάν ρωτήσεις ένα παιδί, τι διαχωρίζει τα αυτοκίνητα από τα αεροπλάνα, πιθανότατα θα απαντήσει ότι τα αεροπλάνα πετούν, ενώ τα αυτοκίνητα όχι, ή ότι τα αεροπλάνα έχουν φτερά ενώ τα αυτοκίνητα όχι. Συνεπώς αυτά τα χαρακτηριστικά είναι σημαντικά. Η πρακτική που εφαρμόζει ο αλγόριθμος Relief είναι η ακόλουθη: εάν επιλέξουμε ένα τυχαίο αυτοκίνητο X, ψάχνει για ένα αυτοκίνητο Y, που να μοιάζει πιο πολύ στο αυτοκίνητο X, και για ένα αεροπλάνο Z επίσης όσο το δυνατόν πιο όμοιο με το αυτοκίνητο X. Ακολούθως εξετάζει τα χαρακτηριστικά στα οποία διαφέρουν, αφού αυτά θα είναι τα σημαντικά για το διαχωρισμό των αυτοκινήτων από τα αεροπλάνα.

Εικόνα 69: *Διαχωρισμός αυτοκινήτων και αεροπλάνων (σημαντικά χαρακτηριστικά: τροχός, φτερά)*

Η ίδια αρχή μπορεί να εφαρμοστεί και σε λιγότερο προφανή προβλήματα. Ο αλγόριθμος Relief είναι βασισμένος στην υπόθεση ότι στιγμιότυπα/δείγματα διαφορετικών κλάσεων μάλλον θα έχουν διαφορετικές τιμές, ενώ στιγμιότυπα/δείγματα της ίδιας κλάσης κατά πάσα πιθανότητα θα έχουν παρόμοιες τιμές [45]. Έτσι για τον εντοπισμό των καλύτερων χαρακτηριστικών εξετάζεται η διαχωριστική ικανότητα των τιμών τους σε κοντινά δείγματα. Το γενικό πλαίσιο του αλγόριθμου ώστε να βρει τα πιο σημαντικά χαρακτηριστικά ενός δείγματος X σε μια κλάση, περιλαμβάνει την εξέταση ενός δείγματος Y, παρόμοιου και ίδιας (ετικέτας) κλάσης με το X, καθώς και ενός δείγματος Z, παρόμοιου αλλά διαφορετικής (ετικέτας) κλάσης από το X. Τα χαρακτηριστικά που έχουν τις ίδιες τιμές στα δείγματα X και Y και διαφορετικές τιμές στα δείγματα X και Z, πιθανότατα, είναι σημαντικά για την ταξινόμηση. Η υπόθεση είναι ότι ένα συναφές χαρακτηριστικό με μεγάλη διακριτική ικανότητα θα μπορεί να διαχωρίσει δύο δείγματα που είναι κοντινά, αλλά ανήκουν σε διαφορετικές κλάσεις. Επομένως, για την εύρεση των συναφών χαρακτηριστικών ενός συγκεκριμένου δείγματος εξετάζονται οι δύο κοντινότεροι γείτονές του: το πιο κοντινό δείγμα από μια αντίθετη κλάση ονομάζεται κοντινότερη αποτυχία (the nearest miss) και το πιο κοντινό δείγμα από την ίδια κλάση ονομάζεται κοντινότερη επιτυχία (the nearest hit).

Ο αλγόριθμος απαρτίζεται από 3 βασικά μέρη:

1. Υπολογισμός κοντινότερων γειτόνων (nearest miss και nearest hit)
2. Υπολογισμός του βάρους ενός χαρακτηριστικού
3. Επιστροφή μια λίστας κατάταξης των χαρακτηριστικών ή ενός συνόλου με τα καλύτερα χαρακτηριστικά σύμφωνα με ένα προκαθορισμένο κατώφλι

Έστω σύνολο δεδομένων εκπαίδευσης S με n δείγματα που αποτελούνται από L χαρακτηριστικά και ανήκουν σε δύο γνωστές κλάσεις. Ακολουθούν τα βασικά βήματα του

αλγορίθμου Relief. Αρχικά, ο αλγόριθμος αρχικοποιεί το διάνυσμα βάρους των χαρακτηριστικών (weight vector), θέτοντας το βάρος κάθε χαρακτηριστικού ίσο με 0. Το επόμενο μέρος επαναλαμβάνεται m φορές για m διαφορετικά δείγματα τα οποία αντλεί ο αλγόριθμος από το σύνολο εκπαίδευσης με επανάθεση (ένα δείγμα κάθε φορά): επιλέγει τυχαία το διάνυσμα χαρακτηριστικών X ενός δείγματος και αναζητεί τους δύο κοντινότερούς του γείτονες: ένα γείτονα από την ίδια κλάση (the nearest hit) και έναν από την άλλη κλάση (the nearest miss). Αυτό επιτυγχάνεται υπολογίζοντας την ολική απόσταση μεταξύ του επιλεγμένου δείγματος και όλων των άλλων δειγμάτων του συνόλου εκπαίδευσης με τη βοήθεια της συνάρτησης diff. Για συνεχείς μεταβλητές μπορεί να χρησιμοποιηθεί η ευκλείδεια απόσταση. Για διακριτές μεταβλητές/χαρακτηριστικά, η απόσταση ενός χαρακτηριστικού μεταξύ δύο δειγμάτων (που χρησιμοποιείται στον υπολογισμό της ολικής απόστασης των δύο δειγμάτων για όλα τα χαρακτηριστικά) ισούται με 0 εάν τα δύο δείγματα για το χαρακτηριστικό αυτό έχουν τις ίδιες τιμές και με 1 εάν έχουν διαφορετικές τιμές. Η διαφορά (diff) στις τιμές ενός διακριτού χαρακτηριστικού k μεταξύ δύο δειγμάτων X και Y, παρουσιάζεται φορμαλιστικά στην ακόλουθη συνάρτηση diff.

$$diff(x_k, y_k) = |x_k - y_k|$$

$$diff(x_k, y_k) = \begin{cases} 0 & , εάν\ x_k, y_k\ τα\ ίδια \\ 1 & , εάν\ x_k, y_k\ διαφορετικά \end{cases}$$

Για συνεχείς μεταβλητές η διαφορά κανονικοποιείται στο διάστημα [0,1]. Η κανονικοποίηση γίνεται υπολογίζοντας το πηλίκο της διαφοράς στις τιμές ενός συνεχούς χαρακτηριστικού μεταξύ δύο δειγμάτων προς το εύρος των τιμών του υπό εξέταση χαρακτηριστικού κατά μήκος όλων των δειγμάτων:

$$diff(x_k, y_k) = \frac{|x_k, -y_k|}{max_k - min_k}\ , \text{πεδίο τιμών για συνεχείς μεταβλητές: } [0,1]$$

, όπου x_k η τιμή του χαρακτηριστικού k του δείγματος X, y_k η τιμή του χαρακτηριστικού k του δείγματος Y , max_k η μέγιστη τιμή του χαρακτηριστικού k αφού εξεταστούν οι τιμές όλων των δειγμάτων, min_k η ελάχιστη τιμή του χαρακτηριστικού k κατά μήκος των δειγμάτων.

Στη συνέχεια για να είναι εφικτός ο υπολογισμός της συνάφειας του κάθε χαρακτηριστικού, ο αλγόριθμος ενημερώνει το διάνυσμα βάρους με βάση τον ακόλουθο τύπο. Το βάρος του χαρακτηριστικού F θα ενημερωθεί ως εξής:

$$W(F) = W(F) - diff(x_F, Nearesthit_F) + diff(x_F, Nearestmiss_F)$$

,όπου x_F η τιμή του χαρακτηριστικού F του δείγματος X, $Nearesthit_F$ η τιμή του χαρακτηριστικού F του κοντινότερου γείτονα ίδιας κλάσης του δείγματος X και $Nearestmiss_F$ η τιμή του χαρακτηριστικού F του κοντινότερου γείτονα διαφορετικής κλάσης του δείγματος X.

Μετά από τις m επαναλήψεις, κάθε στοιχείο του διανύσματος βάρους διαιρείται με τον αριθμό των επαναλήψεων, υπολογίζοντας έτσι τη συνάφεια κάθε χαρακτηριστικού. Το ανανεωμένο διάνυσμα βάρους θα είναι ίσο με $W = \frac{W}{m}$. Η κανονικοποίηση με το m γίνεται έτσι ώστε όλα τα βάρη W[F], που αντιπροσωπεύουν τη συνάφεια, να ανήκουν στο διάστημα [-1,1]. Τα χαρακτηριστικά που επιλέγονται είναι αυτά των οποίων η συνάφεια, δηλαδή η τιμή του βάρους τους, είναι μεγαλύτερη από ένα προκαθορισμένο κατώφλι τ (0 ≤ τ ≤ 1), και τα υπόλοιπα απορρίπτονται.

$$W(F) = \frac{W(F)}{m} \Rightarrow W(F) = W(F) - \frac{diff(x_F, Nearesthit_F)}{m} + \frac{diff(x_F, Nearestmiss_F)}{m}$$

Στην πραγματικότητα ο υπολογισμός του διανύσματος βάρους του χαρακτηριστικού F είναι μια προσέγγιση της διαφοράς των παρακάτω πιθανοτήτων [45]:

$$W(F) = W(F) - P\ (different\ value\ of\ F | nearest\ instance\ from\ same\ class) + P\ (\text{different value of } F | nearest\ instance\ from\ different\ \text{class})$$

Μια μεγάλη τιμή W(F), βάσει του τύπου, υποδεικνύει ένα χαρακτηριστικό με ισχυρή διακριτική ικανότητα αφού τουλάχιστον ένα εκ των ακόλουθων δύο θα ισχύει: για ένα συγκεκριμένο δείγμα, είτε η πιθανότητα της κοντινότερης αποτυχίας του (nearest miss) να εμφανίζει διαφορετική τιμή για το χαρακτηριστικό F θα είναι μεγάλη (υψηλή P(different value of F|nearest instance from different class)), είτε η πιθανότητα της κοντινότερης επιτυχίας του (nearest hit) να εμφανίζει διαφορετική τιμή για το χαρακτηριστικό F θα είναι μικρή (χαμηλή P(different value of F|nearest instance from same class)).

Αφού αντλήσουμε m φορές δείγματα από το σύνολο εκπαίδευσης με επανάθεση, τότε για κάθε χαρακτηριστικό F, η πιθανότητα να έχει διαφορετική τιμή για την πλησιέστερη επιτυχία (nearest hit) του υπό εξέταση δείγματος είναι ίση με τον αριθμό των φορών που παρουσιάστηκε κάποια διαφορά στην τιμή του χαρακτηριστικού F μεταξύ του υπό εξέταση δείγματος και της πλησιέστερης επιτυχίας του προς το συνολικό αριθμό m των φορών που τελέστηκε η διαδικασία. Ομοίως, για κάθε χαρακτηριστικό F, η πιθανότητα να έχει διαφορετική τιμή για την πλησιέστερη αποτυχία (nearest miss) του υπό εξέταση δείγματος είναι ίση με τον αριθμό των φορών που παρατηρήθηκε κάποια διαφορά στην τιμή του χαρακτηριστικού F μεταξύ του υπό εξέταση δείγματος και της πλησιέστερης αποτυχίας, προς τον συνολικό αριθμό m των φορών που έλαβε

χώρα η διαδικασία. Ουσιαστικά, όπως φαίνεται στον παρακάτω τύπο, η κάθε πιθανότητα ισούται με το μέσο όρο των διαφορών στις τιμές του χαρακτηριστικού F μεταξύ δύο δειγμάτων στο σύνολο των m επαναλήψεων.

$$P(different\ value\ of\ F|nearest\ instance\ from\ different\ class)\frac{\sum_{i=1}^{m} diff(X_F, Z_F)}{m}$$

$$P(different\ value\ of\ F|nearest\ instance\ from\ same\ class)\frac{\sum_{i=1}^{m} diff(X_F, Y_F)}{m}$$

,όπου X_F η τιμή του χαρακτηριστικού F στο τυχαία επιλεγμένο δείγματος X, Y_F η τιμή του χαρακτηριστικού F στο δείγμα πλησιέστερης επιτυχίας (nearest hit) και Z_F η τιμή του χαρακτηριστικού F στο δείγμα πλησιέστερης αποτυχίας(nearest miss).

Όταν η διαφορά στις τιμές του χαρακτηριστικού F μεταξύ δύο δειγμάτων είναι μηδενική τότε οι ανωτέρω πιθανότητες παίρνουν το κατώτατο όριο τιμής 0, ενώ όταν υπάρχει διαφορά στις τιμές του χαρακτηριστικού F μεταξύ των ανά περίπτωση δύο δειγμάτων παίρνουν το ανώτατο όριο τιμής 1 (στις συνεχείς μεταβλητές το ανώτατο όριο συναντάται όταν η διαφορά στις τιμές είναι η μέγιστη δυνατή).

Άρα το βάρος του χαρακτηριστικού F, λαμβάνει τιμές από -1 έως 1 : W(F) ϵ [-1,1]. Όταν P(different value of F|nearest instance from different class) = 1 και P(different value of F|nearest instance from same class) = 0, τότε το βάρος του χαρακτηριστικού F παίρνει την ανώτερη τιμή που μπορεί να πάρει (1): W(F) = 1. Σε αυτή την περίπτωση το χαρακτηριστικό έχει τη μέγιστη διαχωριστική ικανότητα που μπορεί να έχει. Οι τιμές του χαρακτηριστικού σε πολύ κοντινά δείγματα που ανήκουν όμως σε διαφορετικές κλάσεις είναι διαφορετικές, ενώ οι τιμές του χαρακτηριστικού σε πολύ κοντινά δείγματα που ανήκουν στην ίδια κλάση είναι ίδιες. Όταν P(different value of F|nearest instance from different class) = 0 και P(different value of F|nearest instance from same class) = 1, τότε το βάρος του χαρακτηριστικού F παίρνει την κατώτερη τιμή που μπορεί να πάρει (-1): W(F) = -1. Σε αυτή την περίπτωση το χαρακτηριστικό έχει μηδενική διαχωριστική ικανότητα.

Όσο πιο μεγάλος είναι ο αριθμός των δειγμάτων m για τα οποία επαναλαμβάνεται η διαδικασία υπολογισμού του βάρους, τόσο πιο αξιόπιστη είναι η προσέγγιση των πιθανοτήτων. Ωστόσο ο αριθμός m, δεν πρέπει να ξεπεράσει τον αριθμό των διαθέσιμων δειγμάτων εκπαίδευσης [45]. Εάν ο αριθμός των διαθέσιμων δειγμάτων εκπαίδευσης είναι μικρός, τότε για καλύτερα αποτελέσματα, το m θα τεθεί ίσο με τον αριθμό αυτό - που είναι και το άνω όριο τιμής που μπορεί να λάβει το m- με αποτέλεσμα η διαδικασία υπολογισμού του βάρους να πραγματοποιηθεί σε όλα τα διαθέσιμα δείγματα.

Ακολούθως, παρουσιάζεται στον *πίνακα 11* ο ψευδοκώδικας του πρωτότυπου αλγόριθμου Relief, όπως αυτός περιγράφεται στο [44]. Θεωρούμε ότι η ταξινόμηση των στιγμιοτύπων γίνεται

σε δύο κλάσεις: τη θετική και την αρνητική. Στον πρωτότυπο αλγόριθμο για τον υπολογισμό της απόστασης μεταξύ του υπό εξέταση χαρακτηριστικού και των γειτόνων του, χρησιμοποιείται η τετραγωνική ευκλείδεια απόσταση (squared euclidean distance), δηλαδή το τετράγωνο της συνάρτησης diff (squared diff), η οποία για διακριτά χαρακτηριστικά είναι πρακτικά ισοδύναμη με την diff [46]. Επομένως, η ενημέρωση του διανύσματος βάρους στον πρωτότυπο Relief γίνεται ως ακολούθως.

$$p = p_1, p_2, \dots, p_L, \quad q = q_1, q_2, \dots, q_L$$

$$d^2(p,q) = (p_1 - q_{1)}{}^2 + (p_2 - q_{2)}{}^2 + \cdots + (p_L - q_{L)}{}^2$$

όπου p και q δύο στιγμιότυπα και p_1 ,p_2, …, p_L και q_1, q_2, …, q_L οι τιμές των L χαρακτηριστικών, και d^2 η τετραγωνική ευκλείδεια απόσταση των στιγμιότυπων p και q. Με βάση την τετραγωνική ευκλείδεια απόσταση υπολογίζεται η squared diff και το διάνυσμα βάρους:

$$diff(p_k, q_k)^2 = (p_k - q_{k)}{}^2$$

$$W_i = W_i - diff(x_i, Nearesthit_i)^2 + diff(x_i, Nearestmiss_i)^2$$

Η συνάρτηση diff εκτός από τον υπολογισμό του διανύσματος βάρους χρησιμοποιείται και για τον υπολογισμό της ολικής απόστασης μεταξύ δύο στιγμιοτύπων για την εύρεση των πλησιέστερων γειτόνων. Η ολική απόσταση total_distance ισούται με το άθροισμα των τετραγώνων των διαφορών των δύο στιγμιοτύπων κατά μήκος όλων των χαρακτηριστικών [46].

$$total\_distance(p,q) = d^2(p,q) = diff(p_1, q_1)^2 + diff(p_2, q_2)^2 + \cdots + diff(p_L, q_L)^2$$

Ο Relief καλεί μια ρουτίνα για την ενημέρωση του διανύσματος βάρους των χαρακτηριστικών W για κάθε δείγμα και τους γείτονές του και ακολούθως το χρησιμοποιεί για να καθορίσει το διάνυσμα μέσου βάρους όλων των χαρακτηριστικών (Relevance vector). Relevance είναι το μέσο διάνυσμα της τιμής: $-(x_i - Nearesthit_i)^2 + (x_i - Nearestmiss_i)^2$ για κάθε χαρακτηριστικό f_i για τα m δείγματα. Αντί να γίνει μια απλή ανανέωση του διανύσματος του βάρους W με την τιμή $\frac{W}{m}$, στον πρωτότυπο αλγόριθμο προτίμησαν να ορίσουν ένα καινούριο διάνυσμα, το διάνυσμα Relevance και το θέτουν ίσο με την τιμή $\frac{W}{m}$. Κάθε στοιχείο του διανύσματος Relevance αντιστοιχεί σε ένα χαρακτηριστικό και δείχνει πόσο συναφές είναι.

Πίνακας 11: Ενδεικτικός ψευδοκώδικας του πρωτότυπου αλγόριθμου Relief με χρήση τετραγωνικής συνάρτησης diff (τετραγωνική ευκλείδεια απόσταση) [44]

```
Algorithm Relief
Input: a training set S (for each training instance a vector of attribute
values and the class value), the number of iterations m, the threshold τ
Output: the vector Relevance of estimations of the qualities of
attributes

Relief(S,m,τ)

    Χώρισε το S σε: S+ = θετικά στιγμιότυπα και
                    S- = αρνητικά στιγμιότυπα
    W=(0,0,… ,0)

    for i = 1 to m {
            Επέλεξε τυχαία ένα στιγμιότυπο X ϵ S
            Επέλεξε τυχαία ένα από τα θετικά στιγμιότυπα που
            είναι πλησιέστερα στο X, Z+ϵ S+
            Επέλεξε τυχαία ένα από τα αρνητικά στιγμιότυπα που
            είναι πλησιέστερα στο X, Z-ϵ S-
            //Σχόλιο: Η εύρεση των πλησιέστερων γειτόνων του X μπορεί να
            //υλοποιηθεί ως εξής: για κάθε πιθανό ζεύγος στιγμιοτύπων του
            //συνόλου εκπαίδευσης S(X και ενός άλλου από τα εναπομείναντα
            //στιγμιότυπα του συνόλου εκπαίδευσης S, έστω Z, υπολογίζεται
            // το άθροισμα των τετραγώνων των διαφορών τους κατά μήκος
            // όλων των χαρακτηριστικών. Αυτή είναι η ολική απόσταση
            // total_diastance. Δηλαδή:
            //
            //for each pair X, Z {
            //   total_distance(X,Z) = 0;
            //   for F = 1 to all_attributes {
            //   total_distance(X,Z) = total_distance(X,Z)+ diff(x_F,z_F)^2;
            //   }
            // }
            //Z+ = Z with min{total_distance(X,Z)} and class(Z)= positive
            //while
            //Z- = Z with min{total_distance(X,Z)} and class(Z)= negative

            if (X είναι θετικό στιγμιότυπο)
                    then Nearesthit = Z+ ; Nearestmiss = Z-
                    else Nearesthit = Z- ; Nearestmiss = Z+
            update-weight(W,X,Nearesthit,Nearestmiss)
    }

    Relevance = 1/m W

    for i = 1 to all_attributes{
             if (relevance_i ≥ τ)
                      then f_i συναφές χαρακτηριστικό
                      else f_i άσχετο χαρακτηριστικό
    }

update-weight(W,X,Nearesthit,Nearestmiss)
    for i = 1 to all_attributes {

            W_i = W_i − diff(x_i, nearesthit_i)^2 + diff(x_i, nearestmiss_i)^2

    }
```

Στον *πίνακα 12* περιγράφεται ένας εναλλακτικός ψευδοκώδικας του βασικού αλγόριθμου Relief με χρήση της απλής diff αντί της squared diff. Εδώ την diff(Attribute,Instance1,Instance2) την ορίσαμε με 3 ορίσματα: το πρώτο είναι το χαρακτηριστικό που μας ενδιαφέρει και τα άλλα δύο τα στιγμιότυπα που εξετάζονται. Η κανονικοποίηση με το m εγγυάται ότι όλα τα βάρη W[F] θα ανήκουν στο διάστημα [-1,1], ωστόσο δεν είναι απαραίτητη εάν το W[F] χρησιμοποιείται για σχετική σύγκριση μεταξύ των χαρακτηριστικών. Σε αυτόν τον αλγόριθμο, η ολική απόσταση μεταξύ δύο στιγμιοτύπων για την εύρεση των πλησιέστερων γειτόνων, ισούται με το άθροισμα των διαφορών των στιγμιοτύπων κατά μήκος όλων των χαρακτηριστικών. Αν και στον πρωτότυπο αλγόριθμο χρησιμοποιείται το τετράγωνο της διαφοράς των στιγμιοτύπων για τον υπολογισμό της ολικής απόστασης και του βάρους, πρακτικά δεν παρατηρείται σημαντική διαφορά ανάμεσα στις δύο μεθόδους [46].

Πίνακας 12: Εναλλακτικός ψευδοκώδικας του βασικού αλγόριθμου Relief με χρήση απλής diff [46]

```
Algorithm Relief
Input: for each training instance a vector of attribute values and the
class value, the number m
Output: the vector W of estimations of the qualities of attributes

for i = 1 to all_attributes {
       W[i] = 0;
}

for i = 1 to m
{
      randomly select an instance Xi;
      find nearest hit Hi and nearest miss Mi;
      //Σχόλιο: Η εύρεση των nearest hit και nearest miss μπορεί να
      //υλοποιηθεί ως εξής: για κάθε πιθανό ζεύγος στιγμιοτύπων του
      //συνόλου εκπαίδευσης (Xi και ενός άλλου από τα εναπομείναντα
      //στιγμιότυπα του συνόλου εκπαίδευσης, έστω Z), υπολογίζεται το
      //άθροισμα των διαφορών τους κατά μήκος όλων των χαρακτηριστικών.
      //Αυτή είναι η ολική απόσταση total_diastance. Δηλαδή:
      //
      //for each pair Xi, Z {
      //   total_distance(Xi,Z) = 0;
      //   for F = 1 to all_attributes {
      //      total_distance(Xi,Z) = total_distance(Xi,Z) + diff(F,Xi,Z);
      //   }
      // }
      // H = Z with min{total_distance(Xi,Z)}and class(Z)= class(Xi)
      // while
      // M = Z with min {total_distance(Xi,Z)}and class(Z)≠ class(Xi)
      for F = 1 to all_attributes{
        W(F) = W(F) + diff(F, Xi,Mi) / m – diff(F, Xi,Hi)/m;
      }
}
```

2.6.4.1 Relief-f

Οι Kononenko et al. [46] [45] έχουν προτείνει κάποιες αναβαθμίσεις στον αλγόριθμο RELIEF συμπεριλαμβανομένου και της γενικοποίησής του σε προβλήματα πολλαπλών κλάσεων. Ο πρωτότυπος αλγόριθμος Relief παρουσιάζει κάποια προβλήματα: δεν μπορεί να αντεπεξέλθει όταν τα δεδομένα είναι ελλιπή ή και θορυβώδη. Εξαιτίας των προβλημάτων του πρωτότυπου αλγόριθμου Relief, ο Kononenko για να τα αντιμετωπίσει δημιούργησε διάφορες επεκτάσεις του Relief. Ο Relief-A προεκτείνει την αναζήτηση των δύο κοντινότερων γειτόνων, που γίνεται στη Relief, σε αναζήτηση των k-κοντινότερων γειτόνων (k-nearest neighbours search), οι Relief-B,C και D επιλύουν με διάφορους τρόπους το πρόβλημα των ελλιπών δεδομένων και οι Relief-E και F ψάχνουν τους κοντινότερους γείτονες από κάθε κλάση, αντί από μια μόνο, κι έτσι εκτιμούν καλύτερα τη διαχωρισιμότητα ενός δείγματος από όλες τις άλλες κλάσεις. Η αναβαθμισμένη έκδοση Relief-f περιλαμβάνει όλες τις προηγούμενες επεκτάσεις και προσαρμόζεται σε διαφόρων ειδών προβλήματα. Περαιτέρω πληροφορίες για τον αλγόριθμο Relief-F, συμπεριλαμβανομένου και ενδεικτικού ψευδοκώδικα, μπορούν να βρεθούν στα [46], [45].

2.7 Επιλογή χαρακτηριστικών με μεθόδους wrapper

Στην κατηγορία των τεχνικών περιτυλίγματος εντάσσονται όλοι οι αλγόριθμοι επιλογής χαρακτηριστικών που χρησιμοποιούν την ακρίβεια ταξινόμησης (ή κάποιο άλλο μέτρο απόδοσης του ταξινομητή) ως κριτήριο αξιολόγησης της καταλληλότητας των υποσυνόλων χαρακτηριστικών.

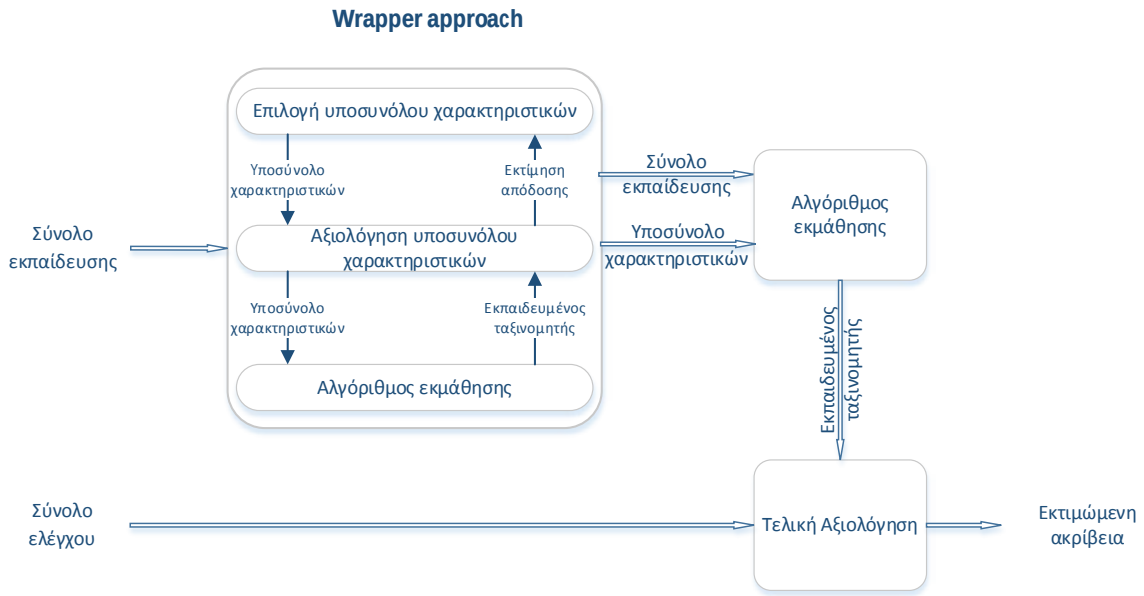
Στην προσέγγιση περιτυλίγματος, ο αλγόριθμος επιλογής υποσυνόλου χαρακτηριστικών υπάρχει ως «περιτύλιγμα» γύρω από τον επαγωγικό αλγόριθμο εκμάθησης του ταξινομητή [47]. Οι αλγόριθμοι περιτυλίγματος αναζητούν το βέλτιστο υποσύνολο χρησιμοποιώντας τον επαγωγικό αλγόριθμο εκμάθησης ως μέρος της συνάρτησης αξιολόγησης των υποσυνόλων χαρακτηριστικών. Ο ίδιος ο αλγόριθμος εκμάθησης αντιμετωπίζεται ως μαύρο κουτί [48].

Στην *εικόνα 70* βλέπουμε τη διαδικασία που ακολουθείται σε προβλήματα ταξινόμησης που χρησιμοποιούν τεχνική περιτυλίγματος.

Πρώτα γίνεται χρήση του αρχικού συνόλου εκπαίδευσης για την επιλογή χαρακτηριστικών, η οποία γίνεται μέσω αξιολόγησης των διάφορων υποσυνόλων με τον αλγόριθμο εκμάθησης του ταξινομητή. Ο αλγόριθμος εκμάθησης χρησιμοποιεί το αρχικό σύνολο δεδομένων εκπαίδευσης, συνήθως χωρίζοντάς το εσωτερικά σε 2 μικρότερα σύνολα, ένα σύνολο εκπαίδευσης και ένα σύνολο ελέγχου, για την εκπαίδευση και τον έλεγχο του ταξινομητή. Για την αξιολόγηση κάθε

πιθανού υποσυνόλου χαρακτηριστικών, αφαιρούνται από τα σύνολα εκπαίδευσης και ελέγχου τα ανάλογα χαρακτηριστικά προτού εισέλθουν στον αλγόριθμο επαγωγής. Το υποσύνολο χαρακτηριστικών με την πιο υψηλή βαθμολογία αξιολόγησης επιλέγεται ως το βέλτιστο υποσύνολο.

Ακολούθως, για την αξιολόγηση του αποτελέσματος της μεθόδου επιλογής χαρακτηριστικών, το βέλτιστο υποσύνολο χαρακτηριστικών χρησιμοποιείται από τον αλγόριθμο επαγωγής για την εκπαίδευση του τελικού ταξινομητή με το αρχικό σύνολο δεδομένων εκπαίδευσης. Αφού γίνει η εκπαίδευση, ο επακόλουθος ταξινομητής αξιολογείται σε ένα ανεξάρτητο σύνολο ελέγχου που δεν είχε χρησιμοποιηθεί κατά τη διάρκεια της αναζήτησης του βέλτιστου υποσυνόλου χαρακτηριστικών [48]. Έτσι διαπιστώνουμε την καταλληλότητα του αποτελέσματος της συγκεκριμένης μεθόδου επιλογής χαρακτηριστικών κι αν θεωρηθεί ότι απαιτείται βελτίωση τότε μπορεί να επαναληφθεί η διαδικασία με κάποια άλλη μέθοδο επιλογής χαρακτηριστικών.



Εικόνα 70: *Η wrapper προσέγγιση για επιλογή υποσυνόλου χαρακτηριστικών σε πρόβλημα ταξινόμησης (τροποποιημένο διάγραμμα από [48])*

Ας σημειωθεί ότι ένας τρόπος να μειωθεί το υπολογιστικό κόστος στις μεθόδους περιτυλίγματος είναι η χρήση παράλληλων επεξεργαστών. Με αυτό τον τρόπο, τα παιδιά κάθε κόμβου μπορούν να αξιολογούνται παράλληλα [48].

Οι μέθοδοι περιτυλίγματος απαιτούν την αναζήτηση όλου του χώρου των πιθανών υποσυνόλων. Για το λόγο αυτό είναι απαραίτητος ο συνδυασμός τους με στρατηγικές αναζήτησης. Συνήθως συνδυάζονται με αλγόριθμους δυναμικού προγραμματισμού, με τους

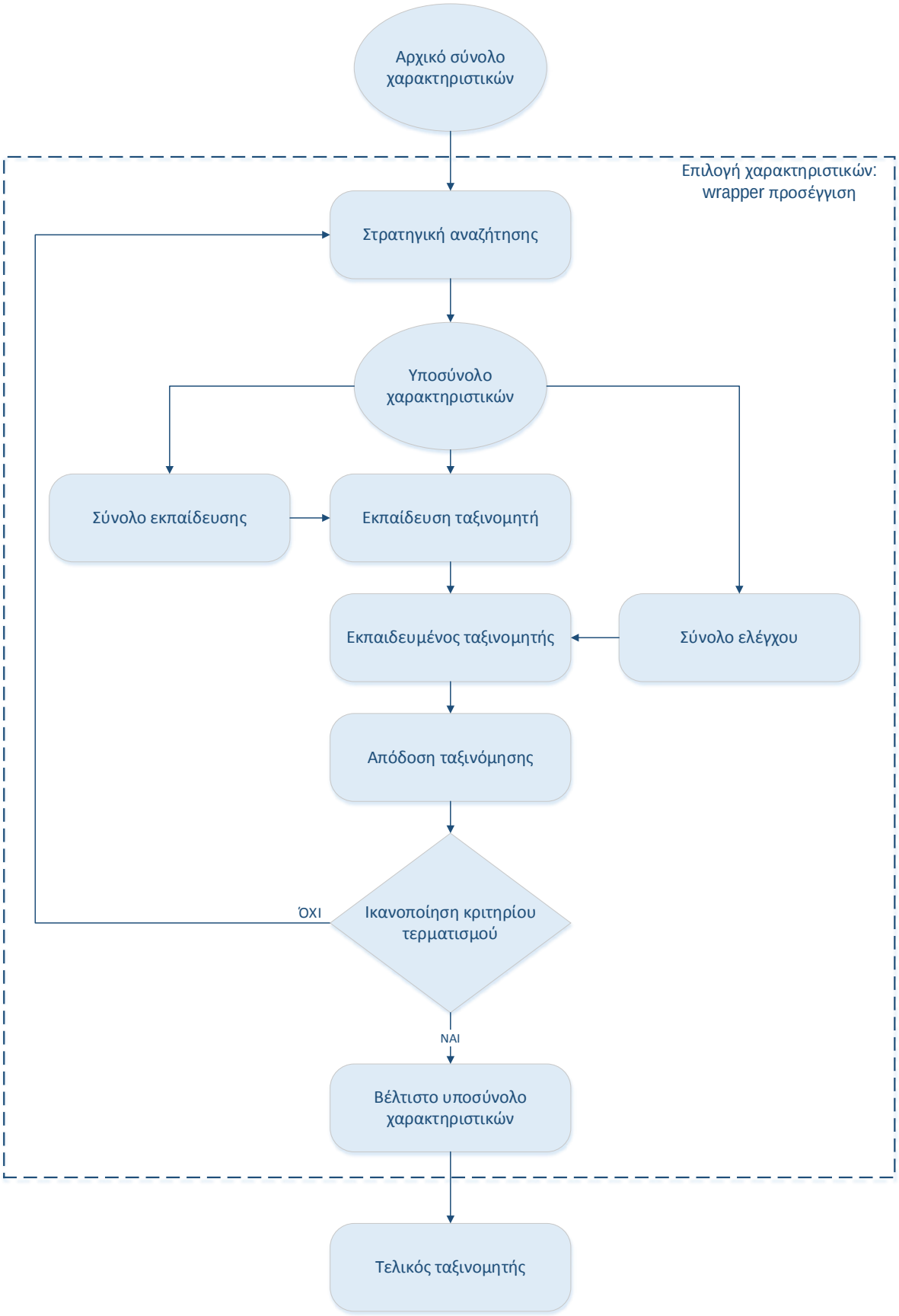
αλγόριθμους αναζήτησης forward selection και backward elimination, ή οποιαδήποτε άλλο αλγόριθμο αναζήτησης. Ο συνδυασμός που μας ενδιαφέρει είναι η wrapper επιλογή χαρακτηριστικών με χρήση γενετικών αλγορίθμων ως στρατηγική αναζήτησης. Ακολούθως παρουσιάζονται κάποιοι βασικοί αλγόριθμοι αναζήτησης.

2.7.1 Στρατηγικές αναζήτησης που συνδυάζονται με wrapper μεθόδους επιλογής χαρακτηριστικών

Μια στρατηγική αναζήτησης απαιτεί ένα χώρο καταστάσεων, μια αρχική κατάσταση, ένα κριτήριο τερματισμού και μια μηχανή αναζήτησης [48]. Κάθε κατάσταση αναπαριστά ένα υποψήφιο υποσύνολο χαρακτηριστικών. Διάφοροι τελεστές καθορίζουν τις διασυνδέσεις μεταξύ των καταστάσεων, τη μετάβαση, δηλαδή, από τη μια κατάσταση στην άλλη. Ο στόχος μιας στρατηγικής αναζήτησης είναι η εύρεση της κατάστασης με την υψηλότερη βαθμολογία αξιολόγησης.

Ακολούθως παρουσιάζονται συνοπτικά μερικές συχνά χρησιμοποιούμενες τεχνικές αναζήτησης όπως π.χ. η προς τα εμπρός ακολουθιακή επιλογή (sequential forward selection) και η προς τα πίσω ακολουθιακή απαλοιφή (sequential backward elimination) που είναι διαφορετικές εκδοχές άπληστης αναζήτησης με διαφορά το σημείο εκκίνησης της αναζήτησης. Οι ακολουθιακές στρατηγικές αναζήτησης προσθέτουν ή αφαιρούν ένα χαρακτηριστικό κάθε φορά από τα διαθέσιμα χαρακτηριστικά. Οι προς τα εμπρός ακολουθιακές στρατηγικές αναζήτησης είναι σχετικά γρήγορες, ενώ οι προς τα πίσω απαιτούν μεγάλο υπολογιστικό χρόνο και δεν είναι εφαρμόσιμες σε όλες τις περιπτώσεις [27].

Στην *εικόνα 71* παρουσιάζεται η διαδικασία επιλογής χαρακτηριστικών χρησιμοποιώντας wrapper προσέγγιση επισημαίνοντας το ρόλο των στρατηγικών αναζήτησης.



Εικόνα 71: *Ρόλος στρατηγικής αναζήτησης σε wrapper επιλογή χαρακτηριστικών*

Οι αλγόριθμοι αναζήτησης που περιγράφονται στη συνέχεια, μπορούν να χρησιμοποιηθούν και σε μεθόδους φιλτραρίσματος, όχι μόνο σε μεθόδους περιτυλίγματος.

2.7.1.1 Sequential Forward selection

Σημείο εκκίνησης είναι το κενό σύνολο χαρακτηριστικών και στη συνέχεια προστίθεται ένα χαρακτηριστικό κάθε φορά έως ότου βρεθεί το βέλτιστο υποσύνολο χαρακτηριστικών. Σε ένα τυπικό βήμα του αλγόριθμου εξετάζονται όλα τα υποσύνολα που προκύπτουν από την προσθήκη ενός χαρακτηριστικού στο τρέχον υποσύνολο. Το χαρακτηριστικό που φέρει τη μεγαλύτερη αύξηση στην απόδοση σύμφωνα με το προκαθορισμένο κριτήριο αξιολόγησης προστίθεται στο τρέχον υποσύνολο. Η επέκταση με την προσθήκη ενός χαρακτηριστικού κάθε φορά συνεχίζεται έως την ικανοποίηση κάποιου κριτηρίου τερματισμού.

Ένα αυστηρό κριτήριο τερματισμού είναι ο τερματισμός του αλγορίθμου όταν κανένα από τα υποσύνολα-παιδιά ενός βήματος δεν βελτιώνει την απόδοση. Αυτό το κριτήριο, όμως, μπορεί να σταματήσει πρόωρα τον αλγόριθμο. Ένα λιγότερο αυστηρό κριτήριο τερματισμού είναι: να συνεχίζονται οι επεκτάσεις του υποσυνόλου χαρακτηριστικών όσο υπάρχει κάποιο υποσύνολο-παιδί σε κάθε βήμα το οποίο δίνει την ίδια απόδοση ή καλύτερη με την έως τότε απόδοση, ενώ ο αλγόριθμος θα τερματιστεί μόνο αν δεν υπάρξει καμία βελτίωση στην απόδοση μετά από ν διαδοχικά βήματα.

Το βασικό μειονέκτημά του αλγόριθμου forward selection είναι ότι σε επόμενο βήμα δεν μπορούν να αφαιρεθούν χαρακτηριστικά που έχουν ήδη προστεθεί σε κάποιο υποσύνολο. Επιπρόσθετα, παγιδεύεται σε τοπικά ελάχιστα.

Ο αλγόριθμος αυτός είναι μέθοδος ευρετικής αναζήτησης και δεν μπορεί να εγγυηθεί τη βελτιστότητα του επιλεγμένου υποσυνόλου. Έστω το βέλτιστο υποσύνολο είναι μεγέθους k και ως σημείο εκκίνησης θεωρούμε το κενό σύνολο. Στην εξαντλητική forward selection εφόσον το k είναι άγνωστο αρχικά, για την εύρεση των k πιο κατάλληλων από N χαρακτηριστικά για επιλογή απαιτούνται $\binom{N}{1} + \binom{N}{2} + \cdots + \binom{N}{k}$ φορές. Συνεπώς η πολυπλοκότητα της εξαντλητικής αναζήτησης με forward selection είναι $O(2^N)$ που δεν είναι αποδοτική [49]. Η sequential forward selection μπορεί να μην εγγυάται τη βελτιστότητα του επιλεγμένου υποσυνόλου χαρακτηριστικών αλλά έχει σημαντικά καλύτερη πολυπλοκότητα.

Η στρατηγική αυτή είναι πιο αποτελεσματική όταν το πλήθος των βέλτιστων χαρακτηριστικών είναι περιορισμένο. Αυτό συμβαίνει επειδή ξεκινά την αναζήτηση από το κενό υποσύνολο.

2.7.1.2 Sequential Backward elimination

Η προς τα πίσω ακολουθιακή απαλοιφή (sequential backward elimination) ή όπως αλλιώς αναφέρεται, η προς τα πίσω ακολουθιακή αναζήτηση (sequential backward selection), έχει προταθεί από τους Marill και Green [31].

Σημείο εκκίνησης είναι το πλήρες σύνολο χαρακτηριστικών, δηλαδή αυτό που περιέχει όλα τα χαρακτηριστικά, και στη συνέχεια αφαιρείται ένα χαρακτηριστικό κάθε φορά έως ότου βρεθεί το βέλτιστο υποσύνολο χαρακτηριστικών. Η προς τα πίσω απαλοιφή (backward elimination) ακολουθεί αντίθετη κατεύθυνση αναζήτησης από αυτή που ακολουθεί η προς τα εμπρός επιλογή (forward selection). Σε κάθε βήμα εξετάζονται όλα τα υποσύνολα που προκύπτουν με αφαίρεση ενός χαρακτηριστικού από το τρέχον υποσύνολο. Στο τέλος του βήματος αφαιρείται το χαρακτηριστικό του οποίου η διαγραφή εξασφαλίζει τη μεγαλύτερη απόδοση ως προς το κριτήριο αξιολόγησης.

Τα κριτήρια τερματισμού παραμένουν τα ίδια, όπως αυτά έχουν περιγραφεί στην προς τα εμπρός επιλογή.

Το βασικό μειονέκτημά του είναι ότι δεν μπορεί σε επόμενο βήμα να επανεκτιμήσει χαρακτηριστικά που έχουν ήδη απορριφθεί, δηλαδή δεν είναι εφικτή η οπισθοδρόμηση. Επιπλέον, παρουσιάζει κι αυτός ο αλγόριθμος, όπως και ο αλγόριθμος forward selection, τον κίνδυνο παγίδευσης σε τοπικά ελάχιστα.

Αυτή η στρατηγική αναζήτησης έχει καλύτερη απόδοση όταν το βέλτιστο υποσύνολο είναι μεγάλο σε μέγεθος επειδή ξεκινά την αναζήτηση από το πλήρες υποσύνολο.

2.7.1.3 Plus l – take away r

Αυτή η μέθοδος είναι ένας συνδυασμός των μεθόδων forward selection και backward elimination: κάθε *l* βήματα forward selection ακολουθούνται από r βήματα backward elimination. Σε μια επανάληψη της μεθόδου προστίθενται *l* χαρακτηριστικά στο τρέχον υποσύνολο, έστω N, χαρακτηριστικών. Ακολούθως αποτιμάται η ποιότητα των υποσυνόλων που προκύπτουν από την αφαίρεση ενός χαρακτηριστικού, επιχειρώντας τον εντοπισμό υποσυνόλου μεγέθους N+*l*-1 που θα έχει υψηλότερη απόδοση από το προηγούμενο υποσύνολο αυτού του μεγέθους. Αυτό γίνεται για r αφαιρέσεις και στη συνέχεια ο αλγόριθμος συνεχίζει πάλι με την προσθήκη *l* χαρακτηριστικών.

Αυτή η στρατηγική αναζήτησης προσπαθεί να ξεπεράσει τα προβλήματα της forward selection και backward elimination οι οποίες δεν μπορούν να προσθέσουν ή να αφαιρέσουν χαρακτηριστικά που είχαν ήδη απορριφθεί σε προηγούμενα βήματα. Αυτό το επιτυγχάνει

λαμβάνοντας υπόψη τις συσχετίσεις ανάμεσα στα χαρακτηριστικά που προστίθενται ή αφαιρούνται από τα υποσύνολα σε επόμενα βήματα. Ένα χαρακτηριστικό που αρχικά μπορεί να εισήχθη στο υποσύνολο επειδή έδωσε υψηλή απόδοση όταν αξιολογήθηκε μεμονωμένα, είναι πιθανό, με τον μετέπειτα συνδυασμό του με άλλα χαρακτηριστικά που προστέθηκαν να μην δίνει τόσο καλή απόδοση. Αντιθέτως, ένα άλλο χαρακτηριστικό που μπορεί μεμονωμένα να μη φαίνεται κατάλληλο, σε συνδυασμό με άλλα χαρακτηριστικά που προστίθενται στη συνέχεια είναι πιθανό να δίνει υψηλότερη απόδοση.

Το βασικό μειονέκτημα αυτής της στρατηγικής είναι ο καθορισμός των παραμέτρων *l* και r, οι οποίες πρέπει να αρχικοποιηθούν σε βέλτιστες τιμές.

2.7.1.4 Best first search

Στις μεθόδους forward selection και backward elimination υπάρχει ένα υποσύνολο το οποίο επεκτείνεται με πρόσθεση ή αφαίρεση χαρακτηριστικών. Από τα παραγόμενα υποσύνολα-παιδιά κάθε επανάληψης, μετά από αξιολόγηση, επιλέγεται το καλύτερο από αυτά και επεκτείνεται ενώ τα άλλα αγνοούνται εντελώς. Στη μέθοδο best first search δεν αγνοείται κανένα πιθανό υποσύνολο: όλα τα υποσύνολα-παιδιά αφού τύχουν αξιολόγησης εισάγονται σε μια λίστα. Σε κάθε επανάληψη βρίσκεται το καταλληλότερο υποσύνολο από τη λίστα και αυτό το υποσύνολο επεκτείνεται. Τα υποσύνολα παιδιά που προκύπτουν, αξιολογούνται κι αυτά και ακολούθως εισάγονται στη λίστα. Το κύριο μειονέκτημα της μεθόδου αυτής, είναι η μεγάλη της πολυπλοκότητα. Το πλεονέκτημά της ότι πραγματοποιεί πιο εξονυχιστική αναζήτηση.

Ως κριτήριο τερματισμού χρησιμοποιείται αυτό που τερματίζει τον αλγόριθμο εφόσον μετά από ν διαδοχικές επαναλήψεις δεν βρεθεί υποσύνολο που να βελτιώνει την απόδοση.

2.7.1.5 Floating search

Μοιάζει με τη μέθοδο Plus l – take away r με τη διαφορά ότι ο αριθμός των backward βημάτων που ακολουθούν τα forward βήματα δεν είναι προκαθορισμένος. Τα backward βήματα συνεχίζονται όσο οι αφαιρέσεις χαρακτηριστικών σχηματίζουν υποσύνολα τα οποία βελτιώνουν την απόδοση.

2.7.1.6 Γενετικοί αλγόριθμοι

Στο πρόβλημα της επιλογής χαρακτηριστικών με χρήση ΓΑ, τα άτομα του πληθυσμού είναι τα υποσύνολα χαρακτηριστικών και αναπαριστώνται από δυαδικά διανύσματα. Κάθε χρωμόσωμα/ άτομο αναπαριστά ένα υποσύνολο χαρακτηριστικών και κάθε γονίδιο αναπαριστά ένα

χαρακτηριστικό. Στο στάδιο της αναπαραγωγής, γονείς είναι δύο υποσύνολα που συνδυάζονται και δίνουν ένα νέο υποσύνολο-παιδί που περιέχει ένα τμήμα των χαρακτηριστικών του ενός υποσυνόλου-γονέα και ένα τμήμα των χαρακτηριστικών του άλλου υποσυνόλου-γονέα. Ως συνάρτηση αξιολόγησης κάθε νέου ατόμου του πληθυσμού συνήθως χρησιμοποιείται η απόδοση του ταξινομητή. Επειδή η ποιότητα ενός υποσυνόλου χαρακτηριστικών καθορίζεται από την απόδοσή του σε σχέση με τη διαδικασία της ταξινόμησης, το ποσοστό σφάλματος ταξινόμησης (classification error rate), χρησιμοποιείται συνήθως ως η αντικειμενική συνάρτηση (fitness function) του ΓΑ. Το σφάλμα ταξινόμησης προκύπτει από ένα συγκεκριμένο ταξινομητή όταν ένα υποσύνολο χαρακτηριστικών χρησιμοποιείται για την εκπαίδευση (training) και τον έλεγχό του (testing).

Οι γενετικοί αλγόριθμοι χρησιμοποιούνται κατά κόρον στην επιλογή χαρακτηριστικών τις τελευταίες δεκαετίες. Είναι μια ενδιαφέρουσα και ελκυστική τεχνική: η προοπτική της εύρεσης μιας εξελισσόμενης υποβέλτιστης λύσης και η ευελιξία για χρήση οποιασδήποτε αναπαράστασης του εκάστοτε προβλήματος σε συνδυασμό με οποιονδήποτε ταξινομητή, γραμμικό ή μη, την καθιστούν ιδιαιτέρως ελκυστική σε ερευνητές που βρίσκονται αντιμέτωποι με πολύπλοκα προβλήματα πολλών μεταβλητών.

Οι ΓΑ όταν αρχικά είχαν εφαρμοστεί σε προβλήματα επιλογής χαρακτηριστικών, χρησιμοποιούνταν ως ανεξάρτητο στάδιο πριν από τον ταξινομητή (filter προσέγγιση) έχοντας ως συνάρτηση αξιολόγησης κάποιο άλλο διαχωριστικό μέτρο αντί την απόδοση του ταξινομητή. Ακολούθως όμως, διαπιστώθηκε ότι μια wrapper προσέγγιση που χρησιμοποιεί τους ΓΑ ως στρατηγική αναζήτησης σε συνδυασμό με τον ταξινομητή είναι προτιμότερη και πιο υποσχόμενη. Σε αυτή την προσέγγιση ο ΓΑ χρησιμοποιεί ως συνάρτηση αξιολόγησης κάθε χαρακτηριστικού τον ίδιο τον ταξινομητή, με επακόλουθο ο ΓΑ να εξελίσσει το υποσύνολο χαρακτηριστικών έτσι ώστε να είναι ειδικά προσαρμοσμένο στον συγκεκριμένο ταξινομητή.

Ακολούθως γίνεται εκτενής παρουσίαση των wrapper τεχνικών επιλογής χαρακτηριστικών με χρήση γενετικών αλγόριθμων.

Κεφάλαιο 3 – Επιλογή χαρακτηριστικών με χρήση Γενετικών Αλγορίθμων

3.1. Εισαγωγή

Η προσέγγιση που περιγράφεται εδώ, περιλαμβάνει τη χρήση γενετικών αλγορίθμων για την αναγνώριση και επιλογή του καλύτερου υποσυνόλου χαρακτηριστικών το οποίο ακολούθως θα τροφοδοτήσει τον ταξινομητή. Τα αποτελέσματα από προηγούμενες χρήσεις τους σε τέτοιου είδους προβλήματα, εισηγούνται ότι οι γενετικοί αλγόριθμοι είναι ένα χρήσιμο εργαλείο επίλυσης δύσκολων προβλημάτων επιλογής χαρακτηριστικών, στα οποία τόσο το μέγεθος του συνόλου χαρακτηριστικών, όσο και η απόδοση του συστήματος, είναι σημαντικοί παράγοντες σχεδίασης [34]. Αν και οι ΓΑ είχαν αρχικά αναπτυχθεί για την επίλυση προβλημάτων βελτιστοποίησης συνάρτησης μπορούν να χρησιμοποιηθούν και στο πρόβλημα της επιλογής χαρακτηριστικών [27] αντιμετωπίζοντάς το ως πρόβλημα βελτιστοποίησης.

Στο προηγούμενο κεφάλαιο έχουμε αναφερθεί στις δύο βασικές κατηγορίες επιλογής χαρακτηριστικών. Στην μια κατηγορία (filter-φιλτραρίσματος) τα χαρακτηριστικά επιλέγονται ανεξάρτητα από την απόδοση του ταξινομητή, ενώ στην άλλη (wrapper-περιτυλίγματος) επιλέγεται κατευθείαν ένα υποσύνολο «d» από τα διαθέσιμα «m» χαρακτηριστικά, με τέτοιο τρόπο ώστε να μην υποβαθμίζεται η απόδοση του ταξινομητή. Αν και οι ΓΑ έχουν χρησιμοποιηθεί ως στρατηγική αναζήτησης και στις δύο προσεγγίσεις επιλογής χαρακτηριστικών, εδώ θα εξετάσουμε τη χρήση των γενετικών αλγόριθμων στην δεύτερη προσέγγιση (wrapper).

Οι περισσότερες μελέτες εφαρμογής των ΓΑ σε προβλήματα επιλογής χαρακτηριστικών κατέληγαν στο συμπέρασμα ότι ο ταξινομητής λειτουργούσε το ίδιο καλά χρησιμοποιώντας το υποσύνολο χαρακτηριστικών που προέκυπτε από την επιλογή χαρακτηριστικών όσο όταν χρησιμοποιούσε ολόκληρο το σύνολο χαρακτηριστικών. Αυτό βέβαια, ενώ στηρίζει τη χρήση επιλογής χαρακτηριστικών, δεν σημαίνει απαραίτητα ότι στηρίζει τη χρήση ΓΑ στην επιλογή χαρακτηριστικών [50]. Ωστόσο οι ΓΑ παρουσιάζουν αρκετά θετικά στοιχεία στα οποία υπερτερούν σε σχέση με κάποιες άλλες τεχνικές.

Το βασικό πρόβλημα με τις υπόλοιπες μεθόδους (π.χ. των ακολουθιακών στρατηγικών αναζήτησης), είναι ότι κάθε βήμα τους προκαθορίζει σε μεγάλο βαθμό τις επακόλουθες επιλογές, αφαιρώντας την πιθανότητα εξερεύνησης κάποιων συνδυασμών χαρακτηριστικών. Είναι λες και αγνοούν εντελώς ένα μεγάλο μέρος του χώρου αναζήτησης [27]. Μια καλή εναλλακτική είναι οι ΓΑ.

Οι ΓΑ είναι γνωστοί για την ικανότητα που έχουν για αποδοτική αναζήτηση σε μεγάλους χώρους για τους οποίους υπάρχουν ελάχιστες πληροφορίες γνωστές εκ των προτέρων [34]. Δεδομένου ότι οι ΓΑ δεν επηρεάζονται όσο άλλοι αλγόριθμοι από το θόρυβο, φαίνεται να αποτελούν μια ελκυστική επιλογή για μια πιο σθεναρή στρατηγική επιλογής χαρακτηριστικών που θα βελτιώνει την απόδοση του ταξινομητή.

Οι ΓΑ συνδυάζουν αποτελεσματικότητα (effectiveness), δηλαδή υψηλή αξιοπιστία εύρεσης – ή προσέγγισης- του καθολικά βέλτιστου υποσυνόλου χαρακτηριστικών, καθώς και αποδοτικότητα (efficiency), δηλαδή υψηλή ταχύτητα σύγκλισης στο καθολικά βέλτιστο υποσύνολο. Η αποτελεσματικότητα και η αποδοτικότητα συχνά είναι αντικρουόμενοι στόχοι. Τεχνικές συστηματικής αναζήτησης προσεγγίζουν το καθολικό βέλτιστο με ακρίβεια (αποτελεσματικές), αλλά ταυτόχρονα έχουν υψηλές απαιτήσεις σε υπολογιστικό κόστος (μη αποδοτικές). Αντίθετα, τεχνικές άμεσης αναζήτησης ενώ είναι γρήγορες, εγκλωβίζονται εύκολα σε τοπικά ακρότατα (χαμηλή αποτελεσματικότητα, αλλά υψηλή αποδοτικότητα σε περίπτωση που δεν εγκλωβιστούν σε τοπικό ακρότατο). Βασικό πλεονέκτημα των ΓΑ σε σχέση με άλλους αλγόριθμους επιλογής χαρακτηριστικών είναι ότι ενώ άλλοι παγιδεύονται σε τοπικά ελάχιστα (ή μέγιστα, ανάλογα με το πρόβλημα) στο χώρο αναζήτησης, οι ΓΑ εξαιτίας της υψηλά στοχαστικής τους φύσης έχουν τη δυνατότητα να απεγκλωβιστούν [51]. Αυτό επιτυγχάνεται μέσω των διαδικασιών αναπαραγωγής: με τον συνδυασμό υποσυνόλων χαρακτηριστικών για τη δημιουργία νέων, καθώς και με την τυχαία μετάλλαξη κάποιων από αυτά, προσδίδεται τυχαιότητα στην εξερεύνηση του χώρου αναζήτησης. Με αυτό τον τρόπο, προηγούμενες λύσεις - που ίσως να είχαν παγιδεύσει τον ΓΑ σε περιοχές τοπικού βέλτιστου - απομακρύνονται, και οι νέες λύσεις κατευθύνουν τον ΓΑ σε ανεξερεύνητες περιοχές του χώρου αναζήτησης.

Μειονέκτημά των ΓΑ είναι η ύπαρξη αρκετών παραμέτρων που επηρεάζουν σημαντικά την επίδοσή τους (π.χ. μέγεθος πληθυσμού, πιθανότητα διασταύρωσης, πιθανότητα μετάλλαξης) και πρέπει να καθοριστούν. Το ζήτημα είναι ότι η ανάθεση τιμών στις παραμέτρους αυτές διαμορφώνεται ανάλογα με το εκάστοτε πρόβλημα. Συνεπώς, για τον καθορισμό τους, για κάθε πρόβλημα απαιτούνται πολλές πειραματικές δοκιμές με διαφορετικές τιμές των παραμέτρων αυτών.

Στον *πίνακα 13* παρουσιάζονται οι πιο σημαντικοί λόγοι χρησιμοποίησης των ΓΑ στο πρόβλημα της επιλογής χαρακτηριστικών.

Πίνακας 13: Λόγοι χρησιμοποίησης ΓΑ στο πρόβλημα της επιλογής χαρακτηριστικών

Λόγοι χρησιμοποίησης ΓΑ στο πρόβλημα της επιλογής χαρακτηριστικών
Ταυτόχρονη εξερεύνηση πολλών διαφορετικών υποσυνόλων στο χώρο αναζήτησης έτσι ώστε ο εντοπισμός των βέλτιστων λύσεων να προκύπτει μέσα από ένα πλήθος διαφορετικών υποσυνόλων χαρακτηριστικών
Αποφυγή, ως ένα βαθμό, του εγκλωβισμού σε τοπικά βέλτιστα υποσύνολα χαρακτηριστικών
Χρησιμοποιούν κωδικοποίηση των δεδομένων του προβλήματος και όχι τα ίδια τα δεδομένα
Απαιτούν τη γνώση μόνο της συνάρτησης καταλληλότητας των υποσυνόλων των χαρακτηριστικών
Σε υπολογιστικώς δύσκολα προβλήματα (με πολυδιάστατο χώρο αναζήτησης), μπορούν να συνδυαστούν με άλλες τοπικές μεθόδους αναζήτησης, αυξάνοντας έτσι την αποτελεσματικότητα και την αποδοτικότητά τους
Δεν επηρεάζονται τόσο πολύ από το θόρυβο
Σχετικά γρήγοροι και αποτελεσματικοί

Για τους παραπάνω λόγους οι ΓΑ έχουν εφαρμοστεί σε πληθώρα προβλημάτων επιλογής χαρακτηριστικών σε διάφορους τομείς.

3.2. Γενικό πλαίσιο εφαρμογής Γενετικών Αλγορίθμων για wrapper επιλογή χαρακτηριστικών

Θεωρούμε ως είσοδο δοθέν αρχικό σύνολο χαρακτηριστικών καθώς και ένα σύνολο εκπαίδευσης με θετικά και αρνητικά παραδείγματα των διάφορων κλάσεων στις οποίες θα πραγματοποιηθεί η ταξινόμηση. Μια διεργασία αναζήτησης - εδώ οι ΓΑ - χρησιμοποιείται για την εξερεύνηση του χώρου όλων των πιθανών υποσυνόλων που προκύπτουν από το δοθέν σύνολο χαρακτηριστικών. Στο πρόβλημα της επιλογής χαρακτηριστικών, τα άτομα (χρωμοσώματα) του πληθυσμού του ΓΑ είναι υποσύνολα χαρακτηριστικών που αναπαριστώνται με διανύσματα. Κάθε γονίδιο ενός χρωμοσώματος είναι ένας δείκτης στο αρχικό σύνολο χαρακτηριστικών. Η όλη διαδικασία υλοποιείται σε τρία βασικά στάδια: στο πρώτο στάδιο (αρχικοποίηση-initialization) δημιουργείται τυχαία ένας αρχικός πληθυσμός ατόμων (δηλαδή πιθανών υποσυνόλων χαρακτηριστικών). Στις περισσότερες εφαρμογές ΓΑ, χρησιμοποιείται δυαδική κωδικοποίηση και έτσι κάθε άτομο του αρχικού πληθυσμού είναι μια συμβολοσειρά της οποίας κάθε bit αρχικοποιείται τυχαία σε 0 ή 1 [52]. Ο αρχικός πληθυσμός σχηματίζει την πρώτη γενεά (generation). Κάθε επανάληψη του αλγόριθμου σχηματίζει μια καινούρια γενεά επιλέγοντας τα πιο κατάλληλα άτομα της προηγούμενης γενεάς. Μετά το στάδιο της αρχικοποίησης, η απόδοση κάθε ενός από τα υποψήφια υποσύνολα χαρακτηριστικών εκτιμάται μέσω μιας συνάρτησης αξιολόγησης του αντίστοιχου μειωμένου χώρου χαρακτηριστικών και συνόλου εκπαίδευσης υπολογίζοντας το καθορισμένο αποτέλεσμα της ταξινόμησης. Εφόσον η ποιότητα κάθε

υποσυνόλου χαρακτηριστικών καθορίζεται από την απόδοσή του σχετικά με τη διεργασία της ταξινόμησης, ως συνάρτηση καταλληλότητας χρησιμοποιείται ένα μέτρο σχετικό με την ταξινόμηση. Τα πιο συνήθη κριτήρια είναι η ακρίβεια ταξινόμησης ή το σφάλμα ταξινόμησης. Αφού τα υποψήφια υποσύνολα χαρακτηριστικών κάθε γενιάς αξιολογηθούν βάσει της αντικειμενικής συνάρτησης αξιολόγησης και έχει ανατεθεί σε κάθε ένα από αυτά μια τιμή καταλληλότητας, οι πιο εύρωστες λύσεις επιλέγονται για αναπαραγωγή (selection operator). Αυτό δίνει τη δυνατότητα στις πιο κατάλληλες λύσεις κάθε γενεάς να επηρεάζουν όλο και περισσότερο τις αλλαγές στον πληθυσμό, έτσι ώστε τελικώς να κυριαρχήσουν οι καταλληλότερες λύσεις [52]. Κατά το στάδιο της αναπαραγωγής εφαρμόζονται στα καλύτερα υποψήφια υποσύνολα διάφοροι γενετικοί τελεστές (διασταύρωση, μετάλλαξη, ελιτισμός) με αποτέλεσμα το σχηματισμό ενός καινούριου πληθυσμού, της επόμενης γενεάς. Με την εφαρμογή του τελεστή διασταύρωσης, δύο υποσύνολα «γονείς» έστω S_1, S_2 συνδυάζονται και δίνουν ένα νέο υποσύνολο «παιδί», που περιέχει ένα μέρος των χαρακτηριστικών του S_1 και ένα μέρος των χαρακτηριστικών του S_2. Η διασταύρωση αποσκοπεί στην παραγωγή νέων λύσεων, οι οποίες θα περιέχουν χρήσιμα κομμάτια και από τα δύο υποσύνολα «γονείς» και θα είναι πιο κατάλληλες λύσεις από αυτά (η ελπίδα κάθε γονέα) [52]. Η διασταύρωση είναι υπεύθυνη για την παραγωγή των περισσότερων νέων λύσεων για αναζήτηση. Ωστόσο, εάν όλες οι λύσεις μοιάζουν, τότε η διασταύρωση χάνει την ικανότητα να παράγει νέες λύσεις και καταλήγει να παράγει τις ίδιες. Για το λόγο αυτό, για να μην αποτελείται όλος ο πληθυσμός από παρόμοιες λύσεις, σε κάθε νέα λύση εφαρμόζεται ο τελεστής της μετάλλαξης (αλλαγή τυχαίων bits). Στο τρίτο στάδιο, ο νέος πληθυσμός που έχει σχηματιστεί, χρησιμοποιείται στην επόμενη επανάληψη του αλγορίθμου και η διαδικασία επαναλαμβάνεται. Ο ΓΑ τελειώνει όταν πληρούται ένα κριτήριο τερματισμού. Ως έξοδος του ΓΑ επιστρέφεται το καλύτερο υποσύνολο χαρακτηριστικών που εντοπίζεται. Το υποσύνολο αυτό, θα είναι το προτεινόμενο σύνολο χαρακτηριστικών που θα χρησιμοποιηθεί ως η πραγματική είσοδος στον ταξινομητή.

Η εφαρμογή ΓΑ σε οποιοδήποτε πρόβλημα προϋποθέτει τον καθορισμό της κατάλληλης αναπαράστασης των χρωμοσωμάτων, καθώς και της κατάλληλης συνάρτησης καταλληλότητας.

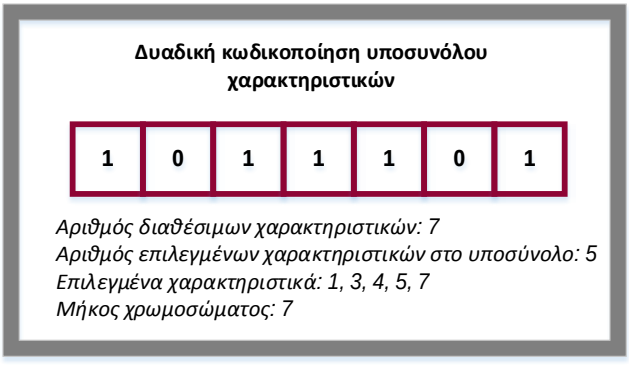
3.2.1 Αναπαράσταση / Κωδικοποίηση του υποσυνόλου χαρακτηριστικών

Στο πρόβλημα της επιλογής χαρακτηριστικών, η βασική παράμετρος που πρέπει να καθοριστεί, είναι η αναπαράσταση του χώρου όλων των πιθανών υποσυνόλων χαρακτηριστικών που προκύπτουν από το δοθέν σύνολο χαρακτηριστικών. Η αναπαράσταση ενός χρωμοσώματος

στο πρόβλημα της επιλογής χαρακτηριστικών, δηλαδή, η αναπαράσταση ενός υποψήφιου υποσυνόλου χαρακτηριστικών, γίνεται συνήθως είτε με δυαδική κωδικοποίηση (binary encoding) είτε με ακέραια κωδικοποίηση στο δεκαδικό σύστημα (integer encoding).

Δυαδική κωδικοποίηση (Binary encoding)

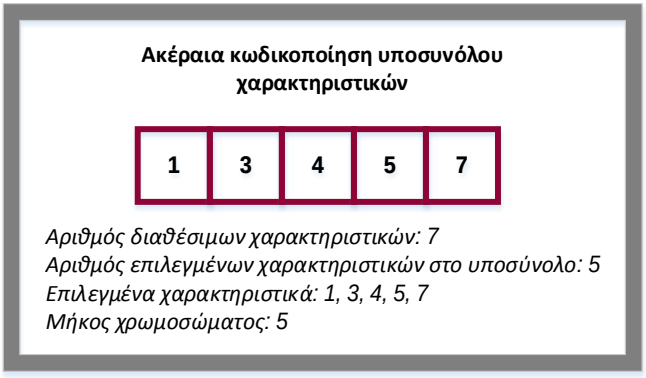
Η πιο συνήθης αναπαράσταση σε προβλήματα επιλογής χαρακτηριστικών είναι η δυαδική (*εικόνα 72*). Με δυαδική αναπαράσταση κάθε χαρακτηριστικό στο υποψήφιο υποσύνολο χαρακτηριστικών αντιστοιχεί σε ένα γονίδιο και κάθε χρωμόσωμα/άτομο αποτελείται από μια συγκεκριμένου μήκους δυαδική συμβολοσειρά που αναπαριστά ένα υποσύνολο χαρακτηριστικών. Ένα χρωμόσωμα μήκους *l* αντιστοιχεί σε ένα δυαδικό *l* διαστάσεων διάνυσμα χαρακτηριστικών X, όπου κάθε bit αντιπροσωπεύει τον αποκλεισμό ή την συμπερίληψη του ανάλογου χαρακτηριστικού. Το μέγεθος του διανύσματος (*l*) είναι ίσο με τον αριθμό όλων των διαθέσιμων χαρακτηριστικών (έστω all_F ο αριθμός όλων των χαρακτηριστικών, τότε all_F = *l*). Εάν $x_i = 0$ τότε το χαρακτηριστικό i δεν περιλαμβάνεται στο υποσύνολο, ενώ εάν $x_i = 1$, τότε υποδηλώνεται ότι το χαρακτηριστικό i περιλαμβάνεται στο υποσύνολο [16]. Ο ΓΑ πειραματίζεται με διάφορες τιμές του *max_F*, όπου *max_F* είναι ο μέγιστος αριθμός των χαρακτηριστικών που επιλέγονται για να συμπεριληφθούν στο υποσύνολο χαρακτηριστικών και λαμβάνει τιμές από [1 έως και (all_F – 1)] [53]. Κάθε ένα από τα χρωμοσώματα έχει ακριβώς max_F σε αριθμό γονίδια με τιμή 1. Για παράδειγμα, σε ένα πρόβλημα με 7 διαθέσιμα χαρακτηριστικά (all_F = 7), το χρωμόσωμα θα έχει μήκος ίσο με 7, με κάθε γονίδιο να λαμβάνει τιμή 0(απόρριψη χαρακτηριστικού) ή 1 (επιλογή χαρακτηριστικού). Με max_F= 3 και all_F= 7, το χρωμόσωμα που συμπεριλαμβάνει στο υποσύνολο μόνο τα χαρακτηριστικά 1,3 και 7 κωδικοποιείται ως 1010001. Το πλεονέκτημα αυτής της κωδικοποίησης είναι ότι οι κλασικοί τελεστές των ΓΑ (binary mutation and crossover) μπορούν εύκολα να εφαρμοστούν σε αυτή την αναπαράσταση χωρίς κάποια μετατροπή. Αυτό εξαφανίζει την ανάγκη για σχεδιασμό καινούριων γενετικών τελεστών ή για μετατροπή της συνήθους μορφής των ΓΑ [51].



Εικόνα 72: *Δυαδική κωδικοποίηση υποσυνόλου χαρακτηριστικών*

Ακέραια κωδικοποίηση (Integer encoding)

Στην ακέραια κωδικοποίηση σε δεκαδικό αριθμητικό σύστημα (*εικόνα 73*), εάν το all_F είναι ο συνολικός αριθμός των διαθέσιμων χαρακτηριστικών, το μήκος του χρωμοσώματος θα είναι ίσο με max_F, όπου max_F είναι ο μέγιστος αριθμός χαρακτηριστικών που πρόκειται να συμπεριλαμβάνονται στο υποσύνολο χαρακτηριστικών και έχει εύρος τιμών [1, (all_F – 1)] [53]. Κάθε γονίδιο μπορεί να λάβει μια τιμή από το εύρος [1, all_F] και κάθε ακέραιος σε αυτό το εύρος αποτελεί δείκτη σε ένα από τα all_F χαρακτηριστικά. Για παράδειγμα, έστω ότι τα διαθέσιμα χαρακτηριστικά ενός προβλήματος είναι 7 (all_F=7). Κάθε γονίδιο στο χρωμόσωμα μπορεί να πάρει ως τιμή έναν ακέραιο εκ των [1,...,7]. Με all_F=7 και max_F = 3, το υποσύνολο χαρακτηριστικών που περιλαμβάνει μόνο τα χαρακτηριστικά 1,3 και 7 κωδικοποιείται ως 137 (ή οποιοδήποτε άλλο συνδυασμό 173, 371, 317, 713, 731). Ο ΓΑ πειραματίζεται με διάφορες τιμές του max_F στο εύρος [1, (all_F – 1)] ώστε να καταλήξει στο πιο κατάλληλο μήκος χρωμοσώματος.



Εικόνα 73: *Ακέραια κωδικοποίηση υποσυνόλου χαρακτηριστικών*

Άλλες αναπαραστάσεις

Η δυαδική αναπαράσταση του χρωμοσώματος, όπου κάθε γονίδιο παίρνει τιμή 0 ή 1, μπορεί να ιδωθεί και από μια άλλη οπτική πλευρά: ως το βάρος 0 ή 1 της σημαντικότητας του κάθε χαρακτηριστικού. Υπό αυτό το πρίσμα θεώρησης, μια άλλη υποσχόμενη προσέγγιση είναι αυτή που έχουν χρησιμοποιήσει στο [54] όπου αντί η αναπαράσταση (το βάρος) να έχει τιμές μόνο 0 ή 1, παίρνει τιμές από ένα μεγαλύτερο εύρος, π.χ. από 0 έως 10, παράγοντας αυτό που περιγράφουν ως μια «στρέβλωση» (“warping”) του χώρου χαρακτηριστικών. Με τον τρόπο αυτό, η αναζήτηση του υποσυνόλου χαρακτηριστικών ανάγεται σε πρόβλημα εύρεσης του διανύσματος με τα σχετικά βάρη των χαρακτηριστικών τα οποία δίνουν τη μέγιστη απόδοση στην ταξινόμηση του συνόλου εκπαίδευσης. Ουσιαστικά, αυτή η αναπαράσταση επιτρέπει στη σχετική σημαντικότητα κάθε χαρακτηριστικού ως προς την ταξινόμηση, να αξιολογηθεί ανάλογα με το βάρος του κάθε χαρακτηριστικού. Τα βάρη που πλησιάζουν το 0, υποδεικνύουν ότι τα αντίστοιχα

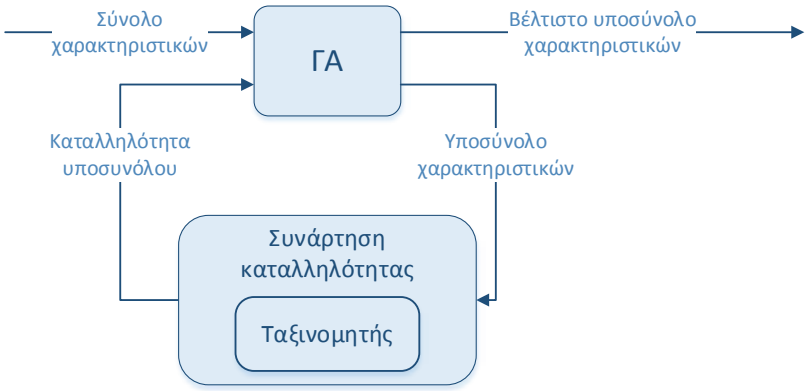
χαρακτηριστικά δεν είναι σημαντικά για τη διαχωριστική διεργασία, ενώ τα βάρη που πλησιάζουν τη μέγιστη τιμή του εύρους, υποδεικνύουν ότι η διαδικασία της ταξινόμησης είναι ευαίσθητη σε μικρές αλλαγές σε αυτά τα χαρακτηριστικά.

Μια άλλη προσέγγιση που έχει προταθεί στο [55] χρησιμοποιεί μια αναπαράσταση γνωστή ως "messy genetic algorithm". Σε αυτού του είδους τους ΓΑ, κάθε γονίδιο αποτελείται από ένα ζεύγος (Gene number, Allele Value) και κάθε χρωμόσωμα είναι ένα σύνολο τέτοιων γονιδίων. Τα χρωμοσώματα μπορεί να έχουν μεταβλητό μέγεθος: μια υποψήφια λύση μπορεί να είναι μη επαρκώς καθορισμένη, δηλαδή ένα δοσμένο γονίδιο να μην αναπαρίσταται καθόλου στη λύση, ή να είναι υπερ-καθορισμένη, δηλαδή ένα γονίδιο να αναπαρίσταται περισσότερες από μια φορές. Για παράδειγμα το ((6,0),(3,1),(3,0),(1,1),(2,0)) είναι ένα χρωμόσωμα με 5 γονίδια. Το χρωμόσωμα αυτό είναι υπερ-καθορισμένο αφού το γονίδιο 3 έχει δύο διαφορετικές τιμές: 0 και 1. Ταυτόχρονα, είναι και μη επαρκώς καθορισμένο, αφού δεν έχει τιμές για όλα τα πιθανά γονίδια: τα γονίδια 4 και 5 (ίσως και άλλα), δεν αναπαρίστανται καθόλου στο χρωμόσωμα. Ένα «messy chromosome" ερμηνεύεται ως εξής: κάθε γονίδιο που είναι καθορισμένο συμπεριλαμβάνεται στο υποσύνολο χαρακτηριστικών, ενώ τα γονίδια που δεν είναι καθορισμένα δεν συμπεριλαμβάνονται.

3.2.2 Αξιολόγηση υποψήφιων υποσυνόλων - Συνάρτηση καταλληλότητας

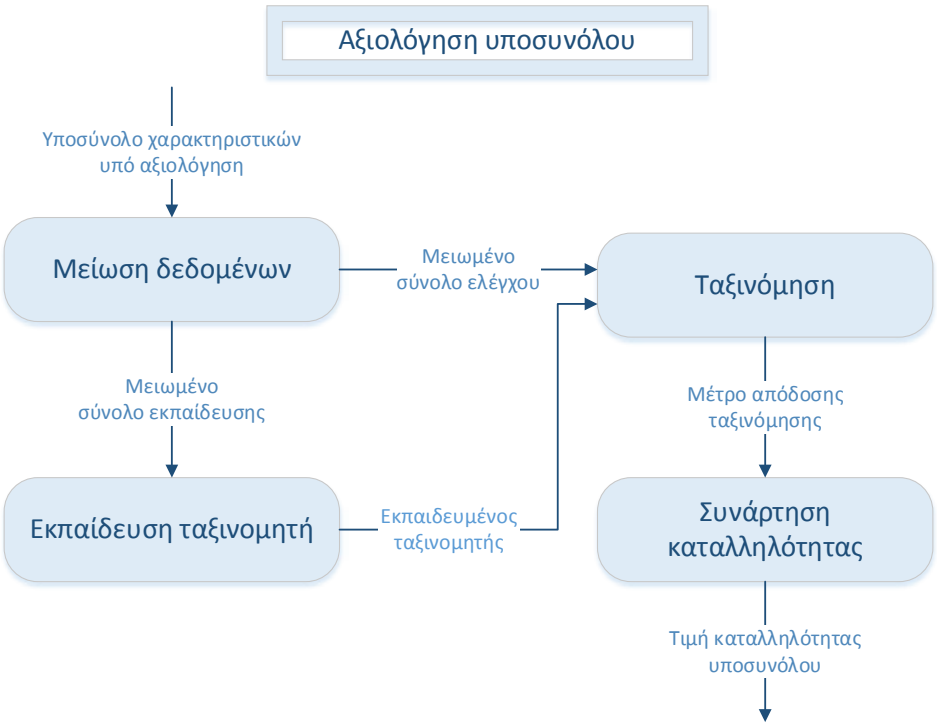
Για μια επιτυχή εφαρμογή των ΓΑ σε οποιοδήποτε είδος προβλήματος, συμπεριλαμβανομένου και αυτού της επιλογής χαρακτηριστικών, πρέπει να καθοριστεί και μια άλλη κρίσιμη παράμετρος, η επιλογή της κατάλληλης συνάρτησης καταλληλότητας (fitness function). Οι συναρτήσεις καταλληλότητας παρέχουν στους ΓΑ πληροφορίες για την καταλληλότητα κάθε ατόμου στον πληθυσμό. Ακολούθως οι ΓΑ χρησιμοποιούν αυτή την πληροφορία για να ανακατευθύνουν τη διαδικασία αναζήτησης ώστε να βελτιωθεί η μέση καταλληλότητα του πληθυσμού [56].

Η απόδοση ενός υποψήφιου υποσυνόλου χαρακτηριστικών εκτιμάται με τη συνάρτηση καταλληλότητας. Ένα υποσύνολο χαρακτηριστικών αξιολογείται βάσει της ικανότητάς του να ταξινομήσει ορθά στις κατάλληλες κλάσεις τα παραδείγματα του συνόλου εκπαίδευσης. Για το λόγο αυτό, είναι ζωτικής σημασίας για την επιτυχία του συστήματος η ακριβής μέτρηση της αποτελεσματικότητας της ταξινόμησης. Στην *εικόνα 74* φαίνεται η λειτουργία της συνάρτησης καταλληλότητας στο πρόβλημα της επιλογής χαρακτηριστικών με χρήση ΓΑ.



Εικόνα 74: *Ρόλος συνάρτησης καταλληλότητας στην επιλογή χαρακτηριστικών με ΓΑ*

Η συνάρτηση καταλληλότητας παίρνει ως είσοδο ένα χρωμόσωμα και επιστρέφει μια αριθμητική αξιολόγηση που αντιπροσωπεύει την καταλληλότητα του υποσυνόλου. Για την αξιολόγηση ενός υποσυνόλου χαρακτηριστικών, πρέπει πρώτα τα αρχικά δεδομένα εκπαίδευσης- τα οποία αποτελούνται από διανύσματα με τιμές για όλα τα χαρακτηριστικά από το αρχικό σύνολο, καθώς και τις αντίστοιχες κλάσεις (για την ακρίβεια ετικέτες κλάσεων) στις οποίες ανήκει κάθε παράδειγμα- να μειωθούν. Αυτό επιτυγχάνεται αφαιρώντας τα χαρακτηριστικά που δεν βρίσκονται στο επιλεγμένο υποσύνολο χαρακτηριστικών από τα διανύσματα του συνόλου εκπαίδευσης. Τα ανανεωμένα μειωμένα διανύσματα του συνόλου εκπαίδευσης ακολούθως χρησιμοποιούνται για την εκπαίδευση του ταξινομητή. Αφού ολοκληρωθεί η εκπαίδευση του ταξινομητή, πρέπει να μειωθούν τα δεδομένα του συνόλου ελέγχου με τον ίδιο τρόπο που μειώθηκαν και τα δεδομένα εκπαίδευσης. Στη συνέχεια, τα μειωμένα δείγματα του συνόλου ελέγχου, με τιμές για κάθε χαρακτηριστικό που περιλαμβάνεται στο υποσύνολο, χρησιμοποιούνται ως διανύσματα εισόδου για να τροφοδοτήσουν τον εκπαιδευμένο ταξινομητή. Κάθε δείγμα ελέγχου που αντιπροσωπεύεται από ένα μειωμένο διάνυσμα χαρακτηριστικών εισάγεται στον ταξινομητή ο οποίος το κατατάσσει σε μια κλάση (επιστρέφει ως έξοδο μια ετικέτα κλάσης). Εάν η κλάση στην οποία το κατατάσσει συμφωνεί με τη δοσμένη κλάση, τότε η ταξινόμηση θεωρείται ορθή. Αφού ταξινομηθούν όλα τα δείγματα του συνόλου ελέγχου, τότε υπολογίζεται το σφάλμα ταξινόμησης (ή κάποιο άλλο μέτρο της απόδοσης του ταξινομητή π.χ. ακρίβεια ταξινόμησης, ειδικότητα, ευαισθησία). Με αυτό τον τρόπο ολοκληρώνεται η αξιολόγηση ενός συγκεκριμένου υποψήφιου υποσυνόλου χαρακτηριστικών ως προς την απόδοση της ταξινόμησης. Στην *εικόνα 75* παρουσιάζεται η παραπάνω διαδικασία.

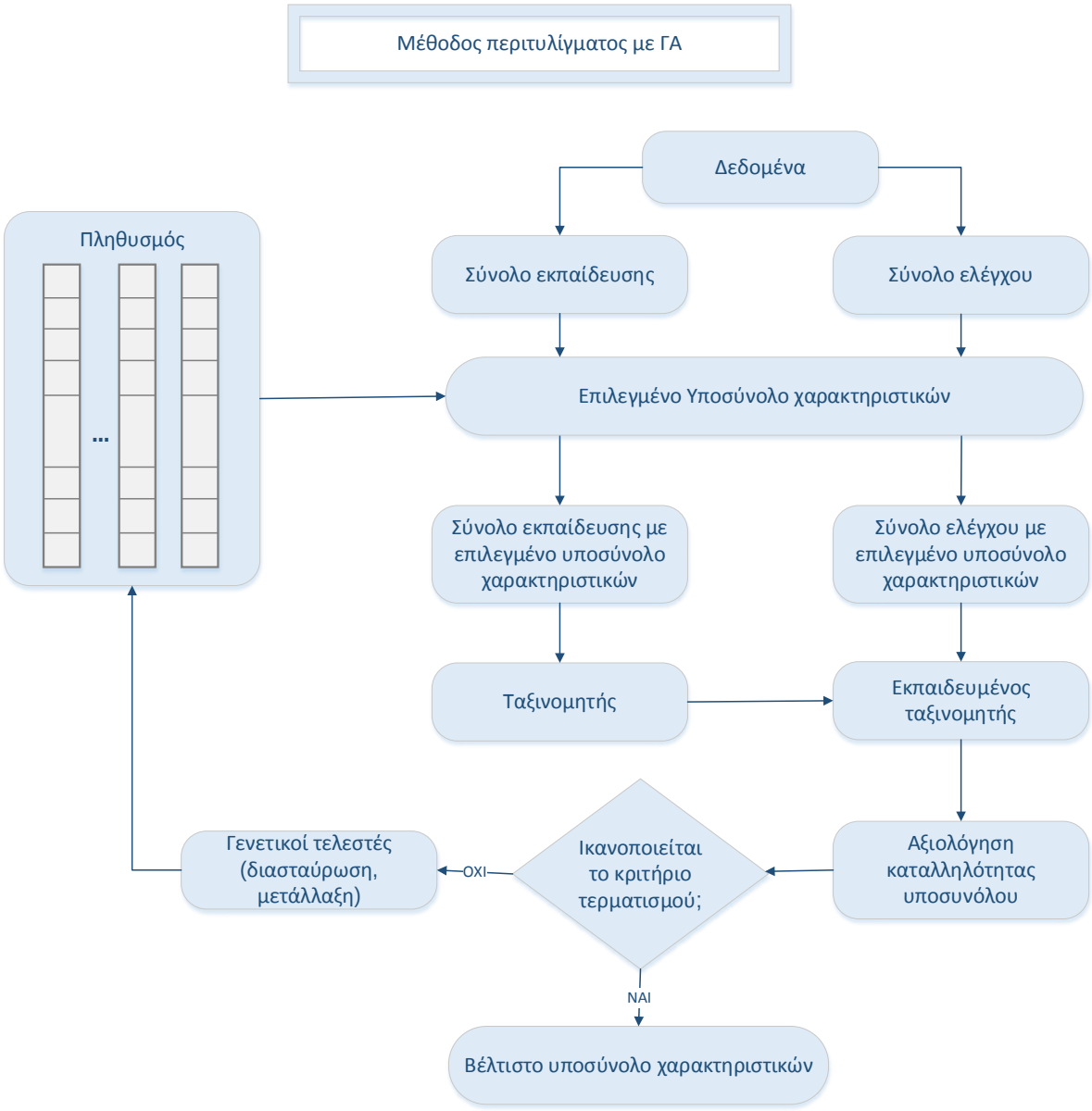


Εικόνα 75: *Αξιολόγηση υποσυνόλου χαρακτηριστικών με χρήση της συνάρτησης καταλληλότητας*

Η συνάρτηση καταλληλότητας εκτός από το κριτήριο σχετικά με την απόδοση ταξινόμησης μπορεί να συμπεριλαμβάνει κι άλλα κριτήρια αξιολόγησης. Μια ευρέως χρησιμοποιούμενη συνάρτηση καταλληλότητας στο πρόβλημα της επιλογής χαρακτηριστικών είναι ο συνδυασμός 2 κριτηρίων: της απόδοσης του ταξινομητή και του αριθμού των χαρακτηριστικών που περιέχει το υπό αξιολόγηση υποσύνολο.

Ας σημειωθεί ότι η συνάρτηση καταλληλότητας πρέπει να σχεδιαστεί με τέτοιο τρόπο ώστε ανεξάρτητα από την είσοδο που θα δεχθεί, να επιστρέφει μια τιμή που να είναι επιτρεπτή από τον ΓΑ. Η τιμή που επιστρέφει η συνάρτηση καταλληλότητας για να είναι επιτρεπτή από το ΓΑ απαιτείται να είναι θετική εξαιτίας της κατασκευής των γενετικών τελεστών του ΓΑ (για να είναι εφικτή η εφαρμογή των γενετικών τελεστών όπως έχουν, διαφορετικά θα είναι αναγκαίο να τροποποιηθούν). Συνεπώς, όποιο/α κριτήριο/α κι αν χρησιμοποιούνται στη συνάρτηση καταλληλότητας, προτείνεται να είναι μαθηματικώς διατυπωμένα με τέτοια μορφή, ώστε η τιμή που θα επιστρέφεται να είναι σε όλες τις περιπτώσεις θετική. Σε αντίθετη περίπτωση που η συνάρτηση καταλληλότητας επιστρέφει και αρνητικές τιμές, θα πρέπει να γίνουν τροποποιήσεις στους γενετικούς τελεστές.

Στην *εικόνα 76* παρουσιάζεται συνολικά μια κλασική wrapper προσέγγιση της χρήσης ΓΑ στο πρόβλημα επιλογής χαρακτηριστικών.



Εικόνα 76: *Μέθοδος περιτυλίγματος για επιλογή χαρακτηριστικών με ΓΑ*

3.3. Σχετική έρευνα για συστήματα που συνδυάζουν Γενετικούς Αλγόριθμους με κάποιο ταξινομητή

Οι ΓΑ έχουν εφαρμοστεί σε προβλήματα επιλογής χαρακτηριστικών από διάφορους ερευνητές σε πολλούς διαφορετικούς τομείς. Ιδιαίτερο ενδιαφέρον παρουσιάζει ο συνδυασμός τέτοιων εφαρμογών, με τεχνητά νευρωνικά δίκτυα (artificial neural networks-ANN, ΤΝΔ) ως ταξινομητή. Ακολούθως, θα αναφερθούμε σε κάποιες από τις εφαρμογές αυτές.

3.3.1 Σχετική μελέτη 1: Εφαρμογή σε ιατρική διάγνωση

Οι Yang και Honavar, το 1998 [57], συνδύασαν έναν ταξινομητή τεχνητών νευρωνικών δικτύων με ένα ΓΑ. Ο ΓΑ υλοποιεί την επιλογή των χαρακτηριστικών που θα χρησιμοποιηθούν για την ταξινόμηση από το τεχνητό νευρωνικό δίκτυο. Το ίδιο το τεχνητό νευρωνικό δίκτυο ενσωματώνεται ως μέρος της αντικειμενικής συνάρτησης αξιολόγησης του ΓΑ.

Κατά τους Yang και Honavar το πρόβλημα επιλογής υποσυνόλου χαρακτηριστικών στο πλαίσιο πολλών πρακτικών εφαρμογών, όπως η ιατρική διάγνωση, μπορεί να αντιμετωπιστεί ως ένα τμήμα ενός προβλήματος βελτιστοποίησης πολλών κριτηρίων. Τα πολλαπλά κριτήρια που πρέπει να βελτιστοποιηθούν, συμπεριλαμβάνουν την ακρίβεια της ταξινόμησης, καθώς και το κόστος και ρίσκο της ταξινόμησης, τα οποία με τη σειρά τους εξαρτώνται από την επιλογή των κατάλληλων χαρακτηριστικών, που θα χρησιμοποιηθούν για την περιγραφή των προτύπων [57]. Οι ίδιοι ακολουθούν μια wrapper προσέγγιση, με τη χρήση ενός ΓΑ σε συνδυασμό με ένα τεχνητό νευρωνικό δίκτυο, για την επιλογή του βέλτιστου υποσυνόλου χαρακτηριστικών.

Για την εκπαίδευση του τεχνητού νευρωνικού δικτύου χρησιμοποίησαν έναν σχετικά γρήγορο, επαγωγικό αλγόριθμο εκμάθησης που χρησιμοποιεί ως μέτρο την απόσταση. Η αναπαράσταση που χρησιμοποιήθηκε για το πρόβλημα ήταν η συνήθης δυαδική αναπαράσταση των n-bits για n χαρακτηριστικά, όπου το 0 υποδεικνύει ότι το αντίστοιχο χαρακτηριστικό δεν συμπεριλαμβάνεται στο υποσύνολο χαρακτηριστικών, ενώ το 1 υποδεικνύει ότι το αντίστοιχο χαρακτηριστικό συμπεριλαμβάνεται. Η συνάρτηση καταλληλότητας (fitness function) σχεδιάστηκε με τέτοιο τρόπο ώστε να συνδυάζει 2 κριτήρια, την ακρίβεια ταξινόμησης (classification accuracy) και το κόστος της συμπερίληψης επιπρόσθετων χαρακτηριστικών στην ταξινόμηση. Η ακρίβεια της ταξινόμησης μπορεί να υπολογιστεί ως το ποσοστό των επιτυχών ταξινομήσεων του συνόλου ελέγχου προς όλο το σύνολο ελέγχου. Για το κόστος ταξινόμησης, μπορούν να χρησιμοποιηθούν διάφορα μέτρα, όπως για παράδειγμα το κόστος μέτρησης της τιμής ενός συγκεκριμένου χαρακτηριστικού, ή αν πρόκειται για ιατρική εφαρμογή, το κόστος διεκπεραίωσης της διαγνωστικής εξέτασης. Επομένως, η συνάρτηση καταλληλότητας λειτουργεί έτσι ώστε να μειώνει τον αριθμό των χαρακτηριστικών που θα συμπεριλαμβάνονται στο βέλτιστο υποσύνολο χαρακτηριστικών. Οι Yang και Honavar τονίζουν τη σημαντικότητα της επιλογής χαρακτηριστικών σε προβλήματα ταξινόμησης που αφορούν ιατρική διάγνωση, όπου διαφορετικές διαγνωστικές εξετάσεις μπορεί να έχουν διαφορετικό κόστος και ρίσκο. Το πρόβλημα με το οποίο ασχολούνται αφορά την επιλογή ενός υποσυνόλου συμπτωμάτων και κλινικών εξετάσεων - η κάθε μια από τις οποίες έχει διαφορετικό οικονομικό κόστος, διαφορετική διαγνωστική αξία και διαφορετικό εμπλεκόμενο κίνδυνο – και αποτελεί μέρος της γενικότερης ιατρικής διαγνωστικής διεργασίας. Για την ταξινόμηση, χρησιμοποιείται ο αλγόριθμος DistAI [58] που είναι ένας γρήγορος και απλός

επαγωγικός αλγόριθμος εκμάθησης τεχνητού νευρωνικού δικτύου για ταξινόμηση. Όσον αφορά τις παραμέτρους του ΓΑ: ορίζουν μέγεθος πληθυσμού ίσο με 50, γενεές 20, πιθανότητα διασταύρωσης 0.6, πιθανότητα μετάλλαξης 0.001 και πιθανότητα επιλογής του καλύτερου ατόμου (πιθανότητα ελιτισμού) ίση με 0.6.

Μέσω δοκιμών, διαπίστωσαν ότι η συνάρτηση καταλληλότητας που συνδυάζει την ακρίβεια και το κόστος ταξινόμησης (μεγιστοποίηση ακρίβειας και ελαχιστοποίηση κόστους), δίνει καλύτερα αποτελέσματα από ότι μια συνάρτηση καταλληλότητας με μόνο κριτήριο την ακρίβεια της ταξινόμησης. Βάσει των αποτελεσμάτων τους, θεωρούν ότι οι ΓΑ είναι μια καλή εναλλακτική προσέγγιση στο πρόβλημα της επιλογής υποσυνόλου χαρακτηριστικών για ταξινόμηση, ιδιαίτερα σε εφαρμογές όπου το κόστος αποτελεί σημαντικό παράγοντα (π.χ. ιατρική διάγνωση, όραση υπολογιστών κτλ).

3.3.2 Σχετική μελέτη 2: Εφαρμογή για διάγνωση Alzheimer βάσει του EEG

Στο [59] έχει προταθεί το 2005 από τους Kim H.T., Kim B.Y., Park, Kim J.W., Hwang, Han και Cho, ένας συνδυασμός ΓΑ και τεχνητού νευρωνικού δικτύου για την υλοποίηση ενός αυτόματου συστήματος αναγνώρισης της ασθένειας Alzheimer. Το σύστημα αναγνώρισης της παρουσίας ή απουσίας της ασθένειας Alzheimer χρησιμοποιεί χαρακτηριστικά που προκύπτουν από αυθόρμητο ηλεκτροεγκεφαλογράφημα (ΗΕΓ, electroencephalogram-EEG) και από ακουστικά προκλητά δυναμικά (auditory event-related potential-ERP) που ηχογραφήθηκαν σε ενιαίο χώρο. Το EEG καταγράφει την εγκεφαλική δραστηριότητα, δηλαδή την αυθόρμητη ηλεκτρική ρυθμική ταλάντωση του εγκεφάλου όταν βρίσκεται σε ηρεμία και δεν δέχεται εξωτερικά ερεθίσματα. Τα ακουστικά προκλητά δυναμικά, προκύπτουν από το EEG, και είναι οι διαφορές δυναμικού που καταγράφονται στη δερματική επιφάνεια του κεφαλιού, οι οποίες προκαλούνται ως απόκριση ή ως προετοιμασία σε κάποιο ακουστικό γεγονός/ερέθισμα το οποίο πραγματοποιείται στο εξωτερικό περιβάλλον (ήχοι, λέξεις κτλ.). Επειδή παρατηρούνται στο επιφανειακό μέρος του κεφαλιού (με τη χρήση ηλεκτροδίων) αποτελούν ένα μη επεμβατικό τρόπο εκτίμησης της εγκεφαλικής λειτουργίας και δραστηριότητας. Τα EEG και ERP αναλύθηκαν και από την ανάλυση αυτή, προέκυψαν διάφορα χαρακτηριστικά, τα οποία αποτέλεσαν την «δεξαμενή» των διαθέσιμων χαρακτηριστικών (118 χαρακτηριστικά από το EEG και 10 από το ERP, σύνολο 128 χαρακτηριστικά). Ο συνδυασμός ΓΑ/ΤΝΔ εφαρμόστηκε για την επιλογή των κυρίαρχων χαρακτηριστικών που απαρτίζουν το μικρότερο δυνατό υποσύνολο χαρακτηριστικών και το πιο αποδοτικό ως προς την ταξινόμηση στις δύο κλάσεις (απουσία/παρουσία Alzheimer). Τα 35 καλύτερα χαρακτηριστικά που επιλέχτηκαν, ακολούθως χρησιμοποιήθηκαν για την τροφοδότηση

του ΤΝΔ για εκπαίδευση και έλεγχο. Το ποσοστό αναγνώρισης της ασθένειας από το ΤΝΔ, με είσοδο τα δεδομένα του συνόλου ελέγχου για αυτά τα 35 χαρακτηριστικά, ήταν 81.9%.

Το ΤΝΔ που χρησιμοποιήθηκε είναι ένα ΤΝΔ πολλαπλών επιπέδων (multi layered perceptron, MLP) και για την εκπαίδευσή του έγινε χρήση του αλγόριθμου της όπισθεν διάδοσης του σφάλματος (error backpropagation).

Ο ΓΑ που χρησιμοποιήθηκε είχε 200 χρωμοσώματα σε κάθε γενεά, μέγιστο αριθμό γενεών 200 (κριτήριο τερματισμού), πιθανότητα διασταύρωσης 0.95, πιθανότητα μετάλλαξης 0.05 και πιθανότητα ελιτισμού 0.001. Για την επιλογή των καλύτερων χρωμοσωμάτων του τρέχοντος πληθυσμού για το στάδιο της αναπαραγωγής, έγινε χρήση της μεθόδου επιλογής της ρουλέτας (roulette wheel method), με πιθανότητα βασισμένη στην τιμή καταλληλότητας κάθε χρωμοσώματος. Ο τελεστής διασταύρωσης για την παραγωγή δύο απογόνων από δύο γονείς υλοποιήθηκε με τη μέθοδο διασταύρωσης ενός σημείου (one-point crossover). Στον τελεστή μετάλλαξης, για την παραγωγή ενός καινούριου χρωμοσώματος, επιλεγόταν τυχαία για μετάλλαξη ένα γονίδιο ενός χρωμοσώματος-γονέα. Ο τελεστής του ελιτισμού αντέγραφε ένα γονέα-χρωμόσωμα στην επόμενη γενεά. Οι τιμές των παραμέτρων του ΓΑ καθορίστηκαν με τη μέθοδο δοκιμής και σφάλματος.

Το χρωμόσωμα ορίστηκε ως μια συμβολοσειρά που απαρτίζεται από τις σταθερές που αντιπροσωπεύουν τους αριθμούς των δεικτών (index numbers) των επιλεγμένων χαρακτηριστικών και από τις σταθερές που αντιπροσωπεύουν τα βάρη του ΤΝΔ. Τα πρώτα 35 γονίδια, δηλαδή, παίρνουν τιμή μια ακέραια σταθερά που αντιπροσωπεύει ένα επιλεγμένο χαρακτηριστικό και τα τελευταία γονίδια παίρνουν τιμή μια σταθερά που αφορά τα βάρη του ΤΝΔ, όπως αυτά καθορίζονται από τον αλγόριθμο οπισθοδιάδοσης του σφάλματος. Για κάθε χρωμόσωμα, το ΤΝΔ εκπαιδευόταν με τα δεδομένα εκπαίδευσης με είσοδο τα χαρακτηριστικά που περιείχε το χρωμόσωμα. Με την εκπαίδευση του ΤΝΔ προσαρμόζονταν και τα βάρη του. Η συνάρτηση καταλληλότητας που εφαρμόστηκε ήταν ίση με τον αντίστροφο του αθροίσματος των μέσων τετραγωνικών σφαλμάτων του ΤΝΔ.

Η επιλογή των 35 κυρίαρχων χαρακτηριστικών έγινε ως εξής: από κάθε γενεά επιλέγεται το χρωμόσωμα με τη μέγιστη τιμή καταλληλότητας. Τα χρωμοσώματα που επιλέχθηκαν από όλες τις γενεές, χρησιμοποιήθηκαν για την επιλογή του βέλτιστου υποσυνόλου χαρακτηριστικών. Για κάθε χαρακτηριστικό, κατέγραψαν τον αριθμό των φορών που επιλέχτηκε από αυτά τα χρωμοσώματα. Ο αριθμός αυτός, αντιπροσώπευε την σημαντικότητα κάθε χαρακτηριστικού για να περιέχεται στην λύση (το τελικό υποσύνολο των κυρίαρχων χαρακτηριστικών). Βάσει της σημαντικότητας, επιλέχθηκαν τα 35 κυρίαρχα χαρακτηριστικά.

Στη συνέχεια, το ΤΝΔ εκπαιδεύτηκε με το σύνολο των δεδομένων εκπαίδευσης (περιέχοντας μόνο τα 35 κυρίαρχα χαρακτηριστικά) έτσι ώστε να προσαρμοστούν τα βάρη του. Έπειτα, αφού

το ΤΝΔ εκπαιδεύτηκε και καθορίστηκαν οι τιμές των βαρών του, εκτιμήθηκε η απόδοσή του, χρησιμοποιώντας το σύνολο των δεδομένων ελέγχου.

Παρατηρήθηκε, ότι η προσέγγιση που ακολουθήθηκε, δηλαδή ο συνδυασμός ΓΑ/ΤΝΔ , ήταν ικανή να εντοπίσει το βέλτιστο υποσύνολο χαρακτηριστικών και να επιδείξει καλή απόδοση στην αναγνώριση ενός δείγματος ως φυσιολογικό EEG ή ως EEG ασθενή με Alzheimer.

3.3.3 Σχετική μελέτη 3: Πρόταση τροποποιημένου Γενετικού Αλγορίθμου με Τεχνητό Νευρωνικό Δίκτυο

Οι Brill, Brown και Martin το 1992 [60] συνδύασαν ένα ΓΑ με ένα ΤΝΔ με διαφορετικό τρόπο από ότι οι προηγούμενες προσεγγίσεις. Επέλεξαν να χρησιμοποιήσουν μια παραλλαγή του κλασικού ΓΑ, τον λεγόμενο GAPE (genetic algorithm with punctuated equilibria). Η εκδοχή ΓΑ που χρησιμοποίησαν ενσωμάτωνε τροποποιημένους γενετικούς τελεστές, και αντί να αποτελείται από ένα μόνο πληθυσμό, αποτελείτο από ένα σύνολο πληθυσμών, οι οποίοι εξελίσσονταν ξεχωριστά ο ένας από τον άλλο, για ένα συγκεκριμένο αριθμό γενεών (μια εποχή), προτού ανταλλάξουν τις καλύτερες λύσεις κάθε πληθυσμού με αυτές από άλλους πληθυσμούς. Με το πέρας κάθε εποχής οι τρέχοντες πληθυσμοί αντάλλαζαν εκ νέου τις καλύτερες τους λύσεις. Ισχυρίζονται, ότι αυτού του είδους ο ΓΑ, έδειχνε να ξεπερνά σε επίδοση τους κλασικούς ΓΑ, με την καθιερωμένη μορφή που γνωρίζουμε, σε διάφορα προβλήματα.

Επιπλέον, για τη συνάρτηση καταλληλότητας του ΓΑ, αντί ΤΝΔ, χρησιμοποίησαν έναν ταξινομητή πλησιέστερου γείτονα (nearest neighbour classifier), επειδή θεώρησαν ότι η εκπαίδευση ενός ΤΝΔ για την αξιολόγηση κάθε χρωμοσώματος είναι απαγορευτικά ακριβή υπολογιστικά. Αντίθετα, ένας ταξινομητής πλησιέστερου γείτονα απαιτούσε μικρότερο χρόνο εκτέλεσης (1-3 λεπτά σε σύγκριση με 15-35 λεπτά) κι έτσι η διαδικασία επιταχυνόταν. Το ΤΝΔ χρησιμοποιήθηκε για την εκπαίδευση και αξιολόγηση του τελικού ταξινομητή. Συγκεκριμένα, το είδος ΤΝΔ που χρησιμοποιήθηκε είναι ένα δίκτυο αντίθετης διάδοσης (counterpropagation). Τα χαρακτηριστικά του βέλτιστου υποσυνόλου που επιλέχτηκαν από τον ΓΑ με τη χρήση του ταξινομητή πλησιέστερου γείτονα, φαίνεται να είναι αποδοτικά όταν χρησιμοποιούνται για την τροφοδότηση του ΤΝΔ αντίθετης διάδοσης με το σύνολο των δεδομένων εκπαίδευσης. Ιδανικά ο ταξινομητής που χρησιμοποιείται στην αξιολόγηση κάθε χρωμοσώματος θα έπρεπε να είναι ο ίδιος με τον τελικό ταξινομητή αφού η χρήση κάποιου διαφορετικού ταξινομητή στην αξιολόγηση προσαρμόζει την επιλογή των καλύτερων χρωμοσωμάτων σε αυτόν. Ως εναλλακτική λύση προτείνουν η αξιολόγηση να γίνεται με ένα γρήγορο ταξινομητή στην αρχή του ΓΑ ώστε να περιοριστούν τα υποψήφια χαρακτηριστικά και αργότερα στις τελευταίες γενεές του ΓΑ η αξιολόγηση να γίνεται με τον τελικό ταξινομητή.

Για τα χρωμοσώματα, χρησιμοποιούν δυαδική αναπαράσταση και για την αξιολόγησή τους χρησιμοποιείται ένας γραμμικός συνδυασμός του σφάλματος (όπως αυτό προκύπτει από τον ταξινομητή πλησιέστερου γείτονα) και του αριθμού των χαρακτηριστικών. Ακολούθως οι βαθμολογίες αξιολόγησης που προκύπτουν, δίνονται στη συνάρτηση καταλληλότητας του ΓΑ , η οποία τις μετατρέπει σε τιμές που μπορούν να χρησιμοποιηθούν από τον ΓΑ. Η συνάρτηση καταλληλότητας που εφαρμόζεται σε ένα χρωμόσωμα χρησιμοποιεί την βαθμολογία αξιολόγησης του χρωμοσώματος, τη μέση βαθμολογία του τρέχοντος πληθυσμού, καθώς και την τυπική απόκλιση των βαθμολογιών του τρέχοντος πληθυσμού. Με αυτό τον τρόπο η συνάρτηση καταλληλότητας είναι ανάλογη του τρέχοντος πληθυσμού, κάτι που είναι χρήσιμο στην εκδοχή του ΓΑ (GAPE) που εφαρμόζουν.

Αν και η χρήση του ταξινομητή πλησιέστερου γείτονα για την αξιολόγηση, μειώνει τον χρόνο εκτέλεσης του ΓΑ, ωστόσο δεν τον μειώνει αρκετά. Ένας ΓΑ πραγματοποιεί πάμπολλες αξιολογήσεις υποψήφιων υποσυνόλων χαρακτηριστικών σε κάθε εκτέλεσή του. Εκτίμησαν ότι το 90% του χρόνου εκτέλεσης του ΓΑ αφορούσε την αξιολόγηση με τον ταξινομητή πλησιέστερου γείτονα. Γι’ αυτό το λόγο προσπάθησαν να επιταχύνουν κι άλλο τη διαδικασία αξιολόγησης, μειώνοντας το χρόνο εκτέλεσης με μια μέθοδο την οποία ονόμασαν δειγματοληψία του συνόλου εκπαίδευσης (training set sampling). Στη μέθοδο αυτή, σε κάθε αξιολόγηση χρησιμοποιούσαν μόνο ένα υποσύνολο του συνόλου εκπαίδευσης. Διαπίστωσαν ότι με αυτή τη μέθοδο τα υποσύνολα χαρακτηριστικών στα οποία κατέληγε ο ΓΑ ήταν εξίσου καλά με αυτά που εντόπιζε αν η αξιολόγηση γινόταν με ολόκληρο το σύνολο εκπαίδευσης. Όταν η αξιολόγηση γίνει με ένα μικρό υποσύνολο του συνόλου εκπαίδευσης, η αξιολόγηση γίνεται πιο γρήγορα αλλά με λιγότερη ακρίβεια και υπάρχει κίνδυνος υπερπροσαρμογής. Με την προτεινόμενη μέθοδο όμως, επειδή σε κάθε γενιά επιλέγεται διαφορετικό υποσύνολο του συνόλου εκπαίδευσης αποφεύγεται η υπερπροσαρμογή και χρησιμοποιείται ολόκληρο το σύνολο εκπαίδευσης. Υπάρχει μόνο ένα πρόβλημα: κάθε γενιά έχει βαθμολογία αξιολόγησης βασισμένη στο δικό της υποσύνολο εκπαίδευσης και κάποια υποσύνολα είναι πιο «εύκολα» από άλλα, με αποτέλεσμα τα άτομα με την υψηλότερη τιμή καταλληλότητας να είναι αυτά που αξιολογήθηκαν με το πιο «εύκολο» υποσύνολο εκπαίδευσης αντί αυτά που αποτελούνται από το καλύτερο υποσύνολο χαρακτηριστικών. Συνεπώς, για να λυθεί το πρόβλημα και να μπορούν να συγκριθούν τα νέα άτομα μιας γενιάς με τα άτομα της προηγούμενης, θα πρέπει η προηγούμενη γενιά να επαναξιολογηθεί χρησιμοποιώντας το καινούριο υποσύνολο του συνόλου εκπαίδευσης. Έτσι, αν και με μικρότερα υποσύνολα εκπαίδευσης, κάθε αξιολόγηση ενός ατόμου γίνεται πιο γρήγορα, απαιτούνται περισσότερες αξιολογήσεις για κάθε γενιά. Ωστόσο ακόμα κι έτσι, η μέθοδος με τα μικρότερα υποσύνολα εκπαίδευσης είναι συνολικά πιο γρήγορη από ότι αν χρησιμοποιείτο ολόκληρο το σύνολο εκπαίδευσης.

Ως τελεστή διασταύρωσης, προτείνουν ένα δικό τους, το διωνυμικό, ο οποίος για κάθε γονίδιο του ατόμου που πρόκειται να δημιουργηθεί, «ρίπτει ένα νόμισμα» με πιθανότητα r για κορώνα. Εάν το αποτέλεσμα της ρίψης είναι κορώνα, τότε το γονίδιο σε αυτή τη θέση λαμβάνεται από τον πρώτο γονέα, αλλιώς λαμβάνεται από το δεύτερο. Τα γονίδια που λαμβάνονται από τον πρώτο γονέα είναι διωνυμικά κατανεμημένα με μέσο r και τα γονίδια που λαμβάνονται από το δεύτερο γονέα είναι επίσης διωνυμικά κατανεμημένα, αλλά με μέσο (1 – r). Εάν το r ισούται με τη σταθερά 0.5, τότε η διασταύρωση που προκύπτει είναι ομοιόμορφη (uniform). Η τιμή της πιθανότητας για κορώνα, δηλαδή η τιμή του r, παράγεται τυχαία, κάθε φορά που εφαρμόζεται ο τελεστής σε ένα ζεύγος γονέων. Με αυτό τον τρόπο ξεπερνούν το πρόβλημα της πόλωσης θέσης (positional bias) που παρουσιάζουν κάποιοι τελεστές, καθώς και το πρόβλημα της πόλωσης κατανομής (distributional bias) (εξηγούνται στο κεφάλαιο 1). Το πρόβλημα της πόλωσης θέσης αποφεύγεται επειδή η απόφαση για κάθε γονίδιο γίνεται ανεξάρτητα από τις αποφάσεις για τα υπόλοιπα γονίδια (κάθε γονίδιο έχει πιθανότητα να λάβει πληροφορία από τον πρώτο γονέα ίση με r). Η πόλωση κατανομής αποφεύγεται επειδή κάθε φορά που ο διωνυμικός τελεστής εφαρμόζεται, η πιθανότητα r επιλέγεται ανεξάρτητα από άλλες εφαρμογές του τελεστή σε άλλα ζεύγη γονέων ($r \in [0,1]$). Κάθε απόγονος μπορεί να έχει μικρή ομοιότητα με τους γονείς του (εάν r τείνει στο 0.5) έως και μεγάλη ομοιότητα (εάν r τείνει στο 1 ή στο 0).

Όσον αφορά τις διάφορες παραμέτρους του ΓΑ: ορίζουν 8 υποπληθυσμούς με 40 άτομα στον κάθε υποπληθυσμό, κάθε εποχή διαρκεί 25 γενεές και στο τέλος κάθε εποχής ανταλλάσσονται 3 άτομα. Η πιθανότητα διασταύρωσης ορίζεται ίση με 0.5 και η πιθανότητα μετάλλαξης με 0.1.

3.3.4 Σχετική μελέτη 4: Συνδυασμός Γενετικού Αλγόριθμου με άλλες μεθόδους επιλογής χαρακτηριστικών

Στο [61], το 2007, προτείνεται από τους Tan, Fu, Zhang και Bourgeois, μια εφαρμογή ΓΑ σε συνδυασμό με κάποιες άλλες τεχνικές επιλογής χαρακτηριστικών. Η ιδέα είναι τα παραγόμενα υποσύνολα χαρακτηριστικών από διαφορετικές μεθόδους επιλογής χαρακτηριστικών να χρησιμοποιηθούν από τον ΓΑ, για την παραγωγή του βέλτιστου υποσυνόλου χαρακτηριστικών.

Κάθε μεμονωμένο κριτήριο επιλογής χαρακτηριστικών προτείνει διαφορετικό υποσύνολο χαρακτηριστικών ως το βέλτιστο, με αποτέλεσμα η απόδοση ενός ταξινομητή, ανάλογα με το κριτήριο επιλογής χαρακτηριστικών στο οποίο βασίζεται, να ποικίλει. Αυτό καθιστά δύσκολη την απόφαση σχετικά με το ποια μέθοδος επιλογής χαρακτηριστικών είναι η καλύτερη για την ταξινόμηση άγνωστων συνόλων δεδομένων. Το πλεονέκτημα της προτεινόμενης προσέγγισης είναι ότι ικανοποιούνται περισσότερα από ένα κριτήρια επιλογής χαρακτηριστικών. Ο στόχος των

ερευνητών, είναι να χρησιμοποιήσουν χρήσιμη πληροφορία από διαφορετικές μεθόδους επιλογής χαρακτηριστικών ώστε να επιλέξουν καλύτερα υποσύνολα χαρακτηριστικών με μικρότερο μέγεθος και πιο υψηλή απόδοση στην ταξινόμηση.

Σε πρώτο στάδιο εφαρμόζονται διάφορες τεχνικές επιλογής χαρακτηριστικών στο σύνολο δεδομένων εκπαίδευσης. Ακολούθως, το υποσύνολο χαρακτηριστικών που προκύπτει από κάθε μια από αυτές τις μεθόδους, χρησιμοποιείται για τη δημιουργία της «δεξαμενής» χαρακτηριστικών του ΓΑ: η «δεξαμενή» χαρακτηριστικών του ΓΑ αντί να αποτελείται από όλα τα αρχικά χαρακτηριστικά αποτελείται μόνο από τα χαρακτηριστικά που προκύπτουν από τα διάφορα κριτήρια επιλογής χαρακτηριστικών. Στο δεύτερο στάδιο, ο ΓΑ θα αναζητήσει το βέλτιστο υποσύνολο χαρακτηριστικών από τη «δεξαμενή» των χαρακτηριστικών.

Κάποιοι αλγόριθμοι επιλογής χαρακτηριστικών παράγουν υποσύνολα χαρακτηριστικών, ενώ κάποιοι άλλοι παράγουν μια λίστα κατάταξης χαρακτηριστικών. Στην δεύτερη περίπτωση, καθορίζεται ένα σημείο το οποίο τεμαχίζει τη λίστα κατάταξης ώστε να προκύψει ένα υποσύνολο χαρακτηριστικών που θα περιέχει τα πιο σημαντικά χαρακτηριστικά (π.χ. επιλέγονται τα 20 καλύτερα χαρακτηριστικά σύμφωνα με τη λίστα κατάταξης).

Σχετικά με την αναπαράσταση των χρωμοσωμάτων του ΓΑ: κάθε άτομο αναπαριστά ένα υποσύνολο χαρακτηριστικών χρησιμοποιώντας την κλασική n-bit δυαδική αναπαράσταση, όπου n ο αριθμός των υποψήφιων χαρακτηριστικών (τιμή 1 σε ένα bit υποδεικνύει επιλογή του αντίστοιχου χαρακτηριστικού, ενώ τιμή 0 αποκλεισμό του).

Η συνάρτηση καταλληλότητας του ΓΑ στοχεύει στη βελτιστοποίηση δύο αντικειμένων: στη μεγιστοποίηση της ακρίβειας ταξινόμησης και στην ελαχιστοποίηση του μεγέθους του υποσυνόλου χαρακτηριστικών. Η συνάρτηση καταλληλότητας είναι έτσι σχεδιασμένη ώστε να δίνει την ελευθερία στο χρήστη να μπορεί να καθορίσει ποιος στόχος από τους δύο (μεγάλη ακρίβεια ταξινόμησης ή μικρό μέγεθος υποσυνόλου) θα έχει μεγαλύτερη προτεραιότητα. Αυτό επιτυγχάνεται με τη χρήση μιας μεταβλητής-βάρους η οποία όταν αυξάνεται δίνει μεγαλύτερη προτεραιότητα στην ακρίβεια ταξινόμησης και μικρότερη στο μέγεθος υποσυνόλου, ενώ όταν μειώνεται δίνει μικρότερη προτεραιότητα στην ακρίβεια ταξινόμησης και μεγαλύτερη στο μέγεθος υποσυνόλου.

Για τον γενετικό τελεστή της επιλογής χρησιμοποιείται η επιλογή ρουλέτας (roulette wheel selection) και για τον τελεστή διασταύρωσης η διασταύρωση ενός σημείου (one-point crossover). Στον τελεστή της μετάλλαξης επιλέγονται τυχαία n bits για αναστροφή.

Οι ερευνητές έδωσαν περισσότερη έμφαση στο να δείξουν πειραματικά ότι το υποσύνολο που προκύπτει από τη χρήση του ΓΑ σε συνδυασμό με διάφορους αλγόριθμους επιλογής χαρακτηριστικών, θα αυξήσει την ακρίβεια ταξινόμησης και θα έχει μικρότερο μέγεθος,

συγκριτικά με οποιοδήποτε από τα υποσύνολα που προκύπτουν από τη μεμονωμένη χρήση του κάθε αλγορίθμου επιλογής χαρακτηριστικών.

Ισχυρίζονται ότι αυτό γίνεται ανεξάρτητα από τον ταξινομητή που θα επιλεγεί ή από τις μεθόδους επιλογής χαρακτηριστικών που θα χρησιμοποιηθούν για τον σχηματισμό της δεξαμενής χαρακτηριστικών του ΓΑ.

Ως αλγόριθμο ταξινόμησης χρησιμοποιούν ένα ταξινομητή μηχανών διανυσμάτων στήριξης (Support Vector Machine - SVM). Εισηγούνται, όμως, ότι μπορούν να χρησιμοποιηθούν διάφοροι άλλοι αλγόριθμοι ταξινόμησης (όπως για παράδειγμα Naïve Bayes, τεχνητά νευρωνικά δίκτυα, δέντρα απόφασης).

Οι αλγόριθμοι επιλογής χαρακτηριστικών που επέλεξαν να χρησιμοποιήσουν για τη δημιουργία της δεξαμενής χαρακτηριστικών του ΓΑ είναι 2 φίλτρα (κριτήριο επιλογής βασισμένο στην εντροπία, στατιστικός έλεγχος υπόθεσης T-statistics) και μια μέθοδος περιτυλίγματος (SVM-RFE). Ωστόσο, υποστηρίζουν ότι η συμπερίληψη μεθόδων επιλογής χαρακτηριστικών που βασίζονται στη συσχέτιση μεταξύ των χαρακτηριστικών θα βελτιώσει την απόδοση του συστήματος.

Για εξοικονόμηση χρόνου καθόρισαν μικρές τιμές για το μέγεθος πληθυσμού και τον αριθμό γενεών του ΓΑ. Η πιθανότητα διασταύρωσης που δοκίμασαν ήταν ίση με 1 και η πιθανότητα μετάλλαξης με 0.001.

Δοκίμασαν τη μέθοδό τους με τρία διαφορετικά σύνολα δεδομένων: δύο σύνολα δεδομένων μικροσυστοιχιών (microarray) για σκοπούς διάγνωσης ασθένειας (έκφραση γονιδιακής πληροφορίας για καρκίνο του παχέος εντέρου στο ένα σύνολο και για καρκίνο του προστάτη στο άλλο) και ένα σύνολο δεδομένων από σήματα ραδιοεντοπιστή (radar) σχετικά με την ιονόσφαιρα. Τα πειραματικά αποτελέσματα έδειξαν ότι η προτεινόμενη προσέγγιση εύρεσης των καλύτερων υποσυνόλων χαρακτηριστικών έδινε υψηλότερη ακρίβεια ταξινόμησης καθώς και μικρότερα σε μέγεθος υποσύνολα, σε σύγκριση με αυτά που έδινε η κάθε μέθοδος επιλογής χαρακτηριστικών ξεχωριστά. Επιπλέον φαίνεται να εντοπίζει κάποια σημαντικά χαρακτηριστικά που είχαν υποτιμηθεί από κάποιες μεθόδους επιλογής χαρακτηριστικών όταν αυτές είχαν εφαρμοστεί μεμονωμένα.

3.3.5 Σχετική μελέτη 5: Γενετικοί αλγόριθμοι για επιλογή χαρακτηριστικών μεγάλης κλίμακας

Στο [33], το 1989, έχει γίνει από τους Siedlecki και Sklansky, μια από τις πρώτες μελέτες σχετικά με την επιλογή χαρακτηριστικών για σχεδίαση ταξινομητή χρησιμοποιώντας ΓΑ. Υποστηρίζουν ότι σύμφωνα με τα πειράματά τους, οι ΓΑ είναι ιδιαίτερα αποδοτικοί ακόμα κι όταν

πρόκειται για μεγάλης κλίμακας προβλήματα, όταν δηλαδή, η διαστασιμότητα του συνόλου των χαρακτηριστικών είναι μεγάλη. Πιο συγκεκριμένα, ως μεγάλης κλίμακας προβλήματα θεωρούν αυτά που έχουν διαστασιμότητα μεγαλύτερη από 20.

Θεωρούν ότι υπάρχουν δύο στόχοι που πρέπει να ικανοποιηθούν στο πρόβλημα της επιλογής χαρακτηριστικών για ταξινόμηση:

1. η εύρεση του υποσυνόλου που όταν τροφοδοτήσει τον ταξινομητή θα δώσει το πιο χαμηλό ποσοστό σφάλματος ταξινόμησης και
2. η αναζήτηση του μικρότερου υποσύνολου για το οποίο το ποσοστό σφάλματος της ταξινόμησης (ή κάποιο άλλο μέτρο απόδοσης) είναι κάτω από μια συγκεκριμένη τιμή κατωφλίου.

Υπό αυτό το πλαίσιο, αντιμετωπίζουν και αυτοί το πρόβλημα επιλογής χαρακτηριστικών ως πρόβλημα βελτιστοποίησης και υποθέτουν ότι αναζητούν το μικρότερο υποσύνολο χαρακτηριστικών για το οποίο η απόδοση της ταξινόμησης δεν κατεβαίνει κάτω από ένα συγκεκριμένο κατώφλι. Ως μέτρο απόδοσης χρησιμοποιούν το σφάλμα του ταξινομητή (η απόδοση είναι αντιστρόφως ανάλογη του σφάλματος άρα αναζητούν τα υποσύνολα που δεν ξεπερνούν ένα συγκεκριμένο κατώφλι). Αρχικά εντοπίζουν όλα τα υποσύνολα τα οποία δίνουν σφάλμα ταξινόμησης κάτω από ένα συγκεκριμένο κατώφλι και ακολούθως από αυτά τα υποσύνολα επιλέγουν το μικρότερο σε μέγεθος.

Η συνάρτηση καταλληλότητας του ΓΑ έχει σχεδιαστεί έτσι ώστε να αναζητεί το υποσύνολο με την μικρότερη βαθμολογία. Η βαθμολογία που αποδίδεται σε κάθε χρωμόσωμα υπολογίζεται ως ο γραμμικός συνδυασμός μιας συνάρτησης «ποινής» (penalty function) και του αριθμού των χαρακτηριστικών που απαρτίζουν το υπό αξιολόγηση χρωμόσωμα (κόστος συμπερίληψης κάθε χαρακτηριστικού που ανήκει στο υποσύνολο ίσο με 1). Οι μεταβλητές που χρειάζονται για τον καθορισμό της συνάρτησης «ποινής» είναι το σφάλμα ταξινόμησης, το κατώφλι μέγιστου σφάλματος ταξινόμησης και το περιθώριο ανοχής. Εάν το σφάλμα ταξινόμησης είναι μικρότερο από το κατώφλι, τότε η συνάρτηση «ποινής» επιστρέφει αρνητική τιμή (όσο το σφάλμα ταξινόμησης πλησιάζει το μηδέν τόσο πιο μικρή τιμή επιστρέφει η συνάρτηση «ποινής»). Για μεγάλες τιμές του σφάλματος ταξινόμησης η συνάρτηση «ποινής» τείνει προς το άπειρο. Με αυτό τον τρόπο τα υποσύνολα με σφάλμα ταξινόμησης κάτω από το κατώφλι λαμβάνουν μικρή «αμοιβή», δηλαδή αρνητική «ποινή», τα υποσύνολα που έχουν σφάλμα ταξινόμησης πάνω από το κατώφλι, αλλά εντός ενός περιθώριου ανοχής, λαμβάνουν μικρή θετική «ποινή» (από 0 έως 1) - δίνοντάς τους τη δυνατότητα να είναι το ίδιο ή και περισσότερο κατάλληλα από υποσύνολα που έχουν μέγεθος κατά 1 μεγαλύτερο - και υποσύνολα των οποίων το σφάλμα ταξινόμησης ξεπερνά και το περιθώριο ανοχής, λαμβάνουν σχετικά ψηλή θετική «ποινή» (πάνω από 1) και δεν μπορούν να ανταγωνιστούν με υποσύνολα που έχουν μέγεθος κατά ένα μεγαλύτερο. Υποσύνολα

με τον ίδιο αριθμό χαρακτηριστικών αξιολογούνται μόνο βάσει του σφάλματος ταξινόμησης: αυτά με πιο χαμηλό σφάλμα ταξινόμησης θεωρούνται πιο κατάλληλα και δίνουν χαμηλότερη βαθμολογία.

Η αναπαράσταση χρωμοσώματος που χρησιμοποιούν είναι η κλασική δυαδική αναπαράσταση: κάθε χρωμόσωμα αποτελείται από d bits (όπου d ο αρχικός αριθμός χαρακτηριστικών) και τιμή 1 σε ένα bit υποδεικνύει επιλογή του αντίστοιχου χαρακτηριστικού, ενώ τιμή 0 αποκλεισμό.

Η μελέτη αυτή αποσκοπούσε στην αξιολόγηση των ΓΑ ως εργαλείο για την επιλογή χαρακτηριστικών σε μεγάλης κλίμακας προβλήματα ταξινόμησης, συγκριτικά με τις μεθόδους εξαντλητικής αναζήτησης, ακολουθιακής αναζήτησης (sequential search) και αναζήτησης διακλάδωσης και περιορισμού (branch and bound search). Η δοκιμή των ΓΑ έγινε σε προσομοιωμένα σύνολα δεδομένων με μοντελοποιημένες συναρτήσεις σφάλματος αλλά και σε πραγματικά σύνολα δεδομένων με πραγματικές συναρτήσεις σφάλματος ταξινόμησης. Το σύνολο πραγματικών δεδομένων περιείχε χαρακτηριστικά που προέκυψαν από ψηφιοποιημένες υπέρυθρες εικόνες πραγματικών χώρων και η ταξινόμηση σε δύο κλάσεις έγινε με τη χρήση 5-NN ταξινομητή (πλησιέστερου γείτονα).

Έδειξαν πειραματικά, ότι η επιλογή χαρακτηριστικών με ΓΑ ξεπερνά σε απόδοση τις άλλες μεθόδους επιλογής που δοκίμασαν (sequential search, branch and bound). Η χρήση των ΓΑ δίνει καλύτερα αποτελέσματα και ταυτόχρονα μειώνει το υπολογιστικό κόστος. Επιπλέον σημειώνουν ότι αφού η αναζήτηση με το ΓΑ γίνεται παράλληλα σε πολλές υποψήφιες λύσεις ενός πληθυσμού, για το ΓΑ είναι πιο πιθανή η εύρεση της βέλτιστης λύσης, από ότι για κάθε άλλη μέθοδο που αξιολογεί και ακολούθως τροποποιεί μια μόνο λύση κάθε φορά.

3.3.6 Σχετική έρευνα 6

Στο [62] οι Zhuo, Zheng, Wang, Li, Ai και Qian παρουσιάζουν μια μέθοδο περιτυλίγματος για υπερφασματικά δεδομένα, που ενσωματώνει ένα ΓΑ και έναν ταξινομητή SVM. Ο στόχος τους είναι διπλός, να βελτιστοποιήσουν το υποσύνολο χαρακτηριστικών, καθώς και τις παραμέτρους του πυρήνα (kernel) του SVM, με απώτερο σκοπό την επίτευξη υψηλότερης ακρίβειας ταξινόμησης. Για την υλοποίηση της εφαρμογής GA-SVM, σχεδιάζουν ένα χρωμόσωμα που αποτελείται από το υποσύνολο χαρακτηριστικών και από τις παραμέτρους του πυρήνα, καθώς και μια συνάρτηση καταλληλότητας που συνδυάζει δύο κριτήρια: υψηλή ακρίβεια ταξινόμησης και μικρό υπολογιστικό κόστος.

Το χρωμόσωμα αποτελείται από 3 μέρη: τις παράμετρους του πυρήνα C και γ και τη μάσκα για τα χαρακτηριστικά. Για την αναπαράστασή τους χρησιμοποιείται δυαδική κωδικοποίηση. Στο

πρώτο και δεύτερο μέρος του χρωμοσώματος, για τη μετατροπή των δυαδικών συμβολοσειρών, που αναπαριστούν τον γενότυπο των παραμέτρων C και γ, σε φαινότυπο χρησιμοποιείται μια συγκεκριμένη εξίσωση. Στο τρίτο μέρος του χρωμοσώματος, όπου αναπαριστάται η μάσκα χαρακτηριστικών, αν το bit έχει τιμή 1 τότε το αντίστοιχο χαρακτηριστικό συμπεριλαμβάνεται στο υποσύνολο, ενώ αν έχει τιμή 0 δεν συμπεριλαμβάνεται. Επειδή ο αριθμός των χαρακτηριστικών για υπερφασματικά δεδομένα είναι πάρα πολύ μεγάλος, η κλασική n-bit δυαδική κωδικοποίηση για τη μάσκα των χαρακτηριστικών (όπου n ο αριθμός όλων των αρχικών χαρακτηριστικών) θα παρήγαγε ένα τεράστιο χώρο λύσεων που θα καθιστούσε ανέφικτη την εύρεση του βέλτιστου υποσυνόλου χαρακτηριστικών σε επιτρεπόμενα πλαίσια χρόνου. Επομένως, υιοθετείται μια νέα εκδοχή δυαδικής αναπαράστασης, η οποία θέτει το μήκος του τρίτου μέρους του χρωμοσώματος διαφορετικό (μικρότερο για την ακρίβεια) από τον αριθμό των αρχικών χαρακηριστικών. Εάν το μήκος του τρίτου μέρους χρωμοσώματος είναι ίσο με nf, τότε nf χρωμοσώματα επιλέγονται τυχαία από όλα τα n χαρακτηριστικά και ταξινομούνται βάσει του αριθμού αναγνώρισής τους. Ακολούθως παράγεται μια συμβολοσειρά από nf-bits, η οποία χρησιμοποιείται ως μάσκα για τα επιλεγμένα nf χαρακτηριστικά. Με αυτό τον τρόπο μειώνεται ο αριθμός των πιθανών λύσεων και η υπολογιστική αποδοτικότητα αυξάνεται, ειδικά στην περίπτωση που το μήκος των χρωμοσωμάτων nf είναι κατά πολύ μικρότερο από τον αριθμό των αρχικών χρωμοσωμάτων n.

Για την συνάρτηση καταλληλότητας χρησιμοποιούνται δύο κριτήρια: η ακρίβεια ταξινόμησης και ο αριθμός των επιλεγμένων χαρακτηριστικών για το υποσυνόλο. Ένα χρωμόσωμα για να έχει υψηλή τιμή καταλληλότητας πρέπει να έχει ψηλή ακρίβεια ταξινόμησης και μικρό αριθμό χαρακτηριστικών.

Τα πειραματικά αποτελέσματα έδειξαν ότι η GA-SVM μέθοδος μείωνε σημαντικά το υπολογιστικό κόστος, ενώ παράλληλα βελτίωνε την ακρίβεια ταξινόμησης εντοπίζοντας το βέλτιστο υποσύνολο χαρακτηριστικών και βελτιστοποιώντας τις παραμέτρους του πυρήνα του SVM ταυτόχρονα.

3.3.7 Άλλες σχετικές έρευνες

Άλλες εφαρμογές, μεταξύ πολλών, όπου γίνεται χρήση ΓΑ στο πρόβλημα της επιλογής χαρακτηριστικών είναι οι [63], [64], [65], [66], [67] [68].

Κεφάλαιο 4 –Καρκίνος του τραχήλου της μήτρας και πληθυσμιακός έλεγχος

4.1 Ο καρκίνος του τραχήλου της μήτρας

Καρκίνος του τραχήλου της μήτρας (Cervical Cancer – CC, Cancer of the Cervix - CxCa) ονομάζεται ο παθολογικός και ανεξέλεγκτος πολλαπλασιασμός των κυττάρων στους ιστούς που επενδύουν τον τράχηλο της μήτρας. Τα κακοήθη καρκινικά κύτταρα αναπτύσσονται, συνήθως, έπειτα από μακρό χρονικό διάστημα εξέλιξης ιστολογικών μεταβολών στα κύτταρα του τραχήλου, που είναι γνωστές ως δυσπλασία. Κατά κανόνα χρειάζεται να περάσουν πολλά χρόνια για να μετατραπούν τα αλλοιωμένα κύτταρα των προκαρκινικών καταστάσεων σε καρκίνο. Προοδευτικά, τα καρκινικά κύτταρα αρχίζουν να αυξάνονται ταχύτατα και να εξαπλώνονται σε μεγαλύτερο βάθος διήθησης μέσα στον τράχηλο και στις περιβάλουσες ανατομικές δομές [69]. Δεν πεθαίνουν στο συνήθη χρόνο που αποπίπτουν τα φυσιολογικά κύτταρα και ως αποτέλεσμα συσσωρεύονται σε ένα σημείο, δημιουργούν ένα όγκο και τέλος εξαπλώνονται στα γειτονικά όργανα (π.χ. στη μήτρα) ή σε απομακρυσμένα σημεία του σώματος (μετάσταση). Ας σημειωθεί ότι, συνήθως, όταν αναφερόμαστε στον όρο καρκίνο εννοούμε το διηθητικό καρκίνωμα (Invasive Cervical Cancer - ICC) που διαδίδεται και στους γειτονικούς ιστούς και όχι τις μη διηθητικές προκαρκινικές καταστάσεις.

Ο καρκίνος του τραχήλου της μήτρας είναι ένα από τα πιο σημαντικά προβλήματα υγείας που μαστίζουν τις γυναίκες. Εκτιμάται ότι σε παγκόσμιο επίπεδο προσβάλλει 500000 γυναίκες ετησίως, εκ των οποίων το 80% είναι κάτοικοι αναπτυσσόμενων χωρών [70]. Το ποσοστό θνησιμότητας από τη νόσο προσεγγίζει το 50% (περίπου 240000 γυναίκες ετησίως).

4.1.1 Επιδημιολογία - Στατιστικά στοιχεία

Ο καρκίνος του τραχήλου της μήτρας, σύμφωνα με την έκθεση του ΠΟΥ (Παγκόσμιος οργανισμός υγείας, World Health Organization-WHO) το 2008, ήταν η δεύτερη πιο συχνή μορφή καρκίνου (μετά τον καρκίνο του μαστού) ανάμεσα στις γυναίκες παγκοσμίως με 493000 περιστατικά και 273000 θανάτους ετησίως (εκ των οποίων στις αναπτυσσόμενες χώρες: 409000 περιστατικά, 234000 θάνατοι) [71]. Η εμφάνιση της νόσου διαφέρει ανάμεσα στις αναπτυγμένες και αναπτυσσόμενες χώρες με τις πλείστες περιπτώσεις να συμβαίνουν στις αναπτυσσόμενες χώρες. Στις αναπτυσσόμενες χώρες αποτελεί την υπ' αριθμόν ένα αιτία νόσου και θανάτου από

καρκίνο των γυναικών και ανιχνεύεται μέσω των συμπτωμάτων σε προχωρημένα στάδια διηθητικού καρκίνου [72].

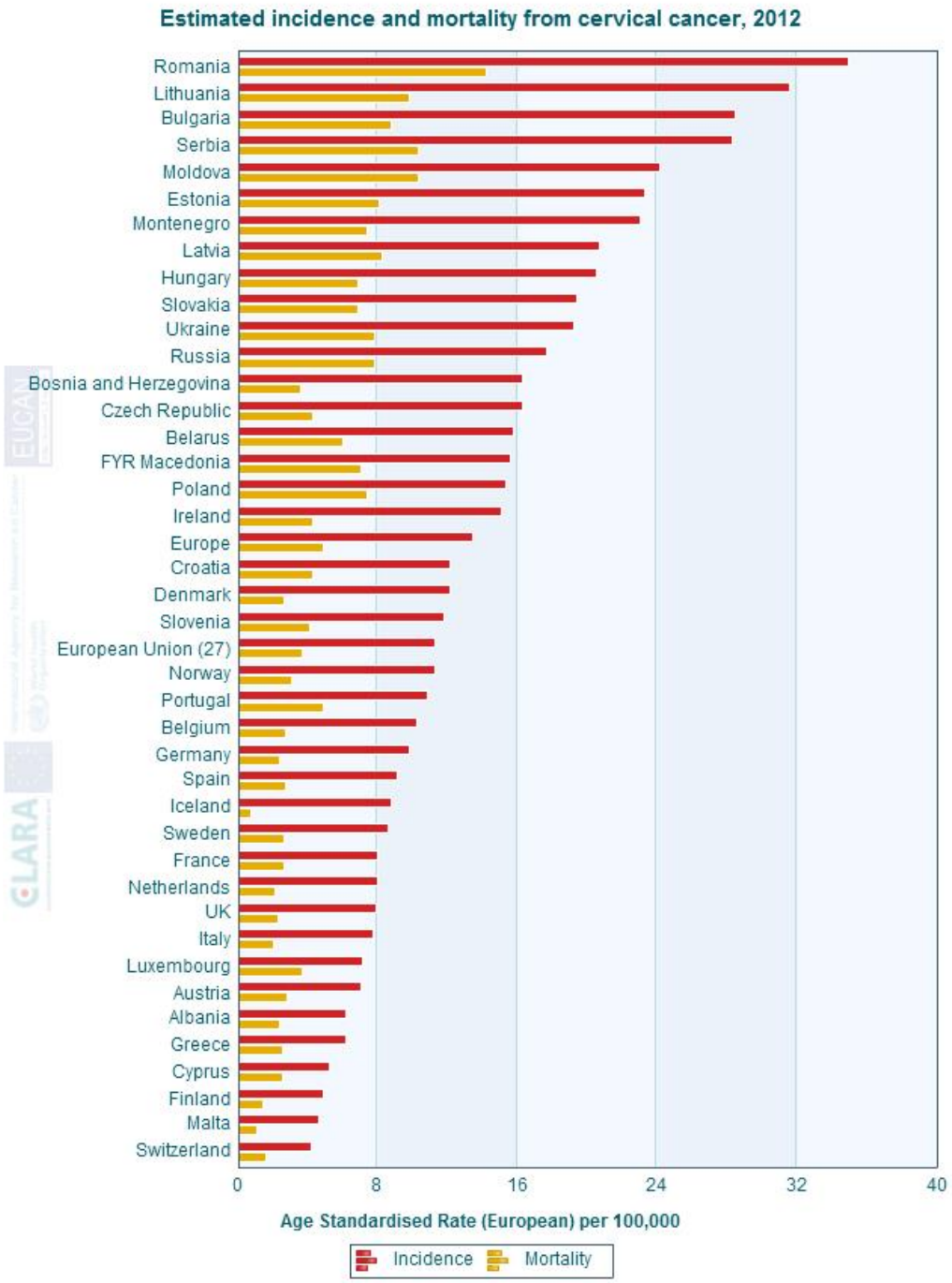
Σύμφωνα με τα στοιχεία του ΠΟΥ στην Ευρώπη κατατάσσεται ως ο δεύτερος σε συχνότητα καρκίνος μεταξύ των γυναικών ηλικίας 15-45 ετών με τις νέες περιπτώσεις να ανέρχονται στις 54323 κάθε χρόνο και τον αριθμό των θανάτων να φθάνει τις 25102 ετήσια [72].

Σύμφωνα με το CDC (Centers for Disease Control and Prevention), το έτος 2012 διαγνώστηκαν 12042 γυναίκες στις Ηνωμένες Πολιτείες Αμερικής με καρκίνο του τραχήλου της μήτρας, ενώ ο αριθμός θανάτων εξαιτίας του καρκίνου του τραχήλου της μήτρας ανέρχεται στις 4074 γυναίκες [73]. Στις Ηνωμένες Πολιτείες Αμερικής ο καρκίνος του τραχήλου της μήτρας παραμένει ο 6^{ος} πιο συχνός διαγνωσθείς τύπος καρκίνου [74].

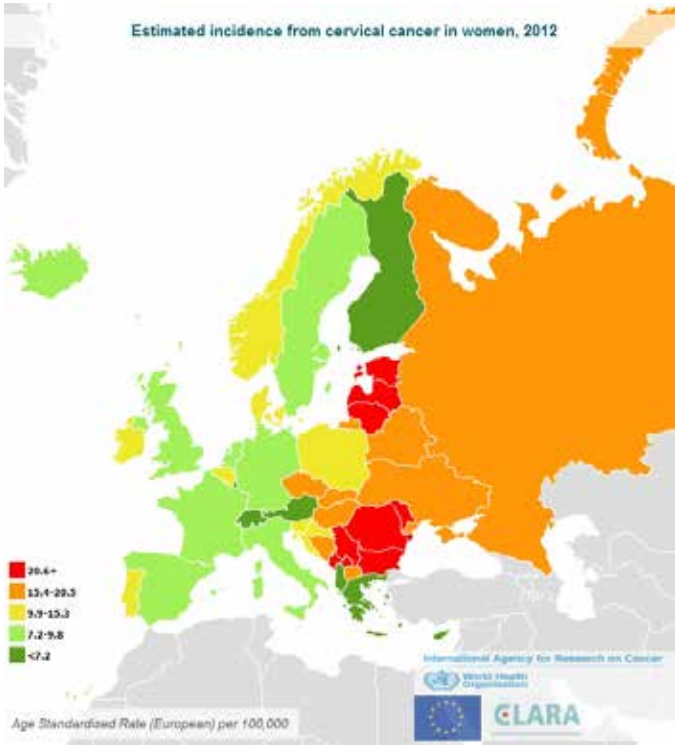
Οι εκτιμήσεις στην Ελλάδα δείχνουν ότι ετήσια διαγιγνώσκονται περίπου 600 γυναίκες με καρκίνο του τραχήλου της μήτρας και 250 πεθαίνουν από τη νόσο επίσης ετήσια [75].

Σύμφωνα με το Αρχείο Καρκίνου που τηρεί η Μονάδα Παρακολούθησης Υγείας, στην Κύπρο κατά τα έτη 1998 έως 2010 έχουν εντοπιστεί και καταχωρηθεί 315 κακοήθη περιστατικά καρκίνου του τραχήλου της μήτρας και 83 περιστατικά ενδοεπιθηλιακού καρκίνου (σύνολο 398), δηλαδή περίπου 28 νέα περιστατικά καρκίνου του τραχήλου της μήτρας εντοπίζονται και καταχωρούνται κάθε έτος στην Κύπρο [76]. Από το 2004 έως και το 2012 έχουν καταγραφεί 68 θάνατοι λόγω καρκίνου του τραχήλου της μήτρας. Μόνο το 2012 παρουσιάστηκαν 31 περιστατικά και η θνησιμότητα αυξήθηκε σε 17 θανάτους [77].

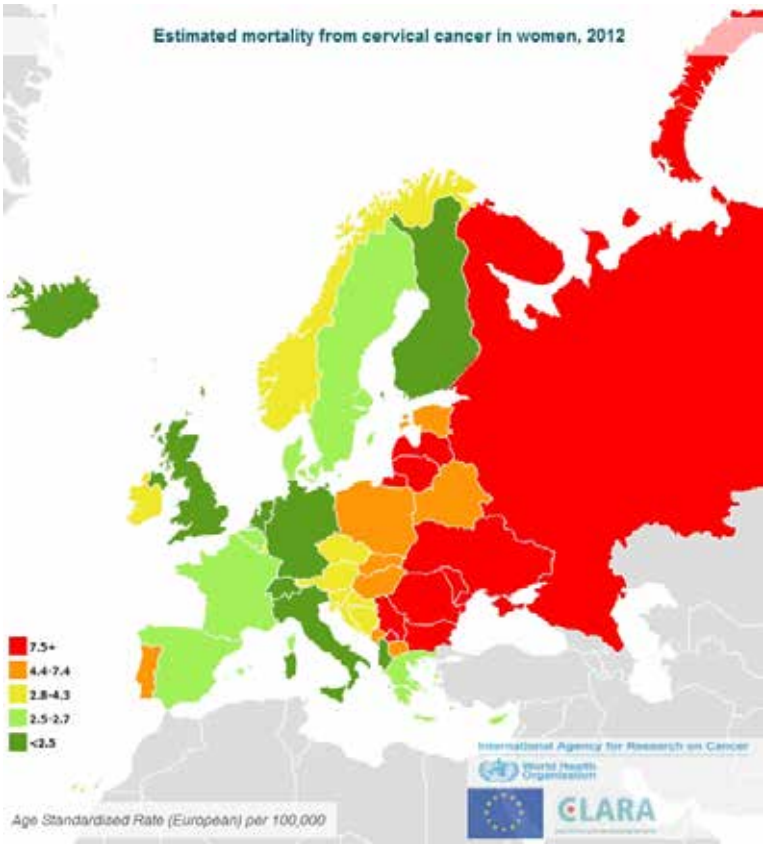
Στην *εικόνα 77* φαίνονται τα στατιστικά στοιχεία για το έτος 2012 αναφορικά με τα ποσοστά (ανά 100000) περιστατικών εμφάνισης καρκίνου του τραχήλου της μήτρας και τα ποσοστά (ανά 100000) θανάτων εξαιτίας του καρκίνου του τραχήλου της μήτρας. Ιδιαίτερο ενδιαφέρον παρουσιάζει το γεγονός ότι τα ποσοστά (ανά 100000) των καταγεγραμμένων περιστατικών σε Κύπρο και Ελλάδα σε σύγκριση με αυτά των υπόλοιπων χωρών είναι αρκετά χαμηλά. Ωστόσο, στην Κύπρο και στην Ελλάδα σε αντίθεση με άλλες χώρες, ο αριθμός των θανάτων σε σχέση με τα περιστατικά εμφάνισης καρκίνου του τραχήλου της μήτρας, είναι μεγάλος (Ελλάδα: περιστατικά 6.2, θνησιμότητα: 2.5, Κύπρος: 5.2, 2.5 [77]).



Εικόνα 77: *Σταθμισμένα με την ηλικία εκτιμώμενα περιστατικά εμφάνισης και θνησιμότητας του καρκίνου του τραχήλου της μήτρας το 2012 σύμφωνα με IARC: EUCAN database [77]*



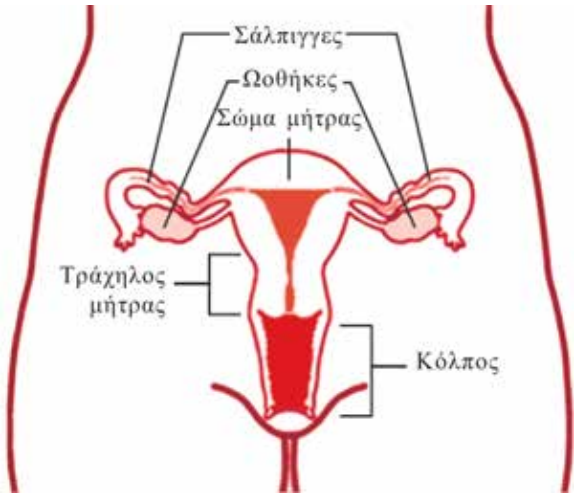
Εικόνα 78: *Σταθμισμένα με την ηλικία εκτιμώμενα περιστατικά εμφάνισης του καρκίνου του τραχήλου της μήτρας το 2012 σύμφωνα με IARC: EUCAN database [77]*



Εικόνα 79: *Σταθμισμένη με την ηλικία εκτιμώμενη θνησιμότητα από καρκίνο του τραχήλου της μήτρας το 2012 σύμφωνα με IARC: EUCAN database [77]*

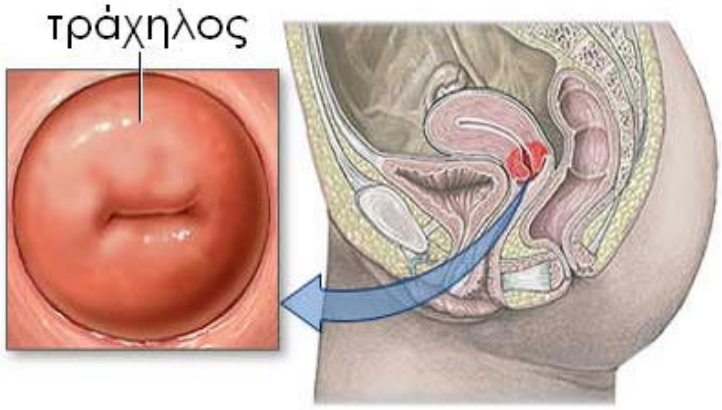
4.1.2 Τράχηλος της μήτρας

Η μήτρα βρίσκεται στο κάτω μέρος της κοιλιακής χώρας της γυναίκας και είναι ένα κοίλο, μυώδες όργανο, απιοειδούς σχήματος (*εικόνα 80*). Το ανώτερο τμήμα της μήτρας λέγεται σώμα της μήτρας και φιλοξενεί το έμβρυο κατά τη διάρκεια της εγκυμοσύνης. Τράχηλος είναι το χαμηλότερο στενό τμήμα της μήτρας το οποίο προβάλλει στον κόλπο και συνδέει τη μήτρα με τον κόλπο. Οδηγεί από την έξοδο της μήτρας στην είσοδο του κόλπου (γεννητική οδός). Ο τράχηλος είναι ένας ινομυώδης σωλήνας (αποτελείται από ινώδη συνδετικό ιστό και λίγες λείες μυϊκές ίνες), που κατά τον τοκετό διαστέλλεται για να επιτραπεί η έξοδος του εμβρύου. Κατά τη γονιμοποίηση, μέσω του τραχήλου, εισέρχονται τα σπερματοζωάρια από τον κόλπο στη κοιλότητα της μήτρας και ακολούθως στις σάλπιγγες.



Εικόνα 80: *Έσω γεννητικά όργανα της γυναίκας [78]*

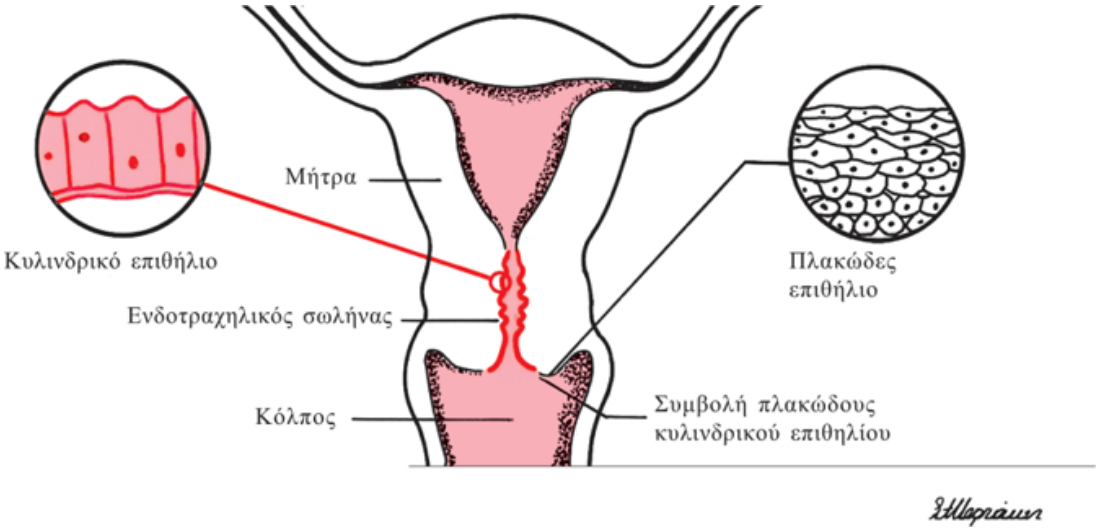
Το τμήμα του τραχήλου της μήτρας που βρίσκεται στον κόλπο ονομάζεται εξωτράχηλος. Έχει κυρτή, στρογγυλή επιφάνεια με μία οπή, το έξω στόμιο του τραχήλου που οδηγεί στον ενδοτραχηλικό αυλό (*εικόνα 81*). Ο ενδοτραχηλικός αυλός έχει μήκος 2-3 cm και συνέχεται με την ενδομητρική κοιλότητα στο ύψος του έσω τραχηλικού στομίου.



Εικόνα 81: *Τράχηλος της μήτρας (κατά την περίοδο κύησης) [79]*

Τα εσωτερικά όργανα καλύπτονται από βλεννογόνους. Η στιβάδα των κυττάρων που σχηματίζουν τους βλεννογόνους, καθώς και την επιδερμίδα του δέρματος, ονομάζεται επιθήλιο και αποτελείται είτε από μια στιβάδα κυττάρων (πλακώδη, κυβοειδή ή κυλινδρικά κύτταρα) είτε από περισσότερες. Στους βλεννογόνους, όπως και στο δέρμα, εντοπίζονται δύο τύποι κυττάρων: τα πλακώδη (επίπεδα) κύτταρα και τα αδενικά κύτταρα. Τα πλακώδη κύτταρα βρίσκονται σε στιβάδες και έχουν προστατευτικό ρόλο, ενώ τα αδενικά κύτταρα είναι αυτά που απαρτίζουν τους αδένες και παράγουν διάφορες εκκρίσεις (στους βλεννογόνους παράγουν βλέννη). Έτσι, τα πλακώδη κύτταρα σχηματίζουν το πλακώδες επιθήλιο και τα αδενικά κύτταρα το αδενικό επιθήλιο.

Ο βλεννογόνος του τραχήλου καλύπτεται από δύο είδη επιθηλίων: το τμήμα του τραχήλου που προβάλλει στον κόλπο καλύπτεται περιφερικά από πολύστοιβο πλακώδες επιθήλιο (εξωτράχηλος), ενώ ο ενδοτραχηλικός σωλήνας και η περιοχή γύρω από το έξω τραχηλικό στόμιο καλύπτεται από αδενικό βλεννοεκκριτικό κυλινδρικό επιθήλιο (*εικόνα 82*). Το αδενικό επιθήλιο παράγει βλέννη μέσα στην οποία κολυμπούν τα σπερματοζωάρια, ανεβαίνοντας προς τα πάνω για να βρουν το ωάριο [78].



Εικόνα 82: *Εγκάρσια διατομή της μήτρας: 2 είδη επιθηλίων που απαρτίζουν τον τραχηλικό βλεννογόνο: κυλινδρικό και πλακώδες [78]*

Υπάρχουν δύο βασικοί τύποι καρκίνων τραχήλου μήτρας ανάλογα με το είδος των κυττάρων από τα οποία προέρχονται. Οι συχνότερες μορφές είναι ο πλακώδης επιθηλιακός καρκίνος (Squamous Cell Cancer - SCC, 80-90% των περιπτώσεων) που αναπτύσσεται στο έξω μέρος του τραχήλου (πλακώδη κύτταρα) και το αδενοκαρκίνωμα (Adenocarcinoma – ADC ή Adeno-Ca, 10-20% των περιπτώσεων) που εμφανίζεται στον ενδοτράχηλο (αδενικά κύτταρα) [80]. Περιστασιακά παρατηρείται μικτό καρκίνωμα που παρουσιάζει χαρακτηριστικά και από πλακώδες καρκίνωμα και από αδενοκαρκίνωμα. Σπάνια, μπορεί να εμφανιστούν και άλλοι τύποι καρκίνου στον τράχηλο της μήτρας (π.χ. λέμφωμα, μελάνωμα, σάρκωμα).

4.1.3 Ταξινόμηση προκαρκινικών αλλοιώσεων του τραχήλου της μήτρας

Ως αποτέλεσμα των κλινικών μελετών σχετικά με τον καρκίνο του τραχήλου της μήτρας και τη θεραπεία του, η ταξινόμηση των προκαρκινικών αλλοιώσεών του έχει αλλάξει αρκετές φορές κατά τον 20^{ο} αιώνα. Σήμερα, συνυπάρχουν τρία παράλληλα συστήματα ταξινόμησης των καρκινικών αλλοιώσεων: το σύστημα που προωθεί ο Παγκόσμιος Οργανισμός Υγείας, το σύστημα Richard και η ταξινόμηση Bethesda.

Το σύστημα που προωθείται από τον Παγκόσμιο Οργανισμό Υγείας και τον Παναμερικανικό Οργανισμό Υγείας (Pan-American Health Organization - PAHO) είναι περιγραφικό της δυσπλασίας/αλλοίωσης. Ο όρος δυσπλασία περιγράφει τις αλλαγές που υφίστανται τα κύτταρα του τραχήλου στη μορφολογία τους. Οι αλλαγές αυτές αναγνωρίζονται με το μικροσκόπιο. Η δυσπλασία εκφράζει την ιστολογική ορολογία της ενδοεπιθηλιακής νεοπλασίας και αν και δεν

είναι καρκίνος, χωρίς θεραπεία, μπορεί να εξελιχθεί σε αρχόμενες μορφές καρκίνου. Ανάλογα με το πάχος του τραχηλικού επιθηλίου που καταλαμβάνεται από νεοπλασματικά κύτταρα και το πόσο αλλοιωμένα εμφανίζονται τα κύτταρα στο μικροσκόπιο, γίνεται διάκριση ανάμεσα σε:

- ήπια (mild) δυσπλασία (παθολογικός πολλαπλασιασμός κυττάρων)
- μέτρια (moderate) δυσπλασία
- σοβαρή (severe) δυσπλασία και
- καρκίνωμα in situ – CIS /Carcinoma In Situ (όλο το πάχος του επιθηλίου έχει αντικατασταθεί από αδιαφοροποίητα νεοπλασματικά κύτταρα) [81]

Ο όρος ενδοεπιθηλιακό καρκίνωμα ή καρκίνωμα in situ αναφέρεται στο προδιηθητικό καρκίνωμα του τραχήλου όπου όλο το πάχος του επιθηλίου έχει καταληφθεί από αλλοιωμένα κύτταρα, τα οποία όμως, δεν έχουν διασπάσει τη βασική μεμβράνη και δεν έχουν εξαπλωθεί σε βαθύτερους ιστούς. Καρκίνωμα του τραχήλου ή διηθητικό καρκίνωμα συμβαίνει όταν αλλοιωμένα κύτταρα έχουν διασπάσει τη βασική μεμβράνη του τραχηλικού επιθηλίου και έχουν διηθήσει τους βαθύτερους ιστούς του τραχήλου ή έχουν μεταφερθεί με τη λεμφική ή αιματική οδό σε άλλους ιστούς και όργανα.

Το σύστημα Richart, αναπτύχθηκε μετά το προαναφερθέν σύστημα και χρησιμοποιείται κυρίως στην κολποσκοπία. Διακρίνει σε τρία στάδια ενδοεπιθηλιακής νεοπλασίας του τραχήλου της μήτρας (ενδοεπιθηλιακή νεοπλασία του τραχήλου της μήτρας είναι ένα άλλο όνομα για τις αλλοιώσεις του τραχήλου της μήτρας ως σύνολο), ανάλογα με τη βαρύτητα και την έκταση των δυσπλαστικών αλλοιώσεων:

- CIN-1
- CIN-2
- CIN-3 [81]

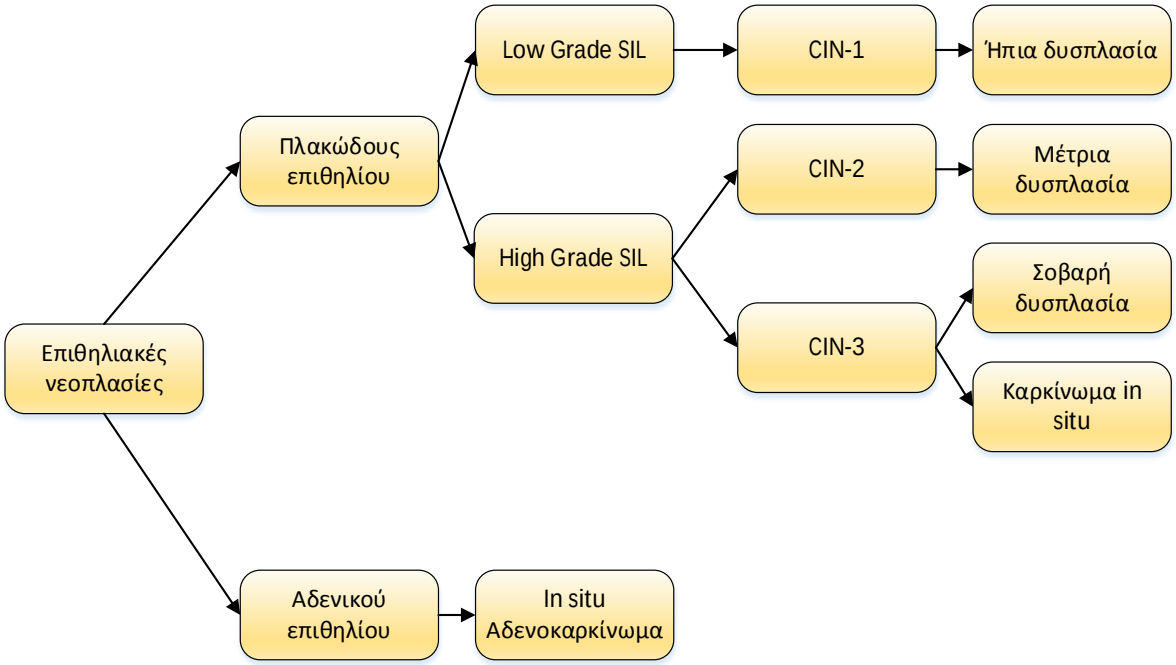
Ο όρος τραχηλική ενδοεπιθηλιακή νεοπλασία (CIN- Cervical Intraepithelial Neoplasia) έχει ως στόχο να δώσει έμφαση στο φάσμα των ανωμαλιών και των αλλοιώσεων και να βοηθήσει στην τυποποίηση της θεραπείας [82]. Ο όρος αυτός περιγράφει ολόκληρο το φάσμα των αλλοιώσεων που υφίστανται τα κύτταρα του τραχηλικού πλακώδους επιθηλίου και περιλαμβάνει όλες τις διαταραχές διαφοροποιήσεως του πλακώδους επιθηλίου που δεν εκπληρώνουν τις προϋποθέσεις του in situ καρκινώματος. Στο καρκίνωμα in situ, τα κακοήθη κύτταρα δεν εμφανίζουν την παραμικρή διαφοροποίηση σε ολόκληρο το πάχος του επιθηλίου, όμως, τουλάχιστον, δεν έχουν εισβάλει σε βαθύτερους ιστούς.

Τέλος, η ταξινόμηση Bethesda (TBS, The Bethesda System) χρησιμοποιείται κυρίως από κυτταροπαθολόγους και κυτταρολόγους. Έχει προταθεί το 1988 και έχει αναθεωρηθεί το 1991 και το 2001. Διακρίνει σε:

- LSIL (Low grade Squamous Intraepithelial Lesion, Χαμηλού βαθμού πλακώδης ενδοεπιθηλιακή αλλοίωση)
- HSIL (High grade Squamous Intraepithelial Lesion, Υψηλού βαθμού πλακώδης ενδοεπιθηλιακή αλλοίωση) [81]

Η πλακώδης ενδοεπιθηλιακή αλλοίωση (SIL) είναι ένας άλλος όρος για την περιγραφή των ανώμαλων αλλοιώσεων των κυττάρων στην επιφάνεια του επιθηλίου του τραχήλου. Ενδοεπιθηλιακή αλλοίωση σημαίνει ύπαρξη αλλοιωμένων κύτταρων στις επιφανειακές στοιβάδες του επιθηλίου του τραχήλου.

Μεταξύ των 3 ταξινομήσεων μπορεί να γίνει μια αντιστοίχηση: η ήπια δυσπλασία αντιστοιχεί στα CIN-1 και LSIL, η μέτρια δυσπλασία στο CIN-2 και σε ένα τμήμα του φάσματος του HSIL και η σοβαρή δυσπλασία μαζί με το καρκίνωμα in situ αντιστοιχούν στο CIN-3 και στο υπόλοιπο τμήμα του φάσματος του HSIL. Η ταξινόμηση των προκαρκινικών αλλοιώσεων του τραχήλου της μήτρας μπορεί να συνοψιστεί στο ακόλουθο διάγραμμα (*εικόνα 83*).



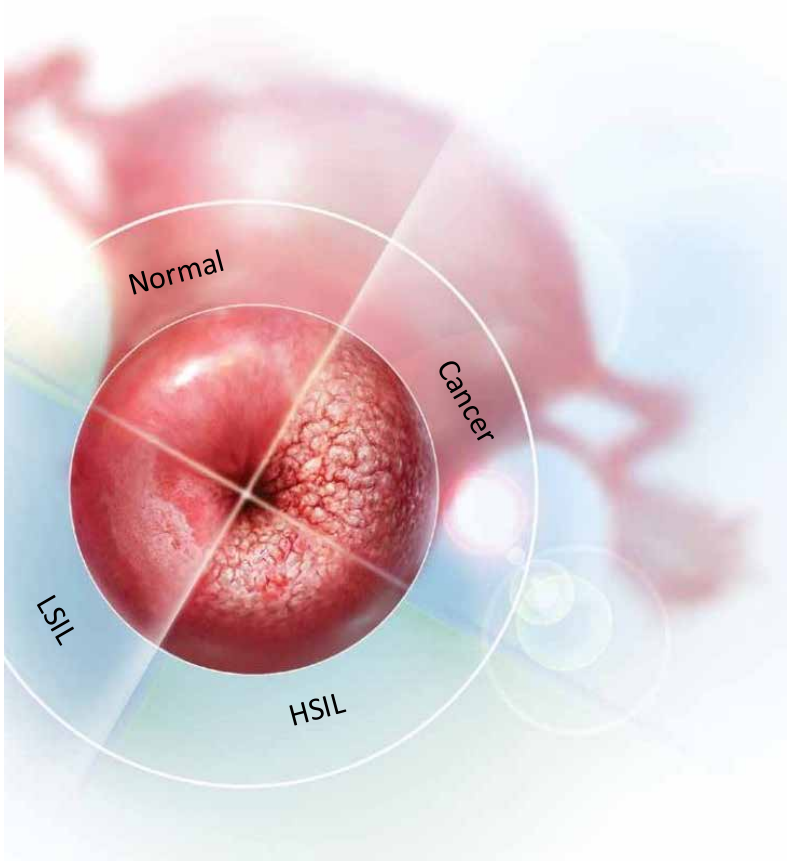
Εικόνα 83: *Ταξινόμηση προκαρκινικών αλλοιώσεων του τραχήλου της μήτρας*

Η αντιστοίχιση μεταξύ των τριών συστημάτων δεν είναι ιδανική και όλες οι ταξινομήσεις περιέχουν κάποιου είδους ασάφεια, με αποτέλεσμα, η ταξινόμηση των ακραίων περιπτώσεων σε κάθε σύστημα να εξαρτάται από τις προτιμήσεις του εκάστοτε ιατρού [81].

Ας σημειωθεί ότι υπάρχουν κυτταρικές αλλοιώσεις οι οποίες δεν έχουν τα μορφολογικά κριτήρια για να μπορούν να χαρακτηρισθούν CIN, SIL ή δυσπλασία και αναφέρονται ως άτυπα

πλακώδη κύτταρα απροσδιόριστης σημασίας (Atypical Cells of Undetermined Significance - ASCUS).

Στην εικόνα 84 παρουσιάζονται φυσιολογικά κύτταρα, οι κατηγορίες δυσπλασίας με βάση την ταξινόμηση Bethesda των προκαρκινικών κυττάρων, καθώς και καρκινικά κύτταρα του τραχήλου της μήτρας.



Εικόνα 84: *Φυσιολογικά, προκαρκινικά (με βάση το σύστημα Bethesda) και καρκινικά κύτταρα του τραχήλου της μήτρας (τροποποιημένη εικόνα από [83])*

4.2 Ιός των ανθρωπίνων θηλωμάτων (Human Papillomavirus) και καρκίνος του τραχήλου της μήτρας

Κατά τον Albert Singer, επίτιμο καθηγητή γυναικολογίας στο UCL, «*Ο τραχηλικός καρκίνος αποτελεί μια πολύ σπάνια συνέπεια μιας ιδιαίτερα συχνής λοίμωξης*» [84].

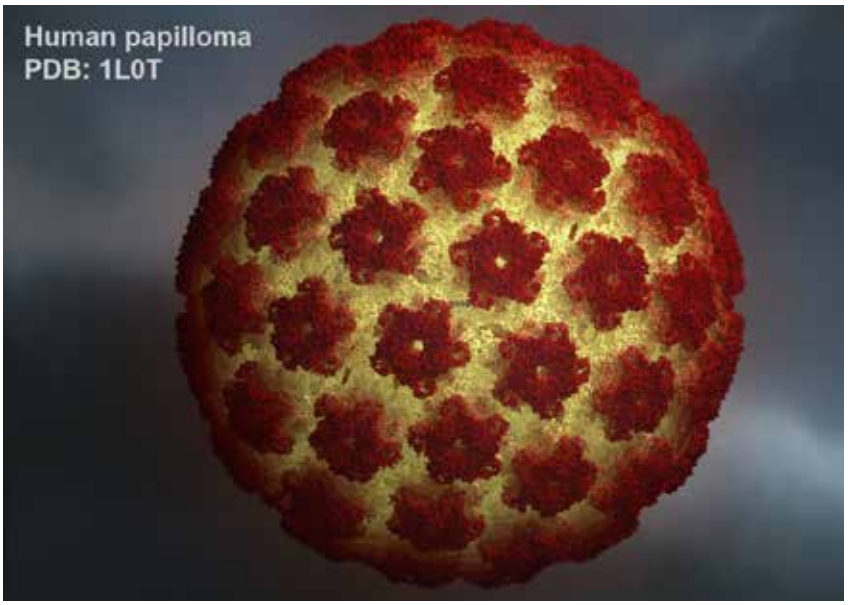
Ο ιός των ανθρωπίνων θηλωμάτων (Human Papilloma Virus - HPV) αποτελεί την πιο κοινή σεξουαλικώς μεταδιδόμενη λοίμωξη σε παγκόσμιο επίπεδο. Ο καρκίνος του τραχήλου της μήτρας θεωρείται ότι προκαλείται από μια λοίμωξη διαφόρων ογκογόνων τύπων HPV και είναι πλέον αποδεκτό ότι σε περισσότερες από 99% των περιπτώσεων, ο καρκίνος του τραχήλου της μήτρας

ξεκινά από μια HPV λοίμωξη [85]. Το ιατρικώς αποδεκτό πρότυπο, που έχει εγκριθεί από την Αμερικανική Κοινωνία Καρκίνου είναι ότι η ασθενής πρέπει να έχει μολυνθεί από τον HPV για να αναπτύξει καρκίνο του τραχήλου της μήτρας [86].

Υποψίες για συσχέτιση ανάμεσα στον HPV και τον καρκίνο του τραχήλου της μήτρας είχαν αρχίσει τη δεκαετία του 1970: ο Dr. Harald zur Hausen το 1976 είχε για πρώτη φορά εισηγηθεί τον HPV ως αρχικό αίτιο του καρκίνου του τραχήλου της μήτρας [87]. Έκτοτε έγιναν πολλές επιδημιολογικές μελέτες για την επιβεβαίωση αυτής της υπόθεσης και έχουν προβληθεί ισχυρά αποδεικτικά στοιχεία για τον αιτιολογικό ρόλο του HPV στη νόσο: η συσχέτιση δείχνει ισχυρή, συνεπής και συγκεκριμένη σε περιορισμένο αριθμό τύπων του ιού [87]. Σε σχεδόν όλες τις βιοψίες καρκίνου του τραχήλου της μήτρας έχει βρεθεί DNA από συγκεκριμένους τύπους HPV, γεγονός που οδηγεί στο συμπέρασμα ότι οι ογκογόνοι HPV τύποι που εκφράζονται σε αυτά τα κύτταρα εμπλέκονται στην αλλοίωσή τους και είναι απαραίτητοι για την εξέλιξή τους προς κακοήθεια. Έτσι, οι διάφορες επιδημιολογικές μελέτες αναδεικνύουν τον HPV ως τον βασικό αιτιολογικό παράγοντα του καρκίνου του τραχήλου της μήτρας: επίμονες HPV λοιμώξεις αποτελούν το πιο σημαντικό παράγοντα κινδύνου για εμφάνιση της νόσου [88]. Το 2008, ο Harald zur Hausen (Γερμανικό κέντρο για την έρευνα του καρκίνου, Χαϊδελβέργη) πήρε βραβείο Νόμπελ Φυσιολογίας/Ιατρικής «για την ανακάλυψή του ότι ο ιός των ανθρωπίνων θηλωμάτων προκαλεί καρκίνο του τραχήλου της μήτρας».

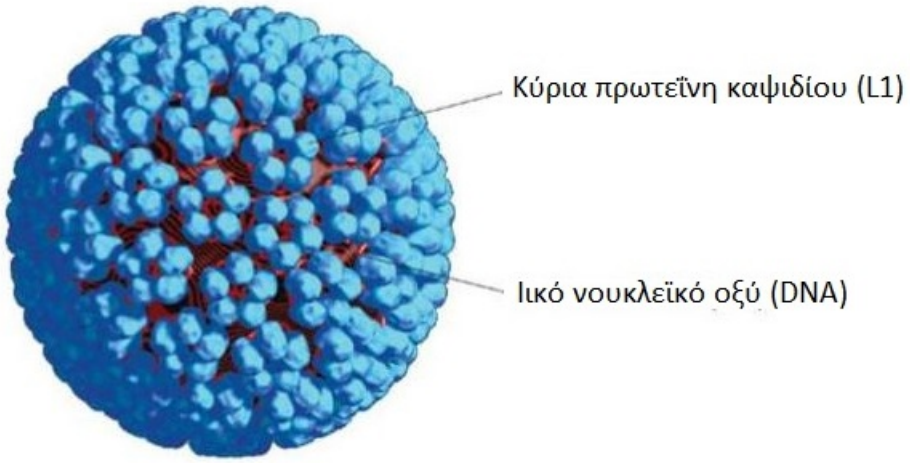
4.2.1 Δομή HPV

Ο ιός των ανθρωπίνων θηλωμάτων (HPV-Human Papillomavirus) ανήκει στην οικογένεια Papova viridae. Σε αντίθεση με τους άλλους PV (Papillomavirus) ιούς, προσβάλλει μόνο τον άνθρωπο. Συγκεκριμένα, προσβάλλει τα επιθηλιακά κύτταρα του ανθρωπίνου σώματος (π.χ. δέρμα, πρωκτογεννητική περιοχή, στοματοφαρυγγική κοιλότητα). Είναι ένας μικρός σε μέγεθος DNA ιός διαμέτρου 52-55 nm. Όπως όλοι οι ιοί, έτσι κι ο HPV αποτελείται από την πρωτεϊνική θήκη (καψίδιο) και το γονιδίωμα. Είναι ένας απλός, χωρίς φάκελο ιός και το γονιδίωμα του πυρήνα του είναι κυκλικό δεσοξυριβοζονουκλεϊκό οξύ διπλής έλικας (ds-DNA, άλλες κατηγορίες ιών έχουν DNA μονής έλικας ή RNA) το οποίο περιβάλλεται από πρωτεϊνικό καψίδιο [87]. Στην *εικόνα 85* φαίνεται το ατομικό μοντέλο του ιού των ανθρωπίνων θηλωμάτων τύπου 16.



Εικόνα 85: *Απεικόνιση ατομικού μοντέλου του σωματιδίου Human papilomavirus τύπου 16 [89][90]*

Οι ιοί είναι οντότητες των οποίων το γενετικό υλικό αποτελείται από τμήμα πυρηνικού (νουκλεϊκού) οξέος, DNA ή RNA, που αναπαράγεται μέσα σε ζώντα κύτταρα και χρησιμοποιεί τον συνθετικό μηχανισμό των κυττάρων ξενιστών προκειμένου να κατευθύνει τη σύνθεση νέων ιικών σωματιδίων τα οποία περιέχουν το ιικό γενετικό υλικό και το μεταφέρουν σε άλλα κύτταρα [91]. Οι απλοί ιοί, όπως ο HPV (*εικόνα 86*), αποτελούνται από ένα μόριο νουκλεϊκού οξέος (DNA ή RNA) που εγκλωβίζεται σε ένα πρωτεϊνικό περίβλημα (καψίδιο). Οι σύνθετοι ιοί περιβάλλονται από φάκελο (πλασματική μεμβράνη του κυττάρου-ξενιστή τροποποιημένη με πρωτεΐνες του ιού).



Εικόνα 86: *Πρωτεϊνικό περίβλημα και γονιδίωμα του σωματιδίου HPV (Τροποποιημένη εικόνα από: [92])*

4.2.2 Γονιδίωμα του HPV

Το γονιδίωμα (ή αλλιώς γένωμα) του HPV είναι κυκλικό, συγκροτείται από διπλή έλικα DNA και έχει μέγεθος 8000 ζεύγεων βάσεων. Μόνο ο ένας από τους δύο κλώνους του DNA έχει τη δυνατότητα μεταγραφής και μετάφρασης. Κάθε έλικα του HPV-DNA απαρτίζεται από 8000 νουκλεοτίδια. Ο συνδυασμός τριών νουκλεοτιδίων ονομάζεται κωδικόνιο. Κάθε σειρά κωδικονίων στο HPV-DNA που κωδικοποιούν την παραγωγή μιας πρωτεΐνης ορίζεται ως πλαίσιο ανοιχτής ανάγνωσης (Open Reading Frame - ORF). Το ιικό γονιδίωμα περιλαμβάνει 6 πλαίσια ανοικτής ανάγνωσης που κωδικοποιούν πρώιμες πρωτεΐνες που εμπλέκονται στην αντιγραφή του ιικού DNA (E1 και E2), στη ρύθμιση της έκφρασης ιικών γονιδίων (E2), τη συγκρότηση του ιού (E4) και την αθανατοποίηση και μεταλλαγή των προσβεβλημένων από τον ιό επιθηλιακών κυττάρων (E5, E6, E7 και HR-HPV), καθώς και 2 πλαίσια ανοιχτής ανάγνωσης όψιμων πρωτεϊνών (όψιμα L1 και L2) που κωδικοποιούν τις δύο πρωτεΐνες καψιδίου [87]. Το γονιδίωμα του HPV εξαρτάται από τις πρωτεΐνες του κυττάρου-ξενιστή για να ολοκληρώσει έναν βιολογικό ιικό κύκλο επειδή δεν έχει κάποιο μηχανισμό σύνθεσης πρωτεϊνών.

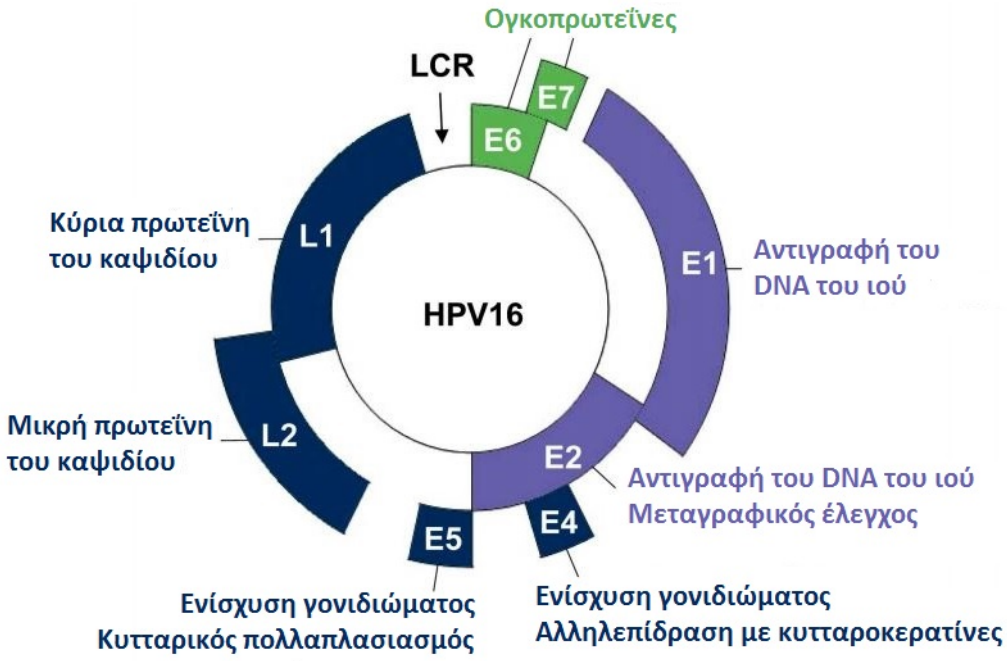
Το ιικό γονιδίωμα του HPV διαιρείται σε 3 περιοχές:

- Πρώιμη περιοχή E: περιλαμβάνει 8 γονίδια E1-E8. Κωδικοποιούν πρωτεΐνες απαραίτητες για την αντιγραφή του ιικού DNA και εκφράζονται αμέσως μετά την αρχική λοίμωξη των κυττάρων από HPV (*πίνακας 14*). Οι E1-E6 είναι υπεύθυνες για την αντιγραφή του ιού ενώ οι E7 και E8 είναι ογκοπρωτεΐνες εξαλλαγής του ιού
- Όψιμη περιοχή L: περιλαμβάνει 2 γονίδια: L1 και L2 τα οποία κωδικοποιούν δομικές πρωτεΐνες για την κατασκευή του καψιδίου και εκφράζονται στα τελικά στάδια του κύκλου του ιού
- Μη κωδική ρυθμιστική περιοχή LCR ή URR ή NCR (Long Control Region ή Upstream Regulatory Region ή Non Coding Region): ρυθμίζει τη μεταγραφή και την έναρξη αντιγραφής του DNA και βρίσκεται ανάμεσα στην E και L περιοχή.

Τα γονίδια διαχωρίζονται σε πρώιμα (Early-E) και όψιμα (Late-L) βάσει του χρόνου εμφάνισής τους κατά τη φυσική εξέλιξη της HPV λοίμωξης. Τα πρώιμα γονίδια εκφράζονται στα βασικά μολυσμένα κύτταρα, ενώ τα όψιμα στα επιφανειακά κύτταρα, όπου συσσωρεύονται τα ιοσωμάτια. Ο ιός προσβάλλει αρχικά τα βασικά κύτταρα, αλλά εκφράζει υψηλά επίπεδα ιικών πρωτεϊνών στις ανώτερες στοιβάδες του επιθηλίου. Στην *εικόνα 87* δίνεται το γονιδίωμα του HPV-16.

Πίνακας 14: Λειτουργία των γονιδίων του HPV: πρώιμες πρωτεΐνες (E) και πρωτεΐνες του καψιδίου (L):

Γονίδιο	Λειτουργία
Πρώιμες πρωτεΐνες (E)	
E1	αντιγραφή DNA (πολλαπλασιασμός ιικού γονιδιώματος), διατήρηση σταθερότητας του επισώματος, μεταγραφική καταστολή των πρωτεϊνών E6 και E7
E2	αντιγραφή του DNA (με το E1), έλεγχος μεταγραφής, διατήρηση σταθερότητας του επισώματος, μεταγραφική ρύθμιση των πρωτεϊνών E6 και E7
E3	άγνωστη
E4	απελευθέρωση ιοσωματίων και καθορισμός τροπισμού των διαφόρων HPV τύπων ως προς κύτταρα/ιστούς, ωρίμανση ιών, αλληλεπίδραση με κυτταροκερατίνες
E5	διαμεμβρανική πρωτεΐνη, μετασχηματιστική ικανότητα: ενισχύει τις ικανότητες μετασχηματισμού των E6 και E7 πρωτεϊνών
E6	πρωτεΐνη με ογκογόνο δράση: μετασχηματιστική ικανότητα/κακοήθη εξαλλαγή κυττάρων, διατήρηση κακοήθους φαινότυπου, έλεγχος μεταγραφής, αλληλεπιδρά με πρωτεΐνη p53
E7	πρωτεΐνη με ογκογόνο δράση: μετασχηματιστική ικανότητα/κακοήθη εξαλλαγή κυττάρων, διατήρηση κακοήθους φαινότυπου, έλεγχος μεταγραφής, αλληλεπιδρά με πρωτεΐνη pRb
E8	άγνωστη
Πρωτεΐνες του καψιδίου (L)	
L1	κύρια πρωτεΐνη του καψιδίου των ώριμων ιικών σωματιδίων
L2	μικρή πρωτεΐνη του καψιδίου, σταθεροποίηση της δομής του καψιδίου



Εικόνα 87: *Γονιδίωμα του HPV-16 (Τροποποιημένη εικόνα από [93])*

Σε καρκινικά κύτταρα και βιοψίες καρκίνου έχει παρατηρηθεί η έκφραση συγκεκριμένων ικών γονιδίων, όπως E6 και E7. Τα γονίδια που διεγείρουν δραστηριότητα πολλαπλασιασμού των κυττάρων είναι τα E5, E6 και E7. Το E5 φαίνεται να έχει σημαντικό ρόλο στα αρχικά στάδια της λοίμωξης: διεγείρει την ανάπτυξη των κυττάρων και εμποδίζει την απόπτωσή τους έπειτα από βλάβη στο DNA. Ωστόσο η παρουσία του E5 δεν είναι υποχρεωτική στα τελευταία στάδια καρκινογένεσης που προκλήθηκε από HPV. Τον πιο σημαντικό ρόλο για την κακοήθη μεταπλασία κατέχουν τα γονίδια E6 και E7 καθώς και οι αντίστοιχές τους πρωτεΐνες. Εκφράζονται επανειλημμένα σε κακοήθη ιστό και αναστέλλοντας την έκφρασή τους αποτρέπεται ο κακοήθης φαινότυπος των καρκινικών κυττάρων του τραχήλου της μήτρας. Οι πρωτεΐνες E6 και E7 είναι ικανές ανεξάρτητα η μια από την άλλη να προκαλέσουν κυτταρική αθανατοποίηση αλλά με μειωμένη αποτελεσματικότητα, όταν, όμως, εκφράζονται και οι δύο μαζί αυξάνεται η κυτταρική αθανατοποίηση και η αλλοίωση που προκαλούν [88]. Τα γονίδια E6 και E7 μόνο των υψηλού κινδύνου τύπων- και όχι των χαμηλού κινδύνου τύπων- οδηγούν σε κυτταρική αθανατοποίηση. Στην παρούσα φάση είναι δύσκολο να καθοριστεί ο ρόλος άλλων πρωτεϊνών του HPV (E1, E2, E4) στη διαδικασία της κακοήθης μετατροπής. Οι πρωτεΐνες L1 και L2 δεν εκφράζονται σε προκαρκινικά και κακοήθη κύτταρα (όμως είναι σημαντικά για την ανάπτυξη εμβολίων).

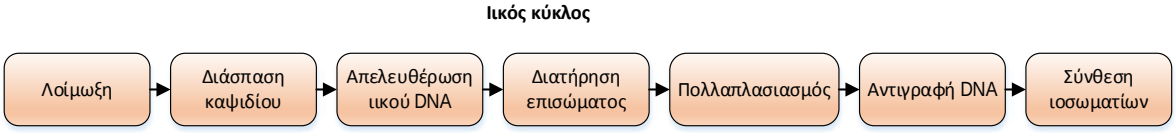
4.2.3 Βιολογικός κύκλος του HPV

Ο βιολογικός κύκλος των ιών παρουσιάζει μια εξωκυτταρική και μια ενδοκυτταρική φάση. Στην εξωκυτταρική του φάση ο ιός ονομάζεται ιικό σωματίδιο ή ιοσωμάτιο και αποτελεί έναν δυνητικά μολυσματικό παράγοντα. Το ιικό σωματίδιο αποτελείται από νουκλεϊκό οξύ σε ένα πρωτεϊνικής φύσης καψίδιο. Στην ενδοκυτταρική φάση αναπαράγεται το νουκλεϊκό οξύ του ιικού σωματιδίου παράγοντας πολλά αντίγραφα του γονιδιώματος και των πρωτεϊνών του περιβλήματος του ιού.

Ο πλήρης βιολογικός κύκλος του HPV περιλαμβάνει 3 στάδια: την ακολουθιακή έκφραση ιικών γονιδίων που οδηγεί στην αντιγραφή του ιικού DNA και στην παραγωγή υψηλά λοιμογόνων ιοσωματίων [87]. Θεωρείται ότι ο κύκλος αναπαραγωγής του HPV ξεκινά με την είσοδο του ιού στα κύτταρα της βασικής μεμβράνης του επιθηλίου [94], ενώ για τη διατήρηση της λοίμωξης, ο ιός στοχεύει τα βλαστοκύτταρα του πλακώδους επιθηλίου (*εικόνα 89*). Για να το πετύχει αυτό, ο HPV προσαρμόζεται στο κύτταρο-ξενιστή και εκμεταλλεύεται τον κυτταρικό μηχανισμό για τους δικούς του σκοπούς.

Η λοίμωξη ξεκινά όταν λοιμογόνα HPV σωματίδια που έχουν εισέλθει στο γεννητικό σύστημα κατά τη σεξουαλική επαφή, φτάσουν στη βασική μεμβράνη του επιθηλίου, όπου προσδένονται και εισέρχονται στα κύτταρα μέσω μικροτραυματισμών. Αφού ο ιός HPV εισέλθει στα κύτταρα, απελευθερώνει νουκλεϊκό οξύ στον κυτταρικό πυρήνα. Στη βασική μεμβράνη, η αναπαραγωγή του ιού θεωρείται μη παραγωγική και ο ιός διατηρείται ως χαμηλός αριθμός επισωματικών αντιγράφων (εξωχρωμοσωμικό επίσωμα: ανεξάρτητο γενετικό στοιχείο επιπρόσθετο του κανονικού γονιδιώματος του κυττάρου), χρησιμοποιώντας το μηχανισμό αναδιπλασιασμού του DNA του κυττάρου-ξενιστή για τη σύνθεση του ιικού DNA (κατά μέσο όρο μια φορά σε κάθε κυτταρικό κύκλο) [94]. Αν και ο κυτταρικός πολλαπλασιασμός είναι απαραίτητος για τη διατήρηση των ιικών επισωμάτων, ο ιός τελικά πρέπει να ενισχυθεί και να «συσκευάσει» τα γονιδιώματά του έτσι ώστε να μπορούν να παραχθούν ιοσωμάτια [95]. Όταν ο ιός εισέλθει στην υπερβασική μεμβράνη του επιθηλίου (superbasal/parabasal/midzone, παραβασική/διάμεση στοιβάδα), ο ιός μεταβαίνει στο μηχανισμό κυλιόμενου κύκλου αντιγραφής (rolling circle replication) του ιικού DNA, αυξάνει τον αριθμό των ιικών αντιγράφων του γονιδιώματος, συνθέτει πρωτεΐνες καψιδίου (L1 και L2) και τέλος δημιουργεί/συναρμολογεί τα ιοσωμάτια [94]. Η σύνθεση των πρωτεϊνών L1 και L2 για το σχηματισμό του πρωτεϊνικού καψιδίου γίνεται αφού γίνει η ενίσχυση του γονιδιώματος. Όταν ολοκληρωθεί η σύνθεση των L1 και L2, το τελικό στάδιο του αναπαραγωγικού κύκλου του HPV απαιτεί τα διάφορα αντίγραφα του ιικού γονιδιώματος να «συσκευαστούν» μέσα στο πρωτεϊνικό καψίδιο ώστε να σχηματιστούν τα ιογενή σωματίδια [95]. Έτσι, μια γενιά ιών σχηματίζεται πάνω στην επιφάνεια του επιθηλίου, οι οποίοι διασκορπίζονται

σε πολλές περιοχές στον ίδιο ξενιστή ή μεταφέρονται σε άλλους ξενιστές. Στην *εικόνα 88* παρουσιάζεται ο βιολογικός κύκλος ενός ιού.

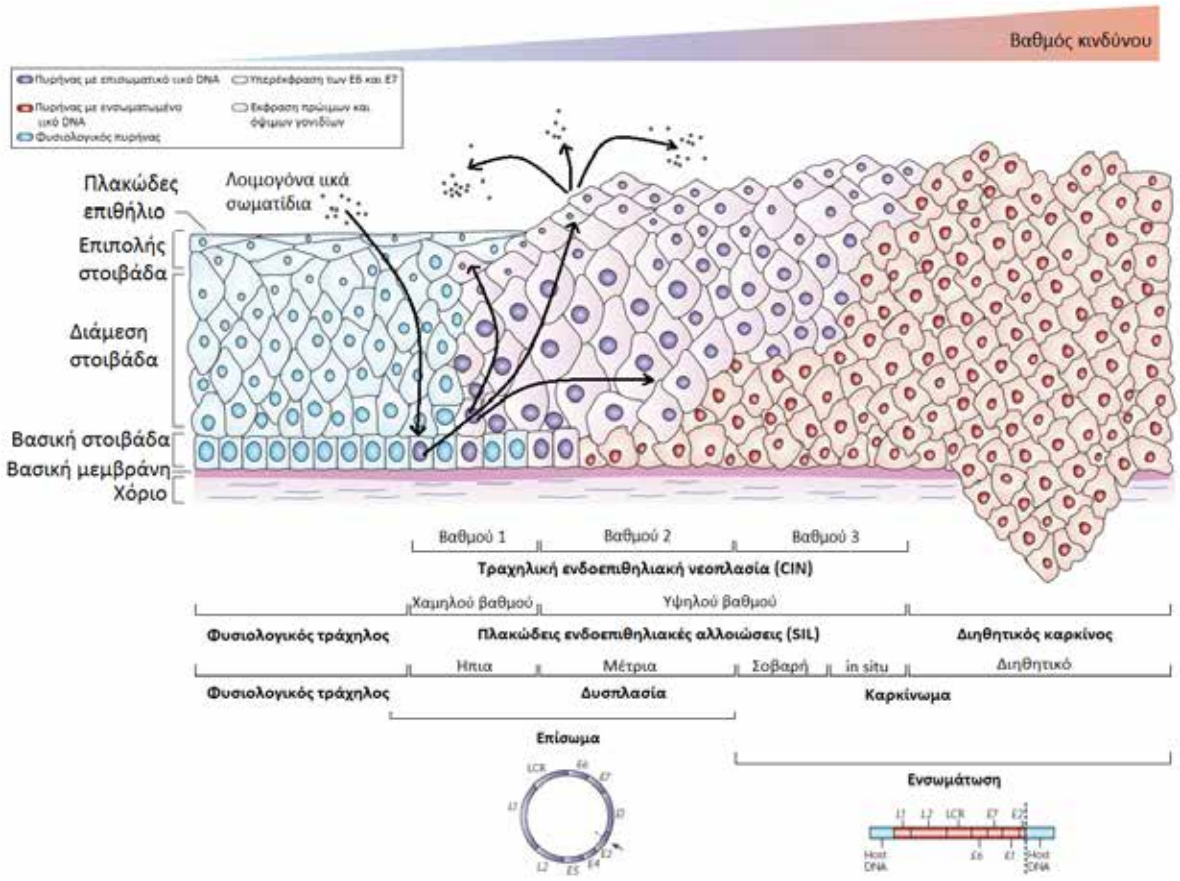


Εικόνα 88: *Τυπικός βιολογικός κύκλος ιού*

Όπως έχει αναφερθεί, όταν η λοίμωξη έχει ξεκινήσει στη βασική κυτταρική μεμβράνη, το ιικό DNA γονιδίωμα αρχικά διατηρείται ως ένα κυκλικό πλήρες επισωματικό αντίγραφο, δηλαδή ως ένα επιπλέον χρωμοσωμικό στοιχείο στα κύτταρα του ξενιστή. Μπορεί να παραμένει στα βασικά κύτταρα σε λανθάνουσα κατάσταση για μεγάλες χρονικές περιόδους ή μπορεί να ενεργοποιηθεί και να ξεκινήσει μια παραγωγική λοίμωξη σε όλα τα επιθηλιακά στρώματα. Σε αυτό το στάδιο ο ιός είναι ικανός να συμπληρώσει το βιολογικό του κύκλο παράγοντας καινούρια αντίγραφα του ιού. Στις καλοήθεις αλλοιώσεις, το γονιδίωμα του ιού βρίσκεται εκτός του κυτταρικού γονιδιώματος και η αντιγραφή του ιού γίνεται σαν ένα εξωχρωμοσωμικό επίσωμα. Στις κακοήθεις αλλοιώσεις, το ιικό γονιδίωμα ενσωματώνεται στο κυτταρικό με γραμμική μορφή (*εικόνα 89*). Για να γίνει αυτό, γίνεται διάσπαση του ιικού γονιδιώματος στα E1 και E2, με αποτέλεσμα την απώλεια αυτών των γονιδίων και συνεπώς και της έκφρασης των πρωτεϊνών που κωδικοποιούν, οι οποίες είναι αναγκαίες για τη ρύθμιση της αντιγραφής και μεταγραφής του ιού (ουσιαστικά, με την ενσωμάτωση, διαταράσσεται ή και διακόπτεται η περιοχή των γονιδίων E1 και E2 με αποτέλεσμα την απώλεια της έκφρασης των πρωτεϊνών E1 και E2). Αυτό παρεμβαίνει στη λειτουργία της πρωτεΐνης E2, η οποία υπό κανονικές συνθήκες, μειώνει την μεταγραφή των γονιδίων E6 και E7 και κατά συνέπεια και την έκφραση των πρωτεϊνών E6 και E7 [94]. Άρα, η ενσωμάτωση του ιικού DNA στο κυτταρικό DNA του ξενιστή καταλήγει στη διατάραξη των επιπέδων των E1 και E2 πρωτεϊνών και τελικά οδηγεί σε υπερέκφραση των δύο ιικών πρωτεϊνών E6 και E7. Αυτές οι πρωτεΐνες σε συνδυασμό με την E5 προωθούν την αθανατοποίηση και μετασχηματισμό των προσβεβλημένων κυττάρων, ιδιαίτερα των τύπων HPV-16 και HPV-18 (η E7 αυξάνει την αθανατοποίηση και η συνεργασία των E6 και E7 προκαλεί το μετασχηματισμό) [87].

Η λειτουργία των E6 και E7 πρωτεϊνών στην παραγωγική φάση της HPV λοίμωξης είναι να ανατρέψει την κανονική ανάπτυξη των κυττάρων και να τροποποιήσει το κυτταρικό περιβάλλον ώστε να διευκολύνει την ιική αντιγραφή [94]. Από διάφορες πειραματικές μελέτες φαίνεται ότι οι αλληλεπιδράσεις των ιικών πρωτεϊνών E6 και E7 με ένα αριθμό κυτταρικών πρωτεϊνών είναι αυτές που επάγουν τον πολλαπλασιασμό και τελικά την αθανατοποίηση και τον κακοήθη

μετασχηματισμό των κυττάρων του ξενιστή [96]. Οι πιο χαρακτηριστικές αλληλεπιδράσεις είναι με τις ογκοκατασταλτικές πρωτεΐνες pRB (πρωτεΐνη ρετινοβλαστώματος, retinoblastoma protein) και p53 (tumor suppressor protein) οι οποίες έχουν κεντρικό ρόλο στον έλεγχο του κυτταρικού κύκλου (οι πρωτεΐνες αυτές εντοπίζονται μεταλλαγμένες σε πολλών ειδών καρκίνους) [96]. Τα παράγωγα των γονιδίων E6 και E7 απορρυθμίζουν τον κυτταρικό κύκλο του ξενιστή δεσμεύοντας και αδρανοποιώντας αυτές τις ογκοκατασταλτικές πρωτεΐνες: η αδρανοποίηση των πρωτεϊνών p53 και pRb αυξάνει το ρυθμό του κυτταρικού πολλαπλασιασμού και οδηγεί σε γονιδιωματική αστάθεια, με συνέπεια τα κύτταρα του ξενιστή να συσσωρεύουν όλο και περισσότερες βλάβες στο DNA (τις οποίες δεν μπορούν να διορθώσουν) και τελικά να μετασχηματίζονται σε καρκινικά κύτταρα [94].



Εικόνα 89: *Εξέλιξη της HPV λοίμωξης σε διηθητικό καρκίνο (Τροποποιημένη εικόνα από [97] [98])*

4.2.4 Ογκογόνος μηχανισμός

Όπως έχουμε δει, ο κύκλος αναπαραγωγής του ιού μέσα στο επιθήλιο μπορεί να διαχωριστεί σε δύο μέρη. Αρχικά το ιικό γονιδίωμα που βρίσκεται στα κύτταρα της βασικής μεμβράνης,

αντιγράφεται γύρω στις 100 φορές και ο αριθμός αυτός διατηρείται με αναδιπλασιασμό επί τακτά χρονικά διαστήματα εντός των αρχικά προσβεβλημένων - αλλά ακόμα ικανών για αναδιπλασιασμό - κυττάρων [96]. Οι ιικές πρωτεΐνες E1 και E2 είναι αναγκαίες για αυτό τον αναδιπλασιασμό του DNA.

Το δεύτερο στάδιο ξεκινά όταν τα βασικά κύτταρα ωθούνται στην υπερβασική μεμβράνη. Τότε, τα προσβεβλημένα επιθηλιακά κύτταρα ενεργοποιούν τον κυτταρικό μηχανισμό άμυνας επανεξετάζοντας την αλληλουχία του DNA πριν από τη διαίρεση [99]. Αυτή η διαδικασία αναθεώρησης του DNA πραγματοποιείται σε ένα σημείο ελέγχου κατά τη διάρκεια της φάσης G1 του κυτταρικού κύκλου με στόχο την παρεμπόδιση της αντιγραφής μεταλλαγμένου DNA και υπεύθυνες για αυτή είναι κάποιες κυτταρικές ογκοκατασταλτικές πρωτεΐνες (κυρίως η pRB και η p53) .

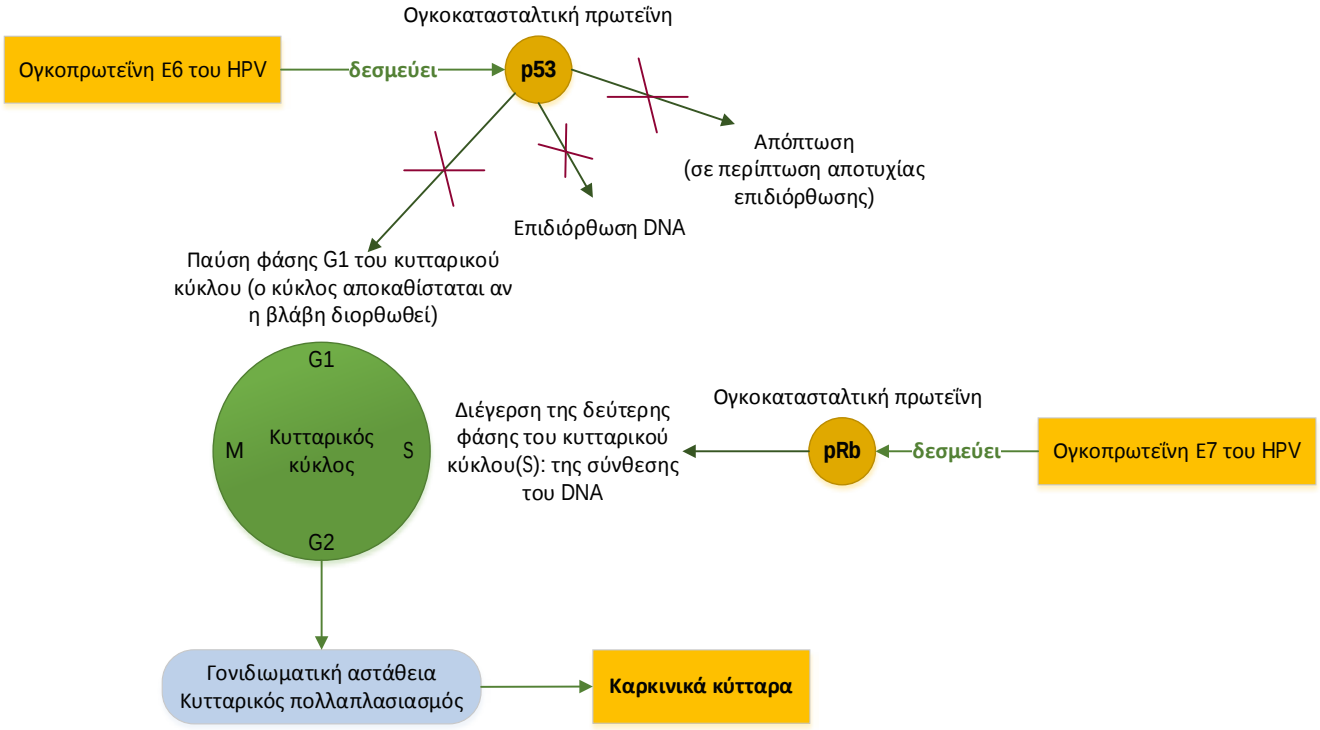
Ο τυπικός κυτταρικός κύκλος αποτελείται από τις φάσεις G1, S (σύνθεση), G2 και M (μίτωση) και είναι μια κυκλική διαδικασία. Η κυτταρική διαίρεση συμβαίνει στη φάση M και η σύνθεση/αντιγραφή του DNA γίνεται στη φάση S. Η φάση G1 είναι η περίοδος ανάμεσα στη φάση M και S, ενώ η G2 ανάμεσα στη φάση S και M. Στις φάσεις G1 και G2 υπάρχουν 2 σημεία ελέγχου του DNA. Με την παρουσία βλάβης στο DNA το γονίδιο p53 ενεργοποιείται και ρυθμίζει τη μετάβαση του κυττάρου από τη φάση G1 στη φάση S του κυτταρικού κύκλου. Προκαλεί, δηλαδή, παροδική παύση του κυτταρικού κύκλου, δίνοντας χρόνο στο κύτταρο είτε να επιδιορθώσει τη βλάβη του DNA με τους μηχανισμούς επιδιόρθωσής του (αναστρέψιμη κατάσταση), είτε να οδηγηθεί σε απόπτωση (κυτταρικός θάνατος).

Στην περίπτωση του HPV, όταν το προσβεβλημένο κύτταρο εντοπίσει το ιικό γονιδίωμα επιχειρεί να διορθώσει το πρόβλημα. Επειδή, όμως, το γονιδίωμα είναι πολύ μεγάλο σε μέγεθος για να εξαλειφθεί, οι πρωτεΐνες pRB και p53 αναγκάζουν το προσβεβλημένο κύτταρο να οδηγηθεί σε «προγραμματισμένο κυτταρικό θάνατο» με απόπτωση έτσι ώστε να μην μπορεί να χρησιμοποιηθεί για την εξάπλωση της λοίμωξης. Κανονικά με αυτό τον μηχανισμό η λοίμωξη περιορίζεται αφού τα προσβεβλημένα κύτταρα οδηγούνται σε κυτταρικό θάνατο. Ωστόσο, οι υψηλού κινδύνου τύποι HPV προστατεύονται από αυτό τον κυτταρικό μηχανισμό άμυνας, συνθέτοντας πρωτεΐνες που τον εμποδίζουν (E6 και E7). Οι διαφορές ανάμεσα στις πρωτεΐνες E6 και E7 των χαμηλού και των υψηλού κινδύνου τύπων είναι μάλλον ποσοτικές παρά ποιοτικές [96]. Η πρόσδεση της E7 στην pRB ενεργοποιεί την έκφραση των πρωτεϊνών που είναι απαραίτητες για τον αναδιπλασιασμό του DNA (διέγερση φάσης S του κυτταρικού κύκλου). Αυτό φυσιολογικά θα οδηγούσε στην απόπτωση των κυττάρων από την πρωτεΐνη p53, αλλά στα προσβεβλημένα κύτταρα η διαδικασία αυτή αντισταθμίζεται από την ιική πρωτεΐνη E6, η οποία στοχεύει την p53 για πρωτεολυτική αποικοδόμηση [96]. Έτσι, τα προσβεβλημένα κύτταρα από υψηλού κινδύνου τύπους προστατεύονται από τον κυτταρικό θάνατο με απόπτωση και μπορούν

να εξακολουθήσουν να χρησιμοποιούνται για την παραγωγή ικών ογκογονιδίων. Η δράση αυτών των ογκογονιδίων επιτρέπει στο μικρό αριθμό των προσβεβλημένων κυττάρων να επεκταθούν, αυξάνοντας τον αριθμό των κυττάρων που εν συνεχεία θα παραγάγουν ικά σωματίδια [95].

Τα φυσιολογικά κύτταρα της υπερβασικής μεμβράνης κανονικά τερματίζουν τον κυτταρικό κύκλο και ξεκινούν τη διαδικασία της τελικής κυτταρικής διαφοροποίησης (διαδικασία κατά την οποία το κύτταρο μετατρέπεται σε ένα πιο εξειδικευμένο) προκειμένου να παραχθεί το προστατευτικό φράγμα που κανονικά παρέχεται από το δέρμα. Στα προσβεβλημένα από HPV κύτταρα, όμως, η φυσιολογική τελική διαφοροποίηση δεν πραγματοποιείται επειδή το σύστημα τερματισμού της εξέλιξης του κυτταρικού κύκλου χάνεται [95].

Ο ογκογόνος μηχανισμός που προκαλείται από μια HPV λοίμωξη παρουσιάζεται στην *εικόνα 90*.



Εικόνα 90: *Ογκογόνος μηχανισμός μέσω HPV λοίμωξης*

Η πιο πάνω διαδικασία αποκλεισμού των p53 και Rb πρωτεϊνών από τις E6 και E7 προκαλεί διάφορα άλλα προβλήματα, ακόμα κι αν δεν οδηγήσει σε κυτταρική αθανατοποίηση. Ως συνέπεια της παρεμπόδισης του μηχανισμού διόρθωσης σφάλματος, το κύτταρο όχι μόνο δεν μπορεί να εξαλείψει το ιικό DNA, αλλά επιπλέον δεν μπορεί ούτε να διορθώσει τα εγγενή σφάλματα στο κυτταρικό DNA με αποτέλεσμα να συσσωρεύει γενετικές μεταβολές. Αν, επιπρόσθετα, έχει παρεμποδιστεί και η διαδικασία απόπτωσης και το κύτταρο δεν μπορεί να

οδηγηθεί σε «κυτταρικό θάνατο», τότε θα μετατραπεί σε ένα αθανατοποιημένο κύτταρο με DNA που θα συνεχίζει να αλλοιώνεται από το DNA του ιού, ή σε ένα κύτταρο με νεοπλασματικό φαινότυπο [99].

Φαίνεται ότι ο μηχανισμός καρκινογένεσης από HPV ξεκινά με την υπερέκφραση των E6 και E7 οι οποίες εμποδίζουν τις p53 και pRb και τελικά αθανατοποιούν το προσβεβλημένο κύτταρο. Ενόψει αυτού, μόνο μια λοίμωξη με υψηλή ποσότητα ιών (υψηλό ιικό φορτίο) θα είναι ικανή να παράξει αρκετές E6 και E7 για να ξεκινήσει την παραπάνω διαδικασία. Όντως, λοιμώξεις με υψηλό ιικό φορτίο, για τις οποίες το ανοσολογικό σύστημα του οργανισμού δεν μπορεί να αντεπεξέλθει και να εξαλείψει την λοίμωξη, έχουν υψηλότερο κίνδυνο για νεοπλασματικό μετασχηματισμό. Ωστόσο, έχουν βρεθεί περιστατικά επίμονων λοιμώξεων με χαμηλό ιικό φορτίο που έχουν παράξει ένα φαινότυπο του όγκου [99].

4.2.5 Στάδια εξέλιξης της λοίμωξης σε διηθητικό καρκίνωμα

Παράγοντες κινδύνου για την παρουσία του HPV είναι ο υψηλός αριθμός σεξουαλικών συντρόφων, η πραγματοποίηση της πρώτης σεξουαλικής επαφής σε μικρή ηλικία, ο τοκετός σε νεαρή ηλικία, η καταστολή και μεταβολή της ανοσοποιητικής κατάστασης, η μη καλή προσωπική υγιεινή, η νεαρή ηλικία και οι ορμονικές επιδράσεις [87]. Εκτός από τον υψηλό αριθμό σεξουαλικών συντρόφων, δεν είναι γνωστό αν οι υπόλοιποι παράγοντες οδηγούν σε αυξημένο κίνδυνο για HPV λοίμωξη ή/και για εμφάνιση δυσπλασιών προκαλούμενων από τον ιό [87].

Τα στάδια εξέλιξης μιας HPV λοίμωξης παρουσιάζονται στην *εικόνα 91*. Μια HPV λοίμωξη ξεκινά με την πρωτολοίμωξη, την οποία ακολουθεί μια φάση επώασης (incubation phase): μετά την είσοδο του ιού στα κύτταρα του επιθηλίου μέσω της κυτταρικής μεμβράνης, το γονιδίωμα του ιού μεταφέρεται στον πυρήνα τους, χωρίς όμως να ενσωματώνεται στα χρωμοσώματά τους (επίσωμα). Η φάση αυτή διαρκεί από 6 εβδομάδες έως 8 μήνες, περίοδο κατά την οποία η λοίμωξη εξαπλώνεται σε μεγάλες επιφάνειες του επιθηλίου του κατώτερου γεννητικού συστήματος και έχει ως συνέπεια μια σταθερή φλεγμονή του επιθηλίου. Αν το επιτρέψει το ανοσοποιητικό σύστημα του οργανισμού, γίνεται μετάβαση στην επόμενη φάση με την εμφάνιση της πρώτης αλλοίωσης.

Σε περίπτωση, κάμψης του αμυντικού συστήματος του ανθρωπίνου οργανισμού από τον ιό σε τοπικό επίπεδο, προκαλείται μετάβαση στη φάση της ενεργού έκφρασης (3-6 μήνες). Σε αυτή τη φάση, ο ιός εισέρχεται στα κύτταρα της βασικής στοιβάδας προκαλώντας υπερπλασία της. Τα διαφοροποιημένα κύτταρα της βασικής στοιβάδας ανέρχονται προς τις επιφανειακές στοιβάδες εξαπλώνοντας τη λοίμωξη και στις υπόλοιπες στοιβάδες του επιθηλίου, όπου παρουσιάζεται

κοιλοκυτταρική ατυπία (είδος αλλοιώσεων). Παράλληλα παρουσιάζεται υπερπλασία του επιθηλίου και των αγγείων του στρώματος, με αποτέλεσμα την εμφάνιση υποκλινικών (μη ορατές με γυμνό μάτι κατά την απλή κλινική εξέταση) ή κλινικών (κλινικά ορατές) αλλοιώσεων.

Ακολούθως, επέρχεται η φάση εξισορρόπησης, κατά την οποία ο οργανισμός αντιδρά ανοσολογικά, αντιρροπώντας τη δραστηριότητα του ιού, με συνέπεια είτε τη μηδενική εμφάνιση νέων αλλοιώσεων είτε την επιβράδυνση του ρυθμού εμφάνισής τους. Με την ανοσολογική άμυνα του οργανισμού οι αλλοιώσεις υποχωρούν έως και 20%.

Τέλος, συνήθως 9 μήνες μετά την εμφάνιση της πρώτης αλλοίωσης, επέρχεται η απώτερη φάση. Στη φάση αυτή η λοίμωξη είτε υποχωρεί μόνιμα, είτε υποτροπιάζει ή εμφανίζει μια συνεχή εξελικτική πορεία. Στην πρώτη περίπτωση, δηλαδή, αν επέλθει κάθαρση του ιού, οι αλλοιώσεις παραμένουν σε διαρκή ύφεση, αλλά το DNA του ιού μπορεί να ανιχνεύεται στους ιστούς επί πολλά χρόνια. Ωστόσο, οι περισσότερες γυναίκες (80%) που έχουν προσβληθεί με ένα συγκεκριμένο τύπο HPV, μετά από περίοδο 18 μηνών δεν επιδεικνύουν καμία ένδειξη αυτού του τύπου και γενικά θεωρείται ότι επαναλοίμωξη με τον ίδιο τύπο ιού δε συνηθίζεται [95]. Στην περίπτωση, που οι αλλοιώσεις υποτροπιάζουν ή συνεχίζουν να βρίσκονται στη φάση της ενεργού έκφρασης, τότε υπάρχει υψηλός κίνδυνος για εμφάνιση νεοπλασιών [100]. Η ανάπτυξη νεοπλασιών υψηλού βαθμού προκαλείται σε γυναίκες οι οποίες δεν μπορούν να επιτύχουν κάθαρση του ιού και διατηρούν επίμονη ενεργή λοίμωξη για χρόνια ή δεκαετίες μετά την αρχική έκθεση [95]. Προηγούμενες μελέτες από διάφορα εργαστήρια εισηγούνται ότι σε κάποια περιστατικά μπορούν να εκδηλωθούν νεοπλασίες υψηλού βαθμού πολύ γρήγορα μετά την αρχική έκθεση [95].

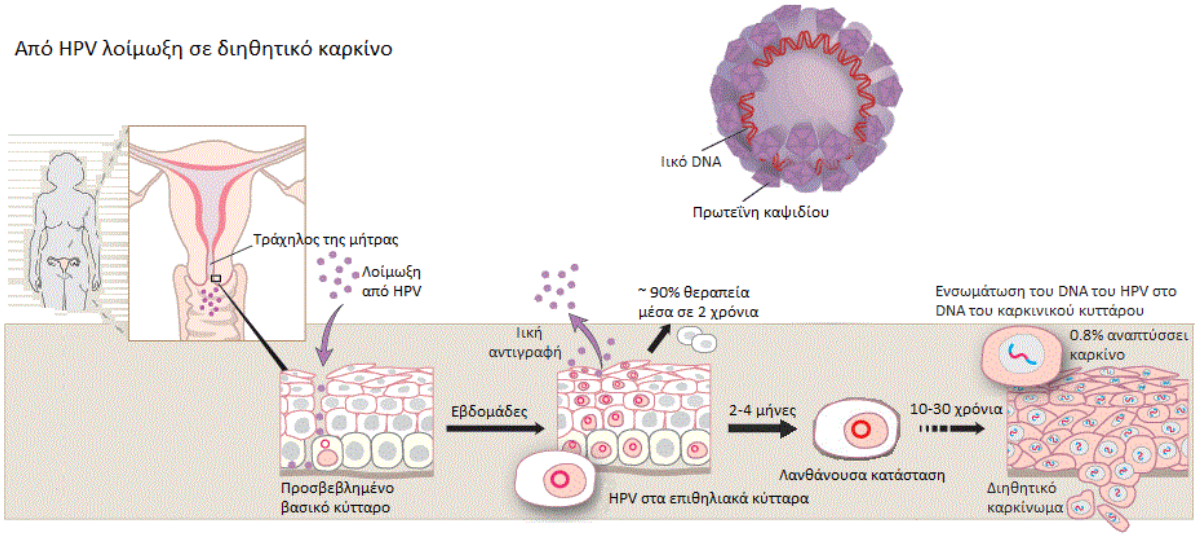
Στάδια εξέλιξης της λοίμωξης

Εικόνα 91: *Στάδια εξέλιξης της λοίμωξης*

Η εξέλιξη της λοίμωξης από τον HPV εξαρτάται από το αν θα ενσωματωθεί ή όχι ο ιός στα χρωμοσώματα των κυττάρων-ξενιστών. Αν δεν ενσωματωθεί, δηλαδή παραμείνει επισωματικός, το γονιδίωμα του ιού αναπαράγεται (αναδιπλασιάζεται) παράλληλα με τον πολλαπλασιασμό των κυττάρων σε περιοχές του επιθηλίου, όπου και εμφανίζεται κοιλοκυτταρική ατυπία. Ως αποτέλεσμα παρουσιάζονται αλλοιώσεις LSIL όπως υποκλινικά κονδυλώματα, ατυπίες και δυσπλασίες ελαφρού βαθμού, οι οποίες είτε μένουν ως έχουν είτε υποχωρούν αυτομάτως. Αν, όμως, το DNA του ιού ενσωματωθεί στα χρωμοσώματα των κυττάρων-ξενιστών, τότε οι κατασταλτικοί παράγοντες που ελέγχουν το μηχανισμό πολλαπλασιασμού των κυττάρων παύουν να είναι σε ισχύ, με άμεσο αποτέλεσμα την εμφάνιση ενός νεοπλασματικού όγκου. Σε όλες σχεδόν τις περιπτώσεις διηθητικού καρκίνου το DNA του ιού είναι ενσωματωμένο. Ωστόσο, σπάνια, παρουσιάζονται περιπτώσεις καρκίνου του τραχήλου της μήτρας που δεν είναι ενσωματωμένο [101].

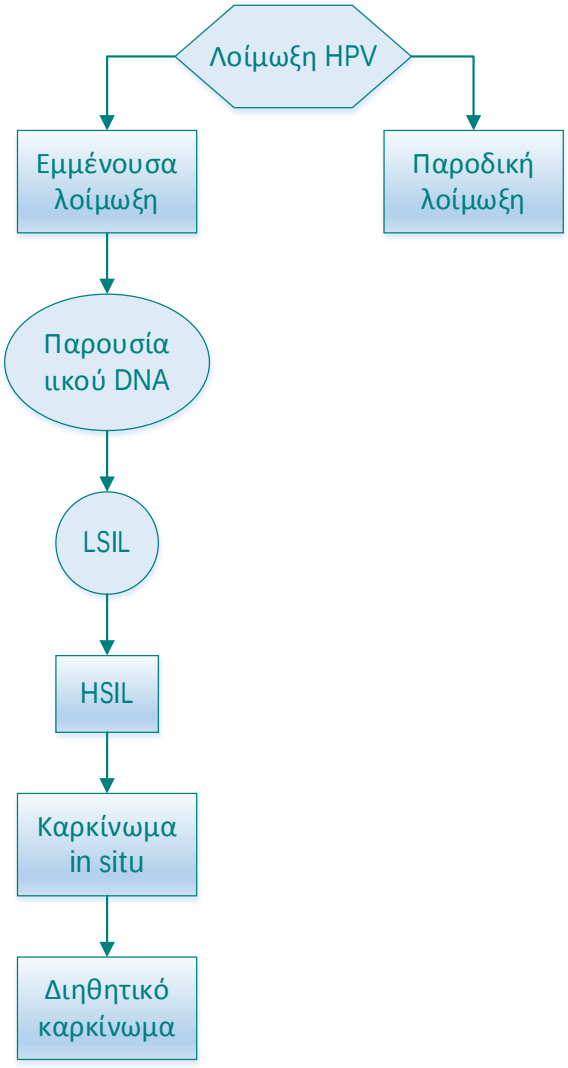
Εκτός από τα διηθητικά καρκινώματα, ενσωματωμένο HPV DNA εντοπίζεται και σε ένα υποσύνολο αλλοιώσεων υψηλού βαθμού, αλλά επίσης και σε κάποιες αλλοιώσεις CIN1. Θεωρείται ότι η ενσωμάτωση είναι ένα πρώιμο στάδιο της καρκινικής εξέλιξης [95] και όντως σε κάποιες αλλοιώσεις CIN1, CIN2 και CIN3 που δείχνουν στοιχεία ενσωμάτωσης παρατηρείται έκφραση του p16 που θεωρείται δείκτης (marker) αυξημένης έκφρασης E7.

Στις παραγωγικές αλλοιώσεις, τα κύτταρα εισέρχονται σε κυτταρικό κύκλο μόνο στα χαμηλά επιθηλιακά στρώματα και επεκτείνονται προς την επιφάνεια του επιθηλίου σε διαφορετικές εκτάσεις ανάλογα με το βαθμό αλλοίωσης και τον τύπο του HPV. Στην περίπτωση των υψηλών κινδύνου τύπων που σχετίζονται με τον καρκίνο του τραχήλου της μήτρας, το σχετικό πάχος των στοιβάδων αυξάνεται ανάλογα με το βαθμό νεοπλασίας, ενώ ο βαθμός της επιθηλιακής διαφοροποίησης μειώνεται [95].



Εικόνα 92: *Από HPV λοίμωξη σε διηθητικό καρκίνο (Τροποποιημένη εικόνα από [102])*

Στην *εικόνα 93* παρουσιάζονται τα στάδια εξέλιξης σε διηθητικό καρκίνο. Η λοίμωξη μεταδίδεται συνήθως μέσω σεξουαλικής επαφής και έχει ως αρχικό αποτέλεσμα πλακώδεις ενδοεπιθηλιακές αλλοιώσεις (SIL – Squamous Intraepithelial Lesion). Οι περισσότερες από αυτές, μετά από διάστημα 6-12 μηνών από την εμφάνισή τους, εξαφανίζονται, πιθανότατα εξαιτίας ανοσολογικής παρέμβασης. Ωστόσο, ένα μικρό ποσοστό εξακολουθεί να υφίσταται και εξελίσσεται σε υψηλού βαθμού SIL, καρκινώματα in situ και τέλος, αν δεν πραγματοποιηθεί χειρουργική επέμβαση, σε πλακώδες καρκίνωμα ή αδενοκαρκίνωμα του τραχήλου της μήτρας [88].



Εικόνα 93: *Στάδια εξέλιξης σε διηθητικό καρκίνωμα*

4.2.6 Τύποι HPV

Ο ιός HPV ταξινομείται σε τύπους ανάλογα με μικροδιαφορές στο γενετικό τους υλικό. Για την περιγραφή τους, απομονώνονται τα DNA γονιδιώματα του ιού και ακολούθως γίνεται σύγκριση της νουκλεοτιδικής ακολουθίας αυτών των γονιδιωμάτων [103]. Η σύνθεση και η αλληλουχία των νουκλεοτιδίων του νουκλεϊκού οξέος είναι χαρακτηριστική για κάθε γονιδίωμα. Διαφορετικά γονιδιώματα του ίδιου ιού (δηλαδή παραλλαγές του βασικού γονιδιώματος του ιού) ονομάζονται στελέχη (isolates) και αναφέρονται και ως γονότυποι ή τύποι. Το 1995, στο International Papillomavirus Workshop, συμφωνήθηκε ότι για τον ορισμό ενός νέου γονότυπου PV (συνεπώς και HPV), απαιτείται να υπάρχει διαφορά άνω του 10% στην αλληλουχία του γονιδίου L1 σε σχέση με τους ήδη καθορισμένους τύπους [104]. Συνεπώς, ως *τύπος* ορίζεται ένα κλωνοποιημένο πλήρους μήκους γονιδίωμα PV του οποίου η L1 αλληλουχία νουκλεοτιδίων του είναι τουλάχιστον 10% ανόμοια από αυτή οποιουδήποτε άλλου τύπου PV [105]. Τα γονιδιώματα HPV εξελίσσονται όσο αργά (ή γρήγορα) εξελίσσονται τα γονιδιώματα των ξενιστών τους [103].

Ο τύπος του ιού καθορίζει το είδος καθώς και τη θέση της αλλοίωσης στο δέρμα: κάθε τύπος προσβάλλει συγκεκριμένο επιθήλιο και έχει το δικό του ογκογόνο δυναμικό. Για το λόγο αυτό, η ανάλυση του ιού HPV σε συγκεκριμένους τύπους έχει μεγάλη ιατρική σημασία. Για παράδειγμα ο τύπος HPV 11 εντοπίζεται σε προκαρκινικές αλλοιώσεις CIN και οι HPV-16 και HPV-18 σε διηθητικά τραχηλικά καρκινώματα [87]. Βάσει των έως τώρα δεδομένων, εκτιμάται ότι υπάρχουν περισσότεροι από 200 στελέχη (isolates) HPV [103]. Ο κάθε τύπος ονομάζεται με ένα αριθμό που αντιπροσωπεύει τη σειρά με την οποία ανακαλύφθηκε. Έως σήμερα, έχουν περιγραφεί πάνω από 189 τύποι PV (papillomavirus), εκ των οποίων πάνω από 120 είναι τύποι HPV (human papilomavirus). Από τους 120 HPV τύπους έχει προσδιοριστεί πλήρως η DNA αλληλουχία των 100 [103]. Οι 40 από αυτούς, προσβάλλουν τη γεννητική και περιπρωκτική περιοχή και είναι σεξουαλικώς μεταδιδόμενοι. Οι υπόλοιποι προσβάλλουν τη στοματική κοιλότητα, το λάρυγγα, το δέρμα κτλ.

Οι διάφοροι τύποι ανάλογα με την ογκογεννητική τους ικανότητα προκαλούν λοιμώξεις HPV υψηλού κινδύνου και χαμηλού κινδύνου. Οι τύποι HPV που εντοπίζονται σε καρκίνους του τραχήλου της μήτρας, ή σε καρκίνους αλλού στο ανώτερο γεννητικό σύστημα, έχουν οριστεί ως τύποι «υψηλού κινδύνου» (high risk) ή ογκογόνοι τύποι. Αντιθέτως, οι τύποι που βρίσκονται κυρίως σε κονδυλώματα των γεννητικών οργάνων και σε μη κακοήθεις ενδοεπιθηλιακές αλλοιώσεις του πλακώδους επιθηλίου (LSIL), χαρακτηρίζονται ως «χαμηλού κινδύνου» (low risk) ή μη ογκογόνοι τύποι [88].

Έχουν γίνει διάφορες μελέτες για τον επακριβή καθορισμό των υψηλού κινδύνου τύπων. Εκτιμάται ότι ο αριθμός των υψηλού κινδύνου τύπων που μπορεί να προκαλέσουν καρκινογένεση

είναι 13-19 εκ των οποίων 11 τύποι (16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58) ταξινομούνται επανειλημμένα ως υψηλού κινδύνου [106]. Ωστόσο, ορισμένοι τύποι από αυτούς (31, 33, 35, 51 και 52), από κάποιους μελετητές κατατάσσονται ως τύποι ενδιάμεσου κινδύνου (intermediate risk) επειδή ανευρίσκονται περισσότερο σε μέτριες και σοβαρές προκαρκινικές δυσπλαστικές αλλοιώσεις από ότι σε καρκινώματα [107] - οι περισσότεροι, όμως, ερευνητές αποφεύγουν αυτό τον όρο.

Οι κυριότεροι τύποι υψηλού κινδύνου που σχετίζονται με κακοήθεις αλλοιώσεις είναι ο HPV-16 και ο HPV-18, οι οποίοι είχαν αρχικά εντοπιστεί το 1983-1984 [87]. Στο 70% των περιπτώσεων εμφάνισης καρκίνου του τραχήλου της μήτρας εντοπίζονται αυτοί οι δύο τύποι [96]. Ο HPV-16 εντοπίζεται σε περισσότερες από 50% των βιοψιών, ενώ ο HPV-18 στο 20% [87]. Ο HPV-16 είναι ο πιο συχνός τύπος στα πλακώδη επιθηλιακά καρκινώματα (εντοπίζεται σε άνω του 50% των SCCs), ενώ ο HPV-18 συσχετίζεται περισσότερο με αδενοκαρκινώματα (εντοπίζεται σε άνω του 50% των αδενοκαρκινωμάτων) [107]. Οι τύποι HPV 16 και 18 φαίνεται να είναι τύποι των οποίων το DNA τους, υπάρχει ενσωματωμένο με το DNA του κυττάρου-ξενιστή [87]. Το 1995, το International Agency for Research on Cancer (IARC), συμπέρανε ότι υπάρχουν επαρκή αποδεικτικά στοιχεία από μελέτες, έτσι ώστε οι τύποι 16 και 18 να ταξινομηθούν ως καρκινογόνοι για τον άνθρωπο, αλλά για άλλους τύπους τα στοιχεία ήταν ελλιπή και ανεπαρκή [106].

Σύμφωνα με το [106] οι HPV τύποι 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 68, 73 και 82 θα πρέπει να θεωρούνται καρκινογόνοι/υψηλού κινδύνου, και οι τύποι 26, 53 και 66 θα πρέπει να θεωρούνται πιθανώς καργκινογόνοι. Στον *πίνακα 15* παρουσιάζονται οι τύποι υψηλού και χαμηλού κινδύνου σύμφωνα με τη μελέτη [88] και στον πίνακα 16 παρουσιάζονται οι χαμηλού και υψηλού κινδύνου τύποι σύμφωνα με τη μελέτη [106]. Παρατηρούμε ότι οι δύο αυτές μελέτες συμφωνούν σε 13 τύπους υψηλού κινδύνου (16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 68).

Πίνακας 15: Υψηλού κινδύνου τύποι HPV σύμφωνα με [88]

Είδος αλλοίωσης	Τύποι HPV	
	Λιγότερο επικρατείς	**Επικρατέστεροι**
Κονδυλώματα	42, 44, 51, 53, 83	6, 11
Ενδοεπιθηλιακές νεοπλασίες	6, 11, 18, 26, 30, 31, 33, 34, 35, 39, 40, 42, 43, 45, 51, 52, 53, 54, 55, 56, 57, 58, 59, 61, 62, 64, 66, 67, 68, 69, 70, 71, 73, 74, 79, 81, 82, 83, 84	16
Καρκίνος του τραχήλου της μήτρας και του ανώτερου γεννητικού συστήματος	6, 11, 18, 31, 33, 35, 39, 45, 51, 52, 54, 56, 58, 59, 66, 68, 69	16

Πίνακας 16: Χαμηλού και υ ψηλού κινδύνου τύποι HPV σύμφωνα με [106]

Επιδημιολογική Ταξινόμηση	HPV τύποι
Υψηλού κινδύνου	16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 68, 73, 82
Ενδιάμεσου κινδύνου (πιθανώς υψηλού κινδύνου)	25, 53, 66
Χαμηλού κινδύνου	6, 11, 40, 42, 43, 44, 54, 61, 70, 72, 81

Σε διάφορες μελέτες παγκόσμιας κλίμακας, δείγματα διηθητικών καρκινωμάτων του τραχήλου της μήτρας έχουν ελεγχθεί για ύπαρξη HPV-DNA και βάσει των θετικών δειγμάτων έχει γίνει κατάταξη των πιο κοινών HPV τύπων. Στον πίνακα 17, παρουσιάζονται οι πιο κοινοί τύποι HPV, κατά φθίνουσα σειρά συχνότητας, που εντοπίστηκαν σε θετικά για HPV-DNA δείγματα διηθητικού καρκίνου του τραχήλου της μήτρας, σύμφωνα με τρεις διαφορετικές μελέτες: Munoz et al. 2004 [108], Li et al. 2011 [109] και de Sanjose et al. 2010 [110]. Οι μελέτες αυτές έγιναν όλες με χρήση δειγμάτων καρκίνου του τραχήλου της μήτρας που λήφθηκαν σε παγκόσμια κλίμακα. Μπορεί να παρατηρηθεί ότι και οι 3 μελέτες συμφωνούν στους 8 πιο κοινούς HPV τύπους με διαφορές στη σειρά κατάταξης (16, 18, 45, 31, 33, 52, 58 και 35). Αυτοί οι 8 τύποι φαίνεται να είναι υπεύθυνοι για περίπου το 90% όλων των περιπτώσεων εμφάνισης διηθητικού καρκίνου του τραχήλου της μήτρας παγκοσμίως [87]. Ιδιαίτερο ρόλο στην πρόκληση καρκίνου του τραχήλου της μήτρας φαίνεται να έχουν οι HPV 16, 18 και 45, αφού εμφανίζονται σε άνω του 90% των δειγμάτων από αδενοκαρκινώματα του τραχήλου της μήτρας που ήταν θετικά ως προς την ύπαρξη HPV-DNA [110].

Πίνακας 17: Οι πιο κοινοί τύποι HPV βάσει 3 διαφορετικών μελετών [108] [109] [110]

Κατάταξη	*Μελέτη [108]: Δεδομένα από IARC N= 3085*	*Μελέτη [109]: Μετα-ανάλυση N=30357*	*Μελέτη [110]: Έρευνα από ICO N=8977*
1	**HPV 16**	**HPV 16**	**HPV 16**
2	**HPV 18**	**HPV 18**	**HPV 18**
3	**HPV 45**	**HPV 58**	**HPV 45**
4	**HPV 31**	**HPV 33**	**HPV 33**
5	**HPV 33**	**HPV 45**	**HPV 31**
6	**HPV 52**	**HPV 31**	**HPV 52**
7	**HPV 58**	**HPV 52**	**HPV 58**
8	**HPV 35**	**HPV 35**	**HPV 35**

9	HPV 59	HPV 59	
10	HPV 56	HPV 39	
11	HPV 39	HPV 51	
12	HPV 51	HPV 56	
13	HPV 73		
14	HPV 68		
15	HPV 66		

Ας σημειωθεί, ότι δεν εξελίσσονται όλες οι λοιμώξεις από τους τύπους 16 και 18 σε καρκίνο. Οι υψηλού κινδύνου τύποι, και ιδιαίτερα ο HPV 16, είναι ευρέως διαδεδομένοι σε όλο τον ανθρώπινο πληθυσμό, καθιστώντας την HPV λοίμωξη μια πολύ συχνή λοίμωξη που όμως μόνο σε κάποιες περιπτώσεις έχει ως συνεπακόλουθο την εμφάνιση καρκίνου [88].

4.2.7 Άλλοι παράγοντες που επηρεάζουν την ανάπτυξη καρκίνου του τραχήλου της μήτρας

Παρά τις έρευνες που συνεχίζονται έως σήμερα δεν έχει αναδειχθεί ένα συγκεκριμένο αίτιο ως αποκλειστικός παράγοντας του καρκίνου του τραχήλου της μήτρας, έχει, όμως, αναδειχθεί ένας συνδυασμός παραγόντων, οι οποίοι παρουσιάζονται συνοπτικά στην παρούσα ενότητα. Εκτός από την άμεση σχέση του καρκίνου του τραχήλου της μήτρας και του ιού HPV, είναι απαραίτητη και η ύπαρξη άλλων παραγόντων για την πρόκληση της νόσου. Η παρουσία του HPV μπορεί να είναι απαραίτητη για την πρόκληση καρκίνου του τραχήλου της μήτρας αλλά δεν είναι επαρκής [96].

Είναι γενικώς αποδεκτό ότι η ογκογένεση εξαιτίας του ιού HPV απαιτεί τη συσσώρευση επιπρόσθετων γενετικών αλλαγών οι οποίες πραγματοποιούνται σε χρόνο κατόπιν της αρχικής λοίμωξης. Ο μέσος όρος ηλικίας των γυναικών με καρκίνο του τραχήλου της μήτρας είναι 50 χρονών (περίπου), ενώ ο μέσος όρος των γυναικών με HSIL είναι 28 χρονών (περίπου), κάτι που υποδηλώνει ότι στις περισσότερες περιπτώσεις προηγείται ένα προκαρκινικό στάδιο μακράς διαρκείας που επιτρέπει τη συσσώρευση δευτερεύοντων γενετικών αλλαγών [95].

Στην εξέλιξη των κυττάρων που έχουν προσβληθεί από HPV σε κακοήθη συνεισφέρουν διάφοροι παράγοντες, κάποιοι άμεσα σχετικοί με τον ιό και κάποιοι όχι (*εικόνα 94*). Είναι πολύ πιθανόν, αν και δεν έχει σαφώς προσδιοριστεί, να επηρεάζουν και κάποιοι γενετικοί και ανοσολογικοί παράγοντες του ξενιστή, καθώς και άλλοι ιογενείς παράγοντες εκτός του τύπου του HPV [96].

Οι πιθανοί παράγοντες μπορούν να ταξινομηθούν σε τρεις κατηγορίες:

- Περιβαλλοντικοί/ εξωγενείς παράγοντες
- Ιογενείς παράγοντες
- Παράγοντες του ξενιστή

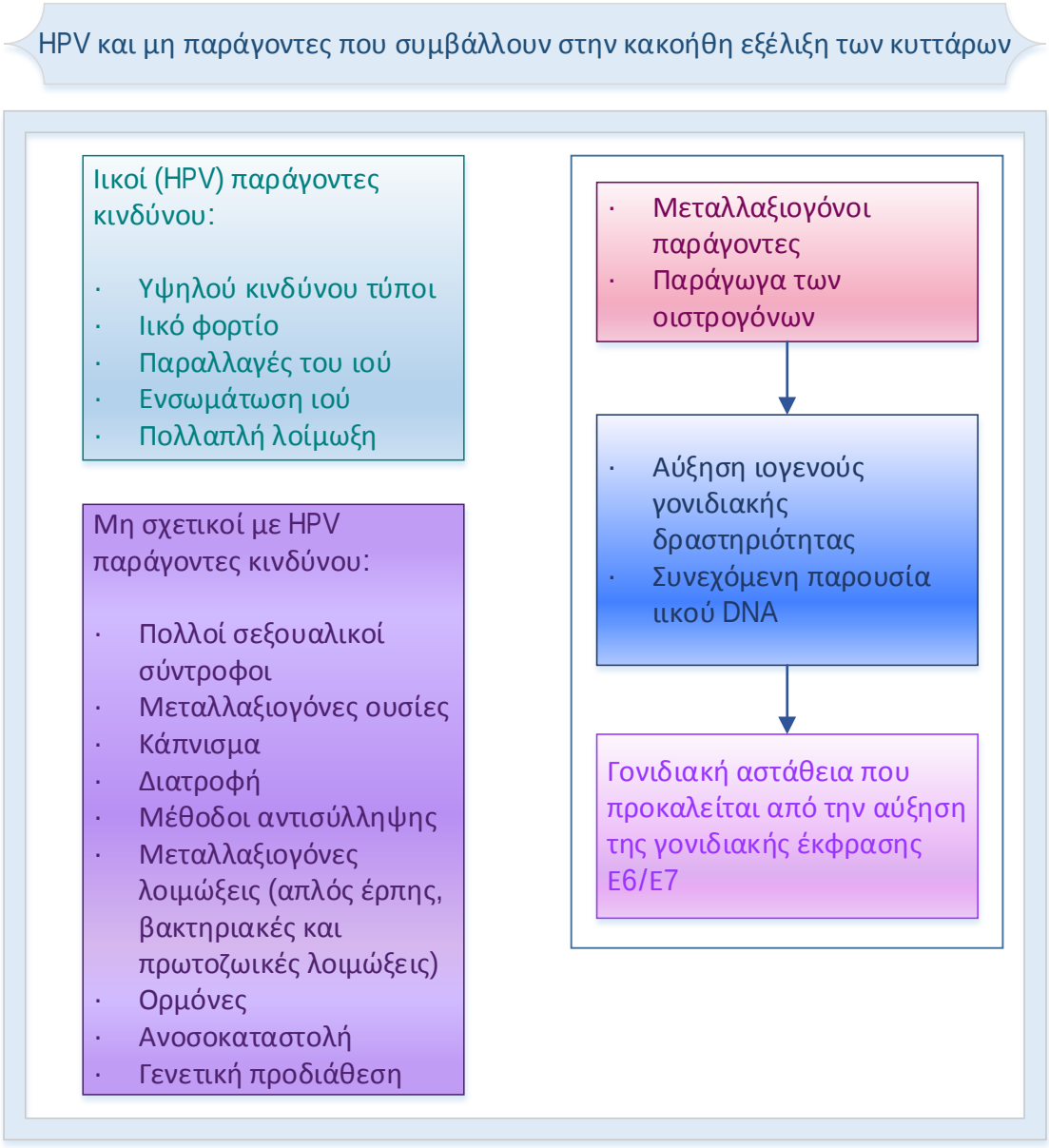
Κάποιοι πιθανοί εξωγενείς παράγοντες είναι η μακροχρόνια χρήση ορμονικών αντισυλληπτικών (από του στόματος χορηγούμενα αντισυλληπτικά χάπια), το κάπνισμα και οι μεταλλαξιογόνες ουσίες, η παράλληλη συνλοίμωξη με άλλα σεξουαλικώς μεταδιδόμενα νοσήματα όπως λοίμωξη από τον ιό ανθρώπινης ανοσοανεπάρκειας (Human Immunodeficiency Virus, HIV), χλαμύδια (Chlamydia Trachomatis, CT) ή/και απλό έρπη τύπου 2 (Herpes Simplex Virus, HSV-2), οι πολλαπλοί σεξουαλικοί σύντροφοι, ο τραυματισμός του τραχήλου της μήτρας, καθώς και ορισμένες διατροφικές ελλείψεις [96] [111] [88]. Άλλος ένας πιθανός εξωγενής παράγοντας που αυξάνει τον κίνδυνο εξέλιξης μιας HPV λοίμωξης σε καρκίνο είναι ο αριθμός των προηγηθέντων κυήσεων και τοκετών [111]. Το κάπνισμα και η μεταλλαξιογόνος δράση των συστατικών του τσιγάρου έχουν δειχθεί να προκαλούν τοπική ανοσολογική καταστολή στα τραχηλικά κύτταρα και μπορεί να συμβάλουν στην επιμονή της HPV λοίμωξης ή/και στην κακοήθη μεταπλασία, όπως συμβαίνει με τον καρκίνο του πνεύμονα [94].

Παραδείγματα ενδεχόμενων ιογενών παραγόντων είναι ο τύπος του HPV (λοίμωξη από συγκεκριμένους τύπους, συνλοίμωξη με άλλους τύπους), οι παραλλαγές του HPV, το ιικό φορτίο και η ενσωμάτωση του ιού [111]. Ο πιο σημαντικός παράγοντας κινδύνου ανάπτυξης διηθητικού καρκινώματος είναι ο τύπος του ιού. Οι παραλλαγές του HPV διαφέρουν σε βιολογικές και χημικές ιδιότητες, καθώς και σε παθογένεια. Η ογκογένεση μιας συγκεκριμένης παραλλαγής HPV φαίνεται να ποικίλει γεωγραφικά και ανάλογα με τη εθνική καταγωγή του υπό μελέτη πληθυσμού. Για τον τύπο HPV-16 έχουν καθοριστεί 5 διαφορετικές παραλλαγές: Ευρωπαϊκή (E, European), Ασιατική (As, Asian), Ασιατική-Αμερικανική (AA, Asian-American), Αφρικανική-1 (Af1, African-1) και Αφρικανική-2 (Af2, African-2). Οι Ασιοαμερικανικές παραλλαγές φαίνεται να έχουν πιο έντονη ογκογενή δραστηριότητα σε σύγκριση με τα Ευρωπαϊκά στελέχη εξαιτίας αυξημένης μεταγραφικής δραστηριότητας [94]. Η παρουσία λοίμωξης από πολλαπλούς τύπους HPV παρατηρήθηκε να αυξάνεται ανάλογα με τη σοβαρότητα της νόσου. Η λοίμωξη πολλαπλών τύπων HPV που έχει παρατηρηθεί είναι συνήθως με δύο διαφορετικούς τύπους, αλλά έχουν βρεθεί δείγματα και με 3, 4 ή 5 τύπους [94].

Παράγοντες σχετικοί με τον ξενιστή που είναι πιθανόν να επηρεάζουν είναι η ανοσοκαταστολή, οι ενδογενείς ορμόνες, γενετικοί παράγοντες όπως το αντιγόνο ανθρωπίνων λευκοκυττάρων και άλλοι παράγοντες του ξενιστή σχετικοί με την ανοσολογική απόκριση του ξενιστή [111]. Επίσης έχει βρεθεί ότι σημαντικός παράγοντας κινδύνου αποτελεί και η γενετική προδιάθεση. Η γενετική κληρονομικότητα θα μπορούσε να επηρεάσει πολλούς παράγοντες οι οποίοι συμβάλλουν στην ανάπτυξη του καρκίνου του τραχήλου της μήτρας,

συμπεριλαμβανομένων την επιδεκτικότητα σε μια HPV λοίμωξη, την ικανότητα κάθαρσης της HPV λοίμωξης και το χρόνο ανάπτυξης της νόσου [94].

Σημαντικός είναι ο ρόλος που παίζουν οι ορμονικοί παράγοντες (τα οιστρογόνα και τα παράγωγά τους) και οι μεταλλαξιογόνες ουσίες. Οι ορμονικοί παράγοντες ενεργοποιούν τον προαγωγέα του HPV και διευκολύνουν την κυτταρική αθανατοποίηση των προσβεβλημένων από HPV κυττάρων, ενώ οι μεταλλαξιογόνες ουσίες, ενισχύουν τη συνεχιζόμενη παρουσία του DNA του HPV και τα αυξημένα επίπεδα ογκογενούς έκφρασης E6 και E7 έχουν ως αποτέλεσμα την αύξηση της γονιδιακής αστάθειας, που με τη σειρά της διευκολύνει την εξέλιξη των προσβεβλημένων κυττάρων σε διηθητικό όγκο [88].



Εικόνα 94: *Παράγοντες που οδηγούν σε κακοήθεια εξαιτίας HPV λοίμωξης*

Παρατηρείται ομοιότητα μεταξύ των παραγόντων που επηρεάζουν την ανάπτυξη προκαρκινικών αλλοιώσεων του τραχήλου της μήτρας και των παραγόντων που προκαλούν καρκίνο του τραχήλου της μήτρας. Η μονογαμία, η καθυστερημένη έναρξη σεξουαλικής δραστηριότητας, η καλή προσωπική υγιεινή και η χρήση μεθόδων αντισύλληψης φραγμού, φαίνεται να βοηθούν στην πρωτογενή πρόληψη (μέτρα ώστε να μην εμφανιστεί η πάθηση) [87].

4.3 Διαγνωστικές εξετάσεις καρκίνου του τραχήλου της μήτρας

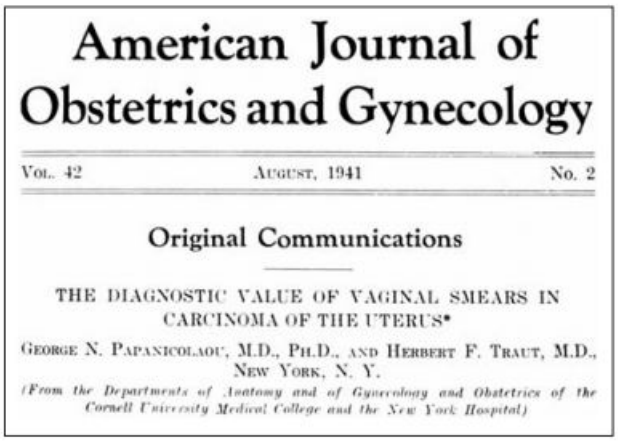
Πριν το 1940, στις ΗΠΑ, ο καρκίνος του τραχήλου της μήτρας είχε προκαλέσει περισσότερους θανάτους από κάθε άλλο καρκίνο (ποσοστό θνησιμότητας 36/100000 γυναίκες) [87]. Στις αρχές του 20ου αιώνα, διάφοροι ειδικοί στον καρκίνο πίστευαν ότι εάν οι γυναίκες συμβουλεύοντας τους γιατρούς τους αμέσως μετά τα πρώτα γυναικολογικά συμπτώματα (π.χ. ακατάσχετη αιμορραγία), τότε θα διαγνωσθούν με καρκίνο σε στάδιο 1 (περιορισμένο στον τράχηλο της μήτρας) και θα έχουν πολύ καλή πιθανότητα να θεραπευτούν. Ωστόσο, αργότερα αντιλήφθηκαν ότι αυτό δεν ήταν ρεαλιστικό σενάριο. Πολλές γυναίκες αν και συμβουλεύτηκαν γρήγορα τους γιατρούς τους, διαγνώσθηκαν με προχωρημένους όγκους (στάδια 2 και 3) και επιπλέον πολλές από αυτές πέθαναν από τη νόσο. Για το λόγο αυτό ήταν αναγκαία η εύρεση ενός τρόπου εντοπισμού του καρκίνου προτού να εμφανιστεί οποιοδήποτε σύμπτωμα [81].

4.3.1 Εξέταση Παπανικολάου (Papanicolaou test)

Το 1928 ο Δρ. Γιώργος Παπανικολάου (1893-1962), ένας Έλληνας παθολόγος που δούλευε στο Νοσοκομείο Νέας Υόρκης (New York Hospital), έγραψε μια μικρή ερευνητική εργασία στην οποία ισχυριζόταν ότι τα τραχηλικά επιχρίσματα (cervical smears) μπορούσαν να αποκαλύψουν κρυμμένο καρκίνο της μήτρας [81]. Ο Δρ. Παπανικολάου είχε εκπαιδευτεί στην ιατρική, αλλά και στη ζωολογία. Κατόπιν έρευνας πολλών χρόνων σε θηλυκά ποντίκια σχετικά με τη δράση του οιστρογόνου (θηλυκή ορμόνη που παράγεται στις ωοθήκες), εντόπισε έναν απλό τρόπο να μετρά την εμφάνιση του θηλυκού κύκλου γονιμότητας (οίστρου) που προκαλείται από την παραγωγή οιστρογόνου (με τη χορήγηση οιστρογόνου σε θηλυκά ποντίκια των οποίων οι ωοθήκες είχαν αφαιρεθεί, επερχόταν οίστρος): μέσω της παρατήρησης αλλαγών στο κολπικό επίχρισμα. Έτσι το τεστ Παπανικολάου έγινε η κλασική μέθοδος μέτρησης του οιστρογόνου στα εμπορικά σκευάσματα. Ακολούθως, επιχείρησε να εφαρμόσει την ίδια μέθοδο και σε «human females" (κατά τον ίδιο), έτσι ώστε να μελετήσει τη δράση του οιστρογόνου στις γυναίκες, ελπίζοντας στη θεραπεία διαταραχών του γυναικείου κύκλου. Τότε, αντιλήφθηκε ότι δεν γνώριζε πως μοιάζουν

τα φυσιολογικά κολπικά επιχρίσματα. Γι' αυτό το λόγο εξέτασε επιχρίσματα από υποθετικά υγιείς γυναίκες, όταν σε κάποιες περιπτώσεις εντόπισε μη φυσιολογικά κύτταρα. Γυναικολογική εξέταση αυτών των γυναικών έδειξε ότι είχαν καρκίνο του τραχήλου της μήτρας σε πρώιμα στάδια. Με αυτό τον τρόπο, ο Δρ. Παπανικολάου ανακάλυψε τυχαία μια απλή μέθοδο διάγνωσης του πρώιμου καρκίνου [81]. Η αρχή του τεστ Παπανικολάου είναι βασισμένη στην παρατήρηση του Δρ. Παπανικολάου ότι η κυτταρολογική εξέταση αποφολιδωμένων κυττάρων από τον τράχηλο της μήτρας μπορούσε να βοηθήσει στην έγκαιρη αναγνώριση γυναικών με προκαρκινικές βλάβες και καρκίνο και έτσι να συμβάλλει στην πρόληψή τους. Για να πειστεί, ωστόσο, η ιατρική κοινότητα για την αναγκαιότητα του Pap test, χρειάστηκαν δύο δημοσιεύσεις. Στη δεύτερή του δημοσίευση, μαζί με τον Traut, το 1941 (*εικόνα 95*), αναφέρει:

«…if by any chance a simple, inexpensive method of diagnosis could be evolved which could be applied to large numbers of women in the cancer-bearing period of life, we would be in a position to discover the disease in its incipiency much more frequently that is now possible.» [112]



American Journal of Obstetrics and Gynecology

VOL. 42 AUGUST, 1941 NO. 2

Original Communications

THE DIAGNOSTIC VALUE OF VAGINAL SMEARS IN CARCINOMA OF THE UTERUS*

GEORGE N. PAPANICOLAOU, M.D., PH.D., AND HERBERT F. TRAUT, M.D., NEW YORK, N. Y.

(From the Departments of Anatomy and of Gynecology and Obstetrics of the Cornell University Medical College and the New York Hospital)

Εικόνα 95: *Το εξώφυλλο της έκδοσης του περιοδικού με το άρθρο που δημοσιεύτηκε το 1941 από τον Παπανικολάου και τον Traut σχετικά με τη διαγνωστική αξία των κολπικών επιχρισμάτων στο καρκίνωμα της μήτρας*

Στο τέλος της δεκαετίας του 1940 αρκετοί γυναικολόγοι υποστήριξαν το συστηματικό προσυμπτωματικό έλεγχο (screening) με χρήση του Pap test σε όλες τις γυναίκες με κίνδυνο ανάπτυξης καρκίνου του τραχήλου της μήτρας [81]. Έτσι μετά τις δημοσιεύσεις των Παπανικολάου και Traut άρχισε να εφαρμόζεται η τεχνική της αποφολιδωτικής κυτταρολογίας. Ο αρχικός στόχος της εξέτασης ήταν ο εντοπισμός των διηθητικών καρκίνων σε πρώιμα στάδια, ώστε να είναι εφικτή η θεραπεία με υστερεκτομή ή ακτινοθεραπεία. Ωστόσο, ακολούθως φάνηκε η χρησιμότητα της εξέτασης και για τη διάγνωση αλλοιώσεων προτού εξελιχθούν σε διηθητικό καρκίνο (carcinoma in situ), καθιστώντας το Pap test μια μέθοδο πρόληψης της νόσου.

Έκτοτε, το τεστ Παπανικολάου ή ΠΑΠ τεστ (Pap test, προηγουμένως γνωστό ως Pap smear) χρησιμοποιείται ευρέως σε παγκόσμιο επίπεδο ως προληπτικό τεστ (screening test) για την ανίχνευση του προδιηθητικού καρκίνου του τραχήλου της μήτρας. Από το 2000, οι επαναλαμβανόμενες προληπτικές εξετάσεις κυτταρολογίας κατά τη διάρκεια της ζωής μιας γυναίκας έχουν μειώσει τη συχνότητα εμφάνισης του καρκίνου τους τραχήλου της μήτρας κατά 75% [87].

4.3.1.1 Κλασικό τεστ Παπανικολάου

Το κλασικό τεστ Παπανικολάου/συμβατική κυτταρολογία (Conventional Cytology- CC ή Conventional Pap smear- CP) είναι διαθέσιμο από το 1950 σε πάρα πολλές χώρες. Η αποτελεσματικότητά του, αν παρέχεται σε οργανωμένα προγράμματα πληθυσμιακού ελέγχου, υπολογίζεται περίπου γύρω στο 80% [87][113][114].

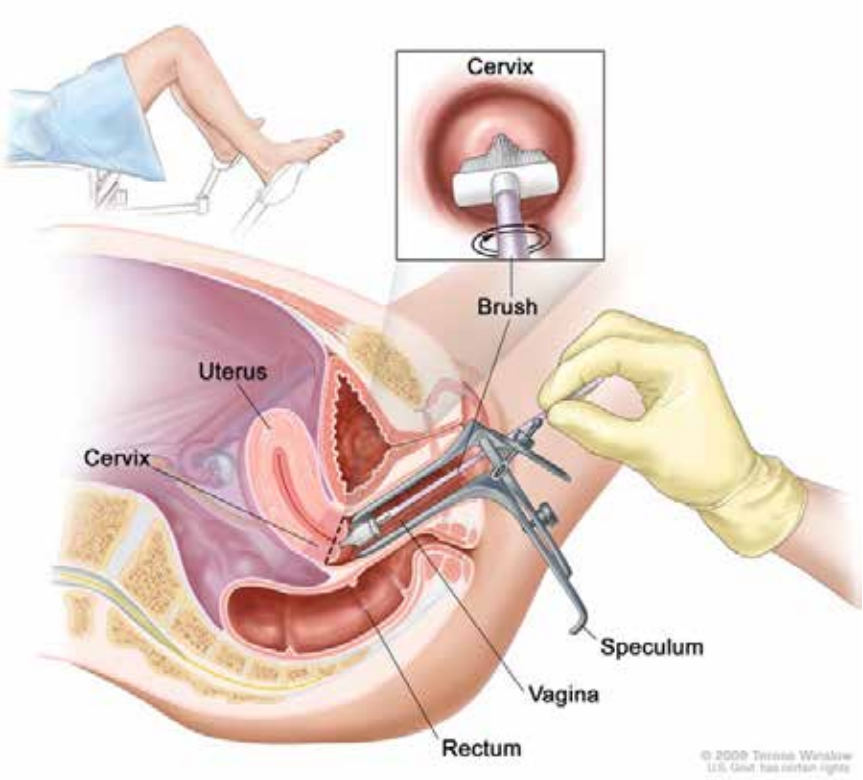
Ως αποτέλεσμα της μακροχρόνιας χρήσης και της ευρείας αποδοχής του κλασικού τεστ Παπανικολάου, έχουν δημιουργηθεί καθιερωμένοι μηχανισμοί ποιοτικού ελέγχου και εκπαίδευσης στη λήψη και αξιολόγησή του. Τα πλεονεκτήματα του κλασικού τεστ Παπανικολάου περιλαμβάνουν, επίσης, το χαμηλό κόστος του και την υψηλή του ειδικότητα. Τα μειονεκτήματά του είναι η μέτρια ευαισθησία του, η αναγκαιότητα ποιοτικού ελέγχου των κυτταρολογικών εργαστηρίων και η μη άμεση διάθεση των αποτελεσμάτων.

Το ΠΑΠ τεστ δεν πρέπει να γίνεται κατά τη διάρκεια εμμηνορρυσίας και πρέπει να πραγματοποιείται αφού περάσουν τουλάχιστον 24 ώρες από την τελευταία σεξουαλική επαφή και τουλάχιστον 48 ώρες από κολπικές πλύσεις και χρήση κολπικών ορμονικών ή φαρμακευτικών σκευασμάτων. Για αξιόπιστο αποτέλεσμα είναι προτιμότερο η λήψη να γίνεται τις ενδιάμεσες μέρες του κύκλου ($14^{η}$ ημέρα).

Κατά τη διενέργεια της εξέτασης η ασθενής τοποθετείται σε θέση λιθοτομίας (*εικόνα 96*). Στην είσοδο του κόλπου τοποθετείται κολποδιαστολέας έτσι ώστε να εκτεθεί ο τράχηλος της μήτρας [115]. Ακολούθως, με τη βοήθεια ειδικού φωτισμού ο ιατρός επισκοπεί τα τοιχώματα του κόλπου και τον τράχηλο για τυχόν ανωμαλίες.

Η συμβατική διαδικασία εξέτασης περιλαμβάνει τη συλλογή δείγματος επιθηλιακών κυττάρων από την επιφάνεια του τραχήλου. Με χρήση ξύλινης σπάτουλας (του Ayre) γίνεται λήψη υλικού από τον εξωτράχηλο και με τη χρήση ειδικής βούρτσας/ψήκτρας γίνεται λήψη υλικού από τον ενδοτράχηλο [115]. Τα κύτταρα επιστρώνονται προσεκτικά σε δύο γυάλινα πλακίδια (αντικειμενοφόρες πλάκες): στο ένα γίνεται επίστρωση με κολπικό και εξωτραχηλικό επίχρισμα και στο άλλο με ενδοτραχηλικό επίχρισμα. Στη συμβατική κυτταρολογία το δείγμα επιστρώνεται απευθείας σε γυάλινη αντικειμενοφόρο πλάκα και σταθεροποιείται/μονιμοποιείται έτσι ώστε να αντέξει για μερικές μέρες. Η μονιμοποίηση του επιχρίσματος γίνεται συνήθως χρησιμοποιώντας

λακ υψηλής περιεκτικότητας σε οινόπνευμα ή με εμβύθηση σε δοχείο Coplin που να περιλαμβάνει ένα από τα ακόλουθα: ίσα μέρη από 95% αιθυλική αλκοόλη και αιθέρα, μόνο 95% αιθυλική αλκοόλη, 100% μεθανόλη ή 85% ισοπροπυλική αλκοόλη [116] [115]. Αφού γίνει η μονιμοποίηση ακολουθεί στέγνωμα των επιχρισμάτων στον αέρα για μια ώρα. Μετά τη μονιμοποίηση και το στέγνωμα τα δείγματα μπορούν να συσκευαστούν. Ως μη αποδεκτά δείγματα θεωρούνται αυτά που δεν περιέχουν κύτταρα από τον ενδοτράχηλο, δείγματα που δεν έχουν μονιμοποιηθεί ή δείγματα σε σπασμένα/θρυμματισμένα πλακίδια. Τα δείγματα μεταφέρονται στο εργαστήριο κυτταρολογίας όπου γίνεται ποιοτική δοκιμή με κυτταρολογική χρώση (Παπανικολάου) και μικροσκόπηση. Το αποτέλεσμα μπορεί να είναι έτοιμο εντός 48 ωρών.



Εικόνα 96: *Δειγματοληψία τραχήλου (εικόνα από [117])*

Η κατηγοριοποίηση των αποτελεσμάτων του ΠΑΠ τεστ, διαφοροποιούνται με την πάροδο του χρόνου και την απόκτηση νέων γνώσεων. Τα διάφορα είδη ταξινόμησης των κυτταρολογικών ευρημάτων παρουσιάζονται ακολούθως.

Αρχικά, δηλαδή από τη δεκαετία του 1940, χρησιμοποιείται η κατάταξη κατά Παπανικολάου και Traut:

Κατηγορία I: δεν παρατηρούνται άτυπα κύτταρα – επιχρίσματα αρνητικά για κακοήθεια με ή χωρίς φλεγμονή

Κατηγορία II: παρατηρείται η παρουσία κυττάρων με χαρακτηριστικά φλεγμονώδους αλλοίωσης αλλά όχι κακοήθειας

Κατηγορία III: παρατηρούνται άτυπα κύτταρα ύποπτα για κακοήθεια

Κατηγορία IV: εικόνα εξαιρετικά ύποπτη για κακοήθεια

Κατηγορία V: εικόνα συμβατή με κακοήθεια (παρατηρούνται καρκινικά κύτταρα)

Αργότερα, το 1968, όταν ο Richard εισήγαγε τον όρο τραχηλική ενδοεπιθηλιακή νεοπλασία (CIN) έτσι ώστε να καλύψει όλο το φάσμα αλλοιώσεων, η κατάταξη κατά Παπανικολάου και Traut αντικαταστάθηκε με το σύστημα Richart, δηλαδή:

CIN-1: ήπια μορφή τραχηλικής ενδοεπιθηλιακής νεοπλασίας

CIN-2: μέτριου βαθμού τραχηλική ενδοεπιθηλιακή νεοπλασία

CIN-3: σοβαρού βαθμού τραχηλική ενδοεπιθηλιακή νεοπλασία – καρκίνωμα in situ

Ακολούθως, από το 1988 έως το 2001, το σύστημα απόδοσης της κυτταρολογικής διάγνωσης που χρησιμοποιήθηκε είναι βασισμένο στο σύστημα Bethesda. Το σύστημα αυτό χρησιμοποιεί τους όρους πλακώδης ενδοεπιθηλιακή αλλοίωση χαμηλού βαθμού (Low Grade Squamous Intraepithelial Lesion- LGSIL ή LSIL) και πλακώδης ενδοεπιθηλιακή αλλοίωση υψηλού βαθμού (High Grade Squamous Intraepithelial Lesion- HGSIL ή HSIL). Η ταξινόμηση που χρησιμοποιήθηκε από το 1988 έως και το 2001 περιγράφεται ακολούθως:

- Κατηγορία WNL (Within Normal Limits): αφορά τα παθολογικά ευρήματα εντός φυσιολογικών ορίων
- Καλοήθεις κυτταρικές αλλοιώσεις, φλεγμονή και αντιδραστικές αλλοιώσεις
- Κατηγορία ASCUS (Atypical Squamous Cells of Undermined Significance): κύτταρα στα οποία έχει βρεθεί κάποια ατυπία, της οποίας όμως, δε μπορεί να προσδιοριστεί με ακρίβεια η σοβαρότητά της, αφήνοντας ανοιχτό το ενδεχόμενο για πιθανή ύπαρξη αλλοιώσεων. Οι τυχόν αλλοιώσεις μπορούν να διαπιστωθούν μέσω κολποσκόπησης με ή χωρίς βιοψία και αναζήτηση ιικού DNA (HPV). Ποσοστό της τάξης του 0.1 % αποδεικνύεται τελικά καρκίνος.
- Κατηγορία AGUS (Atypical Glandular cells of Undetermined significance): κύτταρα στα οποία έχει βρεθεί κάποια ατυπία από αδενικά κύτταρα, η οποία δεν μπορεί να προσδιοριστεί με ακρίβεια. Το ποσοστό του κινδύνου που ελλοχεύει κυμαίνεται από 30% έως 50% βάσει διαφόρων παραγόντων (ηλικία, ιστορικό ασθενούς) και είναι πιθανόν να είναι HSIL, αδενοκαρκίνωμα in situ ή διηθητικός καρκίνος.
- Κατηγορία LSIL: κύτταρα με αλλοιώσεις χαμηλού βαθμού, αλλοιώσεις που οφείλονται σε ιούς HPV, ή ελαφριά δυσπλασία CIN1. Κάποια περιστατικά από αυτά ελλοχεύουν

τον κίνδυνο μετατροπής σε HSIL. Για αυτό το λόγο, οι περιπτώσεις γυναικών μεγαλύτερης ηλικίας χρήζουν περεταίρω διερεύνησης.

- Κατηγορία HSIL: κύτταρα με αλλοιώσεις υψηλού βαθμού. Περιλαμβάνουν δυσπλασία CIN2, δυσπλασία CIN3 και καρκίνωμα in situ. Είναι απαραίτητη η πραγματοποίηση κολποσκόπησης, της οποίας τα ευρήματα σε σημαντικό ποσοστό είναι διηθητικός καρκίνος και σπανίως LSIL.
- Καρκίνωμα πλακωδών κυττάρων
- Καρκίνωμα αδενικών κυττάρων
- Άλλα κακοήθη νεοπλάσματα

Από το 2001 βρίσκεται σε χρήση το αναθεωρημένο σύστημα Bethesda, TBS 2001 [118], σύμφωνα με το οποίο οι κατηγορίες 1 και 2 συνενώθηκαν σε μια κατηγορία: «αρνητικά για ενδοεπιθηλιακή αλλοίωση ή κακοήθεια». Έτσι, οι αντιδραστικές μεταβολές χαρακτηρίζονται πιο καθαρά ως «αρνητικές». Οι αλλοιώσεις ASCUS διαχωρίστηκαν σε δύο κατηγορίες: ASCUS-καλοήθεια αλλοιώσεις και ASC-H (Atypical Squamous Cells with possible HSIL) ή «άτυπα πλακώδη κύτταρα που δεν αποκλείουν την ύπαρξη υψηλού κινδύνου βαθμού πλακώδους ενδοεπιθηλιακής βλάβης». Αν και τα κύτταρα δεν είναι φυσιολογικά, ο κυτταρολόγος δεν είναι σίγουρος για τη σοβαρότητα της αλλοίωσης. Το ASC-H παρουσιάζει μεγαλύτερο κίνδυνο για προκαρκινική κατάσταση. Έτσι το προτεινόμενο σύστημα κατηγοριοποίησης της κυτταρολογικής διάγνωσης διαμορφώνεται ως εξής:

1. Αποτέλεσμα αρνητικό για ενδοεπιθηλιακή αλλοίωση ή καρκίνωμα
2. Επίχρισμα με παρουσία επιθηλιακών κυττάρων με αλλοιώσεις ενδεικτικές ενδοεπιθηλιακής αλλοίωσης ή καρκινώματος

    α) πλακωδών κυττάρων

        - άτυπα πλακώδη κύτταρα:
            - § απροσδιόριστης σημασίας (ASCUS)
            - § δεν αποκλείουν HSIL (ASC-H)
        - LSIL (χαμηλού βαθμού πλακώδης ενδοεπιθηλιακή αλλοίωση): HPV αλλοιώσεις, CIN-1
        - HSIL (υψηλού βαθμού πλακώδης ενδοεπιθηλιακή αλλοίωση): CIN-2, CIN-3, καρκίνωμα in situ (CIS)
        - Καρκίνωμα εκ πλακωδών κυττάρων

    β) αδενικών κυττάρων

        - άτυπα αδενικά κύτταρα (AGC)

- άτυπα αδενικά κύτταρα μάλλον νεοπλασματικά
- ενδοτραχηλικό αδενοκαρκίνωμα in situ
- αδενοκαρκίνωμα

γ) άλλα κακοήθη νεοπλάσματα

Το TBS 2001 αποτελεί το επίσημο παρόν σύστημα για την κυτταρολογική διάγνωση του κολποτραχηλικού επιχρίσματος σε διάφορες χώρες και αποτελεί το αποδεκτό σχέδιο για την έκθεση κυτταρολογικής διάγνωσης, σύμφωνα με τις προτεινόμενες Ευρωπαϊκές Οδηγίες για τη Διασφάλιση Ποιότητας στον Πληθυσμιακό Έλεγχο Καρκίνου Τραχήλου της Μήτρας, έτσι ώστε να ενοποιηθεί η ορολογία.

Η αποτελεσματικότητα του PAP test, όπως και κάθε άλλης διαγνωστικής εξέτασης, δεν είναι καθολική. Αυτό μπορεί να οφείλεται στο μέγεθος (διαμέτρου < 0.5 cm) και στη θέση (ψηλά στον ενδοτραχηλικό σωλήνα) των CIN αλλοιώσεων, στην ανεπιτυχή λήψη δείγματος (απειρία, απουσία κατάλληλων εργαλείων, φτωχό δείγμα με απουσία ενδοτραχηλικών κυττάρων, κακή επίστρωση του επιχρίσματος στην αντικειμενοφόρο πλάκα, φτωχή μονιμοποίηση) και στην υποκειμενικότητα του κυτταρολόγου (ενδεχόμενη εσφαλμένη αξιολόγηση των αποτελεσμάτων). Αυτό καθιστά αναγκαίο τον τακτικό προσυμπτωματικό έλεγχο, καθώς και τη χρήση συμπληρωματικών τεχνικών για ελαχιστοποίηση του ενδεχομένου λανθασμένων αποτελεσμάτων. Στον ακόλουθο πίνακα παρουσιάζεται η διαδικασία ελέγχου ή θεραπείας που ακολουθείται ανάλογα με τα αποτελέσματα του ΠΑΠ τεστ (*πίνακας 18*).

Πίνακας 18: Αποτελέσματα Pap test και προτεινόμενη πορεία ελέγχου-θεραπείας

Αποτέλεσμα Pap test	**Προτεινόμενη πορεία ελέγχου-θεραπείας**
ASC-US	Κολποσκόπηση (κι αν χρειαστεί βιοψία)
ASC-H	Κολποσκόπηση και βιοψία
AGC	Κολποσκόπηση, βιοψία και ενδοτραχηλική διαγνωστική απόξεση
AIS	Κολποσκόπηση, βιοψία και ενδοτραχηλική διαγνωστική απόξεση
LSIL/CIN1/ήπια δυσπλασία	Κολποσκόπηση και βιοψία
HSIL/CIN2 ή CIN3/ μέτρια ή βαριά δυσπλασία/καρκίνωμα in situ	Κολποσκόπηση, βιοψία ή/και ενδοτραχηλική διαγνωστική απόξεση. Θεραπεία με Laser ή LEEP ή κωνοειδή εκτομή

4.3.1.2 Κυτταρολογία υγρής φάσης

Για την προετοιμασία του επιχρίσματος, τα τελευταία χρόνια, εφαρμόζεται η κυτταρολογία υγρής φάσης με λεπτή επίστρωση ή μονοστοιβάδωση (Liquid Based Cytology-LBC). Η τεχνική αυτή αναπτύχθηκε προκειμένου να απαλειφθούν οι κυριότερες αιτίες που περιορίζουν τη διαγνωστική

αξία του κλασικού τεστ Παπανικολάου. Βελτιστοποιώντας τη συλλογή και την προετοιμασία των κυττάρων αποσκοπεί στη μείωση των περιστατικών ψευδώς αρνητικών κυτταρολογικών ευρημάτων.

Η ευαισθησία της συμβατικής κυτταρολογίας για την ανίχνευση του καρκίνου του τραχήλου της μήτρας είναι κάτω από 50%. Οι κύριοι περιορισμοί του συμβατικού Παπ τεστ είναι οι ακόλουθοι: μόνο 20% των κυττάρων που συλλέγονται μεταφέρονται στο πλακίδιο, η σταθεροποίηση είναι ανεπαρκής λόγω του στεγνώματος στον αέρα που έχει ως επακόλουθο τον εκφυλισμό των κυττάρων, τα τυχόν ανώμαλα κύτταρα είναι τυχαία κατανεμημένα, παρατηρείται παρουσία ουσιών που μπορεί να επισκιάσουν τα ενδεχομένως ανώμαλα κύτταρα π.χ. αίμα και φλεγμονώδη κύτταρα [115]. Για να ξεπεραστούν αυτοί οι περιορισμοί έχει προταθεί η κυτταρολογία υγρής φάσης.

Πολλές μελέτες έχουν δείξει ότι η κυτταρολογία υγρής φάσης παρουσιάζει αρκετά πλεονεκτήματα σε σχέση με το συμβατικό τεστ Παπανικολάου επειδή επιτρέπει τη μεταφορά των ληφθέντων κυττάρων σε ένα υγρό μέσο συντήρησης, όπου τα κύτταρα μπορούν να διατηρηθούν σε θερμοκρασία δωματίου για αρκετές εβδομάδες. Με αυτό τον τρόπο η LBC δίνει τη δυνατότητα για επιπρόσθετα HPV tests και παράλληλα μειώνει τα ανεπαρκή δείγματα: με την LBC εξασφαλίζεται περισσότερο καθώς και πιο αντιπροσωπευτικό δείγμα αφού είναι εφικτό να αναπαραχθούν με ευκολία περισσότερα τους ενός δείγματα και στο υλικό που απομένει μπορούν να διενεργηθούν περαιτέρω έλεγχοι π.χ. HPV DNA test σε διφορούμενα δείγματα. Επιπλέον, τα επιχρίσματα κυτταρολογίας υγρής φάσης μικροσκοπούνται πιο γρήγορα και πιο εύκολα σε σχέση με τα επιχρίσματα της συμβατικής κυτταρολογίας, μειώνοντας έως και 40% το συνολικό διαγνωστικό χρόνο του εργαστηρίου.

Η LBC αποσκοπεί στην πλήρη αιμόλυση και βλεννόλυση του υλικού (αφαίρεση αίματος, φλεγμονών και βλέννας από το υλικό) και παραγωγή ενός δείγματος με μειωμένο αριθμό λευκών αιμοσφαιρίων και με αριθμό κυττάρων αντίστοιχο του κυτταρικού πληθυσμού που λήφθηκε από τον τράχηλο. Επειδή στην LBC τα κύτταρα κατανέμονται ομοιόμορφα και τυχαιοποιημένα σε μονοεπίπεδη στοιβάδα, η εκτίμηση και η ερμηνεία του κυτταρολογικού υλικού διευκολύνεται. Επιπλέον η μεταφορά των κυττάρων από το φιαλίδιο με το υγρό συντήρησης στο πλακίδιο είναι πιο αξιόπιστη από ότι στο συμβατικό Παπ τεστ.

Κατά την LBC, αφού τα κύτταρα αφαιρεθούν από την επιφάνεια του τραχήλου με τη χρήση ειδικής συσκευής λήψης κολποτραχηλικού επιχρίσματος σε μορφή βούρτσας, συλλέγονται σε ένα φιαλίδιο με υγρό συντήρησης (η συσκευή λήψης εμβυθίζεται στο υγρό συντήρησης). Η εμβύθιση της συσκευής λήψης στο υγρό συντήρησης επιτρέπει την ανάμιξη και την ομοιόμορφη κατανομή των κυττάρων στο διάλυμα [115]. Το φιαλίδιο, αφού σφραγιστεί, αποστέλλεται στο εργαστήριο κυτταρολογίας. Εκεί, το φιαλίδιο τοποθετείται σε ειδικό μηχάνημα επίστρωσης κυττάρων, στο

οποίο η επίστρωση των κυττάρων μετατρέπεται αυτόματα σε ομοιόμορφη/ομοιογενή μονοεπίπεδη στοιβάδα παρασκευάζοντας ένα επίχρισμα, το οποίο περιέχει αντιπροσωπευτική αναλογία των κυττάρων που ελήφθησαν από τον τράχηλο. Το ειδικό μηχάνημα περιλαμβάνει μια μεμβράνη, η οποία καθώς το υγρό περνά, συγκρατεί τα επιθηλιακά κύτταρα και τους μολυσματικούς οργανισμούς, ενώ επιτρέπει στο αίμα και στα φλεγμονώδη κύτταρα να περάσουν. Το λεπτό κυτταρολογικό υλικό που συλλέγεται στη μεμβράνη μεταφέρεται σε ένα γυάλινο πλακίδιο (σε κυκλική μορφή), όπου και μονιμοποιείται [115]. Έτσι, τα κύτταρα διατηρούνται αναλογικά στην αντικειμενοφόρο πλάκα και μικροσκοπούνται σε καθαρό υπόστρωμα. Το γυάλινο πλακίδιο μπορεί να εξεταστεί χειροκίνητα με μικροσκόπιο είτε με αυτοματοποιημένες τεχνικές. Το εναπομείναν υλικό του φιαλιδίου είναι διαθέσιμο για τυχόν χρήση σε περαιτέρω διαγνωστικές εξετάσεις, που ίσως χρειαστούν για αποσαφήνιση της διάγνωσης. Η αρχική μικροσκόπηση πραγματοποιείται από κυτταρολόγο που έχει εκπαιδευτεί στη ανίχνευση των άτυπων κυττάρων εκ των χιλιάδων φυσιολογικών που υπάρχουν στο δείγμα. Οι γυναίκες, στων οποίων το επίχρισμα εντοπίστηκαν μη φυσιολογικά κύτταρα, παραπέμπονται για επιπρόσθετες εξετάσεις και θεραπεία.

Έως τώρα, έχουν εγκριθεί από τον FDA 3 διαφορετικά συστήματα LBC:

- ThinPrep Processor (Cytyc, Hologic, Boxborough, MA, USA): εγκρίθηκε από το FDA το 1996, διαθέσιμα: ThinPrep 2000 και ThinPrep 3000
- BD SurePath Slide Processor (BD Diagnostics-TriPath, Burlington, NC, USA)): εγκρίθηκε από το FDA το 1999
- MonoPrep Processor (MonoGen, Inc., Chicago, IL, USA) tests): εγκρίθηκε από το FDA το 2006 [119]

Για το σύστημα ThinPrep είναι διαθέσιμες 3 διαφορετικές μηχανές: το ThinPrep 2000 Processor (TP2000), το ThinPrep 3000 Processor (TP3000) και το ThinPrep 5000 Processor (TP5000). Αν και το κάθε σύστημα από τα προαναφερθέντα ακολουθεί τεχνικά διαφορετική προσέγγιση από τα άλλα εγκριθέντα συστήματα, το τελικό προϊόν για όλα είναι το ίδιο: μια γυάλινη αντικειμενοφόρος πλάκα (glass Pap slide) με το κυτταρικό υλικό να είναι ομοιόμορφα κατανεμημένο σε μια μονοστοιβάδα (χωρίς την παρουσία αίματος ή φλεγμονωδών κυττάρων) [119].

Εικόνα 97: *Συμβατική κυτταρολογία και κυτταρολογία υγρής φάσης (τροποποιημένη εικόνα από [120])*

Οι απόψεις για τη μεγαλύτερη αξιοπιστία και εγκυρότητα της κυτταρολογίας υγρής φάσης σε σύγκριση με τη συμβατική κυτταρολογία διίστανται. Πολλές μελέτες υποστηρίζουν την υπεροχή της LBC έναντι του συμβατικού Pap test [121][122][123][124][125][126]. Πιο συγκεκριμένα, αναφέρεται ότι με τη χρήση του ThinPrep Pap test τα ακατάλληλα δείγματα μειώνονται σε ποσοστό άνω του 70%, το ποσοστό ψευδώς θετικών ASCUS διαγνώσεων μειώνεται κατά 25%, το ποσοστό ορθών διαγνώσεων LSIL αυξάνεται κατά 16-37% και το ποσοστό των HSIL αυξάνεται κατά 9-15% [121]. Σύμφωνα με τη μελέτη [122]**,** το SurePath Pap smear βρέθηκε να ξεπερνά το συμβατικό τεστ Παπανικολάου στην ανίχνευση κυτταρολογικών αλλοιώσεων HSIL+ και LSIL+ και επιπλέον φαίνεται να μειώνει τον αριθμό των μη ικανοποιητικών διαγνώσεων.

Παρά τις αναφερόμενες προοπτικές εφαρμογής της κυτταρολογίας υγρής φάσης, το συμβατικό Παπ τεστ συνεχίζει να χρησιμοποιείται ευρέως. Αυτό συμβαίνει εξαιτίας του χαμηλότερου κόστους του και λόγω της ήδη αποδεδειγμένης αξιοπιστίας του που έχει καταδειχτεί από πολλές μελέτες [127][128][129]. Σύμφωνα με τη μελέτη [128], η LBC δεν παρουσιάζει ούτε μεγαλύτερη ευαισθησία ούτε μεγαλύτερη ειδικότητα για την ανίχνευση HSIL+ σε σύγκριση με το CC. Επιπλέον, όσον αφορά το οικονομικό κόστος, σύμφωνα με τη μελέτη [130], το κλασικό τεστ Παπανικολάου είναι οικονομικά πιο συμφέρον από ότι η κυτταρολογία υγρής φάσης. Μόνο υπό συγκεκριμένες συνθήκες, η LBC καθίσταται πιο οικονομικά συμφέρουσα από το CC [130].

Στην Αγγλία και τη Δανία ο προληπτικός πληθυσμιακός έλεγχος γίνεται πλέον με την LBC. Επίσης, φαίνεται ότι η LBC προτιμάται από τη CC στα πλείστα εργαστήρια κυτταρολογίας στις ΗΠΑ [131]. Ωστόσο, η καθολική αντικατάσταση της CC από την LBC απαιτεί περισσότερες τυχαιοποιημένες μελέτες που να αποδεικνύουν τη διαγνωστική της υπεροχή έτσι ώστε να δικαιολογείται το υψηλότερο κόστος της [132].

4.3.1.3 Συσκευές αυτοματοποιημένης σάρωσης υλικού (μέσω υπολογιστών)

Τα τελευταία έτη, παρατηρείται η ολοένα αυξανόμενη χρήση μηχανημάτων αυτοματοποιημένης σάρωσης κυτταρολογικού υλικού, τα οποία διαθέτουν καλύτερη διαγνωστική ακρίβεια όταν αναλύουν επιχρίσματα, στα οποία παρατηρείται σχετικά μικρού βαθμού κυτταρική αλληλοεπικάλυψη. Εξαιτίας του ότι η οπτική εξέταση από κυτταρολόγο είναι χρονοβόρα και έχει υψηλό οικονομικό κόστος, είχαν ξεκινήσει να γίνονται διάφορες απόπειρες αυτοματοποίησης της ανάλυσης του κυτταρολογικού υλικού, από τότε ακόμα που είχε αρχικά προταθεί το τεστ Παπανικολάου [133]. Άλλος ένας λόγος που προέκυψε η ανάγκη για σάρωση υλικού με τη βοήθεια ηλεκτρονικού υπολογιστή, είναι επειδή παρατηρήθηκε ότι ο μονότονος έλεγχος μεγάλου αριθμού φυσιολογικών επιχρισμάτων από τον ίδιο κυτταρολόγο, μπορεί να οδηγήσει σε αδυναμία συγκέντρωσης, με αποτέλεσμα την εσφαλμένη ερμηνεία κάποιων αποτελεσμάτων.

Ειδικά για ένα αποτελεσματικό προσυμπτωματικό πρόγραμμα καρκίνου του τραχήλου της μήτρας, τα ψευδώς αρνητικά και τα ψευδώς θετικά αποτελέσματα κυτταρολογίας πρέπει να περιοριστούν στο ελάχιστο δυνατό αριθμό. Ο έλεγχος που γίνεται χειροκίνητα από κυτταρολόγους είναι μια εντατική και δύσκολη εργασία επειδή ο κυτταρολόγος πρέπει να αναγνωρίσει λίγα μη φυσιολογικά κύτταρα ανάμεσα σε χιλιάδες φυσιολογικά [115]. Η όλη διαδικασία αποκτά υποκειμενικό χαρακτήρα, αφού από κυτταρολόγο σε κυτταρολόγο μπορεί να παρατηρηθούν αποκλίσεις.

Οι αυτοματοποιημένες τεχνικές σάρωσης αποτελούν καλό εργαλείο για να μειωθεί ο φόρτος εργασίας των κυτταρολόγων και τα διαγνωστικά σφάλματα. Ωστόσο, ο ρόλος των συσκευών σάρωσης κυτταρολογικού υλικού μέσω ηλεκτρονικού υπολογιστή είναι βοηθητικός προς τον κυτταρολόγο και δεν τον αντικαθιστούν. Προσφέρουν τη δυνατότητα ελέγχου πολλαπλών επιχρισμάτων ανά ημέρα, μειώνουν τις εσφαλμένες εκτιμήσεις και διευκολύνουν την αναγνώριση των μη φυσιολογικών κυττάρων που ενδεχομένως να μην ανιχνεύονταν σε έναν έλεγχο ρουτίνας με το μικροσκόπιο. Χρησιμοποιώντας τα συστήματα αυτά βελτιώνεται και ο εντοπισμός των εσφαλμένα αρνητικών περιστατικών [119]. Με αυτές τις μεθόδους είναι εφικτή όχι μόνο η αρχική σάρωση των επιχρισμάτων, αλλά και ο έλεγχος ποιότητας (Quality Control-QS) της σάρωσης [115].

Οι αυτοματοποιημένες τεχνικές στηρίζονται στην τεχνολογία τεχνητών νευρωνικών δικτύων και είναι βασισμένες στην υπολογιστική απεικόνιση και αναγνώριση των μη φυσιολογικών κυττάρων του τραχήλου της μήτρας. Αυτές οι τεχνικές απαιτούν τη συλλογή κυτταρολογικού δείγματος σε υγρό μέσο και επεξεργασία του σε αυτοματοποιημένο σύστημα προετοιμασίας, έτσι ώστε να προκύψει μια λεπτή στρώση κυττάρων (κυτταρολογία υγρής φάσης) [115] .

Τα συστήματα αυτοματοποιημένης σάρωσης που είναι εγκεκριμένα από το FDA είναι τα ακόλουθα:

- PAPNET Testing System (Neuromedical Systems, Inc, Suffern, NY): Το PAPNET είναι ένα ημιαυτόματο σύστημα που αποτελείται από δύο φάσεις, τη φάση σάρωσης και τη φάση ελέγχου των τραχηλικών επιχρισμάτων [115]. Για την αναγνώριση των μη φυσιολογικών κυττάρων χρησιμοποιούνται τεχνητά νευρωνικά δίκτυα. Το ίδιο το σύστημα δεν επιχειρεί να διαγνώσει τα ανώμαλα κύτταρα, αντίθετα, επιλέγει και προβάλλει έως και 128 εικόνες μη φυσιολογικών κυττάρων σε οθόνη υψηλής ευκρίνειας με τη βοήθεια υπολογιστή -παρέχοντας επιπλέον τη δυνατότητα μεγέθυνσης- και αφήνει την αξιολόγηση των κυττάρων σε εκπαιδευμένους κυτταρολόγους [134]. Το σύστημα, σύμφωνα με τη μελέτη [134], αναγνωρίζει ανώμαλα κύτταρα σε ποσοστό 97% και μπορεί να βοηθήσει στη διάκριση προκαρκινικής αλλοίωσης από διηθητικό καρκίνωμα. Το αρνητικό του συστήματος PAPNET ήταν ότι απαιτούσε την αποστολή των αντικειμενοφόρων πλακών σε κεντρικές περιοχές αξιολόγησης με σταθμούς σάρωσης, όπου παράγονταν οι εικόνες, οι οποίες, στη συνέχεια, έπρεπε να αποσταλούν πίσω στο εργαστήριο κυτταρολογίας για να αξιολογηθούν σε οθόνες υψηλής ευκρίνειας [119]. Πλέον, έχει σταματήσει να εμπορεύεται στις ΗΠΑ εξαιτίας του υψηλού του κόστους για κάθε μη φυσιολογικό περιστατικό που εντοπίζεται καθώς και λόγω του τρόπου λειτουργίας του [119].
- AutoPap 300 QC (Quality Control) System, Neopath, Inc., Redmont, WA (σε συνδυασμό με τις BD SurePath slides είναι πλέον γνωστό ως BD FocalPoint Slide Profiler, BD Diagnostics-TriPath): Σε αντίθεση με το σύστημα PAPNET, είναι μια αυτόνομη μονάδα που βρίσκεται εντός του εργαστηρίου κυτταροπαθολογίας. Ωστόσο, όπως και το PAPNET, κάνει κι αυτό χρήση ταξινομητών τεχνητών νευρωνικών δικτύων [133]. Χρησιμοποιείται ευρέως στις ΗΠΑ για διάγνωση επιχρισμάτων και έλεγχο ποιότητας. Επιπλέον έχει τη δυνατότητα να αναγνωρίζει τα ανεπαρκή δείγματα. Σε κάθε πλακίδιο με επίχρισμα, με τη χρήση πολλαπλών αλγορίθμων, ανατίθεται μια τιμή που είναι βασισμένη στην πιθανότητα το συγκεκριμένο πλακίδιο να είναι φυσιολογικό, ανεπαρκές ή μη φυσιολογικό (τιμή από 0.0 έως 1.0). Τα κύτταρα ακολούθως, με βάση αυτή την τιμή, κατηγοριοποιούνται σε αυτά που απαιτούν «επισκόπηση» (review) και σε αυτά που απαιτούν «μη επισκόπηση» [115]. Στη συνέχεια, αν το πλακίδιο ανήκει στην κατηγορία που απαιτεί «επισκόπηση», κατατάσσεται σε μια από 5 κατηγορίες: όπου η κατηγορία 1 αντιστοιχεί στην κατηγορία πιο υψηλού κινδύνου και η κατηγορία 5 στην κατηγορία πιο χαμηλού κινδύνου. Με αυτό τον τρόπο γίνεται εκτίμηση του κινδύνου για κάθε πλακίδιο. Τα πλακίδια με τιμή πιθανότητας παρουσίας μη

φυσιολογικών κυττάρων κάτω από το κύριο κατώφλι δεν επισκοπούνται, τα πλακίδια με τιμή πάνω από το κύριο κατώφλι τυγχάνουν επισκόπησης από τον κυτταρολόγο και τα πλακίδια με τιμή πάνω από το κατώφλι ελέγχου ποιότητας (QC) επανασαρώνονται από τον κυτταρολόγο. Τα πλακίδια με τη μεγαλύτερη πιθανότητα παρουσίας μη φυσιολογικών κυττάρων μπαίνουν σε προτεραιότητα. Το σύστημα αυτό παρέχει οπτική επισκόπηση έως και 10 οπτικών πεδίων (Fields Of View - FOV). Συγκριτικά με τον χειροκίνητο έλεγχο από κυτταρολόγο, η αυτόματη σάρωση με το AutoPap βρέθηκε να είναι πιο ευαίσθητη στην αναγνώριση ASCUS και LSIL και το ίδιο ευαίσθητη στον εντοπισμό HSIL [115].

- ThinPrep Imaging System-TIS (Cytic Corporation, Hologic): Το σύστημα αυτό χρησιμοποιεί σάρωση υποβοηθούμενη από ηλεκτρονικό υπολογιστή και ένα αυτοματοποιημένο μικροσκόπιο, έτσι ώστε να εντοπίσει τα κύτταρα ενδιαφέροντος. Το TIS σαρώνει κάθε πλακίδιο και ιεραρχεί τις περιοχές ενδιαφέροντος με βάση το περιεχόμενο του DNA των κυττάρων (τα μη φυσιολογικά κύτταρα τείνουν να έχουν αυξημένες ποσότητες μοριακού DNA, κάτι που μπορεί να εξακριβωθεί με DNA χρώση των κυττάρων). Έτσι, όταν ένας έμπειρος κυτταρολόγος εξετάσει το πλακίδιο, οι περιοχές ειδικού ενδιαφέροντος θα είναι σαφώς σημασμένες για να δεχτούν ερμηνεία. Το TIS είναι μια τεχνολογία βασισμένη στην υπολογιστική απεικόνιση. Αποτελείται από ένα επεξεργαστή εικόνας που είναι συνδεδεμένος σε μια ή περισσότερες συσκευές επισκόπησης (review scopes). Η συσκευή επισκόπησης είναι ένα μικροσκόπιο που είναι συνδεδεμένο με ένα χειριστήριο που επιτρέπει τον εντοπισμό 22 διαφορετικών πεδίων (fields) ενδιαφέροντος. Αυτά τα πεδία σαρώνονται από τον κυτταρολόγο για την αναζήτηση παρουσίας μη φυσιολογικών κυττάρων χρησιμοποιώντας το αυτοματοποιημένο μικροσκόπιο. Σε περίπτωση απουσίας μη φυσιολογικών κυττάρων σε όλα τα πεδία, ο κυτταρολόγος μπορεί να επιβεβαιώσει τις αρνητικές ταξινομήσεις του TIS. Αλλιώς, σε περίπτωση που ο κυτταρολόγος εντοπίσει κύτταρα σε οποιοδήποτε πεδίο που φαίνονται ύποπτα, μπορεί να επανασαρώσει όλο το πλακίδιο και να σημάνει τις ομάδες των μη φυσιολογικών κυττάρων οι οποίες κατόπιν θα επανεξεταστούν από παθολόγο [135]. Ο ThinPrep Imager χρησιμοποιείται με όλα τα επιχρίσματα ThinPrep και βελτιώνει την ευαισθησία των ASCUS και των HSIL [136].

Τα κύρια αυτοματοποιημένα συστήματα σάρωσης είναι το BD FocalPoint Slide Profiler και το PAPNET, με το τελευταίο να έχει σταματήσει να χρησιμοποιείται. Σήμερα, χρησιμοποιούνται διαδραστικά αυτοματοποιημένα συστήματα σάρωσης. Ο σχεδιασμός αυτών των συστημάτων στηρίζεται στην αλληλεπίδραση μεταξύ των κύριων αυτοματοποιημένων συστημάτων σάρωσης

και του κυτταρολόγου [119]. Τα εγκεκριμένα από το FDA συστήματα που χρησιμοποιούνται σήμερα είναι το ThinPrep Imaging System και το BD FocalPoint GS Imaging System (Guided Screening-Καθοδηγούμενη σάρωση). Το BD FocalPoint GS Imaging System είναι ένας συνδυασμός του BD FocalPoint Slide Profiler και του BD FocalPoint GS Review Station.

Κάθε διαδραστικό σύστημα σάρωσης εμπεριέχει ένα σύστημα που σαρώνει τα πλακίδια, επεξεργάζεται τα κυτταρολογικά δεδομένα χρησιμοποιώντας αλγόριθμους απεικόνισης (imaging algorithms) και επισημαίνει στον κυτταρολόγο τα κυτταρικά πεδία (cellular fields) που θεωρεί σημαντικά, κάνοντας αυτόματη μετακίνηση των X-Y αξόνων, με τη βοήθεια αυτοματοποιημένου μικροσκοπίου [119]. Οι σαρωτές και τα μικροσκόπια έχουν ενσωματωμένες διάφορες συσκευές (π.χ. ποντίκι (mouse), διακόπτη ποδιού (foot switch), πληκτρολόγιο (keypad)), οι οποίες επιτρέπουν στον κυτταρολόγο να πλοηγηθεί εύκολα τόσο στην αρχική επισκόπηση των πεδίων ενδιαφέροντος, όσο και σε ολόκληρη την επισκόπηση του πλακιδίου, αν αυτό κριθεί απαραίτητο [119].

Τα διαδραστικά αυτοματοποιημένα συστήματα, διευκολύνουν την εργασία του κυτταρολόγου (ελαττώνουν την κόπωση και αυξάνουν την απόδοσή του), αφού μειώνουν την κυτταρική επιφάνεια που πρέπει να εξεταστεί από τον κυτταρολόγο σε κάθε αρνητικό πλακίδιο στο 70% περίπου. Έτσι, οδηγούν σε συνολική αύξηση της παραγωγικότητας του εργαστηρίου [119]. Ωστόσο, ο ανθρώπινος παράγοντας παραμένει, αφού τα αποτελέσματα συνεχίζουν να στηρίζονται στη διαγνωστική ερμηνεία του κυτταρολόγου. Τα οφέλη που προκύπτουν από αυτά τα συστήματα περιλαμβάνουν αυξημένη ευαισθησία για ανίχνευση της πλακώδους ενδοεπιθηλιακής βλάβης (SIL) και μείωση των ψευδώς αρνητικών περιστατικών. Τέτοια συστήματα φαίνεται να είναι ιδιαίτερα χρήσιμα σε χώρες όπου δεν υπάρχουν άμεσα διαθέσιμοι έμπειροι κυτταρολόγοι [119].

Πλέον, εκτός από το PAP test, χρησιμοποιούνται και κάποιες άλλες τεχνικές ανίχνευσης του καρκίνου του τραχήλου της μήτρας με κύρια τη διαγνωστική εξέταση HPV DNA, η οποία παρουσιάζεται ακολούθως.

4.3.2 HPV DNA test: Ανίχνευση νουκλεϊκού οξέος του ιού HPV

Η ανάγκη βελτίωσης των μεθόδων πρόληψης του καρκίνου του τραχήλου της μήτρας οδήγησε σε μια νέα εξέταση που βασίζεται στην ανίχνευση του ιού (HPV-DNA test ή απλώς HPV testing) και παρουσιάζει μεγαλύτερη ευαισθησία για την ανίχνευση προκαρκινικών αλλοιώσεων του τραχήλου της μήτρας συγκρινόμενο με τη συμβατική κυτταρολογία. Το HPV DNA test χρησιμοποιείται για την ανίχνευση των ογκογόνων τύπων HPV. Στην κλινική πράξη, συνήθως,

χρησιμοποιείται ως συμπληρωματική εξέταση του τεστ Παπανικολάου σε περίπτωση που τα αποτελέσματά του είναι αμφίβολα, δηλαδή, εάν εμφανιστούν κάποια μη φυσιολογικά κύτταρα.

Για την πραγματοποίηση της εξέτασης HPV DNA, απαιτείται, όπως και στο τεστ Παπανικολάου, η λήψη τραχηλικού επιχρίσματος: εισάγεται ένα βουρτσάκι στο έξω τραχηλικό στόμιο, περιστρέφεται για συλλογή βιολογικού υλικού και ακολούθως τοποθετείται σε ειδικό φιαλίδιο και αποστέλλεται στο εργαστήριο. Το φιαλίδιο διατηρείται σε θερμοκρασία 4-8^{o}C. Στις περιπτώσεις που το ΠΑΠ τεστ έχει ληφθεί με τη μέθοδο συλλογής σε υγρό (π.χ. Thin Prep), η εξέταση για αναζήτηση DNA από HPV γίνεται από το υγρό, που έχει μείνει στο φιαλίδιο, μετά τον αποχωρισμό κυττάρων για εξέταση στο μικροσκόπιο από κυτταρολόγο. Αυτό αποτελεί πλεονέκτημα, μια και δεν χρειάζεται να επισκεφθεί η ασθενής για δεύτερη φορά το γιατρό για λήψη νέου υλικού.

Η ανάλυση DNA βασίζεται σε μοριακές τεχνολογίες που μπορούν να ανιχνεύσουν το DNA του ιού HPV σε δείγματα κυττάρων από την περιοχή του τραχήλου της μήτρας. Η HPV λοίμωξη διαγιγνώσκεται κυρίως με μοριακές μεθόδους, διότι η καλλιέργεια του ιού δεν είναι εφικτή και δεν υπάρχουν ούτε διαθέσιμα αξιόπιστα ορολογικά εργαλεία - εξετάσεις που αφορούν την ανάλυση του ορού του αίματος με βάση την ανίχνευση αντισωμάτων στον ορό [137]. Οι ορολογικές αναλύσεις για HPV έχουν περιορισμένη ακρίβεια, επειδή τα αντισώματα προς την HPV λοίμωξη παραμένουν ανιχνεύσιμα για πολλά χρόνια, καθιστώντας τον οροδιαγνωστικό έλεγχο ακατάλληλο για τη διάκριση μεταξύ παρουσών και παρελθοντικών λοιμώξεων. Συνεπώς, η ακριβής διάγνωση μιας HPV λοίμωξης στηρίζεται στην ανίχνευση ιικού νουκλεϊκού οξέος [137]. Ωστόσο, η παρουσία του ιού HPV μπορεί να υποτεθεί από τα μορφολογικά, ορολογικά και κλινικά ευρήματα και ακολούθως να εξακριβωθεί εργαστηριακά με τεχνικές μοριακής βιολογίας [138]. Προς το παρόν, οι μοριακές μέθοδοι ανίχνευσης θεωρούνται η «χρυσή αρχή» (the gold standard) για την ακριβή ανίχνευση και την τυποποίηση του ιού HPV [139].

Η μοριακή ανίχνευση του DNA του ιού HPV μπορεί να γίνει με δύο τρόπους: είτε με μοριακές τεχνολογίες που δεν υφίστανται καμία ενίσχυση, όπως για παράδειγμα η εξέταση ανίχνευσης νουκλεϊκών οξέων με ιχνηθέτες είτε με μοριακές τεχνολογίες που εκμεταλλεύονται τη διαδικασία ενίσχυσης, όπως η αλυσιδωτή αντίδραση πολυμεράσης (PCR – polymerase chain reaction) [140]. Οι τεχνικές ενίσχυσης/πολλαπλασιασμού μπορούν να διαιρεθούν περαιτέρω σε τρεις επιμέρους κατηγορίες: την ενίσχυση στόχου (target amplification) όπου ενισχύονται τα νουκλεϊκά οξέα-στόχος π.χ. PCR, την ενίσχυση σήματος (signal amplification) όπου το σήμα που παράγεται από κάθε ιχνηθέτη ενισχύεται με τη βοήθεια ένωσης/ανιχνευτή και τέλος την ενίσχυση ιχνηθέτη (probe amplification) όπου το ίδιο το μόριο του ιχνηθέτη ενισχύεται μέσω μιας αντίδρασης [140]. Για την ανίχνευση του ιού HPV, έως τώρα, χρησιμοποιούνται τεχνικές ενίσχυσης στόχου και σήματος, καθώς και τεχνικές μη ενίσχυσης. Οι τεχνικές ανίχνευσης του HPV μπορούν να

διαχωριστούν αλλιώς, σε τρεις βασικές κατηγορίες: σε τεχνικές βασιζόμενες στον υβριδισμό, σε τεχνικές βασιζόμενες στην ενίσχυση νουκλεϊκών οξέων (PCR) και σε τεχνικές βασιζόμενες στον συνδυασμό τους.

Από τη δεκαετία του 1970 ως σήμερα έχει αναπτυχθεί μια μεγάλη ποικιλία μεθόδων ανίχνευσης και τυποποίησης του HPV. Στην αγορά, το 2013, υπήρχαν 175 διαφορετικά HPV DNA tests, με, όμως, πολύ λίγα από αυτά να έχουν πιστοποιηθεί από την υπηρεσία FDA (Food and Drug Administration) της Αμερικής ή την αντίστοιχη Ευρωπαϊκή εταιρεία EMEA (European Medicines Evaluation Agency) για κλινική εφαρμογή.

Οι κύριες ενδείξεις εφαρμογής του HPV DNA test στην κλινική πράξη είναι ο πληθυσμιακός έλεγχος (screening) για καρκίνο του τραχήλου της μήτρας, η διαχείριση και αξιολόγηση παθολογικών κυτταρολογικών αποτελεσμάτων χαμηλής ή απροσδιόριστης κλινικής σημασίας (ASCUS/LSIL) και η παρακολούθηση ασθενών μετά τη θεραπεία έτσι ώστε να αποφευχθεί το ενδεχόμενο υποτροπής. Στην περίπτωση του πληθυσμιακού ελέγχου, σύμφωνα με τις τελευταίες κατευθυντήριες γραμμές πρόληψης του καρκίνου του τραχήλου της μήτρας (2014), το HPV DNA test πραγματοποιείται είτε μόνο του είτε σε συνδυασμό με το τεστ Παπανικολάου. Οι επιλογές αυτές έχουν αξιολογηθεί σε μεγάλες επιστημονικές μελέτες, οι οποίες δείχνουν ότι η εξέταση αυτή παρουσιάζει μεγαλύτερη ευαισθησία από το τεστ Παπανικολάου και ο κίνδυνος ανάπτυξης σοβαρής προκαρκινικής αλλοίωσης CIN2,3 ή καρκίνου είναι πολύ μικρότερος, μετά από ένα αρνητικό HPV test παρά μετά από ένα αρνητικό τεστ Παπανικολάου. Τα αποτελέσματα αυτά είναι σημαντικά και διαμορφώνουν νέες προτάσεις για μεταβολές στα συστήματα πρόληψης του καρκίνου του τραχήλου της μήτρας, που ήδη εφαρμόζονται στην Αμερική και σε ορισμένες Ευρωπαϊκές χώρες. Στην περίπτωση της διαχείρισης και αξιολόγησης παθολογικών κυτταρολογικών αποτελεσμάτων χαμηλής ή απροσδιόριστης κλινικής σημασίας (LSIL/ ASCUS), παρατηρείται ότι η πραγματοποίηση του HPV DNA test σε ασθενείς με κυτταρολογική διάγνωση ASCUS προσφέρει μεγάλη ευαισθησία για την ανίχνευση υψηλόβαθμων επιθηλιακών αλλοιώσεων CIN2+, ενώ στα περιστατικά LSIL οι ενδείξεις για την πραγματοποίηση του HPV DNA test είναι λιγότερο σαφείς, λόγω της αυξημένης συχνότητας ανεύρεσης ιών υψηλού κινδύνου σε αυτή την κατηγορία κυτταρολογικών ευρημάτων. Τέλος, στις περιπτώσεις παρακολούθησης ασθενών μετά από θεραπεία, βρέθηκε επίσης, ότι το HPV DNA test είναι περισσότερο ευαίσθητο από το τεστ Παπανικολάου για τη διάγνωση υποτροπών τους πρώτους 24 μήνες μετά τη θεραπεία και μπορεί να χρησιμοποιηθεί σε συνδυασμό με την κολποσκόπηση για την παρακολούθηση αυτών των γυναικών [141].

Ο κίνδυνος ανάπτυξης σοβαρών προκαρκινικών αλλοιώσεων (CIN 2, 3) μετά από ένα αρνητικό HPV DNA test είναι εξαιρετικά χαμηλός. Το HPV DNA test προσφέρει αυξημένη προστασία έναντι της ανάπτυξης καρκίνου του τραχήλου της μήτρας σε σύγκριση με την προστασία που προσφέρει

ένα φυσιολογικό τεστ Παπανικολάου. Εντούτοις, ένα βασικό πρόβλημα της HPV DNA εξέτασης, παρά το γεγονός της υψηλής της ευαισθησίας, είναι η χαμηλή ειδικότητά της ~~για την ανίχνευση CIN2+~~, με χαρακτηριστική την περίπτωση θετικής εξέτασης HPV DNA και αρνητικής κυτταρολογικής εξέτασης. Σχετική μελέτη [142], επισήμανε τον αυξημένο κίνδυνο στις συγκεκριμένες περιπτώσεις, με το 15% από τις 2020 γυναίκες που συμμετείχαν, να αναπτύσσουν τραχηλική αλλοίωση εντός πέντε ετών.

Πλέον ζητούμενο, εκτός από τον προσδιορισμό της παρουσίας του ιού, αποτελεί και η τυποποίηση, δηλαδή ο ακριβής προσδιορισμός του γενότυπου του ιού (genotyping), αφού κάθε τύπος έχει διαφορετικό ογκογενετικό δυναμικό. Η τυποποίηση είναι προφανώς σημαντική, επειδή γνωρίζουμε ότι ανάλογα με τον τύπο ποικίλει και η πιθανότητα καρκινογένεσης π.χ. η επιμένουσα λοίμωξη από HPV-16 έχει σχετικά μεγάλη πιθανότητα καρκινογένεσης.

Το DNA του HPV μπορεί να εντοπιστεί σε επιχρίσματα του τραχήλου και σε δείγματα της βιοψίας με διάφορες μεθόδους, εκ των οποίων συμπληρωματική της κυτταρολογίας είναι η υβριδοποίηση/υβριδισμός in situ (in situ hybridization). Αυτή η μέθοδος στηρίζεται στη χρήση σημασμένων ανιχνευτών που υβριδοποιούνται ειδικά για ενδοκυτταρικό HPV DNA. Αν και η ευαισθησία αυτής της μεθόδου είναι περιορισμένη, επιτρέπει τον εντοπισμό της HPV λοίμωξης στο δείγμα καθώς και συνεντοπισμό από άλλους δείκτες (markers). Η ταυτοποίηση γενότυπων HPV απαιτεί τη χρήση ειδικών ανιχνευτών για συγκεκριμένους τύπους σε πολλαπλά πειράματα in situ υβριδοποίησης. Εναλλακτικά, το HPV-DNA μπορεί να απομονωθεί απευθείας από κλινικά δείγματα και να εντοπιστεί με υβριδισμό southern blot ή dot spot. Ωστόσο, τέτοιες προσεγγίσεις δεν είναι ευαίσθητες, απαιτούν εντατική εργασία και είναι ακατάλληλες για μεγάλης κλίμακας πληθυσμιακό έλεγχο. Ως εκ τούτου, έχουν αναπτυχθεί μέθοδοι ενίσχυσης/πολλαπλασιασμού νουκλεϊνικού οξέος (nucleic acid amplification methods) έτσι ώστε να αυξήσουν τόσο την ευαισθησία όσο και την ειδικότητα της ανίχνευσης του HPV DNA [137].

Για την αναζήτηση του DNA του ιού HPV συνήθως χρησιμοποιούνται το HPV τεστ και η τυποποίηση με PCR. Με το HPV-test προσδιορίζεται αν υπάρχει στον τράχηλο της μήτρας DNA από τους ογκογόνους τύπους HPV, χωρίς, όμως, να δίνεται με ακρίβεια ο συγκεκριμένος τύπος HPV. Αντίθετα, με την PCR μπορεί να εντοπιστεί ο συγκεκριμένος τύπος ή τύποι HPV που υπάρχουν στο δείγμα. Η PCR αποτελεί την πιο ευαίσθητη μέθοδο για την ανίχνευση του DNA του ιού HPV. Με τη μέθοδο αυτή μπορεί να ανιχνευτεί ακόμα και ελάχιστη ποσότητα ιικού DNA (10 DNA). Ωστόσο η PCR, επειδή είναι πολύ ευαίσθητη, δεν παρέχει πληροφορίες για το ιικό φορτίο όπως δίνει το HPV-test, το οποίο μπορεί να ανιχνεύσει 1000-5000 DNA.

4.3.2.1 Τεχνικές πολλαπλασιασμού σήματος

Η πιο διαδεδομένη τεχνική πολλαπλασιασμού/ενίσχυσης σήματος αποτελεί η μη ραδιενεργή τεχνική συγκράτησης υβριδίων ή αλλιώς υβριδισμός σε διάλυμα (Hybrid Capture-HC). Έχει αναπτυχθεί από την Digene Corporation (USA) και αποσκοπεί στην ανίχνευση νουκλεϊκών οξέων, με την ενίσχυση σήματος να βελτιώνει τη ευαισθησία της μεθόδου. Υφίστανται δύο διαφορετικά τεστ στην αγορά, το πρώτης γενιάς Hybrid Capture Tube (HCT) test και το πιο πρόσφατο Hybrid Capture II (HCII). Το HCT test που εγκρίθηκε από την US FDA το Μάιο του 1995, ανιχνεύει την παρουσία των υψηλού κινδύνου τύπων 16, 18, 31, 33, 35, 45, 51, 52 και 56, ενώ το δεύτερης γενεάς HC2 πήρε την έγκριση το Μάρτιο του 1999 και πρόσθεσε στην ομάδα των προς ανίχνευση υψηλού κινδύνου τύπων τους 39, 58, 59 και 68 [143]. Το HC2 είναι η μόνη μέθοδος που έχει πάρει έγκριση για χρήση στον προληπτικό έλεγχο σε γυναίκες άνω των 30 ετών. Το όριο ανιχνευσιμότητας του HC-2 ανέρχεται σε 5000 αντίγραφα ιού ανά δείγμα ή σε 1pg/ml , σε αντίθεση με του HCT που είναι 10pg/ml [137], [139], [140].

Το HC2, μαζί με το PCR (που θα αναλυθεί παρακάτω), είναι οι μέθοδοι που χρησιμοποιούνται κυρίως σήμερα στην κλινική πράξη. Με τη μέθοδο HC2, οι διπλές έλικες DNA των υπό εξέταση κυττάρων, διαχωρίζονται σε απλές έλικες και ακολούθως προστίθενται μικρά σημασμένα τμήματα RNA (probes), τα οποία είναι συμπληρωματικά κάποιων τμημάτων του γονιδιώματος του HPV. Πρακτικά, το υλικό συλλέγεται με ψήκτρα, τοποθετείται σε ειδικό υγρό και υφίσταται υβριδισμό με μείγμα RNA μονόκλωνων ανιχνευτών από 18 διαφορετικούς τύπους. Τα RNA-DNA υβρίδια που δημιουργούνται, ανιχνεύονται με φωτομετρία (χημειοφωταύγεια) αφού πρώτα ακινητοποιηθούν με τη βοήθεια μονοκλωνικών αντισωμάτων. Η HC-2 με τη χρήση δύο διαφορετικών κοκτέιλ ανιχνευτών (probe coctails) μπορεί να ανιχνεύσει την παρουσία 13 συνολικά υψηλού κινδύνου τύπων (16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59 και 68), καθώς και 5 χαμηλού κινδύνου τύπων (6, 11, 42, 43 και 44) [137].

Συγκριτικά με το συμβατικό PAP test, το τεστ HC2 επιδεικνύει υψηλότερη ευαισθησία (88.4 έναντι 77.7%) στην περίπτωση των HSIL, αλλά χαμηλότερη ειδικότητα (89 έναντι 94%) [140]. Η μέθοδος αυτή παρουσιάζει υψηλή ευαισθησία εξαιτίας των ακόλουθων λόγων: οι δεσμοί DNA/RNA είναι πιο σταθεροί από τους δεσμούς DNA/DNA, οι RNA ανιχνευτές χρησιμοποιούνται για υβριδισμό ολόκληρου του γονιδιώματος του ιού και τα μονοκλωνικά αντισώματα που συνδέονται με μόρια αλκαλικής φωσφατάσης έχουν την ικανότητα να αναγνωρίζουν μικρές περιοχές από RNA-DNA υβριδισμούς. Επίσης, κάθε ακινητοποιημένο ένζυμο αλκαλικής φωσφατάσης αντιδρά με μόρια ενός ειδικού χημειοφωτοευαίσθητου υποστρώματος με αποτέλεσμα να παράγεται μια σταθερή ροή φωτονίων, τα οποία καταμετρούνται από ένα ειδικό φωτόμετρο, του οποίου η ένταση του φωτός μεταφράζεται σε ιικό φορτίο.

Το HC2, ωστόσο, παρουσιάζει και κάποιους περιορισμούς: ενώ διακρίνει μεταξύ των ομάδων υψηλού κινδύνου και χαμηλού κινδύνου τύπων, δεν επιτρέπει την τυποποίηση συγκεκριμένων HPV γονοτύπων [137]. Αυτό αποτελεί και το βασικό μειονέκτημα της μεθόδου αυτής σε σχέση με τη μέθοδο PCR. Επίσης το κατώτερο όριο ανιχνευσιμότητας του HC2 (5000 αντίγραφα ιού ανά δείγμα), καθιστά τη μέθοδο αυτή λιγότερο ευαίσθητη από τη PCR. Ένα ακόμα μειονέκτημα της HC2 είναι ο κίνδυνος για διασταυρούμενη αντίδραση μεταξύ των δύο κοκτέιλ ανιχνευτών που μπορεί να μειώσει την κλινική σημασία ενός θετικού αποτελέσματος [137].

4.3.2.2 Τεχνικές ενίσχυσης/πολλαπλασιασμού στόχου νουκλεϊκών οξέων

Οι τεχνικές πολλαπλασιασμού στόχου (target amplification systems) βασίζονται σε μια εργαστηριακή διαδικασία κατά την οποία αντιγράφονται τμήματα της ακολουθίας DNA ενός γονιδίου στόχου, παρέχοντας έτσι συγκεντρωμένα δείγματα μιας συγκεκριμένης γενετικής ακολουθίας. Υπάρχουν διάφορες τεχνολογίες που χρησιμοποιούνται με βάση αυτή την τεχνική. Η αλυσιδωτή αντίδραση της πολυμεράσης (PCR – Polymerase Chain Reaction) είναι η πιο διαδεδομένη από αυτές και η περισσότερο εφαρμόσιμη όσον αφορά την ανίχνευση του ιού HPV [140]. Η PCR χρησιμοποιείται τόσο για ανίχνευση της παρουσίας του HPV όσο και για την τυποποίηση του ιού.

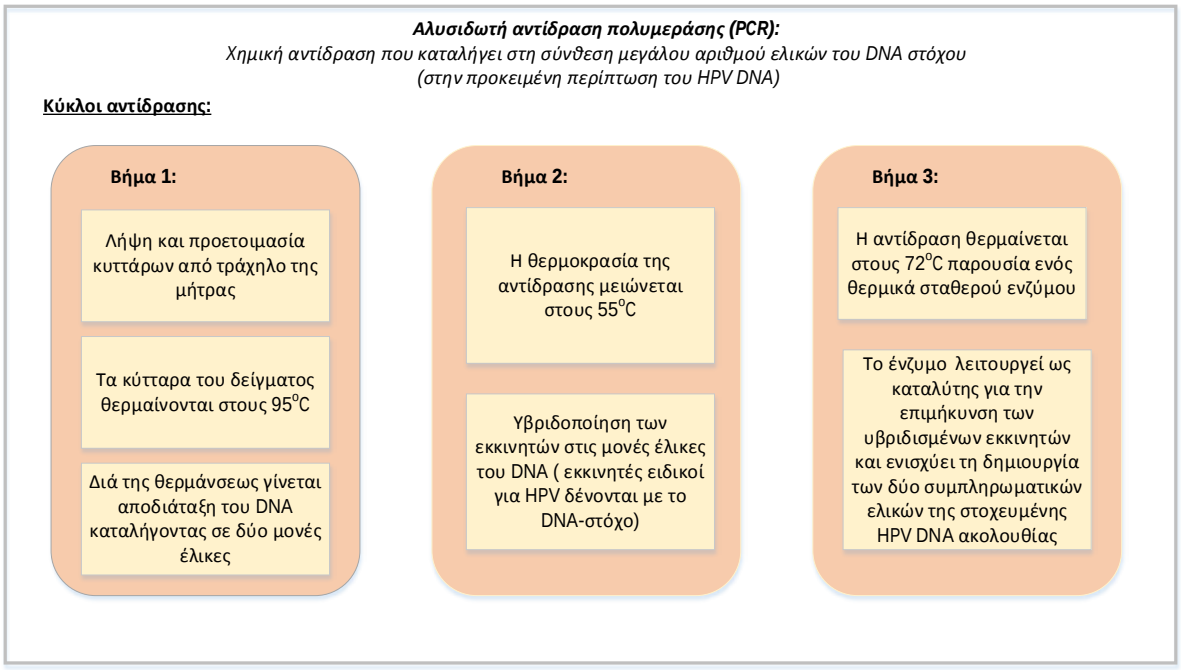
Πρόκειται για μια θερμοκυκλική μέθοδο πολυμεράσης (οι DNA πολυμεράσες είναι μια σειρά από ένζυμα που καταλύουν το πολυμερισμό των δεοξυριβονουκλεοτιδίων σε μια αλυσίδα DNA και έχουν σημαντικό ρόλο στην αντιγραφή του DNA, όπου μια αρχική αλυσίδα DNA χρησιμοποιείται ως «καλούπι» για τη σύνθεση μιας νέας, συμπληρωματικής με τη μητρική αλυσίδα), κατά την οποία, από κατάλληλο βιολογικό υγρό ανιχνεύουμε και πολλαπλασιάζουμε in vitro συγκεκριμένα τμήματα γενετικού υλικού, στην προκειμένη περίπτωση, ιικού HPV DNA , σε διαδοχικούς κύκλους (αλυσιδωτή αντίδραση). Αυτό επιτυγχάνεται με τη χρήση ενός θερμικά σταθερού ενζύμου και ειδικών ολιγονουκλεοτιδικών εκκινητών της αντίδρασης (primers: ειδικά νουκλεοτίδια τα οποία παρουσιάζουν αλληλουχία βάσεων συμπληρωματικά του DNA στόχου), τα οποία προσδιορίζουν την περιοχή ενδιαφέροντος του DNA [139]. Η PCR αποσκοπεί στον πολλαπλασιασμό συγκεκριμένων τμημάτων του DNA έτσι ώστε αυτά να γίνουν ανιχνεύσιμα: το ιικό DNA ενισχύεται αρκετά in vitro έτσι ώστε να παράξει επαρκή ποσότητα του στόχου, η οποία ακολούθως οπτικοποιείται επί πηκτώματος αγαρόζης (agarose gel).

Επιγραμματικά, με την PCR πραγματοποιείται εκλεκτική παραγωγή μεγάλου αριθμού συγκεκριμένου τμήματος HPV-DNA που υπάρχει σε μείγμα νουκλεϊκών οξέων με τη χρήση συνθετικών εκκινητών. Προφανώς, απαραίτητη προϋπόθεση είναι να είναι γνωστή η αλληλουχία των νουκλεοτιδίων του επιθυμητού γονιδίου. Θεωρητικά, η PCR μπορεί να ανιχνεύσει και να

αντιγράψει την ακολουθία-στόχο σε οποιοδήποτε δοθέν δείγμα. Πρακτικά, η βασισμένη σε PCR χρησιμοποιούμενη μέθοδος έχει ευαισθησία να εντοπίσει και να αντιγράψει περίπου 10-100 HPV ιικά γονιδιώματα ανάμεσα σε 100ng κυτταρικού DNA [139].

Η PCR μέθοδος περιλαμβάνει επαναλαμβανόμενους κύκλους ενίσχυσης επιλεγμένων αλληλουχιών νουκλεϊκών οξέων. Απαιτείται η παρουσία ενός ζεύγους συνθετικών ολιγονουκλεοτιδικών εκκινητών που προσδένονται στους δύο κλώνους της αλληλουχίας-στόχου στην πλησιέστερη φυσική απόσταση [144]. Κάθε κύκλος ενίσχυσης αποτελείται από τρία στάδια. Στο πρώτο στάδιο, γίνεται αποδιάταξη της διπλής έλικας του DNA (denaturation) σε υψηλή θερμοκρασία και ενεργοποίηση της πολυμεράσης. Στο δεύτερο στάδιο, πραγματοποιείται η υβριδοποίηση (annealing) των εκκινητών στις μονές έλικες του DNA (οι εκκινητές υβριδοποιούνται με τις συμπληρωματικές αλληλουχίες στο μόριο-στόχο τους). Η υβριδοποίηση των εκκινητών πραγματοποιείται σε χαμηλότερη θερμοκρασία προκειμένου να χρησιμοποιηθούν οι εκκινητές που στοχεύουν την L1 περιοχή του ιικού γονιδιώματος. Στο τρίτο στάδιο, γίνεται επιμήκυνση (elongation) της αντίδρασης, στην οποία η DNA πολυμεράση επιμηκύνει τις αλληλουχίες των υβριδισμένων εκκινητών, δηλαδή συνθέτει τις συμπληρωματικές έλικες στις μονές έλικες του DNA που δημιουργήθηκαν στο πρώτο στάδιο. Πρακτικά με την PCR, οι διπλές έλικες DNA των υπό εξέταση κυττάρων, διαχωρίζονται σε απλές έλικες με θέρμανση και στη συνέχεια προστίθενται σε διάλυμα που περιέχει μικρά σημασμένα τμήματα DNA (primers), τα οποία είναι συμπληρωματικά τμημάτων, συνήθως της L1 περιοχής του γονιδιώματος του HPV.

Στο τέλος κάθε κύκλου, η ποσότητα των προϊόντων της PCR θεωρητικά διπλασιάζεται. Συνεπώς, αφού πραγματοποιηθούν ν κύκλοι ενίσχυσης, θεωρητικά θα προκύψουν 2^{v} αντίγραφα της υπό εξέταση αλληλουχίας DNA - στόχου. Η όλη διαδικασία διεξάγεται σε έναν προγραμματισμένο θερμικό κυκλοποιητή και ολοκληρώνεται μετά από 30-50 κύκλους, με αποτέλεσμα την αύξηση του συνολικού αριθμού των αντιγράφων της αλληλουχίας στόχου [144]. Το αποτέλεσμα γίνεται αντιληπτό με ανοσοφθορισμό ή με ηλεκτροφόρηση των προϊόντων της PCR παρουσία βρωμιούχου αιθιδίου (ή κάποιου άλλου ιχνηθέτη) και οπτικοποίησή τους σε θάλαμο ακτινών UV. Στη συνέχεια μπορεί να γίνει και τυποποίηση, δηλαδή ανίχνευση του ακριβούς τύπου του HPV. Οι κύκλοι της αντίδρασης PCR παρουσιάζονται στην *εικόνα 98*.



Εικόνα 98: *Κύκλοι αλυσιδωτής αντίδρασης πολυμεράσης*

Γενικά, η ανίχνευση του HPV μέσω PCR, μπορεί να πραγματοποιηθεί είτε με ειδικούς εκκινητές σχεδιασμένους για να ενισχύουν αποκλειστικά ένα συγκεκριμένο γονότυπο HPV (type-specific primers), είτε με γενικά ζεύγη PCR εκκινητών (general/consensus PCR primer pairs), σχεδιασμένα για να ενισχύουν ένα πιο ευρύ φάσμα HPV γονοτύπων [139]. Η ανίχνευση HPV DNA σε ένα μόνο δείγμα με χρήση πολλαπλών αντιδράσεων PCR με ειδικούς εκκινητές για συγκεκριμένους γονότυπους είναι χρονοβόρα και έχει μεγάλο κόστος, σε αντίθεση με τη χρήση γενικών εκκινητών, καθιστώντας την τελευταία ως μια πιο βολική επιλογή. Υπάρχουν διάφοροι γενικοί PCR εκκινητές που μπορούν να χρησιμοποιηθούν.

Αρχικά, με τη χρήση γενικών PCR εκκινητών γίνεται η ταυτοποίηση τουλάχιστον 25 HPV τύπων. Ακολούθως λαμβάνει χώρα η διαδικασία για την τυποποίηση των συγκεκριμένων HPV τύπων. Αυτό γίνεται με επώαση του προϊόντος της αντίδρασης PCR με περιοριστικά ένζυμα (RFLP-Restriction/Fragment Length Polymorphism), τα οποία τέμνουν το προϊόν σε διάφορα σημεία, με αποτέλεσμα τη δημιουργία τμημάτων DNA διαφορετικού μεγέθους. Η τυποποίηση πραγματοποιείται με βάση το μέγεθος των τμημάτων DNA που προκύπτουν, καθώς και με τη χρήση ειδικών για κάθε HPV τύπο εκκινητών.

Ένα τεστ που χρησιμοποιεί τη συγκεκριμένη μέθοδο και κυκλοφορεί ευρέως στο εμπόριο είναι το Linear Array HPV Genotyping PCR Test (Roche Diagnostics), το οποίο χρησιμοποιεί την ενίσχυση του DNA στόχου με PCR και υβριδισμό νουκλεϊκού οξέος για την ανίχνευση και ταυτοποίηση 37 διαφορετικών γονότυπων HPV (6, 11, 16, 18, 26, 31, 33, 35, 39, 40, 42 , 45, 51, 52, 53, 54, 55, 56,

58, 59, 61, 62, 64, 66, 67, 68, 69, 70, 71, 72, 73 (MM9), 81, 82 (MM4), 83 (MM7), 84 (MM*), IS39 και CP108) σε κύτταρα τραχήλου. Σαφώς, σε αυτούς τους γονότυπους συμπεριλαμβάνονται και οι 13 γνωστοί με βάση τη σύγχρονη βιβλιογραφία, γονότυποι υψηλού κινδύνου (16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59 και 68).

Η PCR παρουσιάζει αρκετά πλεονεκτήματα σε σχέση με άλλες μεθόδους. Το βασικό της πλεονέκτημα αποτελεί η υψηλή της ευαισθησία: επειδή η PCR μπορεί να εφαρμοστεί ακόμα και σε πολύ μικρές ποσότητες ιικού DNA (10-100ng), καθίσταται κατάλληλη για χρήση σε δείγματα με μικρή περιεκτικότητα ιικού DNA [139]. Άλλο πλεονέκτημα της μεθόδου είναι ο μικρός χρόνος ανίχνευσης του παθογόνου (στην προκειμένη περίπτωση του ιού HPV) σε σχέση με άλλες μεθόδους. Τέλος, παρουσιάζει υψηλή ειδικότητα, αφού ο πολλαπλασιασμός γίνεται αποκλειστικά για συγκεκριμένο DNA. Οποιοδήποτε άλλο γενετικό υλικό δεν πολλαπλασιάζεται και δεν ανιχνεύεται. Έτσι είναι εφικτή η ακριβής τυποποίηση του παθογόνου, δηλαδή στην περίπτωση αυτή, ο προσδιορισμός του συγκεκριμένου τύπου του ιού HPV. Με αυτό τον τρόπο η PCR όχι μόνο ανιχνεύει την παρουσία του HPV στο εξεταζόμενο δείγμα (με αποτέλεσμα αρνητικό/θετικό), αλλά παράλληλα παρέχει και τη δυνατότητα τυποποίησής του στους αντίστοιχους τύπους υψηλής ή χαμηλής ογκογενετικής ικανότητας. Ως αποτέλεσμα, εξαλείφεται η περίπτωση διασταυρούμενων αντιδράσεων ή ψευδώς θετικών/ αρνητικών αποτελεσμάτων. Η PCR, ωστόσο, παρουσιάζει και ένα βασικό μειονέκτημα: αδυνατεί να διακρίνει μεταξύ λανθάνουσας και ενεργού λοίμωξης ή να προσδιορίσει ποσοτικά το ιικό φορτίο. Επίσης, αν και η PCR θεωρείται ως η μέθοδος με τη μεγαλύτερη ευαισθησία και ακρίβεια, απαιτεί εμπειρία από το μοριακό βιολόγο. Άλλο ένα μειονέκτημα της μεθόδου είναι ότι στην παρούσα φάση δεν ενδείκνυται για εφαρμογή σε προγράμματα προληπτικού πληθυσμιακού ελέγχου ευρείας κλίμακας εξαιτίας του προς το παρόν αυξημένου της κόστους (αυτό ενδεχομένως να αλλάξει).

Προς επίλυση των παραπάνω προβλημάτων άρχισε να χρησιμοποιείται η ποσοτική ή αλλιώς PCR πραγματικού χρόνου (quantive PCR/q-PCR, real time PCR/R-T PCR). Με αυτή τη μέθοδο, όπου η ενζυμική αντίδραση παρακολουθείται συνεχόμενα σε πραγματικό χρόνο, επιτυγχάνεται η ανίχνευση του HPV DNA, η τυποποίηση του ιού, ο ποσοτικός προσδιορισμός του ιικού φορτίου καθώς και ο καθορισμός της κατάστασης του ιικού DNA (ενσωματωμένο ή επισωματικό). Επιπλέον, με παράλληλη χρήση R-T PCR που χρησιμοποιεί ως υλικό εκκίνησης ιικό RNA, είναι εφικτή η διάκριση ανάμεσα στην ενεργό και στη λανθάνουσα λοίμωξη, μια και το ιικό RNA εντοπίζεται μόνο στην ενεργό λοίμωξη. Το κόστος εφαρμογής, ωστόσο, παραμένει ανασταλτικός παράγοντας. Στο πρώτο σύστημα PCR πραγματικού χρόνου χρησιμοποιείτο βρωμιούχο αιθίδιο και μια κάμερα CCD για την παρακολούθηση της εξέλιξης των αντιδράσεων ενίσχυσης μέσα σε κλειστό σωλήνα αντίδρασης. Σήμερα, υπάρχουν αρκετοί ιχνηθέτες και πολλές φθορίζουσες χρωστικές που χρησιμοποιούνται στην R-T PCR.

Η μοριακή ανίχνευση του DNA του ιού HPV με την τεχνολογία PCR είναι αποδεδειγμένα από τις πιο ευαίσθητες και αξιόπιστες μη επεμβατικές μεθόδους εντοπισμού ενεργής τραχηλικής HPV λοίμωξης. Η απομόνωση της θερμοσταθερής DNA πολυμεράσης και η εφεύρεση της αλυσιδωτής αντίδρασης πολυμεράσης (PCR), όχι μόνο έχουν απλουστεύσει και επιταχύνει την in vitro ενίσχυση του DNA, αλλά επέτρεψαν την ανάπτυξη μιας σειράς μοριακών εργαλείων, τα οποία είναι χρήσιμα για τη γενετική τυποποίηση των παθογόνων. Η κατοχύρωση της PCR με δίπλωμα ευρεσιτεχνίας οδήγησε στην ανάπτυξη εναλλακτικών τεχνικών που αποτελούν διαφορετικές προσεγγίσεις βελτιστοποίησης της ενίσχυσης των νουκλεϊκών οξέων, όπως για παράδειγμα η NASBA (nucleic acid sequence-based amplification) [144].

4.3.2.3 Τεχνικές μη πολλαπλασιασμού/ άμεσου DNA υβριδισμού

Οι τεχνικές μη πολλαπλασιασμού βασίζονται σε μοριακές διαγνωστικές μεθόδους και συμπεριλαμβάνουν τεχνικές άμεσου υβριδισμού όπως οι Southern blot, dot blot και in situ. Η ενσωμάτωση των μεθόδων άμεσου υβριδισμού σε προγράμματα μαζικού προληπτικού ελέγχου, εμποδίζεται από παράγοντες όπως η χαμηλή ευαισθησία των μεθόδων αυτών, ο μεγάλος χρόνος υλοποίησής τους, η ανάγκη για ειδικά εκπαιδευμένους τεχνικούς, καθώς και η απαίτηση υψηλής τεχνολογίας η οποία κατ' επέκταση αυξάνει το κόστος εξοπλισμού.

Ο DNA υβριδισμός είναι βασισμένος στην αρχή της συμπληρωματικότητας των βάσεων μεταξύ δύο αλυσίδων DNA και στη δυνατότητά τους να διαχωρίζονται κάτω από συνθήκες υψηλής θερμοκρασίας ή pH. Ο ιός HPV τυποποιείται χρησιμοποιώντας σημασμένο DNA ως ανιχνευτή (probe). Με τον τρόπο αυτό δημιουργείται ένα υβρίδιο, δηλαδή ένα δίκλωνο DNA , ο ένας κλάδος του οποίου είναι το υπό ανίχνευση DNA και ο άλλος κλάδος του ο ανιχνευτής. Συνεπώς, αν σε ένα τυχαίο δείγμα υπάρχει HPV DNA, με την προσθήκη του γνωστού μονόκλωνου σημασμένου τμήματος DNA, θα σχηματιστεί ένα υβρίδιο που θα είναι κι αυτό σημασμένο καθιστώντας ανιχνεύσιμη την παρουσία του ιού. Οι ανιχνευτές σημαίνονται με ραδιενεργά νουκλεοτίδια ή με άλλες ενώσεις όπως π.χ. ένζυμα, εκ των οποίων η πιο ευαίσθητη σήμανση είναι η ραδιενεργός. Η μέθοδος του DNA υβριδισμού είναι απλή, σύντομη και εύκολα εφαρμόσιμη. Ακολούθως αναλύονται οι βασικές τεχνικές άμεσου υβριδισμού.

Υβριδισμός Southern blot: Με τη μέθοδο southern blot το κυτταρικό DNA απομονώνεται από το βιολογικό υλικό του τραχήλου της μήτρας (κυτταρολογικό ή ιστολογικό) και ακολούθως υφίσταται επεξεργασία με τη χρήση περιοριστικών ενζύμων. Πιο συγκεκριμένα, το γονιδίωμα του ιού HPV εξάγεται από το δείγμα και η DNA αλυσίδα του διασπάται με τη χρήση των ενζύμων. Ακολούθως, τα παραγόμενα DNA τμήματα διαχωρίζονται με ηλεκτροφόρηση και στη συνέχεια προστίθεται DNA ανιχνευτής.

Η ηλεκτροφόρηση υλοποιείται ως εξής: το παραγόμενο προϊόν από το προηγούμενο στάδιο συνενώνεται με τη μορφή ενός gel το οποίο τροφοδοτείται με ηλεκτρικό ρεύμα (gel electrophoresis) και έχει ως συνέπεια το διαχωρισμό της DNA ακολουθίας σε διαφορετικά τμήματα. Αφού τα DNA τμήματα διαχωριστούν, οδηγούνται σε μεμβράνη νιτροκυτταρίνης και υβριδοποιούνται με ιχνηθέτες (ανιχνευτές) του HPV γονιδιώματος. Οι ιχνηθέτες αναγνωρίζονται συνήθως μέσω ραδιοϊσοτόπων. Τυχόν αναγνώριση συνεπάγεται την παρουσία του ιού στο υπό εξέταση δείγμα [140].

Η μέθοδος southern blot έχει ως βασικό μειονέκτημα ότι για την υλοποίησή της απαιτείται εξειδικευμένο εργαστηριακό περιβάλλον. Επειδή αποτελεί μια χρονοβόρα και δύσχρηστη μέθοδο χρησιμοποιείται μόνο για ερευνητικούς σκοπούς. Η ευαισθησία της μεθόδου κυμαίνεται μεταξύ 0.1-0.01 αντιγράφων ιικού γονιδιώματος ανά κύτταρο και απαιτεί ποσότητα ολικού κυτταρικού DNA της τάξεως των 10ng.

Υβριδισμός Dot blot: Η μέθοδος dot blot αποτελεί παραλλαγή της μεθόδους southern blot. Το DNA σε αυτή την τεχνική αποδιατάσσεται χωρίς τη χρήση ηλεκτροφόρησης και ακολούθως γίνεται υβριδισμός σε δύο φάσεις με τη χρήση μίγματος ανιχνευτών. Η ευαισθησία της μεθόδους αυτής είναι χαμηλή και πλέον δεν χρησιμοποιείται [140].

Υβριδισμός in situ (In Situ Hybridization- ISH): Η μέθοδος αυτή είναι η μόνη που στηρίζεται στην απευθείας ανίχνευση και τυποποίηση HPV-DNA πάνω στα κυτταρικά επιχρίσματα ή σε ιστολογικές τομές. Πρόκειται για μια τεχνική που εφαρμόζει τεχνικές υβριδοποίησης στο άθικτο DNA των υπό λοίμωξη κυττάρων και λαμβάνει χώρα στην αντικειμενοφόρο πλάκα ενός μικροσκοπίου [140]. Αρχικά πραγματοποιείται η προεπεξεργασία του δείγματος, κατά την οποία απομακρύνονται τα κυτταρικά συστατικά τα οποία δεν ανήκουν στο DNA στόχο. Στη συνέχεια, το δείγμα θερμαίνεται έτσι ώστε να επιτευχθεί η αποδόμηση του DNA και εφαρμόζονται ιχνηθέτες οι οποίοι δεσμεύουν το HPV DNA, δεδομένου ότι αυτό είναι παρόν. Ακολούθως εισάγονται αντισώματα που προσδένονται στους ιχνηθέτες και προστίθενται ένζυμα που εάν υπάρχει ο ιός χρωματίζουν το δείγμα. Η αναγνώριση θετικών ή αρνητικών ευρημάτων επιτυγχάνεται οπτικώς μέσω του μικροσκοπίου. Η τεχνική αυτή παρουσιάζει μειωμένη ευαισθησία: 10-20% αντίγραφα του HPV γονιδιώματος ανά κύτταρο. Ένα kit ανίχνευσης για χρήση στον εντοπισμό HPV DNA με την τεχνική υβριδισμού in situ έχει αναπτυχθεί από την Kreatech Biotechnology B.V. [140].

4.3.2.4 Χρήση DNA μικροσυστοιχιών (DNA chips)

Η χρήση DNA μικροσυστοιχιών (micro-arrays) αποτελεί μια σχετικά νέα τεχνική (διεθνής εμπορική διάθεση, 2006) η οποία είναι βασισμένη στο συνδυασμό μεθόδων μοριακής ανίχνευσης με ιχνηθέτες και μικροσυστοιχιών πυριτίου (silicon based chips). Η επιφάνεια του chip καλύπτεται

συνήθως με μια πολύ λεπτή στρώση χρυσού, όπου εφαρμόζονται κατάλληλοι ιχνηθέτες, σχηματίζοντας με αυτό τον τρόπο μια διάταξη που ονομάζεται μικροσυστοιχία. Ο κάθε ιχνηθέτης διαφέρει από τους υπόλοιπους ανάλογα με το DNA-στόχο για τον οποίο έχει σχεδιαστεί να υβριδοποιείται [140].

Τα στάδια της διαδικασίας ανίχνευσης του ιού HPV με αυτή την τεχνική είναι τα ακόλουθα. Το προς ανάλυση δείγμα κυττάρων του τραχήλου της μήτρας προετοιμάζεται για ανάλυση μικροσυστοιχίας και τοποθετείται στην επιφάνεια του chip. Στη συνέχεια εισάγονται εκκινητές που δεσμεύουν τις γενετικές ακολουθίες-στόχους του HPV DNA. Με ένα ειδικό όργανο μετριέται εάν οι DNA-στόχοι στη μικροσυστοιχία έχουν δεσμευτεί. Σε περίπτωση που ανιχνευθεί ότι οι DNA-στόχοι έχουν δεσμευτεί, το δείγμα θεωρείται θετικό για HPV [140].

Η μέθοδος αυτή, στην πραγματικότητα συνδυάζει υβριδισμό PCR υψηλής ευαισθησίας και ειδικότητας και την τεχνολογία των μικροσυστοιχιών, με την αντίδραση να λαμβάνει χώρα στην επιφάνεια του chip. Τα βασικά πλεονεκτήματα που παρουσιάζει είναι η υψηλή διαγνωστική ευαισθησία και ειδικότητα (98.2% και 100% αντίστοιχα), το σύντομο χρονικό διάστημα που απαιτείται για τα αποτελέσματα (48h), η δυνατότητα διαχωρισμού σε πολλαπλή και μονή λοίμωξη και η δυνατότητα εφαρμογής της μεθόδους σε περιβάλλον νοσοκομειακού εργαστηρίου. Αποτελεί μια υποσχόμενη μέθοδο, αλλά ακόμα βρίσκεται σε ερευνητική φάση.

4.3.3 Κολποσκόπηση

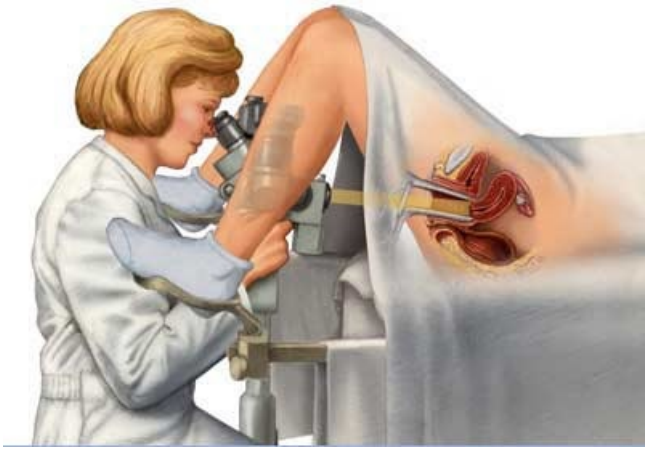
Η κολποσκόπηση αποτελεί την κατεξοχήν μέθοδο διάγνωσης προκαρκινικών αλλοιώσεων ή καρκίνου του τραχήλου της μήτρας (καθώς και οποιασδήποτε βλάβης του κόλπου και του αιδοίου). Πρόκειται για μια μη επεμβατική, ανώδυνη διαδικασία που διαρκεί 15-25 λεπτά. Συνήθως διενεργείται μετά από ανεύρεση παθολογικών κυττάρων στο τεστ Παπανικολάου ή κατόπιν διαπίστωσης κάτι ύποπτου/παρατήρησης ορατών βλαβών στον τράχηλο (με γυμνό μάτι) κατά τη γυναικολογική εξέταση. Χρησιμοποιείται, επίσης, για προεγχειρητικό σχεδιασμό, μετεγχειρητική παρακολούθηση (μετά από επέμβαση/θεραπεία για HSIL ή καρκίνο), για παρακολούθηση προκαρκινικών καταστάσεων, καθώς και σε περίπτωση επιμενουσών φλεγμονών του τραχήλου ή ανεξήγητης κολπικής αιμορραγίας.

Κατά την κολποσκόπηση επισκοπείται με υψηλή λεπτομέρεια το αιδοίο, ο κόλπος και ο τράχηλος. Για την υλοποίησή της από τον γυναικολόγο ιατρό χρησιμοποιείται το κολποσκόπιο, το οποίο μέσω ενός ειδικού τηλεσκοπικού μικροσκοπίου χαμηλής ισχύος και μιας κατάλληλα προσαρτημένης μηχανής μαγνητοσκόπησης, παρέχει δυνατό φωτισμό και διευρύνει έως και 40 φορές το οπτικό πεδίο. Το κολποσκόπιο επιτρέπει στο γιατρό να μπορεί να εντοπίσει τυχόν ανωμαλίες που δεν θα ήταν ορατές με γυμνό μάτι. Η προσάρτηση της μηχανής μαγνητοσκόπησης

γίνεται προκειμένου να παρθεί φωτογραφικό και μαγνητοσκοπικό υλικό από τον κόλπο και τον τράχηλο κατά την εξέταση. Προκειμένου να χωριστούν τα κολπικά τοιχώματα έτσι ώστε να φανεί έυκολα ο τράχηλος, χρησιμοποιείται ένας κολποδιαστολέας. Το κολποσκόπιο τοποθετείται στο άνοιγμα του κόλπου, έξω από αυτόν, σε κατάλληλη θέση ώστε να είναι ορατός ο τράχηλος. Πλέον εκτός από το κλασικό οπτικό κολποσκόπιο χρησιμοποιούνται και υψηλής ανάλυσης ψηφιακά κολποσκόπια, τα οποία αναλύουν μέσω υπολογιστή και ψηφιακής τεχνολογίας την εικόνα του τραχήλου, ανιχνεύοντας βλάβες σε πολύ πρώιμο στάδιο. Με τα ψηφιακά κολποσκόπια γίνεται ψηφιακή χαρτογράφηση του τραχήλου της μήτρας με χρωματικό κώδικα για την αναγνώριση των ύποπτων περιοχών και γίνεται πλήρης καταγραφή και αρχειοθέτηση κάθε εξέτασης για τη βέλτιστη παρακολούθηση της ασθενούς.

Η κολποσκόπηση γίνεται με τη βοήθεια ειδικών χρωστικών διαλυμάτων (οξικό οξύ ή/και Lugol), τα οποία βοηθούν να αναδειχτούν οι τυχόν προ-νεοπλασματικές αλλοιώσεις και να καθοριστεί η βαρύτητα των βλαβών. Ο κόλπος και ο τράχηλος χρωματίζονται με τέτοιο τρόπο ώστε να διαχωριστούν οι τυχόν παθολογικές αλλοιώσεις από το φυσιολογικό επιθήλιο. Συγκεκριμένα, γίνεται εμβροχή της εξεταζόμενης τραχηλικής περιοχής με αραιό διάλυμα οξικού οξέος (3-5%), το οποίο σε περίπτωση που υπάρχουν παθολογικές αλλοιώσεις προκαλεί τήξη των πρωτεϊνών του πυρήνα των κυττάρων και έτσι, εξαιτίας της κατά κανόνα υψηλής πυκνότητας πυρηνικής πρωτεΐνης που εντοπίζεται σε περιοχές όπου υπάρχει αλλοιωμένο επιθήλιο (CIN), αυτές θα αποκτήσουν μια λευκάζουσα απόχρωση (acetowhite) που είναι εύκολα ορατή από τον γυναικολόγο ιατρό. Το φυσιολογικό επιθήλιο διατηρεί την ερυθρά ομαλή εμφάνισή του και μετά τη δράση του οξικού οξέος. Με αυτό τον τρόπο προσδιορίζονται οι τυχόν ανώμαλες περιοχές, οι οποίες δεν είναι εφικτό να γίνουν αντιληπτές με γυμνό μάτι.

Εάν κριθεί αναγκαίο γίνεται επίθιξη με ιωδιούχο διάλυμα Lugol (δοκιμασία κατά Schiller). Η χρώση με Lugol προκαλεί στο φυσιολογικό επιθήλιο μια έντονη φαιά χρώση λόγω του πλούσιου σε γλυκογόνο κυτταροπλάσματος των κυττάρων του. Αντίθετα, το αλλοιωμένο επιθήλιο δεν βάφεται λόγω του ότι τα κύτταρά του στερούνται γλυκογόνου. Προφανώς, οι ιωδιοαρνητικές περιοχές είναι αυτές που προσδιορίζουν τις αλλοιώσεις. Η κολποσκόπηση θεωρείται ικανοποιητική όταν το σύνολο της ζώνης μετάπτωσης είναι ορατό.



Εικόνα 99: *Κολποσκόπηση (εικόνα από [145])*

Σε περίπτωση που ο θεράπων ιατρός το κρίνει σκόπιμο, γίνεται λήψη βιοψιών. Με βάση κάποια μορφολογικά χαρακτηριστικά που υπάρχουν στην επιφάνεια του επιθηλίου, εντοπίζονται οι αλλοιώσεις και επιλέγεται η πιο ύποπτη περιοχή για λήψη δείγματος ιστού. Οι βιοψίες γίνονται με ειδικές λαβίδες χωρίς αναισθησία και χωρίς να προκαλούν πόνο στον ασθενή. Αιμόσταση επιτυγχάνεται με διάλυμα Monsel's ή με νιτρικό άλας. Το τμήμα του ιστού που λαμβάνεται, αποστέλλεται στο παθολογοανατομικό εργαστήριο για ιστολογική ανάλυση. Ο γιατρός που θα κάνει την κολποσκόπηση πρέπει να είναι εξειδικευμένος, έτσι ώστε να μην πάρει βιοψία από εσφαλμένη περιοχή και τα αποτελέσματα να είναι παραπλανητικά. Η ιστολογική εξέταση βιοψιών υπό κολποσκοπικό έλεγχο θεωρείται ως η «χρυσή αρχή» (gold standard) της διάγνωσης όλων των αλλοιώσεων του τραχήλου της μήτρας.

Τα πιθανά σωματικά επακόλουθα της κολποσκόπησης σε περίπτωση βιοψίας περιλαμβάνουν εκκρίσεις που οφείλονται στα υγρά διαλύματα που χρησιμοποιούνται κατά την εξέταση (Monsel ή νιτρικό άλας για έλεγχο μικροαιμορραγίας από τη βιοψία) και παρουσία μικρής ποσότητας αίματος. Επιπλοκές όπως μεγάλη αιμορραγία και λοίμωξη είναι σπάνιες.

Όταν τα αποτελέσματα της κολποσκόπησης δείξουν την παρουσία ανώμαλων κυττάρων στον τράχηλο της μήτρας, τότε η κολποσκόπηση, ανάλογα με τον αριθμό των κυττάρων αυτών και τη σοβαρότητα των ανωμαλιών, μπορεί να συνδυαστεί με μια εκ των ακόλουθων διαδικασιών θεραπείας: αφαίρεση πολύ μικρών δειγμάτων (punch biopsies), χρήση αγκύλης διαθερμίας (Loop Electrical Excision Procedure-LEEP ή σε ονομασία UK: Large Loop Excision of the Transformation Zone - LLETZ) ή κωνοειδή εκτομή (conization). Οι διαδικασίες αυτές αποσκοπούν στην αφαίρεση των ανώμαλων κυττάρων του τραχήλου της μήτρας διατηρώντας όσο περισσότερο υγιή ιστό είναι δυνατό.

4.3.3.1 Ψυχολογικά επακόλουθα της κολποσκόπησης

Έχουν γίνει διάφορες μελέτες σχετικά με τα ψυχολογικά επακόλουθα της κολποσκόπησης. Σύμφωνα με μια από τις πρώτες μελέτες που έγινε το 1986 [146], οι γυναίκες που αποστέλλονται για κολποσκόπηση εξαιτίας αποτελέσματος στο Παπ τεστ που υπάγεται στην κατηγορία «μη φυσιολογικό», καταλαμβάνονται από φόβο για ύπαρξη καρκίνου. Όταν τους ζητήθηκε να αξιολογήσουν το φόβο τους σε μια κλίμακα: ήπιος/μέτριος/σοβαρός, ποσοστό 70% περιέγραψε το φόβο ως σοβαρού βαθμού. Άνω του 70% των ασθενών εξέφρασαν ανησυχίες για τον κίνδυνο απώλειας της μήτρας τους και τις ανάλογες συνέπειες αυτού για τη γονιμότητά τους. Μεταξύ άλλων ανέφεραν και άλλα συμπτώματα, όπως για παράδειγμα, χαμηλή αυτοεκτίμηση, αρνητική εικόνα για το σώμα τους όσον αφορά την περιοχή στο γεννητικό τους σύστημα που παρουσίασε πρόβλημα, σοβαρό βαθμό άγχους, αϋπνία, ευερεθιστικότητα, κρίσεις με κλάμα, θυμό και οργή, αλλά και δυσκολίες στη σεξουαλική τους ζωή και στη σχέση τους με το σύντροφό τους.

Σύμφωνα με μια άλλη μελέτη [147], βρέθηκε ότι οι γυναίκες που παραπέμφθησαν για κολποσκόπηση είχαν μέση βαθμολογία άγχους 51.2 (βάσει μιας κλίμακας βαθμολόγησης του άγχους από 20 έως 80, σύμφωνα με την οποία η συνήθης ενήλικη γυναίκα στις δυτικές κοινωνίες έχει μέση βαθμολογία 35). Αυτό που απασχολούσε περισσότερο τις γυναίκες και επιδείνωνε το άγχος τους ήταν η αναμονή μιας άγνωστης εξέτασης: της κολποσκόπησης. Οι κύριες τους ανησυχίες αφορούσαν το αν η κολποσκόπηση θα είναι μια επίπονη διαδικασία και τι θα γινόταν κατά τη διάρκειά της. Οι φόβοι τους περιλάμβαναν φόβο για πιθανή τραυματική εμπειρία, για πόνο και σωματική ταλαιπωρία, καθώς και για τα πιθανά αποτελέσματα. Η ίδια μελέτη επανελήφθη και από μια άλλη ερευνητική ομάδα σε μεγαλύτερο αριθμό γυναικών και το αποτέλεσμα ήταν 50.2 [148]. Από διάφορες μελέτες προκύπτει ότι τα επίπεδα άγχους πριν από μια κολποσκόπηση κυμαίνονται από 45.2 έως 51.2 [147], [148], [149], [150], [151]. Ας σημειωθεί ότι αυτό το αποτέλεσμα είναι ίσο ή υψηλότερο από ότι το αντίστοιχο που εκτιμάται προεγχειρητικά σε γυναίκες, το βράδυ πριν από μια μείζονα χειρουργική επέμβαση.

Σύμφωνα με το ερωτηματολόγιο μιας μελέτης σχετικά με το άγχος που συνδέεται με την κυτταρολογία και την κολποσκόπηση [152] 17% των γυναικών ένιωθαν αβοήθητες ή ευάλωτες, 14% έβρισκε τη διαδικασία της κολποσκόπησης άβολη, 12% την έβρισκε αναξιοπρεπή και 10% ένιωθε ότι «εισβάλλεται το σώμα τους». Επίσης υπήρχαν ανησυχίες για εύρεση καρκίνου (17%) ή για μελλοντική του εμφάνιση (19%) [149].

Μια άλλη μελέτη [153], καταδεικνύει ως παράγοντες άγχους την ντροπή/αμηχανία και τη δυσφορία σχετικά με τη διαδικασία της κολποσκόπησης, καθώς και την ανησυχία για τις συνέπειες στην υγεία τους, συμπεριλαμβανομένων σεξουαλικών και αναπαραγωγικών ζητημάτων.

Εκτός από τους «ανθρωπιστικούς» λόγους υπάρχουν επίσης και επιστημονικά ευρήματα που υπαγορεύουν την ανάγκη για μείωση του ψυχολογικού άγχους για την κολποσκόπηση [149]. Υπάρχει σχέση μεταξύ του άγχους και της ανάπτυξης καρκίνου του τραχήλου της μήτρας [154] καθώς και συσχέτιση μεταξύ του σωματικού άγχους και της επιδεκτικότητας για καρκίνο του τραχήλου της μήτρας [155]. Επιπρόσθετα, το ψυχολογικό άγχος επηρεάζει τη σωματική υγεία και οδηγεί σε αύξηση των ποσοστών συμβουλευτικής βοήθειας [156]. Εκτός αυτών, το ψυχολογικό άγχος έχει ως αποτέλεσμα τη μειωμένη ικανότητα ανάκλησης και υλοποίησης συμβουλών, καθιστώντας λιγότερο πιθανή τη συμμόρφωση των γυναικών με τις πληροφορίες που τους δίνονται. Κάποιες γυναίκες [10-61%] φτάνουν στο σημείο να μην παραστούν καν στο ραντεβού για κολποσκόπηση [153]. Επομένως, η μείωση του ψυχολογικού άγχους μπορεί να μειώσει και το ποσοστό μη-παρεύρεσης για κολποσκόπηση [157].

4.3.3.2 Άμεση κολποσκόπηση ή κυτταρολογική παρακολούθηση

Μερικές μελέτες [158] έχουν ασχοληθεί με τα ψυχολογικά αποτελέσματα της παραπομπής για άμεση κολποσκόπηση σε σχέση με τον κυτταρολογικό επανέλεγχο σε γυναίκες με «μη φυσιολογικά» αποτελέσματα στο Παπ τεστ. Από τη μια η διαχείριση με κυτταρολογική παρακολούθηση έχει εγείρει τους εξής προβληματισμούς: κάποιες περιπτώσεις με αλλοιώσεις υψηλού βαθμού μπορεί να διαφύγουν εντοπισμού, η ευαισθησία αυτής της μεθόδου είναι περιορισμένη και επιπλέον μπορεί να προκαλέσει άγχος για παρατεταμένη περίοδο [159]. Επίσης, εκτιμάται ότι το 65% των γυναικών με ήπια δυσκαρύωση τελικά καταλήγει να αναφερθεί για κολποσκόπηση [160]. Από την άλλη, η πολιτική της άμεσης παραπομπής για κολποσκόπηση εγείρει προβληματισμούς σχετικά με πιθανή υπερθεραπεία [161], [162], επιπλοκές [163], αρνητικές επιδράσεις σε μια επακόλουθη εγκυμοσύνη [164], υψηλά επίπεδα άγχους [165], [166], περιορισμένους πόρους (οικονομικό κόστος και έμπειρο ιατρικό προσωπικό) [167] και χρόνο αναμονής [167], [168]. Ως επακόλουθο δεν υπάρχει μια σταθερή κοινή πολιτική που να ακολουθείται σε όλες τις χώρες [169]. Μελέτες υποθετικών σεναρίων στην Αμερική εισηγούνται ότι οι γυναίκες προτιμούν την κολποσκόπηση όταν το κυτταρολογικό αποτέλεσμα είναι σοβαρού βαθμού, αλλιώς προτιμούν την κυτταρολογική παρακολούθηση [170], [171].

Η μελέτη [158] σχετικά με τη διαχείριση των γυναικών με χαμηλού βαθμού αλλοιώσεις (nuclear abnormalities or mild dyskaryosis), χώρισε τις συμμετέχουσες γυναίκες τυχαία σε 2 ομάδες: η μια ομάδα έκανε επανέλεγχο με Παπ τεστ κάθε 6 μήνες, ενώ η άλλη παραπεμπόταν άμεσα για κολποσκόπηση (η μελέτη διήρκεσε 3 χρόνια). Σύμφωνα με αυτή τη μελέτη, ήταν αισθητά πολύ πιο υψηλά τα ποσοστά των γυναικών που επιλέχθηκαν (τυχαία) για άμεση κολποσκόπηση, που ανέφεραν πόνο, αιμορραγία, εκκένωση και άλλα επακόλουθα, συγκριτικά με τις γυναίκες που ανήκαν στην ομάδα κυτταρολογικής παρακολούθησης [158]. Επίσης, οι γυναίκες

στην ομάδα της άμεσης κολποσκόπησης ανέφεραν σωματικά επακόλουθα που διήρκησαν μεγαλύτερο χρονικό διάστημα και ήταν πολύ πιο σοβαρά. Σε αυτή τη μελέτη δεν παρατηρήθηκε, ωστόσο, αισθητή διαφορά στα ψυχολογικά επακόλουθα (άγχος και κατάθλιψη) στις 2 ομάδες. Όπως ήταν αναμενόμενο, τα περιστατικά CIN2+ εντοπίστηκαν πιο νωρίς στην ομάδα της άμεσης κολποσκόπησης. Επιπλέον, η πολιτική της άμεσης κολποσκόπησης εντόπιζε περισσότερα περιστατικά CIN2+, καθώς και περισσότερα περιστατικά CIN3+ από ότι η πολιτική της κυτταρολογικής παρακολούθησης (αυτό ήταν πιο προφανές όταν επρόκειτο για περιστατικά ήπιας δυσκαρύωσης). Ωστόσο, παρατηρήθηκε ότι η διαφορά των 2 πολιτικών στον εντοπισμό CIN2+ ήταν μεγαλύτερη από ότι η διαφορά στον εντοπισμό CIN3+, πιθανότατα λόγω αυθόρμητης υποχώρησης μερικών περιστατικών CIN2 που αρχικά εντοπίστηκαν στην ομάδα κυτταρολογικής παρακολούθησης. Αυτό εισηγείται ότι η άμεση κολποσκόπηση ενδεχομένως οδηγεί σε υπερθεραπεία. Τα συμπεράσματα της μελέτης αυτής επιγραμματικά ήταν τα ακόλουθα:

- Η άμεση παραπομπή για κολποσκόπηση εντοπίζει περισσότερα περιστατικά τραχηλικής ενδοεπιθηλιακής νεοπλασίας βαθμού 2 ή υψηλότερου βαθμού αρχικά, αλλά υπάρχει μικρή διαφορά στο συνολικό αριθμό περιστατικών στο διάστημα των 3 ετών.
- Η άμεση κολποσκόπηση οδηγεί στην παράπεμψη μεγάλου αριθμού περιστατικών όπου δεν εντοπίζεται τραχηλική ενδοεπιθηλιακή νεοπλασία βαθμού 2 ή υψηλότερου βαθμού, καθώς και σε περισσότερα προβλήματα λόγω παρενεργειών από ότι η κυτταρολογική παρακολούθηση.
- Η πολιτική παραπομπής των γυναικών με χαμηλού βαθμού αλλοιώσεις του τραχήλου της μήτρας για άμεση κολποσκόπηση, δεν παρέχει σαφές όφελος σε σύγκριση με την κυτταρολογική παρακολούθηση και εκτός αυτού προκαλεί και περισσότερες παρενέργειες.
- Η άμεση κολποσκόπηση, μια πολιτική που ακολουθείται από πολλές χώρες (συμπεριλαμβανομένου και της Αγγλίας), οδηγεί σε υπερθεραπεία, ένα όλο και πιο αναγνωρισμένο πρόβλημα.

Συνοπτικά, κατέληξαν στο συμπέρασμα ότι δεν υπάρχει σαφές όφελος από την πολιτική άμεσης κολποσκόπησης, εφόσον αν και εντοπίζει περισσότερα περιστατικά CIN2+, εντούτοις οδηγεί σε μεγάλο αριθμό παραπομπών που δεν έχουν υψηλού βαθμού τραχηλική ενδοεπιθηλιακή νεοπλασία, σε υπερθεραπεία, επακόλουθες παρενέργειες και καμία αισθητή διαφορά στο ψυχολογικό κόστος σε σχέση με την κυτταρολογική παρακολούθηση.

Μια άλλη μελέτη [172] αναφορικά με την ανάλυση του κόστους-αποτελεσματικότητας των δύο πολιτικών σε περιστατικά ήπιας δυσκαρύωσης, έδειξε ότι η πολιτική άμεσης κολποσκόπησης

και θεραπείας αύξανε το συνολικό κόστος διαχείρισης περιστατικών με ήπια δυσκαρυωτικά επιχρίσματα (κατά 50%). Ωστόσο, το αυξημένο αυτό κόστος αντισταθμιζόταν από την αύξηση του αριθμού των περιστατικών CIN3 που εντοπίζονταν (69% αύξηση). Τα αποτελέσματα από αυτή τη μελέτη έδειξαν ότι σημαντικός αριθμός γυναικών με ήπια δυσκαρυωτικά επιχρίσματα είχε στην πραγματικότητα CIN3 (35%). Επιπρόσθετα, μόνο το 29% των περιστατικών με δυσκαρύωση επανήλθαν στο φυσιολογικό (από μόνα τους, χωρίς οποιαδήποτε εξωτερική επέμβαση). Εκτός αυτών, το 25% των γυναικών, μετά την πάροδο των 2 χρόνων, σταμάτησαν να παρουσιάζονται για κυτταρολογικό επανέλεγχο. Έτσι, σύμφωνα με αυτή τη μελέτη, συνάγεται ότι η κλινική αξιολόγηση με άμεση κολποσκόπηση και θεραπεία είναι πιο αποτελεσματική πολιτική ως προς την συνολική αύξηση της υγείας, αλλά συνεπάγεται χρήση περισσότερων πόρων (οικονομικών και άλλων).

Συνοψίζοντας, φαίνεται να μην υπάρχει ακόμα ξεκάθαρη εικόνα για το ποια πολιτική, άμεσης κολποσκόπησης ή κυτταρολογικού ελέγχου, είναι η βέλτιστη.

4.4 Προληπτικός πληθυσμιακός έλεγχος

Ο στόχος του προληπτικού πληθυσμιακού ελέγχου του καρκίνου του τραχήλου της μήτρας είναι η μείωση των περιστατικών εμφάνισης της νόσου, καθώς και η μείωση της θνησιμότητας από αυτή, μέσω του εντοπισμού και της εφαρμογής κατάλληλης θεραπευτικής αγωγής στις προκαρκινικές αλλοιώσεις [87].

Ο προληπτικός πληθυσμιακός έλεγχος είναι είδος δευτερογενούς πρόληψης. Η δευτερογενής πρόληψη αποτελείται από μέτρα για μια όσο το δυνατόν πιο έγκαιρη διάγνωση μετά την εμφάνιση της πάθησης. Στην πρωτογενή πρόληψη ανήκουν μέτρα που αποσκοπούν στο να αποτρέψουν την εμφάνιση της πάθησης: για παράδειγμα ο εμβολιασμός και η αποφυγή των παραγόντων που μπορεί να προκαλέσουν τη νόσο (κάπνισμα, αυξημένος αριθμός σεξουαλικών συντρόφων σε σύντομο χρονικό διάστημα, αυξημένος αριθμός τοκετών, μικρή ηλικία έναρξης σεξουαλικών επαφών κτλ.).

Οι δύο διαγνωστικές εξετάσεις test PAP και HPV testing, που εφαρμόζονται σήμερα, ανήκουν στη δευτερογενή πρόληψη, δηλαδή στη διαπίστωση της μορφολογικής αλλοίωσης των κυττάρων του επιθηλίου του τραχήλου και την ανακάλυψη και τη διερεύνηση της βαρύτητας της λοίμωξης από τον ιό HPV. Ο δευτερογενής έλεγχος του τραχήλου συμπληρώνεται από την κολποσκόπηση, την βιοψία και τις συμπληρωματικές μοριακές τεχνικές βάσει των οποίων θα αποφασισθεί η παρακολούθηση και ο τρόπος θεραπείας της συγκεκριμένης ασθενούς. Ο προσυμπτωματικός έλεγχος για λοιμώξεις HPV είναι σχετικά αρκετά αποτελεσματικός στη μείωση των περιστατικών

της νόσου παρά το γεγονός ότι η ειδικότητα και η ευαισθησία των διαθέσιμων διαγνωστικών τεχνικών δεν είναι ιδανικές [85].

Ο προσυμπτωματικός έλεγχος (screening) του καρκίνου του τραχήλου της μήτρας βασίζεται κυρίως σε κυτταρολογικές και κολποσκοπικές αναλύσεις. Ο συστηματικός προσυμπτωματικός έλεγχος που γίνεται από έμπειρους κυτταρολόγους και γυναικολόγους έχει οδηγήσει στη δραματική μείωση των περιστατικών της νόσου. Ωστόσο, σε χώρες με λιγότερο έμπειρους κυτταρολόγους, νοσοκόμους ή γυναικολόγους οι ίδιες εξετάσεις έχουν μεγάλη πιθανότητα για false-negative ή false-positive αποτελέσματα [88]. Επομένως, για αποτελεσματικό προσυμπτωματικό έλεγχο μεγάλης κλίμακας, απαιτούνται απλές, αξιόπιστες και φθηνές διαγνωστικές εξετάσεις, ιδιαίτερα σε αναπτυσσόμενες χώρες που έχουν μεγαλύτερα ποσοστά εμφάνισης της νόσου εξαιτίας των μη αποτελεσματικών προσυμπτωματικών ελέγχων [88].

Υπάρχουν δύο είδη μαζικού προληπτικού πληθυσμιακού ελέγχου:

- Οργανωμένος πληθυσμιακός έλεγχος (organized screening)
- Ευκαιριακός πληθυσμιακός έλεγχος (opportunistic screening)

Τα προγράμματα οργανωμένου πληθυσμιακού ελέγχου (Organized Screening Programs - OSP) πρέπει να είναι υψηλής ποιότητας και οι υπηρεσίες πληθυσμιακού ελέγχου πρέπει να υπόκεινται έλεγχο και να παρακολουθούνται από μια ομάδα διαχείρισης απαρτιζόμενη από άτομα που δεν ανήκουν στο πρόγραμμα. Για το πιο ευρύ κοινό, η πιο προφανής διαφορά στα οργανωμένα προγράμματα είναι ότι η γυναίκα λαμβάνει μια πρόσκληση όταν έρθει ο χρόνος για να ελεγχθεί. Με τα οργανωμένα προγράμματα ελέγχου όλοι όσοι συμμετέχουν προσφέρονται τις ίδιες υπηρεσίες, πληροφορίες και υποστήριξη. Συχνά, μεγάλοι αριθμοί ατόμων καλούνται να λάβουν μέρος σε οργανωμένα προγράμματα ελέγχου. Τα οργανωμένα προγράμματα πληθυσμιακού ελέγχου παρέχουν την πιο αποτελεσματική προστασία ενάντια στον καρκίνο του τραχήλου. Τα κύρια στοιχεία ενός οργανωμένου προγράμματος είναι τα εξής:

- Θεσμοθετημένη πολιτική μαζικού πληθυσμιακού ελέγχου με συγκεκριμένο εύρος ηλικιών, διάστημα ελέγχου και συγκεκριμένες διαδικασίες
- Μια ομάδα διαχείρισης υπεύθυνη για την οργάνωση του προγράμματος
- Κατάλληλες υπηρεσίες για την παρακολούθηση και διαχείριση των γυναικών με θετικό αποτέλεσμα σε εξέταση μαζικού ελέγχου
- Ένα εγκατεστημένο πρόγραμμα διασφάλισης ποιότητας
- Μια μέθοδο καταγραφής των περιστατικών καρκίνου στον απευθυνόμενο πληθυσμό

Ο στόχος αυτών των στοιχείων είναι ότι όλες οι γυναίκες θα ελέγχονται τακτικά, ότι οι γυναίκες με θετικό αποτέλεσμα στη διαγνωστική εξέταση/εξετάσεις μαζικού ελέγχου θα

παρακολουθούνται κατάλληλα, ότι το πρόγραμμα θα παρέχει την υψηλότερη δυνατή ποιότητα υπηρεσιών και ότι η ποιότητα των υπηρεσιών θα ελέγχεται ορθά.

Ο ευκαιριακός προσυμπτωματικός έλεγχος εξαρτάται από την πρωτοβουλία της εκάστοτε γυναίκας ή του ιατρού της: μια γυναίκα μπορεί να ζητήσει από το γιατρό της την διενέργεια ελέγχου ή διαγνωστικής εξέτασης ή μπορεί να τα εισηγηθεί ο γιατρός στην εκάστοτε ασθενή. Σε αντίθεση με ένα οργανωμένο πρόγραμμα ελέγχου, ο ευκαιριακός έλεγχος δεν μπορεί να αξιολογηθεί ή να παρακολουθηθεί.

Ο ευκαιριακός προσυμπτωματικός έλεγχος, σε σχέση με τον οργανωμένο παρουσιάζει κάποια μειονεκτήματα. Στον ευκαιριακό προσυμπτωματικό έλεγχο, συνήθως, παρατηρείται υψηλή κάλυψη επιλεγμένων ομάδων του πληθυσμού και παράλληλα χαμηλή κάλυψη άλλων πληθυσμιακών ομάδων χαμηλότερου κοινωνικοοικονομικού επιπέδου και μειονοτήτων. Ο ευκαιριακός έλεγχος, δημιουργεί την τάση οι γυναίκες να ελέγχονται πιο συχνά από ότι είναι απαραίτητο, καθώς και την τάση να ελέγχονται γυναίκες μόνο από υψηλές κοινωνικοοικονομικές ομάδες που έχουν ήδη τακτική πρόσβαση στο σύστημα υγείας και έχουν μικρότερο κίνδυνο για εμφάνιση της νόσου. Οδηγεί έτσι στη συνύπαρξη μεγάλων ποσοστών γυναικών που ελέγχονται αναίτια και συχνά ενώ το ποσοστό του πληθυσμού που χρήζει ελέγχου, συνήθως χαμηλού κοινωνικοοικονομικού προφίλ και ετερογενών χαρακτηριστικών, δεν ελέγχεται επαρκώς. Για το λόγο αυτό η συνολική μείωση περιστατικών καρκίνου του τραχήλου της μήτρας δεν είναι τόσο μεγάλη όσο αυτή που επιτυγχάνεται με τα οργανωμένα προγράμματα. Επιπλέον, ο ευκαιριακός έλεγχος οδηγεί σε κακή σχέση κόστους αποτελεσματικότητας. Παρόλο που ο ευκαιριακός έλεγχος μπορεί να μειώσει τη συχνότητα εμφάνισης της νόσου, είναι λιγότερο αποτελεσματικός από ότι τα οργανωμένα προγράμματα. Επιπρόσθετα, η αποτελεσματικότητά του ευκαιριακού ελέγχου μπορεί να ποικίλει από μια περιοχή σε μια άλλη. Για τους παραπάνω λόγους πρέπει να στοχεύεται ο οργανωμένος πληθυσμιακός έλεγχος και όχι ο ευκαιριακός έλεγχος γυναικών που πραγματοποιείται σε περιβάλλον κλινικής και άπτεται της θέλησης της γυναίκας ή του γιατρού της.

Η Ευρωπαϊκή Ένωση (Ε.Ε.) προτείνει να εφαρμοστούν διαφορετικές λύσεις σε κάθε χώρα και περιοχή, ανάλογα με τους διαθέσιμους πόρους αλλά και την ευρύτερη δομή των υπηρεσιών υγείας, ικανοποιώντας τα εκάστοτε μεθοδολογικά πρότυπα. Στις Ευρωπαϊκές κατευθυντήριες οδηγίες, καθορίζεται με σαφήνεια ότι το πρόγραμμα πληθυσμιακού ελέγχου εκπονείται σε εθνικό επίπεδο και υλοποιείται με βάση συγκεκριμένο κυβερνητικό σχέδιο. Απαιτείται πολιτική υποστήριξη και η απαραίτητη χρηματοδότηση προκειμένου να ξεκινήσει. Είναι σημαντικό το πρόγραμμα να ολοκληρωθεί με το υπάρχον εθνικό σύστημα υγείας και να γίνει αποδεκτό τόσο από τον πληθυσμό όσο και από τους λειτουργούς υγείας που εξαρτώνται οικονομικά από τη λήψη επιχρισμάτων και τη διάγνωσή τους. Με την εισαγωγή του προγράμματος πληθυσμιακού

ελέγχου οι κυβερνητικοί φορείς απαιτείται να μεριμνήσουν προκειμένου *να μην αυξηθεί το κόστος υγείας, λόγω υπερβολικά μεγάλου αριθμού μη απαιτούμενων εξετάσεων και κλινικών πράξεων*.

Η εμπειρία από τα οργανωμένα προγράμματα πληθυσμιακού ελέγχου δείχνει ότι αρκετές γυναίκες που έλαβαν την κυτταρολογική απάντηση ενός μη φυσιολογικού τεστ Παπανικολάου παρουσίασαν αρνητικά ψυχολογικά φαινόμενα, με αποτέλεσμα προβλήματα στη συμμόρφωσή τους με τις κατευθυντήριες οδηγίες του προγράμματος πληθυσμιακού ελέγχου και της παρακολούθησής τους.

Έχει αποδειχτεί ότι πολλές γυναίκες αποφεύγουν να κάνουν το τεστ Παπανικολάου επειδή δεν θέλουν να μπουν στη διαδικασία αναμονής του αποτελέσματος και στην αγωνία που αυτή συνεπάγεται [173], [174], [175], [176], [149]. Οι γυναίκες που υποβάλλονται σε έλεγχο ρουτίνας για καρκίνο του τραχήλου της μήτρας πρέπει να ενημερώνονται εκ των προτέρων για την καρκινογένεση στον τράχηλο της μήτρας, το ρόλο του ιού HPV, τη λογική του μαζικού ελέγχου ρουτίνας, την αξιοπιστία του ΠΑΠ τεστ και την πιθανή ανάγκη για επιπρόσθετες εξετάσεις. Αυτό μπορεί να γίνει και μέσω προκαταρκτικής ενημέρωσης με ενημερωτικά φυλλάδια έτσι ώστε να μειωθεί το άγχος των γυναικών [173], [174], [175], [149]. Σε περίπτωση που τα αποτελέσματα του Παπ τεστ υποδεικνύουν αλλοιώσεις σχετικές με HPV, ο καλύτερος τρόπος πληροφόρησης είναι η προσωπική ενημέρωση από εξειδικευμένο ιατρό. Σύμφωνα με τη μελέτη [177], οι γυναίκες εξέφρασαν την έντονη επιθυμία τους για περισσότερες πληροφορίες σχετικά με τον καρκίνο του τραχήλου της μήτρας και άλλες ασθένειες σχετικές με τον ιό HPV. Υποστήριξαν ότι ο ρόλος του υπεύθυνου ιατρού είναι πολύ σημαντικός ως προς την παροχή υποστήριξης. Επίσης, παρατηρήθηκε μεγάλη εξάρτηση στο στενό περιβάλλον και στο διαδίκτυο για αναζήτηση πληροφοριών, κάτι που αναδεικνύει την ανάγκη για παροχή ακριβής και ορθής πληροφόρησης.

Μια από τις πιο προβληματικές πτυχές του προσυμπτωτικού ελέγχου για καρκίνο του τραχήλου της μήτρας είναι η δυσκολία να καθοριστεί η ακριβής ερμηνεία ενός θετικού αποτελέσματος [81].

Η πιστοποίηση της μοριακής δομής και του γονιδιώματος των πολλαπλών τύπων του ιού HPV, η διάκριση αυτών με την ογκολογική ή μη δυνατότητά τους, καθώς και η αποσαφήνιση του τρόπου δράσης του ιού στα κύτταρα, κατέστησαν δυνατή την ανάπτυξη μοριακών τεχνικών (HPV DNA test, HPV mRNA test, κ.α) με τις οποίες είναι δυνατόν πλέον, να διαπιστώσουμε αν μια γυναίκα έχει μολυνθεί και από ποια συγκεκριμένα στελέχη του ιού, αν ο ιός έχει ενσωματωθεί στα κύτταρα του επιθηλίου ή όχι, καθώς και τη δυνατότητα ανοσολογικής απάντησης του οργανισμού.

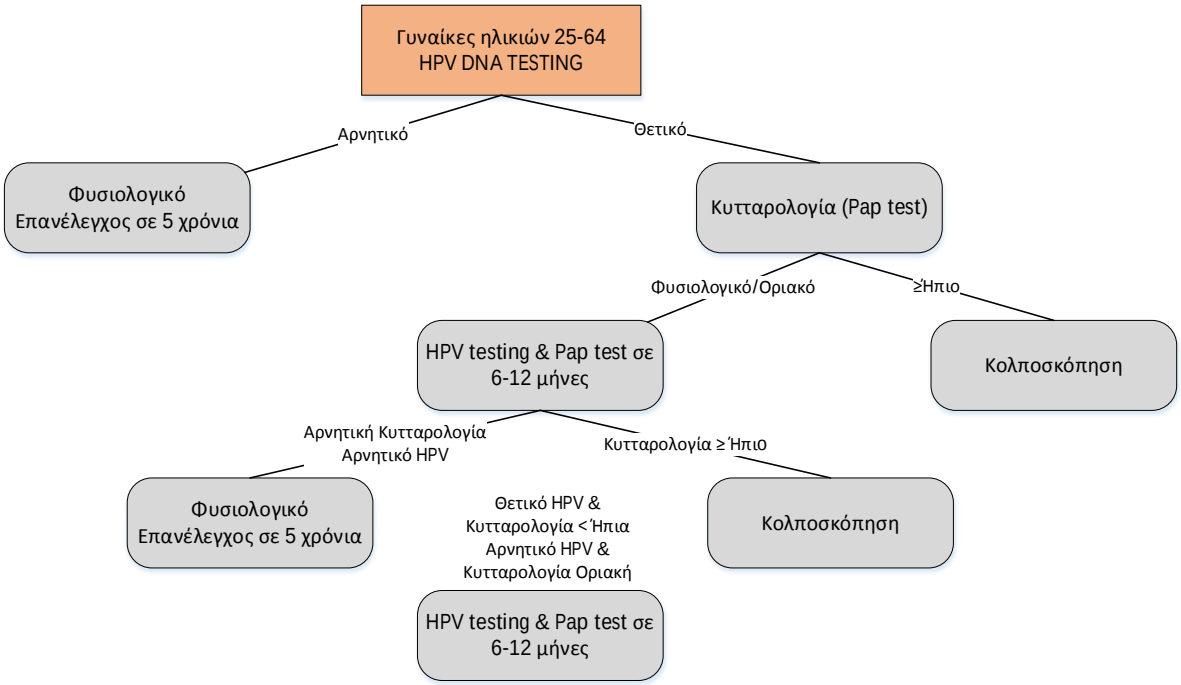
Η παρουσία ή όχι του HPV μπορεί να καθοριστεί με τις ακόλουθες τεχνικές:

- Τεστ Παπανικολάου (PAP test)

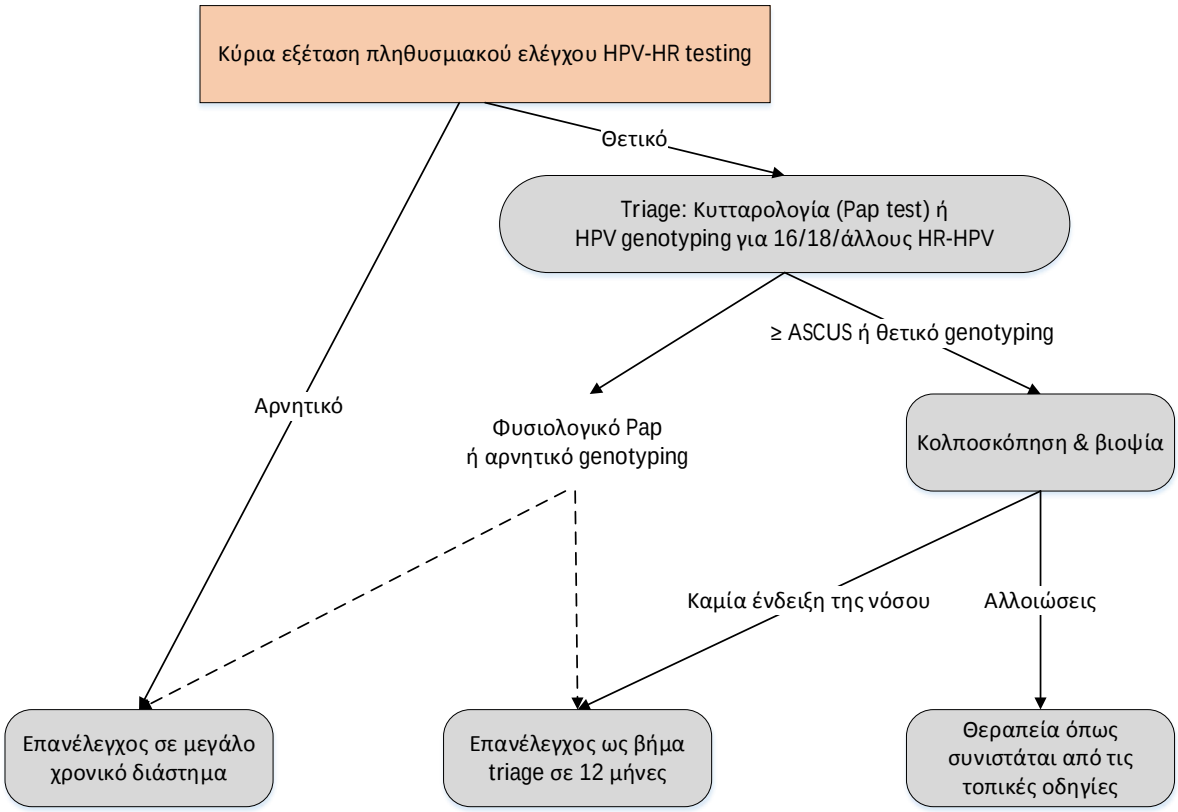
- Βιοψία
- Μοριακές τεχνικές:
    - Τεχνικές HPV DNA μικροσυστοιχιών (DNA microarrays techniques) π.χ. Υβριδοποίηση του DNA (DNA hybridization), Ενίσχυση με Αλυσιδωτή Αντίδραση Πολυμεράσης - PCR (PCR amplification) [87]
    - *mRNA τεχνικές όπως π.χ. NASBA (Nucleic Acid Based Amplification) mRNA test ή Flow Cytometry test που αναγνωρίζει τα ογκογόνα E6/E7*
    - *τεχνικές ανοσοκυτοχημείας (immunocytochemistry techniques) όπως η ανοσοχημική χρώση για υπερέκφραση της p16*

Σήμερα, οι διαγνωστικές εξετάσεις που εφαρμόζονται συνήθως για πληθυσμιακό έλεγχο είναι το test Παπανικολάου και το HPV testing. Το HPV DNA test ως μέθοδος πληθυσμιακού ελέγχου, σε σχέση με το PAP test, παρουσιάζει διάφορα πλεονεκτήματα. Καταρχάς, είναι λιγότερο εκτεθειμένο σε ανθρώπινα λάθη, αφού δεν στηρίζεται στην ανθρώπινη ερμηνεία, η οποία είναι αναγκαία στο Pap test, και επιπλέον απαιτεί μόνο ελάχιστη εκπαίδευση του τεχνικού που θα το πραγματοποιήσει [178]. Εξαιτίας του ότι η απαίτηση για αντιπροσωπευτικό δείγμα από ολόκληρο τον τράχηλο της μήτρας είναι λιγότερο κρίσιμη στο HPV testing, υπάρχει η δυνατότητα η εξέταση να γίνεται σε δείγμα που το έχει αυτοσυλλέξει η ίδια η ασθενής. Με αυτό τον τρόπο, υπάρχει η δυνατότητα αύξησης του ποσοστού κάλυψης του πληθυσμιακού ελέγχου, ειδικά όταν πρόκειται για γυναίκες που διστάζουν να επισκεφτούν εξειδικευμένο ιατρικό προσωπικό για την πραγματοποίηση της εξέτασης. Είναι προφανές, ότι αυτή η επιλογή προτιμάται ευρέως από τις ίδιες τις γυναίκες και οι πρώτες σχετικές μελέτες υποδεικνύουν καλή απόδοση της εξέτασης με χρήση τέτοιων δειγμάτων, αν και ακόμα αναμένεται να γίνουν πολλές βελτιώσεις για τελειοποίηση των συσκευών δειγματοληψίας και των μέσων συλλογής/μεταφοράς [179]. Επίσης, το HPV testing είναι πιο ευαίσθητο στην ανίχνευση CIN2+ και πιο ασφαλές για τη γυναίκα. Επιπρόσθετα, εξαιτίας του γεγονότος ότι για την ανάπτυξη διηθητικού καρκινώματος είναι απαραίτητη η επίμονη λοίμωξη με υψηλού κινδύνου τύπους HPV, το HPV testing επιτρέπει τη διεύρυνση των χρονικών μεσοδιαστημάτων μεταξύ δύο εξετάσεων (π.χ. από 5 χρόνια σε 3) [178] και παράλληλα θα οδηγήσει και σε ελάττωση του κόστους. Ο μόνος προβληματισμός σχετικά με το HPV testing, αφορά την χαμηλότερη του ειδικότητα σε σχέση με το Pap test με αποτέλεσμα το HPV testing να μην μπορεί να διαχωρίσει τις παροδικές από τις επίπονες λοιμώξεις, ενώ μόνο οι επίμονες σχετίζονται με αυξημένο κίνδυνο ανάπτυξης CIN2+ και καρκίνο [179]. Εντούτοις, συνολικά, προβλέπεται ότι θα οδηγήσει στη βελτίωση της ποιότητας και της αποτελεσματικότητας του πληθυσμιακού ελέγχου.

Για τους παραπάνω λόγους, το 2011, το FDA ενέκρινε το HPV test για χρήση σε συνδυασμό με το Pap test ή ως follow-up του Pap test και τελικά το 2014, ενέκρινε το HPV test για χρήση είτε ως συμπληρωματική εξέταση του Pap test είτε ως κύρια εξέταση προληπτικού πληθυσμιακού ελέγχου για τον καρκίνο του τραχήλου της μήτρας, με το Pap test να χρησιμοποιείται ως εξέταση triage για θετικές σε HPV γυναίκες. Ωστόσο, αυτή η πολιτική πληθυσμιακού ελέγχου ακόμα δεν εφαρμόζεται στις πλείστες χώρες. Έως τώρα, στις περισσότερες χώρες, ως κύρια εξέταση του πληθυσμιακού ελέγχου χρησιμοποιείται το Pap test (ή σε κάποιες περιπτώσεις ένας συνδυασμός HPV και Pap test). Αναμένεται, όμως, ότι το μελλοντικό πλάνο προληπτικού πληθυσμιακού ελέγχου θα ακολουθεί τις γραμμές που εισηγούνται οι μελέτες [178], [180], [181]. Προτεινόμενοι αλγόριθμοι από αυτές τις μελέτες παρουσιάζονται στις *εικόνες 100, 101*.

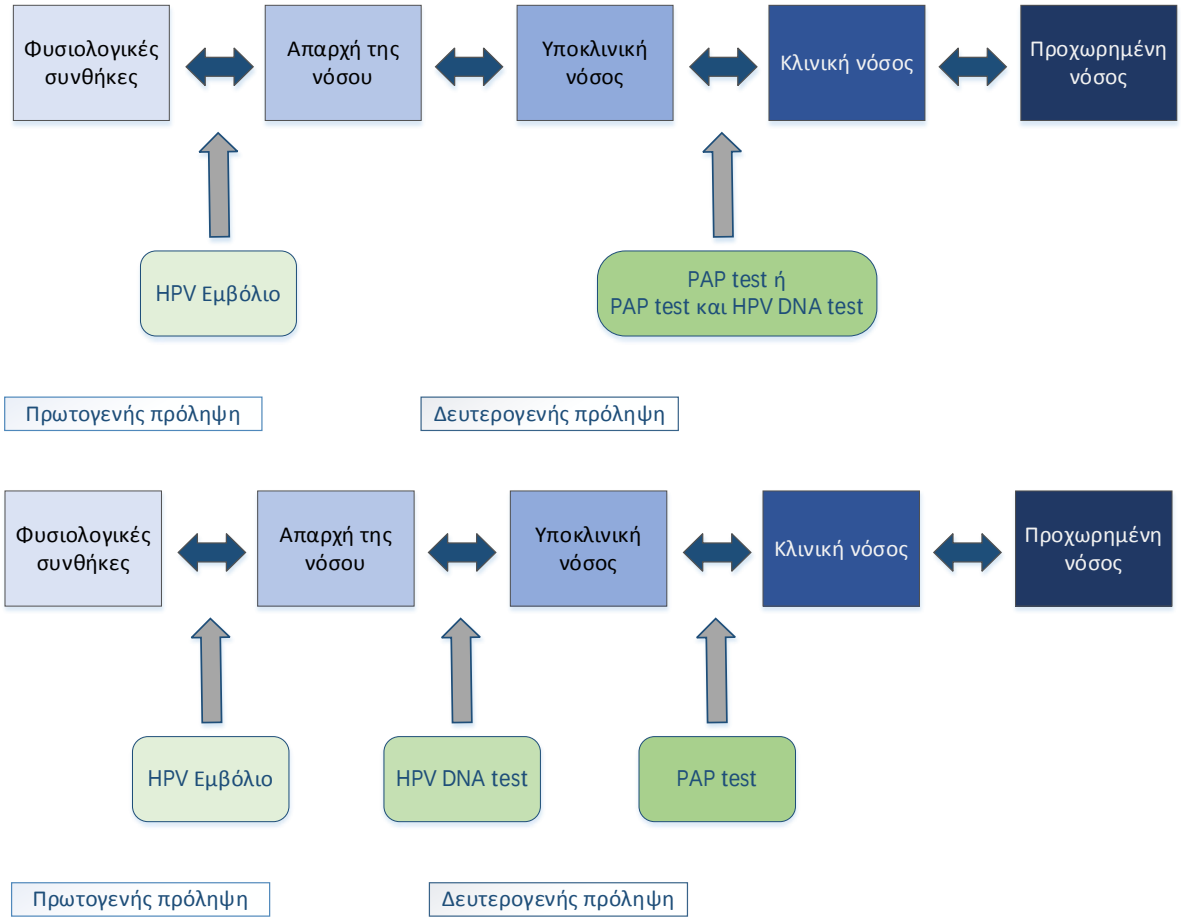


Εικόνα 100: *Προτεινόμενος αλγόριθμος πληθυσμιακού ελέγχου για καρκίνο του τραχήλου της μήτρας (μεταφρασμένη εικόνα από [180])*



Εικόνα 101: *Προτεινόμενος αλγόριθμος πληθυσμιακού ελέγχου για καρκίνο του τραχήλου της μήτρας (μεταφρασμένη εικόνα από [178])*

Το προσεχές μοντέλο πρόληψης του καρκίνου του τραχήλου της μήτρας περιλαμβάνει εμβολιασμό όλων των κοριτσιών και γυναικών ηλικίας 9-15 ετών, έλεγχο από ηλικία 25 ετών και ανά 3-5ετία με HPV test και επί θετικού HPV test, πραγματοποίηση PAP test ή/και κολποσκόπησης (*εικόνα 102*).



Εικόνα 102: *Πρόληψη του καρκίνου του τραχήλου της μήτρας (πάνω: παρόν μοντέλο, κάτω: προσεχές μοντέλο)*

Ωστόσο, σύμφωνα με τη μελέτη [179]**,** έτσι ώστε το HPV testing να αξιοποιήσει πλήρως τις δυνατότητές του, απαιτούνται νέες προσεγγίσεις, με καλύτερη ειδικότητα, είτε ως triage tests για θετικές σε HPV γυναίκες, είτε εάν μπορεί να διατηρηθεί η υψηλή ευαισθησία του HPV DNA test, ως εναλλακτικές κύριες μεθόδους πληθυσμιακού ελέγχου.

Πρόσθετοι βιοδείκτες (biomarkers) αρχίζουν να μελετώνται και να χρησιμοποιούνται τώρα για triage εξετάσεις για τη διαχείριση θετικών σε HPV γυναικών. Εκτός από τον καθορισμό του γονότυπου του HPV (16,18 και άλλων υψηλού κινδύνου τύπων), χρησιμοποιούνται η εξέταση RNA για τις πρωτεΐνες HPV E6 και E7 (με εντοπισμό υπερέκφρασης των E6 και E7 να υποδεικνύει προκαρκινικές ή καρκινικές αλλαγές), μέθοδοι μεθυλίωσης των γονιδίων του ξενιστή και των ιικών γονιδίων, καθώς και νέες κυτταρολογικές μέθοδοι, όπως για παράδειγμα, η κλινική χρήση της ανοσοχημικής χρώσης για ανίχνευση του p16, η οποία εκμεταλλεύεται τη συσχέτιση της υπερέκφρασης του p16 και των τραχηλικών νεοπλασιών. Άλλοι βιοδείκτες που διερευνώνται είναι ο εντοπισμός των πρωτεϊνών E4 και L1. Περαιτέρω πληροφορίες για τους προαναφερθέντες βιοδείκτες δίδονται στο [179].

Το βασικό πρόβλημα των υπάρχοντων διαγνωστικών τεχνικών προληπτικού πληθυσμιακού ελέγχου και triage, ωστόσο, παραμένει: κάθε μεμονωμένη εξέταση παρουσιάζει είτε υψηλή ευαισθησία είτε υψηλή ειδικότητα, αλλά όχι και τα δύο ταυτοχρόνως. Επομένως, σήμερα δεν υπάρχει μια ιδανική μέθοδος για αποδοτικό πληθυσμιακό έλεγχο και διαχείριση των γυναικών με κίνδυνο ανάπτυξης καρκίνου του τραχήλου της μήτρας.

Κεφάλαιο 5 – Υλοποίηση συστήματος

5.1 Παρουσίαση Προβλήματος

Σε πρώτο στάδιο χρησιμοποιούνται 3 διαφορετικές τεχνικές φιλτραρίσματος για τη διαδικασία κατάταξης των HPV τύπων ως προς τη σημαντικότητά τους για τη διάγνωση του καρκίνου του τραχήλου της μήτρας. Οι τεχνικές που χρησιμοποιούνται είναι χρήση της καμπύλης ROC, η μέθοδος mRMR και η μέθοδος RELIF.

Σε δεύτερο στάδιο προτείνεται ένα σύστημα υποστήριξης κλινικών αποφάσεων με χρήση γενετικών αλγορίθμων και του ταξινομητή Naïve-Bayes με στόχο την αναζήτηση του βέλτιστου συνδυασμού χαρακτηριστικών που δίνει τα πιο ισορροπημένα, ως προς ειδικότητα και ευαισθησία, αποτελέσματα.

Ο καρκίνος του τραχήλου της μήτρας (CxCa) κατατάσσεται παγκοσμίως ως ο τρίτος πιο κοινός τύπος καρκίνου και η τέταρτη κύρια αιτία θανάτου από καρκίνο στις γυναίκες. Αυτά τα περιστατικά και οι θάνατοι συμβαίνουν κυρίως στις αναπτυσσόμενες χώρες (πάνω από 85%) λόγω της έλλειψης οργανωμένων προγραμμάτων πληθυσμιακού ελέγχου (organized screening programs) που επιτρέπουν την ανίχνευση των προκαρκινικών αλλοιώσεων του τραχήλου της μήτρας σε πρώιμο στάδιο. Στην πραγματικότητα, οργανωμένος μαζικός πληθυσμιακός έλεγχος για καρκίνο του τραχήλου της μήτρας εφαρμόζεται σε μάλλον αρκετά περιορισμένο αριθμό χωρών. Στις πλείστες ανεπτυγμένες χώρες εφαρμόζεται ευκαιριακός προληπτικός πληθυσμιακός έλεγχος. Ακόμα και σε καλά οργανωμένα προγράμματα ελέγχου και παρά τις προόδους του προληπτικού ελέγχου, ο καρκίνος του τραχήλου της μήτρας παραμένει ένα σοβαρό πρόβλημα της δημόσιας υγείας και στις ανεπτυγμένες χώρες λόγω του σχετικά υψηλού ποσοστού αποτυχίας της ανίχνευσης της νόσου [182]. Οι τραχηλικές ενδοεπιθηλιακές νεοπλασίες (CINs), μια προκαρκινική κατάσταση, είναι πολύ συχνές, με σχεδόν μία στις δέκα γυναίκες να παρουσιάζουν τέτοιου είδους αλλοιώσεις στον προκαταρκτικό έλεγχο.

Ο πληθυσμιακός έλεγχος με χρήση του τεστ Παπανικολάου έχει μειώσει δραματικά τα ποσοστά του καρκίνου του τραχήλου της μήτρας παγκοσμίως. Προς το παρόν, σε πολλές χώρες η πρόληψη του CxCa είναι βασισμένη σε συχνά και επαναλαμβανόμενα τεστ Παπανικολάου, ακολουθούμενα από κολποσκόπηση και εάν χρειαστεί (δηλαδή εφόσον το τεστ Παπανικολάου ή η κολποσκόπηση είναι μη φυσιολογική) από ιστολογική εξέταση στο βιολογικό υλικό της βιοψίας. Ωστόσο, η αξιολόγηση των κυτταρολογικών επιχρισμάτων του τραχήλου της μήτρας αποτελεί μια δύσκολη διεργασία και μπορεί να επιτευχθεί μόνο από πολύ καλά εκπαιδευμένο ιατρικό

προσωπικό (κυτταροπαθολόγους). Ως εκ τούτου, η ερμηνεία τους επηρεάζεται από υποκειμενικούς παράγοντες και είναι επιρρεπής σε διαγνωστικά σφάλματα.

Ο καρκίνος του τραχήλου της μήτρας σχεδόν όλες τις φορές προκαλείται από το ιό των ανθρωπίνων θηλωμάτων (human papillomavirus-HPV), μια από τις συχνότερες σεξουαλικώς μεταδιδόμενες λοιμώξεις. Υπάρχουν άνω των 100 τύπων HPV που μπορούν να προσβάλουν ανθρώπους, ωστόσο, μόνο 14 θεωρούνται εξαιρετικά ογκογόνοι και μπορούν να προκαλέσουν CxCa. Η παρουσία του HPV, εντούτοις, δεν οδηγεί όλες τις φορές στην εμφάνιση της νόσου, αφού η λοίμωξη ενδέχεται να υποχωρήσει εξαιτίας του ανθρώπινου ανοσοποιητικού συστήματος. Οι εξελίξεις στην κατανόηση του ρόλου της HPV λοίμωξης στη φυσική εξέλιξη των νεοπλασιών του τραχήλου της μήτρας, είχαν ως αποτέλεσμα την παράλληλη διενέργεια της εξέτασης HPV DNA μαζί με το τεστ Παπανικολάου [183]. Η εξέταση HPV DNA πλέον χρησιμοποιείται ως βοηθητική εξέταση στο τεστ Παπανικολάου. Επιπρόσθετα, έχει προταθεί η χρησιμοποίησή της για τον προκαταρκτικό έλεγχο. Λόγω αυτού, πολλές αναπτυγμένες χώρες συμπεριλαμβάνουν την εξέταση HPV DNA στις επίσημες κατευθυντήριες γραμμές του πληθυσμιακού ελέγχου για τον CxCa που ακολουθούν.

Έχουν γίνει πολλές μελέτες που επιχειρούν να αναλύσουν το ρόλο της εξέτασης HPV και να τη συγκρίνουν μαζί με το τεστ Παπανικολάου [184], [185], [186]. Οι μελέτες αυτές δείχνουν ότι η απόδοση των δύο αυτών διαγνωστικών εξετάσεων διαφέρει σημαντικά: παρουσιάζουν είτε υψηλή ευαισθησία είτε υψηλή ειδικότητα αλλά όχι και τα δύο ταυτοχρόνως. Επομένως, σήμερα, δεν υπάρχει η ιδανική διαγνωστική εξέταση. Εκτός αυτού, τα αποτελέσματα των σχετικών μελετών επηρεάζονται από την επίπτωση (Γ.1) και τον επιπολασμό (Γ.2) της νόσου καθώς και την παρουσία της HPV λοίμωξης στον υπό μελέτη πληθυσμό, με αποτέλεσμα η εφαρμογή μόνο μιας εξέτασης να προσφέρει προστασία σε ένα βαθμό, αλλά να μην μπορεί να προσδιορίσει με αξιοπιστία τον πραγματικό κίνδυνο κάθε γυναίκας που συμμετέχει σε προγράμματα προσυμπτωματικού ελέγχου του CxCa.

Η μετανάλυση των δημοσιευμένων μελετών [184], [185], [186] δείχνει ότι η ευαισθησία του τεστ Παπανικολάου όταν συνδυάζεται με την εξέταση HPV DNA είναι πιο υψηλή από ότι η ευαισθησία της κάθε μεμονωμένης εξέτασης. Επομένως, οι δύο διαγνωστικές εξετάσεις συμπληρώνουν αποτελεσματικά η μια την άλλη. Από την άλλη, όμως, η ειδικότητα του τεστ Παπανικολάου όταν συνδυάζεται με την εξέταση HPV DNA είναι πιο χαμηλή από ότι η ειδικότητα κάθε μιας μεθόδου ξεχωριστά. Όσον αφορά τη θετική προγνωστική αξία (PPV), έχουν προκύψει αντικρουόμενα ευρήματα: κάποιες μελέτες αναφέρουν παρόμοιες τιμές PPV για κάθε διαγνωστική εξέταση ξεχωριστά, καθώς και για το συνδυασμό τους, ενώ άλλες μελέτες αναφέρουν μικρότερες τιμές PPV για το συνδυασμό των δύο εξετάσεων. Όπως αναμένεται η

αρνητική προγνωστική αξία (NPV) του συνδυασμού των δύο εξετάσεων είναι υψηλή, με πολλές μελέτες μάλιστα να αναφέρουν τιμές σχεδόν ίσες με 100%.

Σήμερα, παρά τις προόδους στον προληπτικό πληθυσμιακό έλεγχο του CxCa, δεν υπάρχει κοινή συναίνεση για τη βέλτιστη διαχείριση των γυναικών με μη φυσιολογικά αποτελέσματα στις διαγνωστικές εξετάσεις. Ένα ποσοστό των γυναικών που, ενώ με βάση την κυτταρολογική εξέταση, έχουν άτυπα πλακώδη κύτταρα απροσδιόριστης σημασίας (ASCUS) ή χαμηλού βαθμού αλλοιώσεις του πλακώδους επιθηλίου (LSIL), μπορεί στην πραγματικότητα να έχουν υψηλού βαθμού ενδοεπιθηλιακή νεοπλασία του τραχήλου (CIN-2 ή CIN-3). Αυτές οι γυναίκες βρίσκονται σε πολύ υψηλό κίνδυνο να αναπτύξουν CxCa. Επομένως, η κατηγορία CIN-2 αποτελεί το κατώφλι απόφασης, πέραν του οποίου το περιστατικό αντιμετωπίζεται χειρουργικά. Αντιθέτως, γυναίκες με κυτταρολογική διάγνωση CIN-1 δεν αντιμετωπίζονται χειρουργικά, εντούτοις, όμως, τυγχάνουν αυστηρής παρακολούθησης. Ωστόσο μπορεί να συμβεί και το αντίστροφο: δεν είναι σπάνιο γυναίκες με κυτταρολογική διάγνωση που τις κατατάσσει στην κατηγορία «υψηλού βαθμού πλακώδους επιθηλιακής νεοπλασίας» (HSIL), στην πραγματικότητα να έχουν CIN-1 ή ακόμα και μια φυσιολογική ιστολογική διάγνωση. Ως εκ τούτου, οι επιλογές διαχείρισης των κυτταρολογικών διαγνώσεων ASCUS ή LSIL, που είναι ευρέως αποδεκτές στην παρούσα φάση, είναι οι ακόλουθες: είτε άμεση πραγματοποίηση κολποσκόπησης, είτε κυτταρολογική επιτήρηση με συχνό επανέλεγχο του περιστατικού με τεστ Παπανικολάου.

Η πολιτική άμεσης παραπομπής για κολποσκόπηση μπορεί εύκολα να έχει ως αποτέλεσμα την υπερφόρτωση των κλινικών κολποσκόπησης, καθώς και την υπερβολική παρέμβαση και/ή υπερθεραπεία σε περίπτωση της παραμικρής μη φυσιολογικής ένδειξης στην κολποσκόπηση. Έτσι, οι γυναίκες εκτίθενται στα σωματικά και ψυχολογικά επακόλουθα της περιττής θεραπείας, η οποία μάλιστα σε εγκυμονούσες γυναίκες εμπεριέχει τον κίνδυνο του πρόωρου τοκετού. Επιπλέον, η προσέγγιση της άμεσης κολποσκόπησης φαίνεται να έχει αξιοσημείωτο ψυχολογικό κόστος στις γυναίκες, οι οποίες αντιδρούν με ανησυχία, φόβο και ενίοτε με πανικό. Η κολποσκόπηση προκαλεί ανησυχία τόσο για την ίδια τη διαδικασία, όσο και για το αποτέλεσμά της.

Από την άλλη πλευρά, τα επαναλαμβανόμενα τεστ Παπανικολάου εμπεριέχουν τον κίνδυνο να μην εντοπιστούν οι HSIL, αυξάνουν τα ποσοστά μη συμμόρφωσης (non-conformance rates), αυξάνουν το κόστος των οργανωμένων προγραμμάτων προσυμπωματικού ελέγχου, αυξάνουν το ψυχολογικό και κοινωνικό φορτίο των γυναικών και τελικά κλονίζουν την αξιοπιστία των οργανωμένων προγραμμάτων προσυμπωματικού ελέγχου.

Κατά συνέπεια, είναι σημαντική η ορθή αναγνώριση των γυναικών που βρίσκονται σε πραγματικό κίνδυνο ανάπτυξης CxCa και ταυτοχρόνως η μείωση αχρείαστων κολποσκοπήσεων και επαναλαμβανόμενων Παπ τεστ. Είναι προφανές, ότι είναι απαραίτητο να καταβληθούν

προσπάθειες για δημιουργία μιας μεθόδου προληπτικού ελέγχου, με κατώφλι την ενδοεπιθηλιακή νεοπλασία του τραχήλου της μήτρας 2ου βαθμού ή άνω (CIN2+), η οποία να παρουσιάζει ταυτόχρονα υψηλή ευαισθησία και υψηλή ειδικότητα. Με βάση αυτή την απαίτηση, σε αυτή τη διπλωματική εργασία, παρουσιάζεται ένα σύστημα υποστήριξης λήψης κλινικών αποφάσεων, το οποίο βασίζεται σε ένα συνδυασμό Γενετικών Αλγορίθμων και Μπεϋζιανής ταξινόμησης, που παρουσιάζει ισορροπημένη ειδικότητα και ευαισθησία στην ανίχνευση CIN2+, συνδυάζοντας τα αποτελέσματα του Pap test και του HPV DNA test.

5.2 Κατάταξη των HPV τύπων με βάση τεχνικές filtering

Έχουν γίνει διάφορες επιδημιολογικές μελέτες με στόχο την κατάταξη των διαφόρων τύπων του ιού HPV ως προς τον εμπλεκόμενο κίνδυνο ανάπτυξης καρκίνου του τραχήλου της μήτρας. Με βάση αυτές τις μελέτες, οι HPV τύποι έχουν χωριστεί σε τύπους υψηλού κινδύνου και σε τύπους χαμηλού κινδύνου.

Στην παρούσα εργασία, γίνεται χρήση τεχνικών φιλτραρίσματος (filter techniques) για την κατάταξη των HPV τύπων με βάση τη σημαντικότητά τους για την ανάπτυξη ενδοεπιθηλιακής νεοπλασίας του τραχήλου της μήτρας 2ου βαθμού ή άνω (CIN2+). Οι μέθοδοι αυτές, μπορεί να εντοπίσουν συσχετίσεις που ενδεχομένως, αλλιώς, να μην ήταν προφανείς. Οι τύποι που κατατάσσονται πρώτοι είναι αυτοί που φαίνεται να φέρουν τον μεγαλύτερο κίνδυνο για ανάπτυξη CIN2+.

Η κατάταξη των HPV τύπων αντιμετωπίζεται ως πρόβλημα επιλογής χαρακτηριστικών για δυαδική ταξινόμηση περιστατικών σε κλάσεις με βάση το κατώφλι της ενδοεπιθηλιακής νεοπλασίας του τραχήλου της μήτρας 2ου βαθμού ή άνω (CIN2+). Ως σύνολο πιθανών χαρακτηριστικών θεωρούμε τους 35 HPV γονότυπους: [6, 11, 16, 18, 26, 31, 33, 35, 39, 40, 42, 43, 44, 45, 51, 52, 53, 54, 56, 58, 59, 61, 62, 66, 68, 70, 71, 72, 73, 81, 82, 83, 84, 85, 89]. Η κατάταξη γίνεται με ένα συνδυασμό των ακόλουθων μεθόδων φιλτραρίσματος: καμπύλη ROC, μέθοδος mRMR και μέθοδος RELIEF.

5.2.1 Εφαρμογή καμπύλης ROC

Πίνακας 19: Κατάταξη τύπων HPV με χρήση της καμπύλης ROC

Σειρά κατάταξης	*Τύπος HPV*
1	HPV-16
2	HPV-18
3	HPV-53
4	HPV-45
5	HPV-39
6	HPV-42
7	HPV-84
8	HPV-31
9	HPV-59
10	HPV-40
11	HPV-70
12	HPV-33
13	HPV-35
14	HPV-51
15	HPV-82
16	HPV-43
17	HPV-68
18	HPV-73

5.2.2 Εφαρμογή μεθόδου mRMR

Ο πίνακας των βέλτιστων 18 χαρακτηριστικών που προέκυψαν από την εφαρμογή της μεθόδους mRMR παρατίθεται ακολούθως.

Πίνακας 20: Κατάταξη τύπων HPV με χρήση της μεθόδου mRMR

Σειρά κατάταξης	*Τύπος HPV*
1	HPV-58
2	HPV-42
3	HPV-16
4	HPV-6
5	HPV-33
6	HPV-52
7	HPV-18
8	HPV-56
9	HPV-31

10	HPV-45
11	HPV-53
12	HPV-51
13	HPV-35
14	HPV-59
15	HPV-44
16	HPV-62
17	HPV-66
18	HPV-73

5.2.3 Εφαρμογή μεθόδου RELIEF

Όπως έχει αναφερθεί, ο αλγόριθμος Relief βασίζεται στην αναζήτηση των δύο κοντινότερων γειτόνων ενός γνωρίσματος, από τους οποίους ο ένας βρίσκεται στην ίδια κατηγορία με το γνώρισμα (nearest hit) και ο άλλος σε διαφορετική (nearest miss). Μαθηματικά, εκφράζεται μέσω της διαφοράς:

$$W(F) = \mathrm{P}\,(\text{different value of } F | nearest\ instance\ from\ different\ \text{class}) - P\,(different\ value\ of\ F |\ nearest\ instance\ from\ same\ class)$$

Συνεπώς, όσο μεγαλύτερη είναι η τιμή της διαφοράς τόσο μεγαλύτερη είναι η διακριτική ισχύς του γνωρίσματος. Τα αποτελέσματα της εφαρμογής του αλγορίθμου παρατίθενται στον ακόλουθο πίνακα.

Πίνακας 21: Κατάταξη τύπων HPV με χρήση της μεθόδου RELIEF

Σειρά κατάταξης	***Τύπος HPV***
1	HPV-16
2	HPV-31
3	HPV-51
4	HPV-6
5	HPV-56
6	HPV-35
7	HPV-83
8	HPV-53
9	HPV-59

10	HPV-18
11	HPV-39
12	HPV-82
13	HPV-52
14	HPV-66
15	HPV-61
16	HPV-11
17	HPV-68
18	HPV-73

5.2.4 Συνδυασμός των παραπάνω μεθόδων

Τα επιμέρους αποτελέσματα των τριών προαναφερθεισών μεθόδων (χρήση καμπύλης ROC, mRMR, RELIEF) παρουσιάζονται στον ακόλουθο πίνακα.

Πίνακας 22: Κατάταξη τύπων HPV από τις προαναφερθείσες μεθόδους

Σειρά κατάταξης	***ROC***	***mRMR***	***RELIEF***
1	HPV-16	HPV-58	HPV-16
2	HPV-18	HPV-42	HPV-31
3	HPV-53	HPV-16	HPV-51
4	HPV-45	HPV-6	HPV-6
5	HPV-39	HPV-33	HPV-56
6	HPV-42	HPV-52	HPV-35
7	HPV-84	HPV-18	HPV-83
8	HPV-31	HPV-56	HPV-53
9	HPV-59	HPV-31	HPV-59
10	HPV-40	HPV-45	HPV-18
11	HPV-70	HPV-53	HPV-39
12	HPV-33	HPV-51	HPV-82
13	HPV-35	HPV-35	HPV-52
14	HPV-51	HPV-59	HPV-66
15	HPV-82	HPV-44	HPV-61
16	HPV-43	HPV-62	HPV-11
17	HPV-68	HPV-66	HPV-68
18	HPV-73	HPV-73	HPV-73

Από τη συνολική επισκόπηση των αποτελεσμάτων μπορεί να παρατηρηθεί ότι τα αποτελέσματα διαφέρουν ανάλογα με τη μέθοδο που χρησιμοποιείται, με το συνδυασμό τους να οδηγεί δυνητικά σε πιο αντιπροσωπευτικά αποτελέσματα. Ακολούθως παρουσιάζεται η κατάταξη

όπως προκύπτει από ένα συνδυασμό των αποτελεσμάτων των παραπάνω μεθόδων. Ο συνδυασμός αυτός, υλοποιήθηκε με τη βαθμολόγηση κάθε κατάταξης και τον υπολογισμό της μέσης βαθμολογίας για κάθε χαρακτηριστικό (τύπο HPV). Η τελική κατάταξη διαμορφώνεται βάσει της μέσης βαθμολογίας των χαρακτηριστικών και καταγράφεται στον πίνακα που ακολουθεί.

Πίνακας 23: Τελική κατάταξη τύπων HPV μέσω συνδυασμού των προαναφερθούσων μεθόδων

Σειρά κατάταξης	***Τύπος HPV***
1	HPV-16
2	HPV-18
3	HPV-31
4	HPV-53
5	HPV-51
6	HPV-35
7	HPV-59
8	HPV-66
9	HPV-6
10	HPV-42
11	HPV-45
12	HPV-56
13	HPV-39
14	HPV-33
15	HPV-52
16	HPV-82
17	HPV-68
18	HPV-73

5.3 Σύστημα ταξινόμησης που συνδυάζει Pap test και HPV DNA test: Εύρεση του βέλτιστου συνδυασμού χαρακτηριστικών με χρήση Γενετικών Αλγορίθμων

5.3.1 Κλινικά δεδομένα

Για τους σκοπούς της παρούσας διπλωματικής, χρησιμοποιήθηκαν ανώνυμα δεδομένα. Τη συλλογή των δεδομένων ηγείτο το τμήμα κυτταρολογίας της Ιατρικής Σχολής του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών (Πανεπιστημιακό Γενικό Νοσοκομείο «Αττικόν»). Η Επιτροπή Ηθικής και Δεοντολογίας του ιδρύματος ενέκρινε τη μελέτη, καθώς και όλες τις

διαδικασίες. Όλες οι γυναίκες συμμετέχοντες υπέγραψαν ένα έντυπο έγγραφης συγκατάθεσης ασθενούς κατόπιν ενημέρωσης, που επιτρέπει την ανώνυμη χρήση των δεδομένων τους για σκοπούς έρευνας.

Τα δεδομένα που συλλέχθησαν συμπεριλάμβαναν αποτελέσματα από τις ακόλουθες εξετάσεις: HPV DNA τεστ, ΠΑΠ τεστ και ιστολογική εξέταση εφόσον γινόταν, καθώς και στοιχεία του ασθενή: δημογραφικά στοιχεία και στοιχεία ταυτοποίησης. Τα δεδομένα αποθηκεύτηκαν σε μια βάση δεδομένων και αποτελούνται από 740 περιστατικά με πλήρη δεδομένα για όλες τις εξετάσεις (χωρίς ελλιπή στοιχεία). Τα δεδομένα αυτά των 740 περιστατικών, εξήχθησαν από τη βάση και χρησιμοποιήθηκαν ως ανώνυμα για περαιτέρω ανάλυση.

Κάθε παράδειγμα των δεδομένων αντιστοιχεί σε ένα περιστατικό γυναίκας και αποτελείται από τα αποτελέσματα του ΠΑΠ τεστ και του HPV DNA τεστ. Για κάθε μια από τις 740 γυναίκες, δημιουργήθηκε ένα σύνολο χαρακτηριστικών που αποτελείτο από 40 δείκτες, οι οποίοι προέκυψαν από τα αποτελέσματα των εξετάσεων.

Το βιολογικό υλικό συλλέχθηκε σε φιαλίδια ThinPrep (Κυτταρολογία υγρής φάσης - LBC). Το HPV DNA τεστ πραγματοποιήθηκε με χρήση του CLART HUMAN PAPILLOMAVIRUS 2 test, το οποίο εντοπίζει ταυτοχρόνως 35 διαφορετικούς τύπους HPV, υψηλού ή χαμηλού κινδύνου.

Τα αποτελέσματα της εξέτασης HPV DNA εκφράστηκαν ως 35 διαφορετικές μεταβλητές, μια για κάθε γονότυπο HPV DNA (6, 11, 16, 18, 26, 31, 33, 35, 39, 40, 42, 43, 44, 45, 51, 52, 53, 54, 56, 58, 59, 61, 62, 66, 68, 70, 71, 72, 73, 81, 82, 83, 84, 85, 89), που λαμβάνουν τιμή θετική ή αρνητική. Εκτός από αυτές τις 35, χρησιμοποιούνται και κάποιες επιπρόσθετες μεταβλητές που επίσης εκφράζουν τα αποτελέσματα της εξέτασης HPV DNA: για παράδειγμα, προστέθηκε μια μεταβλητή που εκφράζει την ύπαρξη υψηλού κινδύνου τύπων-HR (16, 18, 26, 31, 33, 35, 39, 45, 51, 52, 53, 56, 58, 59, 66, 68, 70, 73, 82, 85), μια άλλη που εκφράζει την ύπαρξη χαμηλού κινδύνου τύπων-LR (6, 11, 40, 42, 43, 44, 54, 61, 62, 71, 72, 81, 83, 84, 89), που επίσης λαμβάνουν τιμή αρνητική ή θετική.

Τα 40 χαρακτηριστικά που χρησιμοποιήθηκαν για κάθε περιστατικό παρουσιάζονται στον *πίνακα 24*.

Πίνακας 24: Περιγραφή του συνόλου χαρακτηριστικών

Χαρακτηριστικό	Περιγραφή	Εύρος τιμών
ΠΑΠ τεστ	Το αποτέλεσμα της κυτταρολογικής εξέτασης εκφρασμένη σύμφωνα με το TBS 2001.	1: WNL 2: ASCUS 3: LSIL 4: HSIL 5: Cancer
Συστοιχίες HPV DNA για τους γονότυπους: 6, 11, 16, 18, 26, 31, 33, 35, 39, 40, 42, 43, 44, 45, 51, 52, 53, 54, 56, 58, 59, 61, 62, 66, 68, 70, 71, 72, 73, 81, 82, 83, 84, 85, 89	Η ύπαρξη ή όχι κάθε ξεχωριστού γονότυπου σύμφωνα με την εξέταση HPV DNA.	0 , εάν ο συγκεκριμένος γονότυπος δεν έχει εντοπιστεί, 1, εάν ο συγκεκριμένος γονότυπος έχει εντοπιστεί
HPV DNA	Θετικό, εάν ένας ή περισσότεροι γονότυποι έχουν εντοπιστεί από την εξέταση HPV DNA (αλλιώς αρνητικό).	0 , εάν δεν έχει εντοπιστεί ούτε και ένας γονότυπος (αρνητικό τεστ), 1, εάν ένας ή περισσότεροι γονότυποι έχουν εντοπιστεί (θετικό τεστ)
HR-HPV DNA	Θετικό, εάν ένας ή περισσότεροι γονότυποι υψηλού κινδύνου (16, 18, 26, 31, 33, 35, 39, 45, 51, 52, 53, 56, 58, 59, 66, 68, 70, 73, 82, 85) έχουν εντοπιστεί από την εξέταση HPV DNA (αλλιώς αρνητικό).	0 , εάν δεν έχει εντοπιστεί ούτε και ένας γονότυπος υψηλού κινδύνου, 1, εάν έστω και ένας γονότυπος υψηλού κινδύνου έχει εντοπιστεί
VHR-HPV DNA	Θετικό, εάν ένας ή περισσότεροι γονότυποι πολύ υψηλού κινδύνου (16, 18, 31, 33, 45) έχουν εντοπιστεί από την εξέταση HPV DNA (αλλιώς αρνητικό).	0 , εάν δεν έχει εντοπιστεί ούτε και ένας γονότυπος πολύ υψηλού κινδύνου, 1, εάν έστω και ένας γονότυπος πολύ υψηλού κινδύνου έχει εντοπιστεί
LR-HPV DNA	Θετικό, εάν ένας ή περισσότεροι γονότυποι χαμηλού κινδύνου (6, 11, 40, 42, 43, 44, 54, 61, 62, 71, 72, 81, 83, 84, 89) έχουν εντοπιστεί από την εξέταση HPV DNA (αλλιώς αρνητικό).	0 , εάν δεν έχει εντοπιστεί ούτε και ένας γονότυπος χαμηλού κινδύνου, 1, εάν έστω και ένας γονότυπος χαμηλού κινδύνου έχει εντοπιστεί

Για τον τελικό καθορισμό της κατηγορίας στην οποία ανήκει κάθε περιστατικό χρησιμοποιούνται τα αποτελέσματα της ιστολογικής εξέτασης, εφόσον είχε γίνει, διαφορετικά χρησιμοποιούνται τα αποτελέσματα της κολποσκόπησης.

Οι συμμετέχουσες γυναίκες είχαν αναφερθεί για κολποσκόπηση είτε επειδή είχαν μη φυσιολογικό τεστ Παπανικολάου, είτε επειδή, αν και είχαν φυσιολογικό τεστ Παπανικολάου, είχαν εθελοντικά συμμετάσχει στη μελέτη και αποδέχτηκαν κολποσκοπική εξέταση (καθώς και τη χρήση του βιολογικού υλικού τους για εξέταση).

Στις γυναίκες, των οποίων το αρνητικό ΠΑΠ τεστ ακολουθήθηκε από αρνητική κολποσκόπηση, δεν έγινε βιοψία και θεωρούνται ως κλινικά αρνητικά περιστατικά. Σε περιστατικά που ήταν κλινικώς αρνητικά λόγω αρνητικής κυτταρολογίας και αρνητικής κολποσκοπίας, δεν πάρθηκαν βιοψίες, αφού δεν θα ήταν ηθική η λήψη βιολογικών δειγμάτων για ιστολογική εξέταση.

Εάν το ΠΑΠ τεστ αποκάλυπτε κατηγορία ASCUS ή πιο υψηλή κυτταρολογική κατηγορία (ASCUS+) και κατά την κολποσκόπηση παρατηρείτο ορατή αλλοίωση, τότε γινόταν βιοψία του τραχήλου της μήτρας. Για τα περιστατικά που έχουν ιστολογικό αποτέλεσμα από τη βιοψία, ως καθοριστικό αποτέλεσμα, χρησιμοποιείται αυτό της ιστολογικής διάγνωσης κι όχι των άλλων των τεστ.

Τα αποτελέσματα της κυτταρολογικής εξέτασης για κάθε γυναίκα ερμηνεύονται σύμφωνα με το σύστημα ταξινόμησης Bethesda (σύστημα TBS 2001), το οποίο κατατάσσει κάθε περίπτωση σε μια από τις ακόλουθες κατηγορίες:

- Εντός φυσιολογικών ορίων (within normal limits-WNL)
- Άτυπα κύτταρα του πλακώδους επιθηλίου απροσδιόριστης σημασίας (ASCUS)
- Χαμηλού βαθμού ενδοεπιθηλιακή αλλοίωση πλακώδους επιθηλίου (LSIL)
- Υψηλού βαθμού ενδοεπιθηλιακή αλλοίωση πλακώδους επιθηλίου (HSIL)
- Πλακώδης επιθηλιακός καρκίνος (SCC) ή αδενοκαρκίνωμα (Adeno-Ca)

Για την ιστολογική διάγνωση χρησιμοποιείται το σύστημα Richard, σύμφωνα με το οποίο κάθε περιστατικό κατατάσσεται σε μια από τις παρακάτω κατηγορίες (σε αύξουσα σειρά ως προς τη σοβαρότητα):

- Κλινικώς αρνητικά
- Αρνητικά (καμία ένδειξη κακοήθειας)
- Ενδοεπιθηλιακή νεοπλασία βαθμού 1 (CIN-1)
- Ενδοεπιθηλιακή νεοπλασία βαθμού 2 ή 3 (CIN-2/3)
- Καρκίνος (CxCa): Πλακώδες επιθηλιακό καρκίνωμα (SCC) ή αδενοκαρκίνωμα (Adeno-Ca)

Η κατανομή των 740 περιστατικών παρουσιάζεται στον *πίνακα 25*.

Πίνακας 25: Κατανομή περιστατικών

Αποτέλεσμα Ιστολογίας	*Αποτέλεσμα του ΠΑΠ τεστ*					Συνολικά
	WNL	**ASCUS**	**LSIL**	**HSIL**	**CxCa**	
Κλινικώς αρνητικά	196	0	0	0	0	196 (26.5%)
Αρνητικά	35	60	22	5	0	122 (16.5%)
CIN-1	31	66	142	22	0	261 (35.3%)
CIN-2/3	3	13	27	93	0	136 (18.4%)
CxCa	0	1	2	7	15	25 (3.4%)
Συνολικά	265	140	193	127	15	740

5.3.2 Επισκόπηση προβλήματος: Εργαλείο Γενετικού αλγόριθμου σε συνδυασμό με ταξινομητή Naïve-Bayes για την επιλογή χαρακτηριστικών

Ο σκοπός αυτής της μελέτης είναι η δημιουργία ενός συστήματος ταξινόμησης το οποίο θα συνδυάζει αποτελεσματικά τα αποτελέσματα του ΠΑΠ τεστ και του HPV-DNA τεστ. Ωστόσο, επικεντρωνόμαστε, όχι τόσο στην ακρίβεια της ταξινόμησης, αλλά στη δημιουργία ενός συστήματος το οποίο θα οδηγεί σε πιο ισορροπημένα αποτελέσματα όσον αφορά την ευαισθησία και την ειδικότητα. Για την επίτευξη του ισορροπημένου αυτού αποτελέσματος, απαιτείται η εύρεση ενός υποσυνόλου χαρακτηριστικών, το οποίο όταν τροφοδοτηθεί ως είσοδος στον ταξινομητή, θα ικανοποιήσει το στόχο αυτό. Επομένως, το πρόβλημα που αντιμετωπίζουμε είναι περισσότερο πρόβλημα επιλογής χαρακτηριστικών παρά πρόβλημα ταξινόμησης.

Όπως έχουμε αναφέρει, το πρόβλημα επιλογής υποσυνόλου χαρακτηριστικών μπορεί να θεωρηθεί ως ένα πρόβλημα αναζήτησης, στο οποίο όμως, η εξαντλητική αναζήτηση για τον δοσμένο αριθμό διαθέσιμων χαρακτηριστικών [40], είναι υπολογιστικά απαγορευτική. Συνεπώς, μια ευριστική αναζήτηση, όπως οι ΓΑ, είναι μια πιο κατάλληλη προσέγγιση για την επίλυση του συγκεκριμένου προβλήματος.

Στην παρούσα θέση, έχει υιοθετηθεί ένας συνδυασμός Γενετικού Αλγορίθμου και Bayesian ταξινομητή για την υλοποίηση της επιλογής υποσυνόλου χαρακτηριστικών. Ο Γενετικός Αλγόριθμος και ο ταξινομητής Naïve-Bayes (NB) ενσωματώνονται ακολουθώντας μια wrapper προσέγγιση.

5.3.3 Επιλογή του ταξινομητή Naïve Bayes

5.3.3.1 Απλοϊκός ταξινομητής κατά Bayes (Naïve Bayes)

Ο απλοϊκός ταξινομητής κατά Bayes (Naïve Bayes - NB) είναι ένας πιθανοτικός αλγόριθμος που μοντελοποιεί τις πιθανολογικές σχέσεις μεταξύ του συνόλου των χαρακτηριστικών και της

μεταβλητής της κλάσης. Ένας Μπεϋζιανός ταξινομητής, δοθέντος των τιμών των χαρακτηριστικών ενός νέου παραδείγματος, υπολογίζει τις δεσμευμένες πιθανότητες για όλες τις πιθανές κλάσεις. Με αυτό τον τρόπο αποφασίζει σε ποια κλάση ανήκει αυτό το παράδειγμα: θα το κατατάξει στην κλάση με τη μεγαλύτερη μεταγενέστερη πιθανότητα κλάσης.

Η ταξινόμηση κατά Bayes στηρίζεται στην υπόθεση πως τα στιγμιότυπα του υπό εξέταση προβλήματος ακολουθούν κατανομές πιθανοτήτων και με βάση τις κατανομές αυτές και τα παρατηρηθέντα δεδομένα μπορούν να προκύψουν οι βέλτιστες αποφάσεις. Η μάθηση κατά Bayes απαιτεί τον υπολογισμό πολλών τιμών πιθανοτήτων. Όταν ο υπολογισμός των πιθανοτήτων αυτών δεν υπολογίζεται με ακρίβεια, η προγενέστερη γνώση, που είναι είτε εμπειρική είτε προέρχεται από παλιότερες υποθέσεις, συνδυάζεται με τα παρατηρηθέντα στοιχεία για να καθορίσει την τελική πιθανότητα μιας υπόθεσης. Στην μπεϋζιανή μάθηση, κάθε παράδειγμα μπορεί να αυξήσει ή να μειώσει την εκτιμώμενη πιθανότητα ότι μια υπόθεση/πρόβλεψη είναι ορθή. Οι ακριβείς ταξινομητές κατά Bayes είναι δύσκολο να εξασφαλιστούν εκτός κι αν τα δεδομένα εκπαίδευσης καλύπτουν πλήρως το χώρο των χαρακτηριστικών και δεν υπάρχει καθόλου θόρυβος. Η δυσκολία που παρουσιάζεται στον ακριβή υπολογισμό των τιμών των πιθανοτήτων έχει οδηγήσει σε μια απλουστευμένη εκδοχή της μάθησης κατά Bayes, του απλοϊκού ταξινομητή Bayes (Naïve Bayes), όπου γίνεται η παραδοχή ότι η τιμή ενός χαρακτηριστικού μιας δεδομένης κλάσης είναι ανεξάρτητη από τις τιμές των υπόλοιπων χαρακτηριστικών. Με αυτή την υπόθεση εξασφαλίζονται αξιόπιστες εκτιμήσεις των υπό συνθήκη πιθανοτήτων ακόμα και για πολύ μικρά σύνολα δεδομένων. Ο Naive Bayes ταξινομητής συνήθως χρησιμοποιείται σε προβλήματα με προκαθορισμένες κλάσεις.

Η μάθηση Bayes βασίζεται στο θεώρημα Bayes για την υπό συνθήκη πιθανότητα.

Τύπος του θεωρήματος Bayes (Bayes' theorem) σε απλή μορφή: Σύμφωνα με το θεώρημα του Bayes, η δεσμευμένη πιθανότητα (conditional probability) του γεγονότος A δοθέντος του B, δηλαδή η πιθανότητα να συμβεί το A δεδομένου ότι έχει συμβεί/ ότι θα συμβεί το B είναι ίση με:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Το θεώρημα Bayes μπορεί να διατυπωθεί και σε μορφή κατάλληλη για τυχαίες μεταβλητές αντί για ενδεχόμενα, σε όρους τυχαίων μεταβλητών με πυκνότητες που συμβολίζονται με *f*, ως εξής:

$$f(\theta|x) = \frac{f(\theta)f(x|\theta)}{\int f(\theta)f(x|\theta)d\theta}$$

Η $f(\theta)$ καλείται προγενέστερη/εκ των προτέρων κατανομή (a-priori distribution), η $f(\theta|x)$ καλείται μεταγενέστερη/εκ των υστέρων κατανομή (a-posteriori distribution), η $f(x|\theta)$ καλείται

πιθανοφάνεια και η *f*(x) καλείται σταθερά κανονικοποίησης. Με ολοκλήρωση ως προς θ ο παρονομαστής είναι συνάρτηση ως προς x και συνεπώς για δεδομένες παρατηρήσεις x, ο παρονομαστής είναι σταθερά (σταθερά κανονικοποίησης). Άρα, η μεταγενέστερη κατανομή είναι ανάλογη της προγενέστερης κατανομής πολλαπλασιαζόμενης με τη συνάρτηση πιθανοφάνειας.

Συμπέρασμα Bayes: προέρχεται από τη μεταγενέστερη πιθανότητα ως μια συνέπεια δύο προηγούμενων παραγόντων, αυτού της προηγούμενης πιθανότητας και αυτού της συνάρτησης πιθανότητας που προέρχονται από ένα στατιστικό μοντέλο για τα στοιχεία που παρατηρούμε. Το συμπέρασμα Bayes υπολογίζει την προηγούμενη πιθανότητα σύμφωνα με το θεώρημα Bayes:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Η h χαρακτηρίζεται ως υπόθεση, η P(h) καλείται προγενέστερη πιθανότητα της υπόθεσης h (a-priori probability), η P(h|D) μεταγενέστερη πιθανότητα (a-posteriori probability), η P(D|h) συνάρτηση πιθανότητας των στοιχείων D δεδομένης της υπόθεσης h και η P(D) προγενέστερη πιθανότητα των δεδομένων D. Η μεταγενέστερη πιθανότητα μιας υπόθεσης καθορίζεται από έναν συνδυασμό της προγενέστερης πιθανότητας της υπόθεσης και τη συμβατότητα των παρατηρούμενων στοιχείων με την υπόθεση, δηλαδή οι τιμές του P(h|D) επηρεάζεται μόνο από τους παράγοντες P(h) και P(D|h).

Έστω Y η μεταβλητή κλάσης και X το διάνυσμα των n χαρακτηριστικών, δηλαδή X = {X1, X2, ..., Xn}, όπου Xi μια τυχαία μεταβλητή που υποδηλώνει το χαρακτηριστικό i του X. Ο Μπεϋζιανός ταξινομητής καλείται να προσεγγίσει τη συνάρτηση f: X → Y, ή ισοδύναμα την P(Y|X).

Εφαρμόζοτας το θεώρημα Bayes, η δεσμευμένη πιθανότητα P(Y=yi|X) αναπαριστάται ως εξής:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \Rightarrow$$

$$P(Y = y_i|X = x_k) = \frac{P(X = x_k|Y = y_i)P(Y = y_i)}{P(X = x_k)} = \frac{P(X = x_k|Y = y_i)P(Y = y_i)}{\sum_j P(X = x_k|Y = y_j)P(Y = y_j)}$$

όπου ym αντιστοιχεί στην m πιθανή τιμή της μεταβλητής της κλάσης Y (σε ένα δυαδικό πρόβλημα η Y μπορεί να πάρει μόνο 2 τιμές, 0 ή 1), xk υποδηλώνει την k πιθανή τιμή του διανύσματος X (π.χ. για boolean διάνυσμα X 5 χαρακτηριστικών: 01001) και το άθροισμα στον παρονομαστή καλύπτει όλες τις τιμές που λαμβάνει η μεταβλητή Y (δηλαδή για όλες τις πιθανές κλάσεις).

Συνεπώς για τον υπολογισμό της P(Y|X = xk) για κάθε νέο παράδειγμα xk , ένας τρόπος είναι η εκτίμηση των πιθανοτήτων P(X|Y) και P(Y). Μια καλή εκτίμηση της P(Y) μπορεί να επιτευχθεί με σχετικά λίγα παραδείγματα εκπαίδευσης. Ωστόσο, μια ακριβής εκτίμηση της P(X|Y) απαιτεί πολύ περισσότερα παραδείγματα. Αυτό συμβαίνει επειδή πρέπει να υπολογιστούν όλες οι παράμετροι $\theta_{ij} \equiv P\left(X = x_i \middle| Y = y_j\right)$, όπου, για παράδειγμα αν το Y είναι boolean μεταβλητή και το X είναι διάνυσμα boolean τιμών, το i παίρνει 2n πιθανές τιμές, μια για κάθε πιθανό διάνυσμα τιμών X και

το j παίρνει 2 τιμές, 0 ή 1. Επομένως πρέπει να υπολογιστούν περίπου 2n+1 παράμετροι. Αυτό καθιστά υπολογιστικά απαγορευτική την ακριβή εκτίμηση της P(X|Y). Ο ταξινομητής Naïve Bayes επιλύνει το πρόβλημα αυτό: μειώνει την πολυπλοκότητα της εκτίμησης του P(X|Y) θεωρώντας ότι ισχύει μια υπόθεση.

Υπό συνθήκη ανεξαρτησία (conditional independence): Έστω τυχαίες μεταβλητές X, Y και Z. Η X καλείται υπό συνθήκη ανεξάρτητη από τη Y, δοθέντος της Z, εάν και μόνο εάν η κατανομή της πιθανότητας που διέπει τη X είναι ανεξάρτητη από την τιμή της Y, δοθέντος του Z. Δηλαδή:

$$(\forall i, j, k) P\left(X = x_i \middle| Y = y_j, Z = z_k\right) = P(X = x_i | Z = z_k)$$

Ο ταξινομητής Naïve Bayes βασίζεται στο θεώρημα Bayes: υπολογίζει τη δεσμευμένη πιθανότητα ως προς την κλάση (class-conditional probability) θεωρώντας ότι όλα τα χαρακτηριστικά X1, X2, ..., Xn είναι υπό συνθήκη ανεξάρτητα μεταξύ τους, δεδομένου του Y [187]. Ουσιαστικά, η υπόθεση είναι ότι δοθέντος της κλάσης στην οποία ανήκει το παράδειγμα, η πιθανότητα να συμβούν τα X1, X2,..,Xn, ισούται με το γινόμενο των πιθανοτήτων για κάθε ξεχωριστό χαρακτηριστικό. Δηλαδή, για X που περιέχει n χαρακτηριστικά, τα οποία είναι υπό συνθήκη ανεξάρτητα μεταξύ τους, δοθέντος Y, ισχύει:

$$P(X_1 \dots X_n | Y) = \prod_{i=1}^{n} P(X_i | Y)$$

Με αυτή την υπόθεση, απλοποιείται δραστικά η αναπαράσταση της δεσμευμένης πιθανότητας P(X|Y), καθώς και η εκτίμησή της από τα δεδομένα εκπαίδευσης. Αντί να υπολογιστεί η δεσμευμένη πιθανότητα δοθέντος του Y για κάθε πιθανό συνδυασμό X, υπολογίζεται μόνο η δεσμευμένη πιθανότητα του κάθε Xi δοθέντος του Y. Έτσι ο αριθμός των παραμέτρων που πρέπει να υπολογιστούν μειώνεται: για μια καλή εκτίμηση της πιθανότητας $P\left(X = x_{ik} \middle| Y = y_j\right)$ αρκούν 2n παράμετροι. Συνεπώς, με την προσέγγιση αυτή μειώνεται κατά πολύ η πολυπλοκότητα και για να επιτευχθεί μια καλή εκτίμηση της δεσμευμένης πιθανότητας P(X|Y) δεν απαιτείται μεγάλο σε μέγεθος σύνολο εκπαίδευσης.

Ο ταξινομητής, αφού εκπαιδευτεί, για κάθε νέο παράδειγμα X που πρέπει να ταξινομηθεί επιστρέφει την κατανομή της πιθανότητας για όλες τις τιμές του Y. Η πιθανότητα ότι το Y θα λάβει την k πιθανή τιμή του, σύμφωνα με το θεώρημα Bayes, αναπαρίσταται ως εξής:

$$P(Y = y_k | X_1 \dots X_n) = \frac{P(Y = y_k) P(X_1 \dots X_n | Y = y_k)}{\sum_j P(Y = y_j)\, P(X_1 \dots X_n | Y = y_j)}$$

, όπου το άθροισμα υπολογίζεται για όλες τις πιθανές τιμές yj του Y.

Με την εισαγωγή της υπόθεσης ότι τα Xi είναι υπό συνθήκη ανεξάρτητα δοθέντος του Y, η παραπάνω εξίσωση μετατρέπεται σε:

$$P(Y = y_k | X_1 \dots X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

Αυτή είναι και η βασική εξίσωση για τον ταξινομητή Naïve Bayes. Δοθέντος ενός νέου παραδείγματος Χνέο = {X1...Xn}, αυτή η εξίσωση δείχνει πώς θα υπολογιστεί η πιθανότητα το Y να πάρει οποιαδήποτε τιμή (στο εύρος τιμών του: σε δυαδικό πρόβλημα 0 ή 1) , δοθέντος των τιμών των χαρακτηριστικών του νέου παραδείγματος και με τη χρήση των κατανομών P(Y) και P(X|Y) που έχουν εκτιμηθεί από τα δεδομένα εκπαίδευσης. Άρα, για τον υπολογισμό της πιο πιθανής τιμής του Y (της κλάσης στην οποία θα ταξινομηθεί το νέο παράδειγμα) από τον NB ταξινομητή, ακολουθείται ο παρακάτω κανόνας ταξινόμησης:

$$Y^{νέο} \leftarrow arg\max_{y_k} P(Y = y_k) \prod_i P(X_i^{νέο} | Y = y_k)$$

Επιγραμματικά, ο Naïve Bayes ταξινομητής χρησιμοποιεί τα δεδομένα εκπαίδευσης για να εκτιμήσει τις παραμέτρους μιας κατανομής πιθανότητας, υποθέτοντας ότι τα χαρακτηριστικά είναι υπό συνθήκη ανεξάρτητα δεδομένης της κλάσης. Ακολούθως ο NB ταξινομητής, κατατάσσει ένα νέο παράδειγμα, υπολογίζοντας, με τον κανόνα Bayes, τις μεταγενέστερες πιθανότητες να ανήκει το παράδειγμα αυτό σε κάθε κλάση και ταξινομεί το δείγμα αυτό στην κλάση που επιφέρει τη μέγιστη μεταγενέστερη πιθανότητα. Για να εκτιμήσει τη μεταγενέστερη πιθανότητα για κάθε κλάση, πρέπει να υπολογιστούν οι προγενέστερες πιθανότητες των κλάσεων και των κατανομών των χαρακτηριστικών. Η προγενέστερη πιθανότητα μιας κλάσης μπορεί να υπολογιστεί εκτιμώντας την πιθανότητα κλάσης από το σύνολο εκπαίδευσης. Όσον αφορά τις κατανομές των χαρακτηριστικών, γίνεται μια υπόθεση ως προς την κατανομή που ακολουθείται ή παράγονται μη-παραμετρικά μοντέλα των χαρακτηριστικών από το σύνολο εκπαίδευσης.

5.3.3.2 Πολυωνυμική πολυμεταβλητή Naïve Bayes ταξινόμηση

Όπως έχει αναφερθεί παραπάνω, η χρησιμοποίηση της Naive Bayes ταξινόμησης, απαιτεί τον καθορισμό των προγενέστερων κατανομών των χαρακτηριστικών. Ανάλογα με τα πρόβλημα επιλέγεται διαφορετική κατανομή. Για διακριτά χαρακτηριστικά οι πιο ευρέως χρησιμοποιούμενες κατανομές είναι η πολυωνυμική κατανομή και η κατανομή Bernoulli. Για κατηγορικά χαρακτηριστικά, όπως αυτά που υπάρχουν σε αυτό το πρόβλημα, η πιο κατάλληλη κατανομή είναι η πολυμεταβλητή πολυωνυμική κατανομή.

Με τη χρήση της πολυμεταβλητής πολυωνυμικής κατανομής (multivariate multinomial distribution), η μέθοδος αρχικά καταγράφει τα διακριτά κατηγορικά επίπεδα κάθε

χαρακτηριστικού. Κάθε συνδυασμός χαρακτηριστικού/κλάσης αντιστοιχεί σε μια ξεχωριστή, ανεξάρτητη πολυωνυμική τυχαία μεταβλητή. Για κάθε συνδυασμό χαρακτηριστικού/κλάσης ξεχωριστά, η μέθοδος μετρά τα παραδείγματα κάθε κατηγορικού επιπέδου και υπολογίζει για όλα τα επίπεδα του χαρακτηριστικού, την πιθανότητα το χαρακτηριστικό f στην κλάση C να έχει επίπεδο L. Με αυτό τον τρόπο, υπολογίζει ένα ξεχωριστό σύνολο πιθανοτήτων για το σύνολο των επιπέδων ενός χαρακτηριστικού για κάθε συνδυασμό χαρακτηριστικού/κλάσης. Οι εξαρτημένες από την κλάση, πολυωνυμικές τυχαίες μεταβλητές σχηματίζουν μια πολυμεταβλητή πολυωνυμική τυχαία μεταβλητή.

5.3.3.3 Πλεονεκτήματα Naïve Bayes ταξινομητή

Η Naïve Bayes ταξινόμηση παρουσιάζει πολλά πλεονεκτήματα. Καταρχάς, είναι απλή, εύκολη στη χρήση και αποτελεσματική. Ιδιαίτερα σε προβλήματα με απλές συσχετίσεις δεδομένων ή όταν παρέχεται ένα σύνολο υποθέσεων που απλοποιούν την κατασκευή του μοντέλου, δίνει καλά αποτελέσματα σε μικρό χρονικό διάστημα. Επιπλέον, παρουσιάζει ανθεκτικότητα απέναντι στο θόρυβο και μπορεί να χρησιμοποιηθεί και σε προβλήματα με ελλιπή δεδομένα απλά παραλείποντας τις αντίστοιχες πιθανότητες. Εκτός αυτών, πειραματικές μελέτες εισηγούνται ότι εκπαιδεύεται πιο γρήγορα από ότι οι πλείστοι αλγόριθμοι ταξινόμησης: για την εκπαίδευση του ταξινομητή είναι αρκετό ένα μόνο πέρασμα των δεδομένων εκπαίδευσης (γραμμική πολυπλοκότητα O(n), όπου n ο αριθμός των παραδειγμάτων εκπαίδευσης). Η υπόθεση της υπό συνθήκη ανεξαρτησίας ως προς την κλάση απλοποιεί την εκπαίδευση αφού μπορεί να εκτιμηθεί η μονοδιάστατη υπό συνθήκη πυκνότητα για κάθε χαρακτηριστικό ξεχωριστά. Αν και η υπό συνθήκη ανεξαρτησία ως προς την κλάση μεταξύ των χαρακτηριστικών δεν ισχύει γενικώς, οι διάφορες μελέτες δείχνουν ότι πρακτικά αυτή η υπόθεση φέρει καλά αποτελέσματα [187]. Επίσης, η υπόθεση αυτή επιτρέπει στον NB ταξινομητή να εκτιμήσει την κατανομή των παραμέτρων που απαιτούνται για την ταξινόμηση χρησιμοποιώντας λιγότερα δεδομένα εκπαίδευσης από ότι άλλοι ταξινομητές. Αυτό καθιστά την NB ταξινόμηση ιδιαίτερα αποτελεσματική για δεδομένα εκπαίδευσης που περιέχουν πολλά χαρακτηριστικά. Επιπρόσθετα, ο NB είναι λιγότερο ευαίσθητος σε αποκλίνουσες τιμές επειδή ακολουθεί πιθανοτική προσέγγιση.

Οι πιο σημαντικοί λόγοι επιλογής του NB ως ταξινομητή στο συγκεκριμένο πρόβλημα είναι οι ακόλουθοι:

- Ο NB είναι ένας μη παραμετρικός ταξινομητής και έτσι δεν απαιτείται να πραγματοποιηθεί επιλογή παραμέτρων εντός του ΓΑ. Αυτό είναι υψίστης σημασίας αφού η διαδικασία επιλογής παραμέτρων του ταξινομητή θα αύξανε κατά πολύ την πολυπλοκότητα του προβλήματος αναζήτησης.

- Ο NB θεωρείται ως ένας από τους πιο γρήγορους ταξινομητές, τόσο για εκπαίδευση όσο και για ταξινόμηση νέων δεδομένων, κάτι που είναι εξαιρετικά σημαντικό στη χρήση του σε έναν αλγόριθμο ευριστικής αναζήτησης.
- Ο NB υπολογίζει τις μεταγενέστερες πιθανότητες κάθε κλάσης, γεγονός πολύ χρήσιμο στο συγκεκριμένο πρόβλημα, αφού αυτές οι πιθανότητες μπορούν να χρησιμοποιηθούν για την εκτίμηση του κινδύνου μιας γυναίκας (risk assessment odd) να έχει CIN2+.

5.3.4 Υλοποίηση

Η ανάπτυξη του ΓΑ έχει γίνει χρησιμοποιώντας την πλατφόρμα MATLAB®. Η λεπτομερής υλοποίηση περιγράφεται ακολούθως:

Αρχικοποίηση (Initialization):

Δημιουργείται ένας αρχικός πληθυσμός από τυχαία παραχθέντα χρωμοσώματα. Ο πληθυσμός αυτός περιέχει υποψήφιες λύσεις του προβλήματος, δηλαδή πιθανά υποσύνολα χαρακτηριστικών. Μεγάλοι σε μέγεθος πληθυσμοί μπορεί να βελτιώνουν την ικανότητα αναζήτησης του ΓΑ, αλλά παράλληλα αυξάνουν τον υπολογιστικό χρόνο που απαιτείται για κάθε γενεά. Λαμβάνοντας αυτό υπόψη, ο αριθμός των ατόμων που αποτελούν τον αρχικό πληθυσμό, καθώς και τον πληθυσμό κάθε γενεάς, καθορίστηκε ίσος με 1200.

Αναπαράσταση Χρωμοσώματος

Για την αναπαράσταση των πιθανών υποσυνόλων χαρακτηριστικών χρησιμοποιείται ακέραια κωδικοποίηση (integer encoding): κάθε χρωμόσωμα ορίζεται ως ένα διάνυσμα που αποτελείται από ένα σύνολο ακεραίων, οι οποίοι αντιπροσωπεύουν δείκτες στα επιλεγμένα χαρακτηριστικά. Στην ακέραια κωδικοποίηση, το μήκος του χρωμοσώματος είναι ίσο με n, όπου n είναι ο αριθμός των χαρακτηριστικών που πρόκειται να επιλεχθούν για να σχηματίσουν το υποψήφιο υποσύνολο χαρακτηριστικών.

Στην προκειμένη περίπτωση, κάθε χρωμόσωμα είναι κωδικοποιημένο ως ένα διάνυσμα n ακέραιων αριθμών που αναπαριστά ένα συνδυασμό n χαρακτηριστικών, τα οποία έχουν επιλεχθεί από τα 40 διαθέσιμα χαρακτηριστικά. Επομένως, κάθε γονίδιο λαμβάνει μια ακέραια τιμή από το διάστημα [1,..., 40] και κάθε ακέραιος αριθμός αποτελεί δείκτη σε μία από τις 40 μεταβλητές. Το μήκος του χρωμοσώματος (n) μπορεί να καθοριστεί πριν από την έναρξη του ΓΑ (προφανώς δεν μπορεί να ξεπερνά τον αριθμό των διαθέσιμων χαρακτηριστικών, n<40). Τα 40 διαθέσιμα χαρακτηριστικά με τους δείκτες τους παρουσιάζονται στον *πίνακα 26.*

Πίνακας 26: Διαθέσιμα χαρακτηριστικά (Feaure pool)

Δείκτης (Index)	**Χαρακτηριστικό (Feature)**
1	PAP test
2	HPV 6
3	HPV 11
4	HPV 16
5	HPV 18
6	HPV 26
7	HPV 31
8	HPV 33
9	HPV 35
10	HPV 39
11	HPV 40
12	HPV 42
13	HPV 43
14	HPV 44
15	HPV 45
16	HPV 51
17	HPV 52
18	HPV 53
19	HPV 54
20	HPV 56
21	HPV 58
22	HPV 59
23	HPV 61
24	HPV 62
25	HPV 66
26	HPV 68
27	HPV 70
28	HPV 71
29	HPV 72
30	HPV 73
31	HPV 81
32	HPV 82
33	HPV 83
34	HPV 84

35	HPV 85
36	HPV 89
37	VHR HPV
38	HR HPV
39	LR HPV
40	HPV DNA test (θετικό/αρνητικό)

Για παράδειγμα, το χρωμόσωμα μήκους 8 που παρουσιάζεται στην *εικόνα 103*, αναπαριστά το υποσύνολο που αποτελείται από τα χαρακτηριστικά με δείκτες τους 1, 35, 17, 5, 8, 40 , 11 και 22, Βάσει του πίνακα 26, το υποσύνολο αυτό περιλαμβάνει: ΠΑΠ τεστ, HPV-85, HPV-52, HPV-18, HPV-33, HPV DNA test, HPV-40 και HPV-59.

Εικόνα 103: *Κωδικοποίηση χρωμοσωμάτων – Παραδείγματα*

Συνάρτηση Καταλληλότητας (Fitness function)

Μια συνάρτηση καταλληλότητας χρησιμοποιείται για να αξιολογηθεί εάν ένα άτομο είναι «κατάλληλο» για να επιβιώσει. Η συνάρτηση καταλληλότητας επιστρέφει μια τιμή καταλληλότητας για κάθε υπό αξιολόγηση χρωμόσωμα, η οποία μετρά την απόδοση του υποψήφιου υποσυνόλου χαρακτηριστικών.

Στο πρόβλημα αυτό, όπως έχει ήδη παρουσιαστεί, η ακρίβεια δεν είναι τόσο σημαντική όσο η ειδικότητα και η ευαισθησία όσον αφορά τη διαγνωστική αξία. Για το λόγο αυτό, ως συνάρτηση καταλληλότητας επιλέχθηκε ο δείκτης Youden's J statistic (ή αλλιώς Youden's index). Ο Youden's index [188] είναι ένας στατιστικός δείκτης, ο οποίος συλλαμβάνει την απόδοση μιας διαγνωστικής εξέτασης συνδυάζοντας την ευαισθησία και την ειδικότητα. Θεωρείται ως ένα καθολικό μέτρο της απόδοσης μιας διαγνωστικής εξέτασης και χρησιμοποιείται για τη σύγκριση μεταξύ διαφορετικών εξετάσεων. Ο ΓΑ στην προσπάθειά του να μεγιστοποιήσει τον Youden's index, στην πραγματικότητα επιχειρεί να εντοπίσει τον συνδυασμό χαρακτηριστικών που οδηγεί

στην πιο ισορροπημένη ευαισθησία και ειδικότητα. Ο Youden's index παρουσιάζεται στην εξίσωση (1).

$$J = Sensitivity + Specificity - 1 \qquad (1)$$

Η τιμή καταλληλότητας κάθε υποσυνόλου χαρακτηριστικών υπολογίζεται με τη χρησιμοποίηση ενός ταξινομητή Naïve-Bayes και την εφαρμογή της τεχνικής διασταυρωμένη επικύρωση 5 τμημάτων (5-fold cross validation). Εφόσον ο αριθμός των τμημάτων είναι 5, προκύπτουν 5 διαφορετικές τιμές για κάθε διαγνωστικό μέτρο. Επομένως, για κάθε διαγνωστικό μέτρο (Ευαισθησία, Ειδικότητα, PPV, NPV) εκτιμάται ο μέσος όρος των 5 αυτών τιμών.

Επιλογή (Selection)

Αφού έχει ανατεθεί μια τιμή καταλληλότητας σε κάθε χρωμόσωμα του τρέχοντος πληθυσμού, τα πιο κατάλληλα χρωμοσώματα επιλέγονται για αναπαραγωγή. Τα επιλεγμένα άτομα καλούνται «γονείς» (parents) και θα χρησιμοποιηθούν στο επόμενο στάδιο (αναπαραγωγή) για να σχηματίσουν απογόνους (offspring). Ως συνάρτηση για τον τελεστή της επιλογής χρησιμοποιείται η tournament selection μεταξύ 4 ατόμων.

Δημιουργία πληθυσμού της επόμενης γενεάς

Για τη δημιουργία του πληθυσμού της επόμενης γενεάς, θα χρησιμοποιηθούν τρεις βασικοί γενετικοί τελεστές: ελιτισμού, διασταύρωσης και μετάλλαξης. Όλοι οι γενετικοί τελεστές έχουν τροποποιηθεί έτσι ώστε να μπορούν να χρησιμοποιηθούν με τα ακέραια κωδικοποιημένα χρωμοσώματα.

Ελιτισμός (Elitism)

Το καλύτερο έως τώρα άτομο του πληθυσμού μπορεί να μην επιλεχθεί να επιβιώσει στην επόμενη γενεά. Ωστόσο, αυτό το άτομο μπορεί να αποτελεί μιας υψηλής ποιότητας λύση, ή ακόμα και το ολικό βέλτιστο και ως επακόλουθο να μην περιλαμβάνεται στον τελικό πληθυσμό. Για να αποφευχθεί αυτό, εφαρμόζεται ο τελεστής του ελιτισμού. Σε κάθε γενεά, τα άτομα με τις πιο υψηλές τιμές καταλληλότητας, αντιγράφονται όπως έχουν, χωρίς τροποποιήσεις, κατευθείαν στην επόμενη γενεά. Με αυτό τον τρόπο εξασφαλίζεται ότι η τελική λύση θα είναι η ολικά βέλτιστη λύση. Ο τελεστής του ελιτισμού εφαρμόζεται σε 120 άτομα του πληθυσμού (elite count = 120), δηλαδή στο 10% του πληθυσμού: σε κάθε γενεά, 120 χρωμοσώματα του τρέχοντος πληθυσμού με τις πιο υψηλές τιμές καταλληλότητας, αντιγράφονται όπως έχουν, χωρίς τροποποιήσεις, κατευθείαν στον πληθυσμό της επόμενης γενεάς.

Αναπαραγωγή

Σε αυτό το στάδιο, οι υπόλοιπες λύσεις του πληθυσμού της επόμενης γενεάς, παράγονται μέσω της διασταύρωσης και της μετάλλαξης. Οι τελεστές αυτοί είναι σχεδιασμένοι έτσι ώστε ιδιότητες των γονέων να αναπαράγονται στους απογόνους.

Διασταύρωση (Crossover)

Η συνάρτηση που χρησιμοποιείται για τον τελεστή της διασταύρωσης είναι η ομοιόμορφη διασταύρωση (uniform crossover), με ποσοστό μίξης 0.5 (exchange probability). Η πιθανότητα διασταύρωσης τίθεται ίση με 0.7, τιμή η οποία καθορίζει το ποσοστό κάθε πληθυσμού, εκτός των παιδιών που προκύπτουν από τον ελιτισμό, που δημιουργούνται από τον τελεστή της διασταύρωσης. Επομένως, τα παιδιά που δημιουργούνται από διασταύρωση θα είναι ίσα με: 0.7 * (population size - elite count) = 0.7 * (1200 - 120) = 756. Τα υπόλοιπα παιδιά του πληθυσμού δημιουργούνται με μετάλλαξη. Συνοψίζοντας, με αυτή την τιμή της πιθανότητας διασταύρωσης, σε κάθε γενεά θα υπάρχουν 120 παιδιά από ελιτισμό (10% ολόκληρου του πληθυσμού), 756 παιδιά από διασταύρωση (63% ολόκληρου του πληθυσμού) και 324 παιδιά από μετάλλαξη (27% ολόκληρου του πληθυσμού).

Η συνάρτηση διασταύρωσης είναι τροποποιημένη με τρόπο που να εξασφαλίζει ότι κάθε γονίδιο ενός χρωμοσώματος έχει μια ξεχωριστή και διακριτή τιμή, έτσι ώστε κανένα χαρακτηριστικό να μην εμφανίζεται δύο φορές σε ένα χρωμόσωμα.

Μετάλλαξη (Mutation)

Ακολούθως λαμβάνει χώρα η μετάλλαξη, η οποία διενεργείται με τη συνάρτηση της ομοιόμορφης μετάλλαξης. Η ομοιόμορφη μετάλλαξη εφαρμόζεται στα γονίδια των χρωμοσωμάτων που έχουν επιλεχθεί για μετάλλαξη με πιθανότητα μετάλλαξης ίση με 0.2.

Η ομοιόμορφη μετάλλαξη είναι η πιο απλή και ευρέως χρησιμοποιούμενη συνάρτηση μετάλλαξης στους ΓΑ. Αντικαθιστά ένα γονίδιο με βάση μια χαμηλή πιθανότητα που καλείται πιθανότητα μετάλλαξης. Με πιθανότητα μετάλλαξης 0.2, κάθε γονίδιο έχει 20% πιθανότητα να μεταλλαχθεί. Ο τελεστής της μετάλλαξης αλλάζει την τιμή του τυχαία επιλεγμένου γονιδίου, με μια τυχαία ακέραια τιμή, διαφορετική από αυτή που έχει, που ανήκει στο διάστημα [1,40]. Ωστόσο, είναι τροποποιημένη με τέτοιο τρόπο, έτσι ώστε να εξασφαλίζεται ότι κάθε γονίδιο σε ένα χρωμόσωμα να έχει μοναδική τιμή (κανένα γονίδιο στο ίδιο χρωμόσωμα να μην έχει την ίδια τιμή με ένα άλλο).

Η μετάλλαξη είναι μια στοχαστική διαδικασία αλλαγής μιας ή περισσότερων τιμών μιας υποψήφιας λύσης, π.χ. τυχαία αντικατάσταση ενός ή περισσότερων γονιδίων (χαρακτηριστικών) ενός συγκεκριμένου χρωμοσώματος με άλλα. Οι ΓΑ, λόγω της στοχαστικής φύσης της

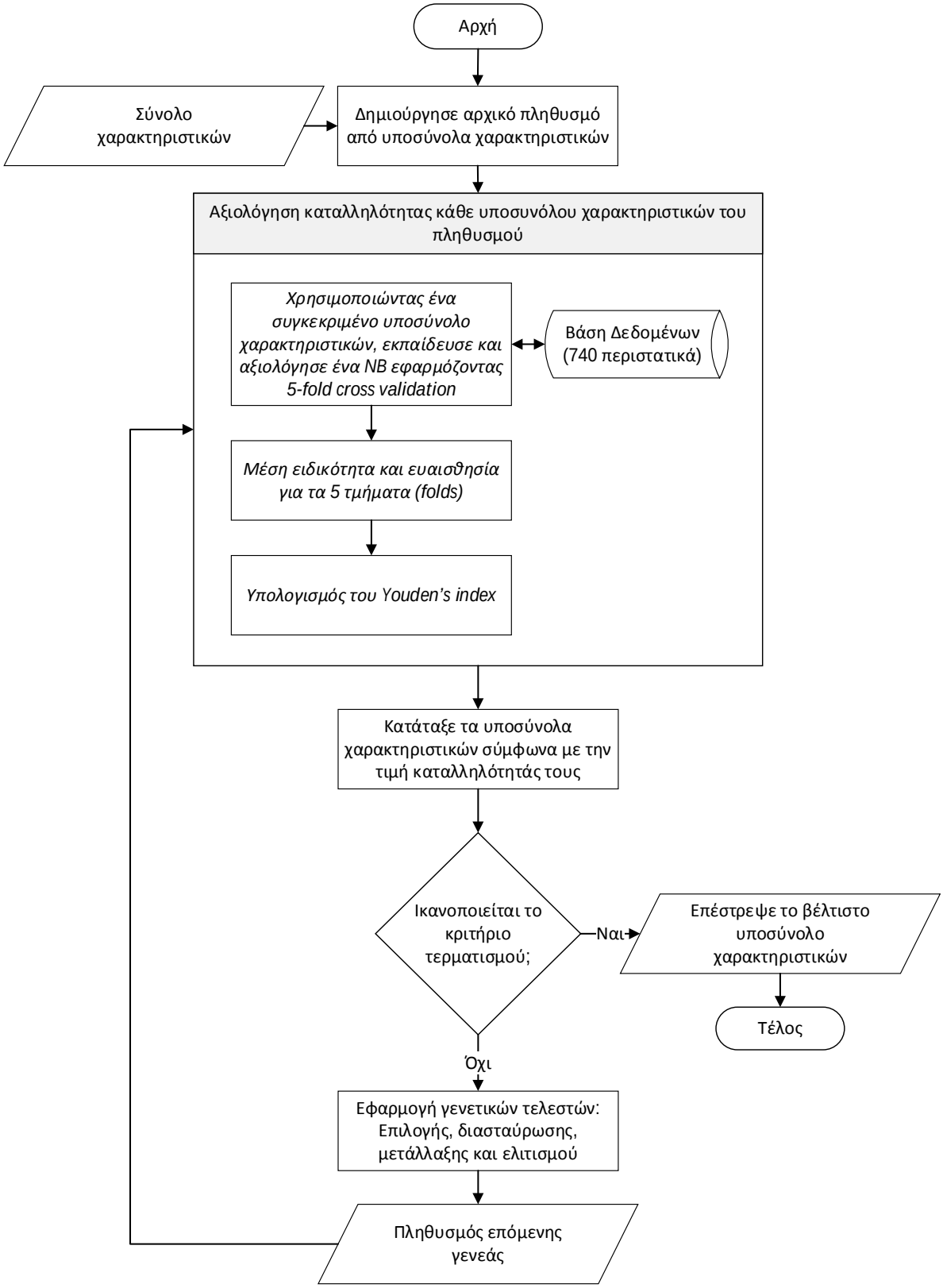
μεταλλαγής, μπορούν να δραπετεύσουν από τοπικά ελάχιστα στο χώρο αναζήτησης, εκεί που άλλοι αλγόριθμοι μπορεί να εγκλωβιστούν. Ο στόχος της μετάλλαξης είναι η εξερεύνηση του προηγουμένως απρόσιτου χώρου αναζήτησης. Δρα ως ένας τελεστής διατάραξης του πληθυσμού και εισάγοντας νέα πληροφορία στον πληθυσμό, επιτρέπει την ποικιλομορφία στις επόμενες γενεές.

Κριτήριο Τερματισμού (Termination criterion)

Τέλος, το κριτήριο τερματισμού που χρησιμοποιείται, είναι ένας συνδυασμός δύο συμπληρωματικών κριτηρίων. Ο ΓΑ τερματίζεται:

- εάν έχει συμπληρωθεί ο μέγιστος αριθμός γενεών που τίθεται ίσος με 50
- εάν η βέλτιστη τιμή καταλληλότητας παραμένει σταθερή για ένα συγκεκριμένο αριθμό γενεών (stall generations) που τίθεται ίσος με 12

Όπως έχει αναφερθεί, στην ακέραια κωδικοποίηση, το μήκος ενός χρωμοσώματος είναι ίσο με τον αριθμό των χαρακτηριστικών που επιλέγονται για να σχηματίσουν τον υποψήφιο υποσύνολο χαρακτηριστικών. Επομένως, με στόχο την εύρεση του πιο κατάλληλου υποσυνόλου χαρακτηριστικών, ο ΓΑ εκτελείται ξεχωριστά για κάθε διαφορετικό μήκος χρωμοσώματος (2...39), δηλαδή, για υποσύνολα 2 έως 39 χαρακτηριστικών. Τελικά, ο ΓΑ επιστρέφει το καλύτερο υποσύνολο χαρακτηριστικών για κάθε διαφορετικό μήκος χρωμοσώματος. Έτσι, μπορεί να καθοριστεί το βέλτιστο μήκος χρωμοσώματος, το οποίο παράγει το υποσύνολο χαρακτηριστικών με την πιο υψηλή τιμή καταλληλότητας για το συγκεκριμένο πρόβλημα. Ακολούθως, παρουσιάζεται το διάγραμμα ροής του ΓΑ.

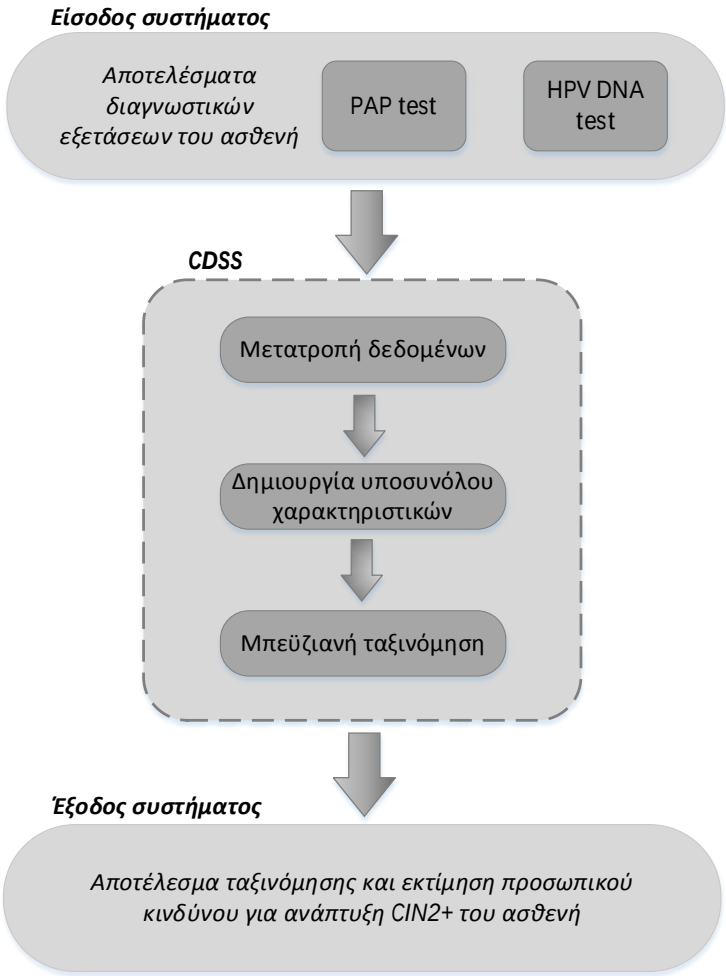


Εικόνα 104: *Διάγραμμα ροής του ΓΑ*

5.3.5 Αρχιτεκτονική του συστήματος υποστήριξης κλινικών αποφάσεων

Ο ταξινομητής Naïve Bayes σε συνδυασμό με το υποσύνολο χαρακτηριστικών που απέδωσε καλύτερα, χρησιμοποιούνται μαζί για τη δημιουργία ενός συστήματος υποστήριξης λήψης κλινικών αποφάσεων (Clinical Decision Support System- CDSS). Το σχηματικό διάγραμμα του προτεινόμενου συστήματος υποστήριξης λήψης κλινικών αποφάσεων παρουσιάζεται στην *εικόνα 105*. Όπως απεικονίζεται στην *εικόνα 105*, τα αποτελέσματα των διαγνωστικών εξετάσεων κάθε γυναίκας χρησιμοποιούνται ως είσοδος στο CDSS. Αρχικά, οι ιατρικές πληροφορίες μετατρέπονται σε δεδομένα κατάλληλα για επεξεργασία από τον NB ταξινομητή, δηλαδή χρησιμοποιούνται για τη δημιουργία του πλήρους συνόλου δεδομένων (40 χαρακτηριστικά). Από αυτό το σύνολο χαρακτηριστικών, επιλέγουμε τα χαρακτηριστικά που αντιστοιχούν στο καλύτερο υποσύνολο χαρακτηριστικών, το οποίο προέκυψε από την εφαρμογή ΓΑ-NB, έτσι ώστε να δημιουργηθεί το διάνυσμα εισόδου του NB ταξινομητή. Ακολούθως, το διάνυσμα εισόδου προωθείται στον NB ταξινομητή, ο οποίος επιστρέφει το αποτέλεσμα της ταξινόμησης, καθώς και τις μεταγενέστερες πιθανότητες κάθε κλάσης, δηλαδή την πιθανότητα για CIN1- και την πιθανότητα για CIN2+. Με αυτό τον τρόπο το CDSS παρέχει εκτίμηση του κινδύνου ανάπτυξης CIN2+ εξατομικευμένα σε κάθε γυναίκα.

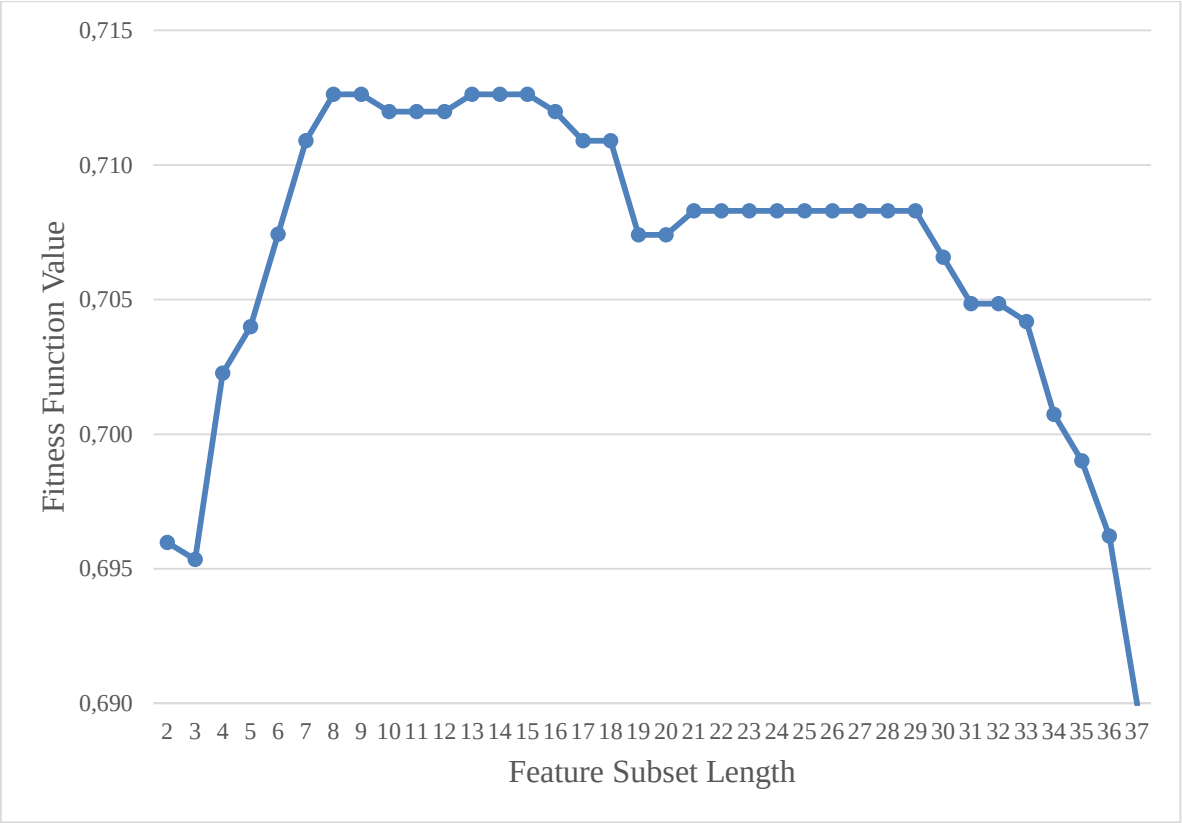
Το CDSS έχει αναπτυχθεί με τη χρήση της πλατφόρμας MATLAB® και έχει ενταχθεί στο προηγουμένως κατασκευασμένο web-based πληροφοριακό σύστημα CxCaDSS που έχει παρουσιαστεί στο [189].



Εικόνα 105: *Αρχιτεκτονική συστήματος υποστήριξης κλινικών αποφάσεων*

5.3.6 Αποτελέσματα

Η *εικόνα 106* απεικονίζει τις τιμές καταλληλότητας των καλύτερων υποσυνόλων χαρακτηριστικών για κάθε διαφορετικό μήκος χρωμοσώματος.



Εικόνα 106: *Οι τιμές καταλληλότητας των καλύτερων υποσυνόλων χαρακτηριστικών για κάθε διαφορετικό μήκος υποσυνόλου χαρακτηριστικών*

Όπως φαίνεται στην εικόνα 106, αρχικά, η τιμή της συνάρτησης καταλληλότητας αυξάνεται με την αύξηση του αριθμού των χαρακτηριστικών που χρησιμοποιούνται για το υποσύνολο. Αυτό παρατηρείται να συμβαίνει έως ότου η συνάρτηση καταλληλότητας πάρει τη μέγιστη τιμή της για υποσύνολο που αποτελείται από 8 χαρακτηριστικά. Για υποσύνολα χαρακτηριστικών με περισσότερα από 8 χαρακτηριστικά (εκτός των υποσυνόλων, 9, 13, 14 και 15 χαρακτηριστικών που έχουν την ίδια τιμή καταλληλότητας με το υποσύνολο 8 χαρακτηριστικών), η απόδοση πέφτει όλο και περισσότερο, με την τιμή της συνάρτησης καταλληλότητας να μειώνεται με την αύξηση του αριθμού χαρακτηριστικών που χρησιμοποιούνται για το σχηματισμό του υποσυνόλου. Συνεπώς, συμπεραίνεται ότι η προσθήκη περισσότερων χαρακτηριστικών (> 9) δεν παρέχει χρήσιμη πληροφορία στον ταξινομητή: μόνο 8 από τα 40 χαρακτηριστικά φαίνεται να περιέχουν σημαντική πληροφορία, η οποία να οδηγεί στην ταξινόμηση με τα πιο ισορροπημένα μέτρα ευαισθησίας και ειδικότητας. Τα υπόλοιπα χαρακτηριστικά μπορούν να θεωρηθούν άσχετα ή περιττά όσον αφορά το συγκεκριμένο πρόβλημα. Στον *πίνακα 27*, παρουσιάζεται το υποσύνολο χαρακτηριστικών, που αποτελείται από 8 χαρακτηριστικά, το οποίο παρουσίασε την πιο υψηλή τιμή καταλληλότητας ολικά.

Η εφαρμογή αυτή, επιβεβαιώνει την «κατάρα της διαστασιμότητας» και αποτελεί ένα κλασικό παράδειγμα των συνεπειών της. Η χρήση άσχετων ή περιττών χαρακτηριστικών αυξάνει τη διαστασιμότητα του χώρου χαρακτηριστικών και ως αποτέλεσμα και την πολυπλοκότητα του προβλήματος ταξινόμησης χωρίς οποιοδήποτε κέρδος στην απόδοση. Επιπλέον, όπως αναφέρεται στην «κατάρα της διαστασιμότητας», για σταθερό αριθμό δειγμάτων εκπαίδευσης, καθώς η διαστασιμότητα αυξάνεται, η ικανότητα προβλεψιμότητας του ταξινομητή μειώνεται από ένα μέγιστο σημείο και μετά, εξαιτίας του φαινομένου της υπερπροσαρμογής (overfitting). Παρατηρώντας την *εικόνα 106*, μπορεί να εξαχθεί το συμπέρασμα ότι ο ταξινομητής αντιμετωπίζει πρόβλημα υπερπροσαρμογής για υποσύνολα που αποτελούνται από περισσότερα των 16 χαρακτηριστικών. Επιπρόσθετα, αξίζει να σημειωθεί ότι τιμή καταλληλότητας που επιτυγχάνεται όταν χρησιμοποιούνται μόνο 2 χαρακτηριστικά για το σχηματισμό του υποσυνόλου, είναι πολύ καλύτερη σε σύγκριση με τις τιμές που επιτυγχάνονται όταν χρησιμοποιούνται άνω των 36 χαρακτηριστικών. Τα εν λόγω ευρήματα, αναδεικνύουν τη σημαντικότητα της διεργασίας της επιλογής χαρακτηριστικών για την κατασκευή ενός συστήματος ταξινόμησης.

Πίνακας 27: Το υποσύνολο χαρακτηριστικών με την πιο υψηλή τιμή καταλληλότητας

Χαρακτηριστικά
Pap test
HPV DNA (θετικό ή αρνητικό)
VHR HPV
HPV 31
HPV 58
HPV 62
HPV 66
HPV 6

Ο *πίνακας 28* παρουσιάζει την απόδοση του προτεινόμενου συστήματος συγκριτικά με τις διαγνωστικές εξετάσεις που εφαρμόστηκαν, ως προς την ευαισθησία (sensitivity-SN), την ειδικότητα (specificity-SP), τη θετική προγνωστική αξία (positive predictive value-PPV), την αρνητική προγνωστική αξία (negative predictive value-NPV) και τον δείκτη Youden (YI), όσον αφορά την ανίχνευση CIN2+. Η απόδοση του CDSS που παρουσιάζεται, είναι η απόδοση ενός NB ταξινομητή, ο οποίος χρησιμοποιεί το καλύτερο υποσύνολο χαρακτηριστικών (*πίνακας 27*) ως είσοδο και εφαρμόζει την τεχνική 5-fold cross validation (τα δοθέντα μέτρα απόδοσης ισούνται με το μέσο όρο των τιμών που προκύπτουν από όλα τα folds).

Συγκριτικά με το Pap test και το HPV DNA test (λαμβάνοντας υπόψη διαφορετικά κατώφλια θετικής διάγνωσης (positivity thresholds) για τις εξετάσεις αυτές), το προτεινόμενο σύστημα

παρήγαγε τα πιο ισορροπημένα αποτελέσματα ως προς την ευαισθησία και την ειδικότητα. Επιπλέον, στην κατάταξη των διαγνωστικών εξετάσεων ως προς τον Youden's index, ο οποίος δίνει ίσο βάρος στην ευαισθησία και την ειδικότητα, το CDSS κατετάγη υψηλότερα, ξεπερνώντας τις υπόλοιπες διαγνωστικές εξετάσεις.

Πίνακας 28: Απόδοση % του Pap test, HPV DNA test και του συστήματος GA-NB στην ανίχνευση του CIN2+

	SN	***SP***	***PPV***	***NPV***	***YI***
Pap test (cut-off ASCUS+)	98.1	45.3	33.3	98.9	43.4
Pap test(cut-off LSIL+)	89.4	67.0	43.0	96.0	56.4
Pap test(cut-off HSIL+)	71.4	95.3	81.0	92.3	66.7
HPV DNA test	91.9	61.5	39.9	96.5	53.4
HR HPV DNA	89.4	67.4	43.2	95.8	56.8
VHR HPV DNA	74.5	83.9	56.3	92.2	58.5
HPV 16/18 (16 ή 18)	56.5	88.6	58.0	88.0	45.1
Pap test (cut-off ASCUS+) or HPV 16/18	98.8	44.4	33.1	99.2	43.1
Pap test (cut-off ASCUS+) and HPV 16/18	55.9	89.5	59.6	87.9	45.4
GA-NB CDSS	85.1	86.2	63.1	95.4	71.3

Κεφάλαιο 6 - Συμπεράσματα & Προοπτικές για Μελλοντικές Επεκτάσεις

6.1 Συμπεράσματα

Όπως έχει ήδη επισημανθεί προηγουμένως, κίνητρο της παρούσας διπλωματικής εργασίας αποτέλεσε η ανάγκη για βελτίωση της απόδοσης των διαγνωστικών μεθόδων του καρκίνου του τραχήλου της μήτρας που πλήττει μεγάλο αριθμό γυναικών. Το παραγόμενο πλήθος των αποτελεσμάτων των διαγνωστικών εξετάσεων αυξάνεται σε επίπεδα που καθίσταται εξαιρετικά δύσκολη η συσχέτιση των δεδομένων και η εξαγωγή ορθών και χρήσιμων συμπερασμάτων από αυτά. Η αξιολόγηση όλο και περισσότερων διαγνωστικών εξετάσεων από τους ιατρούς, απαιτεί τη σύνθετη επεξεργασία πολύ μεγάλου πλήθους δεδομένων, εργασία η οποία είναι χρονοβόρα και δύσκολα πραγματοποιείται από τον άνθρωπο. Οι διαγνωστικές εξετάσεις (PAP test και HPV DNA test) για τον καρκίνο του τραχήλου της μήτρας παρουσιάζουν διαφορετικά χαρακτηριστικά ευαισθησίας και ειδικότητας, γεγονός που αυξάνει σημαντικά το βαθμό δυσκολίας της αξιολόγησης των αποτελεσμάτων των εξετάσεων αυτών.

Η εξέταση Παπανικολάου, παρά το γεγονός ότι αποτελεί την πιο επιτυχή μέθοδο πρόληψης του καρκίνου, στην παρούσα φάση, συναγωνίζεται με την εξέταση HPV DNA. Προφανώς, το HPV DNA test δεν υποκαθιστά το Pap test, ωστόσο, ως συμπληρωματική εξέταση, μπορεί να βελτιώσει τη διαγνωστική απόδοση. Η ανάγκη χρήσης του HPV DNA test ως συμπληρωματικό του Pap test πηγάζει από πληθώρα λόγων. Εν πρώτοις, το Pap test απαιτεί έμπειρους επαγγελματίες υγείας (κυτταροπαθολόγους ή κυτταροτεχνολόγους) για την ανάλυση των γυάλινων αντικειμενοφόρων πλακών, μέσω του μικροσκοπίου, έτσι ώστε να μη γίνει παρερμηνεία του δείγματος. Η διαδικασία αυτή απαιτεί χρόνο και είναι πολύ ευαίσθητη σε ανθρώπινα λάθη. Επιπροσθέτως, σε ορισμένες περιπτώσεις, το τεστ Παπανικολάου, μπορεί να θεωρηθεί ως «μη ικανοποιητικό» για αξιολόγηση λόγω ακατάλληλης δειγματοληψίας του κολπικού επιχρίσματος, δηλαδή, λόγω ανεπαρκούς συλλογής κυττάρων (άλλος λόγος μπορεί να είναι η παρείσφρυση διαφορετικής ουσίας π.χ. αίματος, που επικαλύπτει τα κύτταρα, με αποτέλεσμα τα ίδια τα κύτταρα να μην μπορούν να προσδιοριστούν με σαφήνεια). Εάν η δειγματοληψία κυττάρων μιας γυναίκας δεν πραγματοποιηθεί ορθά και προκύψουν ανεπαρκή δείγματα, τότε η γυναίκα αυτή θα πρέπει να επισκεφτεί ξανά τις ιατρικές εγκαταστάσεις για να επαναλάβει τη διαδικασία λήψης βιολογικού υλικού, κάτι που είναι κουραστικό για την ασθενή και συνάμα δαπανηρό. Από την άλλη πλευρά, το HPV DNA test μπορεί να γίνει σε batches, αποτελεί μια λιγότερο ευαίσθητη εξέταση σε

ανθρώπινα λάθη, απαιτεί λιγότερο έμπειρο προσωπικό και μπορεί τελικά να στοιχίζει λιγότερο από ότι το Pap test. Σήμερα, οι αναπτυσσόμενες χώρες επιλέγουν να εφαρμόζουν σε τακτά χρονικά διαστήματα μόνο μία από τις δύο εξετάσεις. Αντίθετα, στις προηγμένες οικονομικά χώρες, η πραγματοποίηση και των δύο εξετάσεων είναι συνηθισμένη και δικαιολογημένη ως προς τα οικονομικά της υγείας. Επιπλέον, η παράλληλη πραγματοποίηση και των δύο διαγνωστικών εξετάσεων, διευκολύνεται ακόμα περισσότερο αφού και οι δύο εξετάσεις μπορούν να γίνουν με χρήση βιολογικού υλικού το οποίο λαμβάνεται σε μία μόνο επίσκεψη.

Σήμερα, περιστατικά με κυτταρολογική διάγνωση HSIL, αποστέλλονται άμεσα για κολποσκόπηση, μια και η εξέταση Παπανικολάου παρουσιάζει πολύ υψηλή ειδικότητα όταν θεωρούμε ως κατώφλι θετικής διάγνωσης την κατηγορία HSIL+ (cut-off HSIL+). Είναι σαφώς ξεκάθαρη η διαδικασία που θα ακολουθηθεί για τη διαχείριση των ασθενών σε αυτές τις περιπτώσεις. Ωστόσο, το ουσιαστικό ερώτημα αφορά τον βέλτιστο τρόπο διαχείρισης των γυναικών με κυτταρολογική διάγνωση ASCUS ή LSIL, καθώς και των γυναικών που είναι θετικές σε στελέχη HPV υψηλού κινδύνου (HR HPV). Εστιάζοντας σε αυτό το πρόβλημα, μας ενδιαφέρουν περισσότερο τα μέτρα απόδοσης που χρησιμοποιούν ως κατώφλι θετικής διάγνωσης τις κατηγορίες ASCUS+ ή LSIL+ για την εξέταση Παπανικολάου και την ύπαρξη στελεχών HPV υψηλού ή πολύ υψηλού κινδύνου όσον αφορά την εξέταση HPV DNA.

Σύμφωνα με τα αποτελέσματα (πίνακας 28), η εξέταση Παπανικολάου έχει πολύ υψηλή ευαισθησία όταν ως κατώφλι χρησιμοποιείται η κατηγορία ASCUS+ (98.1%) ή η κατηγορία LSIL+ (89.4%), ενώ η εξέταση HPV DNA έχει υψηλότερη ειδικότητα όταν κατώφλι αποτελεί η ύπαρξη τύπων HPV υψηλού (67.4%) ή πολύ υψηλού (83.9%) κινδύνου (HR ή VHR HPV). Ωστόσο, η υψηλή ευαισθησία που παρουσιάζει η εξέταση Παπανικολάου με κατώφλι την κατηγορία ASCUS+ ή LSIL+ έρχεται με κόστος τη χαμηλή ειδικότητα (45.3% για ASCUS+ και 67.5% για LSIL+). Παρόμοια και η εξέταση HPV DNA, ενώ παρουσιάζει υψηλότερη ειδικότητα με κατώφλι την ύπαρξη HR ή VHR HPV τύπων, εντούτοις παρουσιάζει χαμηλή ευαισθησία (89.4% για HR HPV και 74.5% για VHR HPV).

Συνοψίζοντας, και οι δύο προσεγγίσεις εξασφαλίζουν υψηλή τιμή σε ένα από τα δύο διαγνωστικά μέτρα, με κόστος, όμως, το άλλο: η εξέταση Παπανικολάου (με cut-off ASCUS+ ή LSIL+) έχει πολύ υψηλή ευαισθησία και χαμηλή ειδικότητα και η εξέταση HPV DNA (με cut-off HR ή VHR HPV) έχει υψηλότερη ειδικότητα και πιο χαμηλή ευαισθησία. Αυτό αντανακλάται σε περιττές παραπομπές για κολποσκόπηση (εξαιτίας της υψηλής ευαισθησίας του PAP test) και σε θετικά περιστατικά που όμως δεν εντοπίζονται (εξαιτίας της υψηλής ειδικότητας του HPV DNA test), αντίστοιχα. Προφανώς, μια προσέγγιση που συνδυάζει τις δύο αυτές διαγνωστικές εξετάσεις έχει ενδιαφέρον εξαιτίας μιας ενδεχομένως πιο ισορροπημένης απόδοσης. Έχουν προταθεί απλοί αλγόριθμοι που συνδυάζουν την εξέταση Παπανικολάου και την εξέταση HPV DNA, με σκοπό να εφαρμοστούν ως μέθοδοι διαλογής (triage) για τη διάκριση των γυναικών που

θα ωφεληθούν περισσότερο από επιπρόσθετες διαγνωστικές εξετάσεις όπως την άμεση παραπομπή για κολποσκόπηση και διαγνωστικό επανέλεγχο από τις γυναίκες που μπορούν να παρακολουθηθούν με λιγότερο επεμβατικές μεθόδους. Παραδείγματα τέτοιων αλγορίθμων χρησιμοποιούν λογικούς τελεστές για το συνδυασμό των δύο εξετάσεων: με χρήση του λογικού AND, κριτήριο διαλογής αποτελεί η συνθήκη να έχουν θετικό αποτέλεσμα και οι δύο εξετάσεις, ενώ με χρήση του λογικού OR, αρκεί μία από τις δύο εξετάσεις να είναι θετική. Ωστόσο, οι αλγόριθμοι αυτοί δεν είναι αποτελεσματικοί επειδή η απόδοση αυτών των συνδυασμών εξακολουθεί να μην είναι βέλτιστα ισορροπημένη. Όπως παρουσιάζεται στον πίνακα 4, ο συνδυασμός της προσέγγισης που παρουσιάζει την καλύτερη ευαισθησία - 98.1% (Pap test με κατώφλι ASCUS+) και της προσέγγισης με την καλύτερη ειδικότητα - 88.6% (HPV 16/18), με χρήση του λογικού τελεστή OR, οδηγεί σε αύξηση της ευαισθησίας - 98.8%, ενώ με χρήση του λογικού τελεστή AND, οδηγεί σε αύξηση στην ειδικότητα – 89.5%. Ωστόσο, το κέρδος που επιτυγχάνεται στην ευαισθησία γίνεται σε βάρος της ειδικότητας και αντιστρόφως. Παρόμοια αποτελέσματα παρουσιάζονται και στο [190], όπου έχουν εξερευνηθεί περισσότεροι συνδυασμοί.

Σε σύγκριση με τις διαγνωστικές εξετάσεις και τους συνδυασμούς τους, η προτεινόμενη αρχιτεκτονική, η οποία χρησιμοποιεί το Youden's index για αξιολόγηση, παράγει τα πιο ισορροπημένα αποτελέσματα όσον αφορά την ειδικότητα και την ευαισθησία.

Τα ισορροπημένα, αυτά, αποτελέσματα καθιστούν το προτεινόμενο σύστημα ως μια πιθανώς χρήσιμη μέθοδο για ένα εργαστήριο κυτταροπαθολογίας ή μια γυναικολογική κλινική, και θα μπορούσε να εξυπηρετήσει αποτελεσματικά στη μείωση του φόρτου των κολποσκοπικών κλινικών και στην ορθολογική διαχείριση γυναικών που βρίσκονται σε πραγματικό κίνδυνο για ανάπτυξη καρκίνου του τραχήλου της μήτρας.. Το σύστημα αυτό, παρέχοντας εκτιμήσεις κινδύνου για την ανάπτυξη CIN2+, μπορεί να χρησιμοποιηθεί ως μια «τρίτη» γνωμάτευση, για γυναίκες που πρόκειται να παραπεμφθούν για κολποσκόπηση εξαιτίας κυτταρολογικής διάγνωσης ASCUS ή LSIL, ή εξαιτίας θετικού αποτελέσματος σχετικά με την ύπαρξη υψηλού κινδύνου τύπων HPV. Με αυτό τον τρόπο, αυτό το CDSS μπορεί να καθοδηγήσει την προσωπική διαχείριση της ασθενούς και τις αποφάσεις σχετικά με την πορεία της θεραπείας της και να μειώσει τις μη αναγκαίες κολποσκοπήσεις και θεραπείες. Ενσωματώνοντας το CDSS στο διαδικτυακό πληροφοριακό σύστημα CxCaDSS [189], αυτό μπορεί να χρησιμεύσει ως ένα διαδικτυακό σύστημα υποστήριξης κλινικής απόφασης για ιατρούς και ερευνητές για την εξατομικευμένη διαχείριση των γυναικών με μη φυσιολογικά αποτελέσματα διαγνωστικών εξετάσεων.

6.2 Προοπτικές για Μελλοντικές Επεκτάσεις

Τα τελευταία έτη έχει σημειωθεί σημαντική πρόοδος στον τομέα της πληροφορικής τεχνολογίας η οποία επιτρέπει την ανάπτυξη συστημάτων Βιοπληροφορικής για υποστήριξη της κλινικής διάγνωσης και λήψη αποφάσεων θεραπευτικής αγωγής που βασίζονται σε εξατομικευμένα στοιχεία των ασθενών. Τα κλινικά συστήματα υποστήριξης αποφάσεων, στηριζόμενα σε μαθηματικά εργαλεία μοντελοποίησης και επεξεργασίας δεδομένων, μπορούν να χρησιμοποιηθούν για την υποστήριξη της διάγνωσης, τη δημιουργία μοντέλων για την πιθανότητα εμφάνισης μιας ασθένειας ή την αξιολόγηση της αποτελεσματικότητας νέων διαγνωστικών εξετάσεων και θεραπευτικών μεθόδων. Τα κλινικά συστήματα υποστήριξης αποφάσεων βασίζονται σε μαθηματικές μεθόδους ανάλυσης και επεξεργασίας δεδομένων και μπορούν να συνδυάσουν με μη-γραμμικό τρόπο ένα σύνολο διαφορετικών δεδομένων, όπως ατομικά στοιχεία ασθενών, αποτελέσματα εξετάσεων, στοιχεία αποτελεσματικότητας των θεραπευτικών μεθόδων, παράγοντες κινδύνου, επιδημιολογικά δεδομένα, κ.α. Τα συστήματα αυτά μπορούν να εξάγουν κρυμμένες πληροφορίες υψηλής κλινικής αξίας από μεγάλα σύνολα δεδομένων. Παρέχουν έτσι στους ιατρούς τη δυνατότητα εξατομικευμένων αξιολογήσεων και συστάσεων για κάθε ασθενή, καθώς και την έγκαιρη και ορθή λήψη αποφάσεων.

Στον καρκίνο του τραχήλου της μήτρας, τα κλινικά συστήματα υποστήριξης αποφάσεων μπορούν να υποστηρίξουν τους ιατρούς σε τομείς όπως στη σχεδίαση βελτιωμένων πρωτοκόλλων πληθυσμιακού ελέγχου, στη βελτίωση της διαχείρισης γυναικών με παθολογικό τεστ Παπανικολάου, καθώς και στην εξατομικευμένη παρακολούθηση και λήψη αποφάσεων για γυναίκες που έχουν υποβληθεί σε θεραπεία.

Συγκεκριμένα, η προσέγγιση με χρήση ΓΑ, αποτελεί μια ευέλικτη μέθοδο για μελλοντική χρήση. Για παράδειγμα, μπορεί να χρησιμοποιηθεί για τη βελτιστοποίηση άλλων διαγνωστικών μέτρων απόδοσης, όπως το PPV, ανάλογα με τις απαιτήσεις και τη στρατηγική που εφαρμόζεται στο σύστημα υγείας. Επιπλέον, μελλοντικό ερευνητικό στόχο μπορεί να αποτελέσει η εφαρμογή της μεθοδολογίας που πραγματοποιήθηκε σε περισσότερα και διαφορετικά περιστατικά ώστε να αξιολογηθεί πληρέστερα η συμβολή της στη διαγνωστική διαδικασία και παράλληλα να ερευνηθούν και άλλες μέθοδοι τεχνητής νοημοσύνης οι οποίες ενδεχομένως να μπορέσουν να επιτύχουν πιο υψηλή απόδοση.

Παράρτημα Α

Α.1 Διαγνωστικά μέτρα

Α.1.1 Ευαισθησία

Ακολούθως παρουσιάζονται οι πιθανές καταστάσεις.

Πίνακας 29: Πιθανές καταστάσεις

	Παρουσία νόσου	Απουσία νόσου
Δοκιμασία/Εξέταση θετική	TP	FP
Δοκιμασία/Εξέταση αρνητική	FN	TN

Η ευαισθησία (sensitivity), γνωστή και ως true positive rate, μετρά την αναλογία των θετικών περιστατικών που αναγνωρίστηκαν ορθά (πιθανότητα ορθής θετικής διάγνωσης). Μια εργαστηριακή εξέταση θεωρείται ευαίσθητη όταν μπορεί να εντοπίσει ορθά τα θετικά αποτελέσματα. Αλλιώς, η ευαισθησία μπορεί να οριστεί ως το ποσοστό των ασθενών που νοσούν και έχουν θετική την εν λόγω δοκιμασία.

Ευαισθησία (SN/TPR): $SN = TPR = \frac{TP}{TP+FN}$ (Α.1)

Α.1.2 Ειδικότητα

Η ειδικότητα (specificity/true negative rate) μετρά την αναλογία των αρνητικών περιστατικών τα οποία αναγνωρίστηκαν ορθά (πιθανότητα ορθής αρνητικής διάγνωσης). Μια εργαστηριακή εξέταση θεωρείται ειδική όταν μπορεί να εντοπίσει ορθά τα αρνητικά αποτελέσματα. Η ειδική εξέταση, έχει όσο το δυνατόν λιγότερα ψευδώς θετικά αποτελέσματα. Αλλιώς, η ειδικότητα μπορεί να οριστεί ως το ποσοστό των ασθενών που δεν νοσούν και έχουν αρνητική την εν λόγω δοκιμασία.

Ειδικότητα (SP/TNR): $SP = TNR = \frac{TN}{TN+FP}$ (Α.2)

Α.1.3 Θετική Διαγνωστική Προβλεπτική Αξία

Η θετική διαγνωστική προβλεπτική αξία (ΘΠΑ- PPV) είναι η πιθανότητα ενός περιστατικού να είναι όντως παθολογικό όταν η δοκιμασία αποβαίνει θετική.

Θετική διαγνωστική προβλεπτική αξία (ΘΠΑ- PPV): $PPV = \frac{TP}{TP+FP}$ (Α.3)

Α.1.4 Αρνητική Διαγνωστική Προβλεπτική Αξία

Η αρνητική διαγνωστική προβλεπτική αξία (ΑΠΑ-NPV) είναι η πιθανότητα ένα άτομο να είναι φυσιολογικό (δηλαδή να μην ασθενεί από την υπό μελέτη νόσο) όταν η δοκιμασία είναι αρνητική.

Αρνητική διαγνωστική προβλεπτική αξία (ΑΠΑ-NPV): $NPV = \frac{TN}{TN+FN}$ (Α.4)

Παράρτημα Β

Β.1 Βασικές Αρχές Πιθανοτήτων

Β.1.1 Δεσμευμένη Πιθανότητα (Conditional Probability)

Η πιθανότητα P(B|A) στην εξίσωση (Β.2) είναι η δεσμευμένη πιθανότητα (ή αλλιώς η υπό συνθήκη πιθανότητα) του Β δοθέντος του Α και ορίζεται από τον τύπο:

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \quad \text{(Β.1)}$$

Β.1.2 Κανόνας του Bayes (Bayes Rule)

Από την εξίσωση (Β.2) το θεώρημα Bayes γράφεται ως:

$$P(B|A)P(A) = P(A|B)P(B) \quad \text{(Β.2)}$$

Β.1.3 Κανόνας τομής

$$P(A \cap B) = P(A|B)P(B) \quad \text{(Β.3)}$$

Β.1.4 Κανόνας ένωσης

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad \text{(Β.4)}$$

Β.1.5 Στατιστική Ανεξαρτησία (Statistical Independence)

Δύο ή περισσότερες τυχαίες μεταβλητές x και y είναι στατιστικά ανεξάρτητες αν και μόνο αν:

$$P(x, y) = P(x)P(y) \quad \text{(Β.5)}$$

ΠΑΡΑΡΤΗΜΑ Γ

Γ.1 Άλλοι χρήσιμοι ορισμοί

Γ.1.1 Επίπτωση

Επίπτωση είναι η συχνότητα εμφάνισης νέων κρουσμάτων της νόσου, δηλαδή ο αριθμός των αρχικά υγιών ατόμων που εμφάνισαν τη νόσο σε ένα ορισμένο χρονικό διάστημα προς τον ολικό αριθμό των υγιών ατόμων την ίδια χρονική περίοδο.

$$Επίπτωση = \frac{\text{Αριθμός αρχικά υγιών ατόμων που εμφάνισαν τη νόσο σε χρόνο } \Delta t}{\text{Ολικός αριθμός υγιών ατόμων την ίδια χρονική περίοδο } \Delta t} \quad (Γ.1)$$

Γ.1.2 Επιπολασμός

Επιπολασμός είναι η συχνότητα εμφάνισης και εξάπλωσης μιας υπάρχουσας νόσου στο γενικό πληθυσμό. Ο επιπολασμός είναι ίσος με τον αριθμό των ατόμων με το κλινικό συμβάν προς τον ολικό αριθμό ατόμων την ίδια χρονική στιγμή.

$$Επιπολασμός = \frac{\text{Αριθμός ατόμων με το κλινικό συμβάν}}{\text{Ολικός αριθμός ατόμων την ίδια χρονική στιγμή}} \quad (Γ.2)$$

Αρκτικόλεξα

ΓΑ: Γενετικοί Αλγόριθμοι

GA: Genetic Algorithms

BBH: Building Block Hypothesis

ΣΙ: Συνάρτηση Ικανότητας

Cs: Chromosomes

UX: Uniform crossover

LUX: Half Uniform crossover

FS: Feature Selection

FSS: Feature Subset Selection

IFR: Individual Feature Ranking

ROC: Receiver Operating Characteristic

TPR: True Positive Rate

TNR: True Negative Rate

FPR: False Positive Rate

FNR: False Negative Rate

TP: True Positive

TN: True Negative

FP: False Positive

FN: False Negative

SN: Sensitivity

SP: Specificity

mRMR: minimum Redundancy and Maximum Relevance

RELIEF: RELevance In Estimated Features

ANN: Artificial Neural Networks

ΤΝΔ: Τεχνητά Νευρωνικά Δίκτυα

EEG: Electroencephalogram

ΗΕΓ: Ηλεκτροεγκεφαλογράφημα

ERP: Event Related Potential

MLP: Multi Layered Perceptron

GAPE: Genetic Algorithm with Puncuated Equilibria

SVM: Support Vector Machine

k-nn: k nearest neighbors (classifier)

CC: Cervical Cancer

CxCa: Cancer of the Cervix

SCC: Squamous Cell Carcinoma

ADC: Adenocarcinoma

Adeno-Ca: Adenocarcinoma

ICC: Invasive Cervical Cancer

WHO: World Health Organization

ΠΟΥ: Παγκόσμιος Οργανισμός Υγείας

PAHO: Pan-American Health Organization

CDC: Centers for Disease Control and Protection

HPV: Human Papilloma Virus

PV: Papilloma Virus

URR: Upstream Regulatory Region

NCR: Non Coding Region

LCR: Long Control Region

CIN: Cervical Intraepithelial Neoplasia

SIL: Squamous Intraepithelial Lesion

LSIL: Low grade Squamous Intraepithelial Lesion

HSIL: High grade Squamous Intraepithelial Lesion

CIS: Carcinoma In Situ

TBS: The Bethesda System

IARC: International Agency for Research on Cancer

ICO: Institut Català d'Oncologia (Information Centre on HPV and Cancer)

DNA: Deoxyribonucleic Acid

RNA: Ribonucleic Acid

ORF: Open Reading Frame

HIV: Human Immunodeficiency Virus

CT: Clamydia Trachomatis

HSV: Herpes Simplex Virus

pRb: πρωτεΐνη ρετινοβλαστώματος

E: European

As: Asian

AA: Asian-American

Af1: African-1

Af2: African-2

WNL: Within Normal Limits

AGUS: Atypical Glandular cells of Undetermined Significance

ASCUS: Atypical Squamous Cells of Undetermined Significance

LBC: Liquid Based Cytology

TBS: The Bethesda System

SN: Sensitivity

SP: Specificity

PPV: Positive Predictive Value

NPV: Negative Predictive Value

NB: Naïve-Bayes

CDSS: Clinical Decision Support System

YI: Youden's Index

OSP: Organized Screening Programs

FDA: Food and Drug Administration

EMEA: European Medicines Evaluation Agency

Βιβλιογραφία

[1] Charles Darwin, "The Origin of Species on the basis of Natural Selection", 1859
[2] Stuart J.Russell and Peter Norvig, "Artificial Intelligence: A Modern Aprroach", Prentice Hall, 1995
[3] Melanie Mitchell, "An introduction to genetic algorithms", (5th edition) The MIT press, Cambridge, MA, 1999
[4] John H. Holland, "Adaptation in Natural and Artificial Systems", Ann Arbor: University of Michigan Press, 1975
[5] Goldberg D.E., Genetic algorithms in search, optimization and machine learning, Addison Wesley, Reading, 1989
[6] P. Fleming and R.C Purshouse, "Evolutionary algorithms in control systems engineering: a survey", Control Engineering Practice, 10: pp1223-1241, 2002
[7] Lowen R. and Verschoren A., Mathematical Modelling: Theory and application, Foundations of generic optimization Volume2: Applications of fuzzy control, genetic algorithms and neural networks, Springer, 2008
[8] Kumar R. , Novel Encoding Scheme in Genetic Algorithms for Better Fitness, International Journal of Engineering and Advanced Technology, Vol.1 , 6, 2012
[9] Patvichaichod S., Application of hybrid encoding genetic algorithms on pickup and delivery traveling salesman problem with traffic conditions, Journal o Computer Science, vol.7 , 5 , pp605-610, 2011
[10] H.Kitano, "Designing neural networks using genetic algorithms with graph generation system, Complex Systems, 4, 1990
[11] F.Rothlauf, "Representations for Genetic and Evolutionary Algorithms", Springer, 2006
[12] Smith M.J., Evolutionary Genetics, Oxford University Press, 1998
[13] Haupt R.L. and Haupt S.E., "Practical genetic algorithms", Wiley-Interscience, 2004
[14] K.Jebari and M. Madiafi, "Selection methods for Genetic Algorithms", International Journal of Emerginh Sciences 3, 4: pp333-334, 2013
[15] Baeck T., Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms, Oxford University Press, New York, 1996
[16] Sivanadam S.N. and Deepa S.N., Introduction to Genetic Algorithms, Springer, 2008
[17] De Jong K.A., Spears W.M., "A formal analysis of the role of multi-point crossover in genetic algorithms", Annals of Mathematics and Artificial Intelligence, 5, pp1-26, 1992
[18] De Jong K.A., Spears W.M., "An analysis of multi-point crossover", Foundations of Genetic Algorithms, Morgan Kaufmann Publishers, 1991
[19] Yang S., "Adaptive Crossover in Genetic Algorithms Using Statistics Mechanism, Artificial Life, 8, Massachusetts Institute of Technology, 2003
[20] Schaffer J.D.,Caruna R.A., Eshelman L.J.,Das R., "A study of control parameters affecting online performance of genetic algorithms for function optimization", Proceeding of the 3rd International Conference on Genetic Algorithms and their applications, 1989
[21] Koumousis V.K. and Katsaras C., "The effect of population variation in genetic algorithms- The saw-tooth GA", IEEE Transactions on Evolutionary Computation, 10, 1, pp19-28, 2006

[22] Vasconcelos J.A., Ramirez J.A., Takahashi R.H.C., Saldanha R.R., "Improvements in genetic algorithms", Magnetics, IEEE Transactions, vol. 37, 5, pp3414-3417, 2001
[23] Λυκοθανάσης Σ., «Γενετικοί Αλγόριθμοι και Εφαρμογές», vol.3, ΕΑΠ, 2000
[24] Michalewicz Z. , Genetic Algorithms + Data Structures = Evolution Programs, Springer, 1992
[25] John H. Holland, "Genetic Algorithms", Scientific American, 267(1):pp66-72, 1992
[26] Mitchell M. and Forrest S., Relative building block Fitness and the building-block hypothesis, Foundations of Genetic Algorithms, D. Whitley, vol.2, pp109-126, 1993
[27] Leardi R., Boggia R., Terrile M., "Genetic Algorithms as a strategy for feature selection", Journal of Chamometrics, vol.6, pp267-281, 1992
[28] Guyon I., Eliseff A., "An introduction to variable and feature selection", Journal of machine learning research, vol. 3, pp 1157-1182, 2003
[29] Chao Sima, E.R. Dougherty, "The peaking phenomenon in the presence of feature-selection", Pattern Recognition Letters, vol. 29, pp1667-1674, 2008
[30] Liu H., Motoda H., Feature selection for knowledge Discovery and data mining, Springer, 1998
[31] Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection", International Joint Conference on Artificial Intelligence (IHCAI), 1995
[32] Yang J., Honavar V., "Feature subset selection using genetic algorithm", IEEE Intelligent Systems vol. 13, pp 44-49, 1998
[33] Siedlecki W., Sklansky J., "A note on genetic algorithms for large-scale feature selection", Pattern Recognition Letters, vol. 10, pp335-347, 1989
[34] Vafaie H., De Jong K., "Genetic algorithms as a tool for feature selection in machine learning", Proceedings of the 4th International Conference on Tools with Artificial Intelligence, Arlington, IEEE Society Press, pp200-203, 1992
[35] Srivastava S., Joshi N., Gaur M., "A review paper on feature selection methodologies and their applications", International Journal of Engineering and Development, vol. 7, pp57-61, 2013
[36] Guyon I., Gunn S., Nikravesh M., Zadeh L.A., Feature Extraction: Foundations and applications, Springer, 2008
[37] Saeys Y., Inza I., Larranaga P., "A review of feature selection techniques in bioinformatics", Bioinformatics, vol. 23, pp 2507-2517, 2007
[38] Dash M., Liu H., "Consistency-based search in feature selection", Artificial Intelligence, Elsevier, vol 151, pp 155-176, 2003
[39] Dua S., Chowriappa P., Data mining for Bioinformatics, CRC Press, 2012
[40] Theodoridis S., Koutroumbas K., Pattern Recognition, 4th ed., London: Elsevier: Academic Press, 2009
[41] Lasko T.A., Bhagwat J.G.,Zou K.H.,Lucila O.M., "The use of receiver operating characteristic curves in biomedical informatics", Journal of Biomedical Informatics, Elsevier, vol. 38, pp404-415,2005
[42] Fawcett T., "An introduction to ROC analysis", Pattern Recognition Letters, Elsevier, vol. 27, pp861-874, 2006
[43] Ding C., Peng H., "Minimum redundancy feature selection from microarray gene expression data", 2nd IEEE Computer Society Bioinformatics Conference (CSB), pp523-529, 2003 / Ding C., Peng H., "Minimum redundancy feature selection from microarray gene expression data", Journal of Bioinformatics and Computational Biology, vol. 03, pp185-206 , 2005

[44] Kira K., Rendell L.A., "The feature selection problem: traditional methods and a new algorithm", AAAI-92 Proceeding, AAAI Press, pp129-134, 1992
[45] Kononenko Igor, "Estimating Attributes: Analysis and Extensions of Relief", In Proceedings of the European conference of Machine Learning, Catana, Italy, Springer-Verlag, New York, pp171-182, 1994
[46] Kononeko Igor et al., "Overcoming the myopia of inductive learning algorithms with RELIEFF", Applied Intelligence, pp39-55, 1997
[47] John G.H., Kohavi R., Pfleger K., "Irrelevant Features and the Subset Selection Problem", Machine Learning: Proceeding of the 11th International Conference, pp121-129, Morgan Kaufmann Publishers, 1994
[48] Kohavi R., John G.H., "Wrappers for feature subset selection", Artificial Intelligence, vol. 97, pp273-324, 1997
[49] Liu H., Motoda H., Computational methods of feature selection, CRC Press, 2007
[50] Halliman J,"Feature selection and classification in the diagnosis of cervical cancer", In Chambers L.D. (ed), The practical Handbook of Genetic Algorithms: Applications, 2nd Edition, CRC Press, 2000
(51) Vafaie H., De Jong K., "Robust feature selection algorithms", Proceedings of the International Conference on Tools with AI, Boston, IEEE Computer Society Press, pp356-364, 1993
[52] Chang E.I., Lippman R.P., "Using Genetic Algorithms to Improve Pattern Classification Performance", Advances in neural information processing systems, vol. 3, pp797-803, 1991
[53] Karegowda A.G., Jayaram M.A., Manjunath A.S., Shama V.T., "GA based dimensionality reduction for improving performance of k-means clustering: a case study for categorization of medical dataset", International Journal of Soft. Computing, vol. 7 (5), pp249-255, 2012
[54] Punch W.F., Goodman E.D., Pei M., Chia-Shun L., Hovland P., Enbody R., "Further research on feature selection and classification using genetic algorithms", Proceeding of the 5th International Conference on Genetic Algorithms, Champaign III, pp.557-564, 1993
[55] Whitley D., Beveridge J.R., Gierra-Salcedo C., Graves C., "Messy genetic algorithm for subset feature selection", Proceedings of the 7th International Conference on Genetic Algorithms, 1997
[56] Vafaie H., De Jong K., "Improving the performance of a rule unduction system using Genetic algorithms", Proceedings of the 1st International Workshop on Multistrategy Learning, 1991
[57] Yang J., Honavar V., "Feature Subset Selection Using A Genetic Algorithm", Computer Science Technical Reports, 1997
[58] Yang J., Parekh R., Honavar V., "DistAI: An Inter-pattern Distance-based Constructive Learning Algorithm", Technical Report ISU-CS-TR 97-05, Artificial Intelligence Research Group, Department of Computer Science, Iowa State University, 1997
[59] Kim H.T., Kim B.Y., Park E.H., Kim J.W., Hwang E.W., Han S.K., Cho S., "Computerized recognition of Alzheimer disease-EEG using genetic algorithms and neural network", Future Generation Computer Systems, vol. 21, pp1124-1130, 2005
[60] Brill F.Z., Brown D.E., Martin W.N., "Fast genetic selection of features for neural network classifiers", IEEE Transactions on Neural Networks, vol.3, no.2, pp324-328, 1992

[61] Tan F., Fu X., Zhang Y., Bourgeois A. G., "A genetic-algorithm based method for feature subset selection", Soft Computing – A Fusion of Foundations, Methofologies and Applications, vol.12, no.2, pp111-120, Springer-Verlag, 2007
[62] Zhuo L., Zheng J., Wang F., Li X., Ai B., Qian J., "A genetic algorithm based wrapper feature selection method for classification of hyperspectral images using support vector machine", The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 37,pp 397-402, 2008
[63] Chtioui Y., Bertrand D., Barba D., "Feature Selection by a Genetic Algorithm. Application to Seed Discrimination by Artificial Vision", Journal of the Science of Food and Agriculture, vol. 76, pp77-86, 1998
[64] Li S., Wu H., Wan D., Zhu J., "An effective feature selection method for hyperspectral image classification based on genetic algorithm and support vector machine", Knowledge-Based Systems, vol. 24, pp40-48, 2011
[65] Pernkopf F., O'Leary P., "Feature Selection for Classification Using Genetic Algorithms with a Novel Encoding", Computer Analysis of Images and Patterns, vol. 2124, pp161-168, 2001
[66] Rejer I., "Genetic Algorithms in EEG Feature Selection for the Classification of Movements of the Left and Right Hand", Proceedings of the 8th International Conference on Computer Recognition Systems CORES, vol. 226, pp579-589, 2013
[67] Kalapanidas E., Avouris N.**, "**Feature Selection using a Genetic Algorithm applied on an Air Quality Forecasting Problem", 3rd BESAI, Proc. ECAI, 2002
[68] Tan K.C., Teoh E.J., Yu Q., Goh K.C., "A hybrid evolutionary algorithm for attribute selection in data mining", Expert Systems with Applications, vol. 36, pp8616-8630, 2009
[69] URL: http://www.bestrong.org.gr/el/cancer/typesofcancer/cervicalcancer/ (Retrieved: 16 September 2015)
[70] Khan F.A., Biotechnology in Medical Sciences, chapter 4, pp101, CRC Press, 2014
[71] World Cancer Report 2008, World Health Organization (WHO), International Agency for Research on Cancer (IARC), edited by Peter Boyle and Bernard Levin, 2008, Available at url: http://www.iarc.fr/en/publications/pdfs-online/wcr/index.php (Retrieved: 16 September 2015)
[72] Athens Medical Society, Archives of Hellenic Medicine, Available at url: www.mednet.gr/archives (Retrieved: 16 September 2015)
[73] U.S. Cancer Statistics Working Group, United States Cancer Statistics: 1999-2012 Incidence and Mortality Web-based Report, Atlanta: Department of Health and Human Services, Centers for Disease Control and Prevention (CDC) and National Cancer Institute: 2015, Available at URL: http://www.cdc.gov/cancer/cervical/statistics/, (Retrieved: 16 September 2015)
[74] Johns Hopkins Medicine, The Sydney Kimmel Comprehensive Cancer Center, John Hopkins Cervical Dysplasia Centers, url: http://www.hopkinsmedicine.org/kimmel_cancer_center/centers/cervical_dysplasia/ (Retrieved: 16 September 2015)
[75] Αρεταίειο Νοσοκομείο, Β' Μαιευτική και Γυναικολογική Κλινική, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, Available at url: http://www.aretaieio-obgyn.com/el/hpv.html?start=1, (Retrieved: 16 September 2015)
[76] Ministry of Health of the Republic of Cyprus, Health Monitoring Unit, Cyprus Cancer Registry, Available at url: http://www.moh.gov.cy/Moh/MOH.nsf/All/82B40CE2FEE19D7AC22579C6002CBFAD?OpenDocument , (Last accessed: 16 September 2015)

[77] European Code Against Cancer (EUCAN) national estimates 2012, ECO database, International Agency for Research on Cancer (IARC), European Cancer Observatory, World Health Organization, URL: http://eu-cancer.iarc.fr/EUCAN/CancerOne.aspx?Cancer=25&Gender=2, (Retrieved: 16 September 2015)
[78] URL: http://mortakis.hpvinfocenter.gr/index.php/cin-123-lsil-hsil/o-traxilos-tis-mitras-stoixeia-anatomias, (Retrieved: September 2015)
[79] A.D.A.M., URAC (American Accreditation HealthCare Commision), URL: www.adamimages.com
[80] MD Anderson Cancer Center, The University of Texas, http://www.mdanderson.org/patient-and-cancer-information/cancer-information/cancer-types/cervical-cancer/index.html, (Retrieved: 16 September 2015)
[81] Löwy I., A woman's disease: The history of cervical cancer, Oxford University Press, 2011
[82] Demay M.R., Practical Principles of Cytopathology, American Society for Clinical Pathology Press, 2007
[83] Illustration: Joe Gorman, OBG Management, vol. 25, pp42, 2013
[84] Μιχαήλ Γ., «Οι προκαρκινικές αλλοιώσεις και ο καρκίνος του τραχήλου της μήτρας στην Ελλάδα», Σεμινάριο Γυναικολογικού Καρκίνου, 2008
[85] Monsonego J., Bosch F.X., Coursaget P., Cox J.T., Franco E., Frazer I., Sankaranarayanan R., Schiller J., Singer A., Wright T., Kinney W., Meijer C., Linder J., "Mini Review: Cervical Cancer Control, Priorities and New Directions", International Journal of Cancer (IJC), vol. 108, pp329-333, 2004
[86] Saslow D., Solomon D., Lawson H.W., Killackey M., Kulasingam S.L., Cain J., Garcia A.R., Moriarty A.T., Waxman A.G., Wilbur D.C., Wentzensen N., Downs L.S., Spitzer M., Moscicki A.B., Franco E.L., Stoler M.H., Schiffman M., Castle P.E., Myers E.R., "American Cancer Society, American society for Colposcopy and Cervical Pathology, and American Society for Clinical Pathology Screening Guidelines for the Prevention and Early Detection of Cervical Cancer", American Journal of Clinical Pathology, vol. 137, pp516-542, 2012
[87] Borruto F., De Ridder M. (Eds), HPV and Cervical Cancer: Achievements in Prevention and Future Prospects, Springer, 2012
[88] zur Hausen H., "Papillomaviruses and cancer: from basic studies to clinical application", Nature Reviews Cancer, vol. 2, pp.342-350, 2002
[89] ICTV 8th Report, Images by Jean-Yves Sgro, 2004, Available at: http://www.virology.wisc.edu/virusworld (Retrieved: 18 September 2015)
[90] Modis Y., Trus B.L., Harrison S.C., "Atomic model of the papillomavirus capsid", The EMBO Journal, vol. 21, pp4754-4762, 2002
[91] Luria S.E., Darnell J.E., General virology, New York: Wiley, 1967
[92] Physicians Research Network (prn), url: http://www.prn.org/index.php/provider_resources/prn_artwork (Retrieved 16 of September 2015)
[93] D'Abramo C.M., Archambault J., "Small Molecule Inhibitors of Human Papillomavirus Protein – Protein Interactions", The Open Virology Journal, vol. 5, pp80-95, 2011
[94] Gomez D.T., Santos J.L., "Human papillomavirus infection and cervical cancer: pathogenesis and epidemiology", Communicating Current Research and Educational Topics and Trends in Applied Microbiology, ed. A. Mendez-Vilas, Spain: Formatex Research Center, pp680-688, 2007
[95] Doorbar J., "Molecular biology of human papillomavirus infection and cervical cancer", Clinical Science, vol. 110, pp525-541, 2006

[96] Munoz N., Castellsague X., de Gonzalez A.B., Gissmann L., "HPV Vaccines and Screening in the Prevention of Cervical Cancer, Chapter 1: HPV in the etiology of human cancer", Vaccine, vol. 24, Suppl. 3, ppS1-S10, 2006
[97] Kellof G.J., Sigman C.C., "Assessing intraepithelial neoplasia and drug safety in cancer-preventive drug development", Nature Reviews Cancer, vol. 7, pp508-518, 2007
[98] Woodman C.B.J., Collins S.I., Young L.S., "The natural history of cervical HPV infection: unresolved issues", Nature Reviews Cancer, vol. 7, pp11-22, 2007
[99] Alba A., Cararach M., Rodriguez-Cerdeira C., "The Human Papillomavirus (HPV) in Human Pathology: Description, Pathogenesis, Oncogenic Role, Epidemiology and Detection Techniques", The Open Dermatology Journal, vol. 3, pp90-102, 2009
[100] Moscicki A.B., Shiboski S., Broering J., Powell K., Clayton L., Jay N., Darragh T.M., Brescia R., Kanowitz S., Miller S.B., Stone J., Hanson E., Palefsky J., "The natural history of human papillomavirus infection as measured by repeated DNA testing in adolescent and young women", The Journal of Pediatrics, vol. 132, pp277-284, 1998
[101] Hudelist G., Manavi M., Pischinger K.I., Watkins-Riedel T., Singer C.F., Kubista E., Czerwenka K.F., "Physical state and expression of HPV DNA in benign and dysplastic cervical tissue: different levels of viral intergration are correlated with lesion grade", Gynecologic Oncology, vol. 92, pp873-880, 2004
[102]http://www.nobelprize.org/nobel_prizes/medicine/laureates/2008/illpres.html (Retrieved: 15 October 2015)
[103] Bernard H.U., "The clinical importance of the nomenclature, evolution and taxonomy of human papillomaviruses", Journal of Clinical Virology, 2005
[104] de Villiers E.M., Fauquet C., Broker T.R., Bernard H.U., zur Hausen H., "Classification of papilomaviruses", Virology, vol. 324, pp17-27, 2004
[105] Bernard H.U., Burk R.D., Chen Z., van Doorslaer K., zur Hausen H., de Villiers E.M., "Classification of Papilomaviruses (PVs) Based on 189 PV Types and Proposal of Taxonomic Amendments", Virology, vol. 401, pp70-79, 2010
[106] Munoz N., Bosch X., de Sanjose S., Herrero R., Castellsague X., SHh K.V., Snijders P.J.F., Meijer C.J.L.M., "Epidemiologic Classification of Human Papillomavirus Types Associated with Cervical Cancer", The New England Journal of Medicine, vol. 348, pp518-527, 2003
[107] Milde-Langosch K., Riethdorf S., Loning T., "Association of human papillomavirus infection with carcinoma of the cervix uteri and its precursor lesions: theoretical and practical implications", Virchows Archiv: an international journal of pathology, vol. 437(3), pp227-233, 2000
[108] Munoz N., Bosch F.X., Castellsague X., Diaz M., de Sanjose S., Hammouda D., Shah K.V., Meijer C.J., "Against which human papillomavirus types shall we vaccinate and screen? The International perspective", International Journal of Cancer (IJC), vol. 111, pp278-285, 2004
[109] Li N., Franceschi S., Howell-Jones R., Snijders P.J., Clifford G.M., "Human papillomavirus type distribution in 30848 invasive cervical cancers worldwide: variation by geographical region, histological type and year of publication", International Journal of Cancer (IJC), vol. 128, pp927-935, 2011
[110] de Sanjose S., Quint W.G., Alemany L., Geraets D.T., Klaustermeier J.E., Lloveras B., Tous S., Felix A., Bravo L.E., Shin H.R., Vallejos C.S., de Ruiz P.A., Lima M.A., Guimera N., Clavero O., Alejo M., Llombart-Bosch A., Cheng-Yang C., Tatti S.A., Kasamatsu E., Iljazovic E., Odida M., Prado R., Seoud M., Grce M., Usubutun A., Jain A., Suarez G.A., Lombardi L.E., Banjo A., Menendez C., Domingo E.J., Velasco J., Nessa A., Chichareon S.C., Qiao Y.L., Lerma E., Garland

S.M., Sasagawa T., Ferrera A., Hammouda D., Mariani L., Pelayo A., Steiner I., Oliva E., Meijer C.J., Al-Jassar W.F., Cruz E., Wright T.C., Puras A., Liave C.L., Tzardi M., Agorastos T., Garcia-Barriola V., Clavel C., Ordi J., Andujar M., Castellsague X., Sanchez G.I., Nowakowski A.M., Bornstein J., Munoz N., Bosch F.X., "Human papillomavirus genotype attribution in invasive cervical cancer: a retrospective cross-sectional worldwide study", Lancet. Oncol., vol. 11, pp1048-1056, 2010

[111] Castellsague X., Munoz N., "Chapter 3: Cofactors in Human Pappilomavirus Carginogenesis – Role of Parity, Oral Contraceptives, and Tobacco Smoking", Journal of the National Cancer Institute Monographs, vol. 31, pp20-28, 2003

[112] Papanicolaou G.N., Traut H.F., "The diagnostic value of vaginal smears in carcinoma", American Journal of Obstetrics and Gynecology, vol. 42, pp193-206, 1941

[113] Arbyn M., Anttila A., Jordan J., Ronco G., Schenck U., Segnan N., Wiener H., Herbert A., von Karsa L., "European Guidelines for Quality Assurance in Cervical Cancer Screening. Second Edition—Summary Document", Annals of Oncology, vol.21, pp448-458, 2010

[114] International Agency for Research on Cancer, World Health Organization, Cervix Cancer Screening, IARC Handbook of Cancer Prevention. Vol. 10. IARC Press, 2005

[115] Prabhu T.R.B, A Practical Approach to Cervical Cancer Screening Techniques, Jaypee Brothers Medical Publishers (P), 2015

[116] Mohan H., Pathology Practical Book, 2nd edition, Jaypee Brothers Medical Publishers (P), 2007

[117]http://www.womenshealth.gov/publications/our-publications/fact-sheet/pap-test.html (Retrieved: November 2015)

[118] Solomon D., Davey D., Kyrman R. et al., "The 2001 Bethesda System: Terminology for Reporting Results of Cervical Cytology", JAMA, vol. 287, pp2114-2119, 2002

[119] Pantanowitz L., Hornish M., Goulart R.A., "The impact of digital imaging in the field of cytopathology", CytoJournal, vol. 6, 2009

[120] http://www.biodyne-emea.com/lbc.php (Retrieved: November 2015)

[121] Καρακίτσος Π., «Κυτταρολογία των τραχηλικών ενδοεπιθηλιακών αλλοιώσεων», Κολποσκόπηση και Παθολογία του Κατώτερου Γεννητικού Συστήματος της Γυναίκας, Εκδόσεις Πασχαλίδης, 2010

[122] Fremont-Smith M., Marino J., Griffin B., Spencer L., Bolick D., "Comparison of the SurePath liquid-based Papanicolaou smear with the conventional Papanicolaou smear in a multisite direct-to-vial study", vol. 102, pp269-279, 2004

[123] Zhu J., Norman I., Elfgren K., Gaberi V., Hagmar B., Hjerpe A., Andersson S., "A comparison of liquid-based cytology and Pap smear as a screening method for cervical cancer.", Oncology Reports, vol.18(1), pp157-160, 2007

[124] Linder J., Zahniser D., "The ThinPrep Pap test. A review of clinical studies", Acta Cytologica, vol. 41(1), pp30-38, 1997

[125] Strander B., Andersson-Ellström A., Milsom I., Rådberg T., Ryd W., "Liquid-based cytology: evaluation of effectiveness, cost-effectiveness, and application to present practice", Cancer, vol. 111, pp285-291, 2007

[126] Corkill M., Knapp D., Hutchinson M.L., "Improved Accuracy for Cervical Cytology with the ThinPrep Method and the Endocervical Brush-Spatula Collection Procedure", J Low Genit Tract Dis., vol. 2(1), pp12-16, 1998

[127] Coste J., Cochand-Priollet B., de Cremoux P., Le Gales C., Isabelle C., Vincent M., Labbe S., Vacher-Lavenu M., Vielh P., "Cross sectional study of

conventional cervical smear, monolayer cytology, and human papillomavirus DNA testing for cervical cancer screening", BMJ, vol. 326:733, 2003
[128] Arbyn M., Bergeron C., Klinkhamer P., Martin-Hirsch P., Siebers A.G., Bulten J., "Liquid compared with conventional cervical cytology: a systematic review and meta-analysis", Obstetrics and Gynecology, vol. 111(1), pp167-177, 2008
[129] Siebers A.G., Klinkhamer P.J., Grefte J.M., Massuger L.F., Vedder J.E., Beijers-Broos A., Bulten J., Arbyn M., "Comparison of liquid-based cytology with conventional cytology for detection of cervical cancer precursors: a randomized controlled trial", JAMA, vol. 302(16), pp1757-1764, 2009
[130] de Bekker-Grob E.W., de Kok I.M., Bulten J., van Rosmalen J., Vedder J.E., Arbyn M., Klinkhamer P.J., Siebers A.G., van Ballegooijen M., "Liquid-based cervical cytology using ThinPrep technology: weighing the pros and cons in a cost-effectiveness analysis", Cancer Causes Control. 2012 vol. 23(8), pp1323-1331, 2012
[131] Gibb R.K., Martens M.G., "The Impact of Liquid-Based Cytology in Decreasing the Incidence of Cervical Cancer", MedReviews, Rev Obstet Gynecol, vol. 4 (1), 2011
[132] World Health Organization, "Comprehensive Cervical Cancer Control. A guide to essential practice", WHO, 2006.
[133] Bengtsson E., Patrik Malm P., "Screening for Cervical Cancer Using Automated Analysis of PAP-Smears", Computational and Mathematical Methods in Medicine, 2014
[134] Koss L.G., Lin E., Schreiber K., Elgert P., Mango L., "Evaluation of the PAPNET cytologic screening system for quality control of cervical smears.", Am J Clin Pathol., vol. 101, pp220-229, 1994
[135] Chivukula M., Saad R.S., Elishaev E., White S., Mauser N., Dabbs D.J., "Introduction of the Thin Prep Imaging System (TIS): experience in a high volume academic practice", Cytojournal, vol. 8, pp4-6, 2007
[136] Davey E., d'Assuncao J., Irwig L., Macaskill P., Chan S.F., Richards A., Farnsworth A., "Accuracy of reading liquid based cytology slides using the ThinPrep Imager compared with conventional cytology: prospective study", BMJ, vol. 335, 2007
[137] Molijn A., Kleter B., Quint W., van Doorn L.J., "Molecular diagnosis of human papillomavirus (HPV) infections", Journal of Clinical Virology, vol. 32 (1), pp43-51, 2005
[138] Abreu A.L., Souza R.P., Gimenes F., Consolaro M.E., "A review of methods for detect human Papillomavirus infection", Vorology Journal, vol. 9, 2012
[139] Zaravinos A., Mammas I.N., Sourvinos G., Spandidos D.A., "Molecular detection methods of human papillomavirus (HPV)", The International Journal of Biological Markers, vol. 24(4), pp215-222, 2009
[140] Malloy C., Sherris J., Herdman C., "HPV DNA testing: Technical and Programmatic Issues for Cervical Cancer Prevention in Low-Resource Setting", 2000
[141] Paraskevaidis E., Arbyn M., Sotiriadis A., Diakomanolis E., Martin-Hirsch P., Koliopoulos G., Makrydimas G., Tofoski J., Roukos D.H., "The role of HPV DNA testing in the follow-up period after treatment for CIN: a systematic review of the literature.", Cancer Treatment Reviews, vol. 30, pp205-211, 2004
[142] P.E. Castle, S. Wacholder, M.E. Sherman, A.T. Lorincz, A.G. Glass, D.R. Scott, B.B. Rush, F. Demuth, M. Schiffman, "Absolute risk of a subsequent abnormal pap among oncogenic human papillomavirus DNA-positive, cytologically negative women", Cancer, vol. 95(10), pp2145-2151, 2002

[143] Arney A., Bennett K.M, "Molecular Diagnostics of Human Papillomavirus", Labaratory Medicine, vol. 41 (9), pp523-530, 2010
[144] Kyriazis I.D., Kambouris M.E., Poulas K., Patrinos G.P., "Molecular techniques for the detection and characterization of microorganisms", Archives of Hellenic Medicine, vol. 31 (1), pp23-40, 2014
[145] Mayo Foundation for Medical Education and Research http://www.mayoclinic.org
[146] Beresford J., Gervaize P., "The emotional impact of abnormal Pap smears on patients referred for colposcopy", Colposc Gynecol Laser Surg, vol. 2(2), pp83-87, 1986
[147] Marteau T., Walker P., Giles J., Smail M., "Anxieties in women undergoing colposcopy", Br J Obstet Gynaecol, vol. 97, pp859-861, 1990
[148] Freeman-Wang T., Walker P., "Psyxhological aspects of colposcopy", European Academy of Gynecological Cancer Book Series- Course Book on Colposcopy, Primed-X press, pp166-169, 2003
[149] Rogstad K.E., "The psychological impact of abnormal cytology and colposcopy", BJOG: an International Journal of Obstetrics and Gynaecology, vol. 109, pp364-368, 2002
[150] Kincey J., Statham S., McFarlane T., "Women undergoing colposcopy: their satisfaction with communicating health knowledge and level of anxiety", Health Educ J, vol. 50, pp70-72, 1991
[151] Richardson P.H., Doherty I., Wolfe C.D.A., Carman N., Chamberlain F., Holtom R., Raju K.S., "Evaluation of cognitive-behavioural counselling for the distress associated with an abnormal cervical smear result", Br J Health Psychol, vol. 1, pp327-338, 1996
[152] Bennets A., Irwig L., Oldenburg B., Simpson J.M., Mock P., Boyes A., Adams K., Weisberg E., Shelley J., "PEAPS-Q: a questionnaire to measure the psychosocial effects of having an abnormal PAP smear", J Clin Epidemiol, vol. 48, pp1235-1243, 1995
[153] Shinn E., Basen-Engquist K., Le T., Hansis-Diarte A., Bostic D.S, Martinez-Cross J., Santos A., Follen M., ""Distress after an abnormal Pap smear result: scale development and psychometric validation", Prev Med., vol. 39(2), pp404-412, 2004
[154] Goodkin K., Antoni M.H., Blaney P.H., "Stress and hopelessness in the promotion of cervical intraepithelial neoplasia in invasive squamous cell carcinoma of the cervix", J Psychosex Res, vol. 30, pp67-76, 1986
[155] Antoni M.H., Goodkin K., "Host moderator variables in the promotion of cervical neoplasia", J Psychosex Res, vol. 32, pp327-338, 1988
[156] Totman R., "Mind, Stress and Health", London: Souvenir Press, 1990
[157] Lerman C., Miller S.M., Scarborough R., Hanjani P., Nolte S., Smith D., "Adverse psychologic consequences of positive cytologic cervical screening", Am J Obstet Gynecol, vol. 165, pp658-662, 1991
[158] TOMBOLA Group**,** "Cytological surveillance compared with immediate referral for colposcopy in management of women with low grade cervical abnormalities: multicentre randomised controlled trial", BMJ, 2009
[159] Bentley E., Cotton S.C., Cruickshank M.E., Duncan I., Gray N.M., Jenkins D., Little J., Neal K., Philips Z., Russell I., Seth R., Sharp L., Waugh N., Trial of Management of Borderline and Other Low-Grade Abnormal Smears (TOMBOLA) Group, "Refining the management of low-grade cervical abnormalities in the UK National Health Service and defining the potential for human papillomavirus testing: a commentary on emerging evidence", J Low Genit Tract Dis, vol. 10(1), pp26-38, 2006

[160] Johnson N., Sutton J., Thornton J.G., Lilford R.J., Johnson V.A., Peel K.R., "Decision analysis for best management of mildly dyskaryotic smear", Lancet, vol. 342, pp91-96, 1993
[161] Etherington I.J., Luesley D.M., Shafi M.I., Dunn J., Hiller L., Jordan J.A., "Observer variability among colposcopists from the West Midlands region", Br J Obstet Gynaecol, vol. 104, pp1380-1384, 1997
[162] Cardenas-Turanzas M., Follen M., Benedet J.L., Cantor S.B., ""See-and-treat strategy for diagnosis and management of cervical squamous intraepithelial lesions", Lancet Oncol, vol. 6, pp43-50, 2005
[163] Dunn T.S., Killoran K., Wolf D., "Complications of outpatient LLETZ Procedures", J Reprod Med Obstet Gynecol, vol. 49, pp76-78, 2004
[164] Kyrgiou M., Koliopoulos G., Martin-Hirsch P., Arbyn M., Prendiville W., Paraskevaidis E., "Obstetric outcomes after conservative treatment for intraepithelial or early invasive cervical lesions: systematic review and meta-analysi", Lancet, vol. 367:, pp489-498, 2006
[165] Marteau T.M., "Psychology and screening—narrowing the gap between efficacy and effectiveness", Br J Clin Psychol, vol. 33, pp1-10, 2006
[166] Jones M.H., Singer A., Jenkins D., "The mildly abnormal cervical smear: patient anxiety and choice of management", J R Soc Med, vol. 89, pp257-260, 1996
[167] Eggington S., Hadwin R., Brennan A., Walker P., "Modelling the impact of referral guideline changes for mild dyskaryosis on colposcopy services in England", Sheffield: NHS Cancer Screening Programmes, 2006
[168] Raffle A.E., "Cervical Screening", BMJ, 2004
[169] Scheungraber C., Kleekamp N., Schneider A., "Management of lowgrade squamous intraepithelial lesions of the uterine cervix", Br J Cancer, vol. 90, pp975-978, 2004
[170] Ferris D.G., Kriegel D., Cote L., Litaker M., Woodward L., "Women's triage and management preferences for cervical cytologic reports demonstrating atypical squamous cells of undetermined significance and low-grade squamous intraepithelial lesions", Arch Fam Med, vol. 6, pp348-353, 1997
[171] Melnikow J., Kuppermann M., Birch S., Chan B.K.S., Nuovo J., "Management of the low-grade abnormal pap smear: what are women's preferences?", J Fam Pract, vol. 51, pp849-855, 2002
[172] Flannelly G., Campbell M.K., Meldrum P., Torgerson D.J., Templeton A., Kitchener H.C., "Immediate colposcopy or cytological surveillance for women with mild dyskaryosis: a cost effectiveness analysis.", Journal of Public Health Medicine, vol. 19(4), pp419-423, 1997
[173] Kahn J.A., Slap G.B, Bernstein D.I., Kollar L.M, Tissot A.M., Hillard P.A., Rosenthal S.L., "Psychological, behavioral and interpersonal impact of human Papillomavirus and Pap test results", J Womens Health (Larchmt), vol. 14 (7), pp650-659, 2005
[174] Maissi E., Marteau T.M., Hankins M., Moss S., Legood R., Gray A., "Psychological impact of human papillomavirus testing in women with borderline or mildly dyskaryotic cervical smear test results: cross sectional questionnaire study", BMJ, vol. 328(7451), 2004
[175] McCaffery K., Irwig L., "Australian women's needs and preferences for information about human Papillomavirus in cervical screening", J Med Screen, vol. 12(3), pp132-141, 2005
[176] Schwartz M., Savage W., George J., Emohare L., "Women's knowledge and experience of cervical screening: a failure of health education and medical organization", Community Med, vol. 11(4), pp279-289, 1989

[177] Monsonego J., Cortes J., da Silva D.P., Jorge A.F., Klein P., "Psychological impact, support and information needs for women with an abnormal Pap smear: comparative results of a questionnaire in three European countries". BMC Womens Health, vol. 11(18), 2011
[178] Tota J.E., Ramana–Kumar A.V., El-Khatib Z., Franco E.L., "The road ahead for cervical cancer prevention and control", Current Oncology, vol. 21(2), ppe255-264, 2014
[179] Cuzick J., Bergeron C., von Knebel Doeberitz M., Gravitt P, Jeronimo J., Lorincz A.T., Meijer C., Sankaranarayanan R., Snijders P., Szarewski A. "New technologies and procedures for cervical cancer screening", Vaccine, vol. 30 (5), ppF107-116, 2012
[180] Cuzick J., Arbyn M., Sankaranarayanan R., Tsu V., Ronco G., Mayrand M.H., Dillner J., Meijer C.J., "Overview of human papillomavirus-based and other novel options for cervical cancer screening in developed and developing countries.", Vaccine, vol. 26(10), ppK29-41, 2008
[181] Richardson L.A., Tota J., Franco E.L., "Optimizing technology for cervical cancer screening in highresource settings", Expert Rev Obstet Gynecol., vol. 6(3), pp343-353, 2011
[182] Leyden W.A., Manos M.M., Geiger A.M., Weinmann S., Mouchawar J., Bischoff K., Yood M.U., Gilbert J., Taplin S.H., "Cervical Cancer in Women with Comprehensive Health Care Access: Attributable Factors in the Screening Process", Natl Cancer Inst, vol. 97(9), pp675-683, 2005
[183] Gomez-Roman J.J., Echevarria C., Salas S., Gonzalez-Moran M.A., Perez Mies B., Garcia-Higuera I., Nicolas Martinez M., Val-Bernal J.F., "A Type-Specific Study of Human Papillomavirus Prevalence in Cervicovaginal Samples in Three Different Spanish Regions", APMIS, vol. 117(1), pp22-27, 2009
[184] Mayrand M.H., Duarte-Franco E., Rodrigues I., Walter S.D., Hanley J., Ferenczy A., Ratnam S., Coutlée F., Franco E.L., Canadian Cervical Cancer Screening Trial Study Group, "Human Papillomavirus DNA Versus Papanicolaou Screening Tests for Cervical Cancer", N Engl J Med, vol. 357(16), pp1579-1588, 2007
[185] Cuzick J., Arbyn M., Sankaranarayanan R., Tsu V., Ronco G., Mayrand M.H., Dillner J., Meijer C.J., "Overview of Human Papillomavirus-Based and Other Novel Options for Cervical Cancer Screening in Developed and Developing Countries", Vaccine, vol.26, Suppl 10, ppK29-41, 2008
[186] Naucler P., Ryd W., Tornberg S., Strand A., Wadell G., Elfgren K., Radberg T., Strander B., Forslund O., Hansson B.G., Hagmar B., Johansson B., Rylander E., Dillner, J., "Efficacy of Hpv DNA Testing with Cytology Triage and/or Repeat Hpv DNA Testing in Primary Cervical Cancer Screening", J Natl Cancer Inst, vol. 101(2), pp88-99, 2009
[187] Mitchell T. M., Machine Learning, McGraw-Hill, 1997
[188] Youden W.J., "Index for rating diagnostic tests", Cancer, vol. 3, pp32-35, 1950
[189] Bountris P., Kotronoulas G., Tagaris T., Haritou M., Spathis A., Karakitsos P., Koutsouris D., "Cxcadss: A Web-Based Clinical Decision Support System for Cervical Cancer", 6th European Conference of the International Federation for Medical and Biological Engineering, Springer International Publishing, 2015
[190] Bountris P., Haritou M., Pouliakis A., Margari N., Kyrgiou M., Spathis A., Pappas A., Panayiotides I., Paraskevaidis E.A., Karakitsos P., and Koutsouris D.D., "An Intelligent Clinical Decision Support System for Patient-Specific Predictions to Improve Cervical Intraepithelial Neoplasia Detection", Biomed Res Int, pp341-483, 2014