ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΕΡΓΑΣΤΗΡΙΟ ΜΙΚΡΟΚΥΜΑΤΩΝ ΚΑΙ ΟΠΤΙΚΩΝ ΙΝΩΝ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
ΤΜΗΜΑ ΙΑΤΡΙΚΗΣ

**Σχεδιασμός ανάπτυξη και εφαρμογή συστήματος
υποστήριξης της διάγνωσης επιχρισμάτων θυρεοειδούς
δεδομένων βιοψίας με λεπτή βελόνη FNA με χρήση
εξελιγμένων μεθόδων εξόρυξης δεδομένων**

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΣΤΗ
ΒΙΟΪΑΤΡΙΚΗ ΤΕΧΝΟΛΟΓΙΑ

ΕΜΜΑΝΟΥΗΛ Α. ΖΟΥΛΙΑΣ

ΑΘΗΝΑ, Μάρτιος 2012

Σχεδιασμός, ανάπτυξη και εφαρμογή συστήματος υποστήριξης της διάγνωσης επιχρισμάτων θυρεοειδούς δεδομένων βιοψίας με λεπτή βελόνη –FNA με χρήση εξελιγμένων μεθόδων εξόρυξης δεδομένων

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

EMMANOYHΛ A. ZOYΛΙΑΣ

Συμβουλευτική Επιτροπή : Γεώργιος Ματσόπουλος, Επίκουρος Καθηγητής Ε.Μ.Π.
Νικόλαος Ουζούνογλου, Καθηγητής Ε.Μ.Π.
Σοφία Τσελένη – Μπαλαφούτα, Καθηγήτρια Ιατρικής
Σχολής Πανεπιστημίου Αθηνών

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την **8η Μαρτίου 2012**

.....
Γεώργιος Ματσόπουλος
Επ. Καθηγητής, ΣΗΜ&ΜΥ, ΕΜΠ

.....
Σοφία Τσελένη – Μπαλαφούτα
Καθηγήτρια Ιατρικής Σχολής
Πανεπιστημίου Αθηνών

.....
Νικόλαος Ουζούνογλου
Καθηγητής, ΣΗΜ&ΜΥ, ΕΜΠ

.....
Δημήτριος – Διονύσιος Κουτσούρης
Καθηγητής, ΣΗΜ&ΜΥ, ΕΜΠ

.....
Ανδρέας – Γεώργιος Σταφυλοπάτης
Καθηγητής, ΣΗΜ&ΜΥ, ΕΜΠ

.....
Κωνσταντίνος Σιέττος
Επίκουρος Καθηγητής ΣΕΜΦΕ, ΕΜΠ

.....
Ερρίκος Βεντούρας
Καθηγητής, ΤΙΟ, ΤΕΙ Αθήνας

Αθήνα, Μάρτιος 2012

.....
ΕΜΜΑΝΟΥΗΛ ΖΟΥΛΙΑΣ

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

Copyright © Εμμανουήλ Ζούλιας, 2012

Με επιφύλαξη παντός δικαιώματος. All rights reserved.

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

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

Περίληψη

Σκοπός της παρούσας διδακτορικής διατριβής είναι η ανάπτυξη ενός ολοκληρωμένου συστήματος υποστήριξης της διάγνωσης (Decision Support System - DSS) με χρήση μεθόδων εξόρυξης δεδομένων για την ταξινόμηση επιχρισμάτων βιοψίας με λεπτή βελόνα (Fine Needle Aspiration - FNA). Δύο κατηγορίες επιλέχθηκαν για τα δείγματα FNA: καλοήθεια και κακοήθεια. Το σύστημα αυτό αποτελείται από τις ακόλουθες βαθμίδες: 1) συλλογής δεδομένων, 2) επιλογής δεδομένων, 3) εύρεσης κατάλληλων χαρακτηριστικών, 4) εφαρμογής ταξινόμησης με χρήση μεθόδων εξόρυξης δεδομένων. Επίσης, βασικός στόχος της παρούσας διδακτορικής διατριβής ήταν η βελτίωση της ορθής ταξινόμησης των ύποπτων επιχρισμάτων (suspicious), για τα οποία είναι γνωστή η αδυναμία της μεθόδου FNA να τα ταξινομήσει. Το σύστημα εκπαιδεύτηκε και ελέγχθηκε σε σχέση με το δείγμα για το οποίο είχαμε ιστολογικές επιβεβαιώσεις (ground truth). Για περιπτώσεις οι οποίες χαρακτηρίστηκαν ως μη κακοήθεις από την FNA, και για τις οποίες δεν είχαμε ιστολογικές επιβεβαιώσεις, το δείγμα προέκυψε από την συνεκτίμηση και άλλων κλινικών, εργαστηριακών και απεικονιστικών εξετάσεων.

Στα πλαίσια της παρούσας διδακτορικής διατριβής συλλέχθηκαν εξετάσεις FNA θυρεοειδούς από το Εργαστήριο Παθολογοανατομίας του Α' Τμήματος Παθολογίας της Ιατρικής Σχολής του Πανεπιστημίου Αθηνών. Δεδομένου ότι το εν λόγω εργαστήριο λειτουργεί και σαν κέντρο αναφοράς, σημαντικός αριθμός των δειγμάτων εστάλησαν εκεί και από άλλα Εργαστήρια Παθολογοανατομίας για επανέλεγχο. Το αρχειακό υλικό ήταν πολύ καλά ταξινομημένο σε χρονολογική σειρά αλλά ήταν σε έντυπη μορφή. Αρχικά πραγματοποιήθηκε η ανάλυση απαιτήσεων για τη δομή και το σχεδιασμό της βάσης δεδομένων. Με βάση τα στοιχεία από την τεκμηριωμένη διάγνωση σχεδιάστηκε και αναπτύχθηκε προηγμένο σύστημα για την κωδικοποίηση και αρχικοποίηση των δεδομένων. Με τη βοήθεια του σχεδιασμού και ανάλυσης απαιτήσεων αναπτύχθηκε και υλοποιήθηκε η βάση δεδομένων στην οποία αποθηκεύτηκαν τα δεδομένα προς επεξεργασία. Παράλληλα, με το σχεδιασμό της βάσης έγινε και η προεργασία για το σχεδιασμό και την ανάλυση απαιτήσεων του γραφικού περιβάλλοντος εισαγωγής στοιχείων. Λαμβάνοντας υπόψη ότι το σύστημα θα μπορούσε να χρησιμοποιηθεί και πέρα από τα πλαίσια της παρούσας διδακτορικής διατριβής λήφθηκε μέριμνα ώστε να παρέχεται ένα φιλικό και ευέλικτο προς το χρήστη περιβάλλον.

Σύμφωνα με τη μεθοδολογία προσέγγισης η οποία ακολουθήθηκε προηγήθηκε στατιστική ανάλυση των 9.102 συλλεχθέντων δειγμάτων FNA ως προς τα κυτταρολογικά χαρακτηριστικά τους και τις διαγνώσεις. Οι κυτταρολογικές διαγνώσεις των συγκεκριμένων δειγμάτων συσχετίστηκαν με τις ιστολογικές διαγνώσεις, στοχεύοντας στον υπολογισμό της πιθανής επίδρασης και συμβολής κάθε κυτταρολογικού χαρακτηριστικού σε μια ορθή ή ψευδή κυτταρολογική διάγνωση, έτσι ώστε να προσδιοριστούν οι πιθανές πηγές λανθασμένης διάγνωσης. Τα δείγματα τα οποία περιείχαν μόνο αίμα ή πολύ λίγα θυλακειώδη κύτταρα χωρίς κολλοειδές θεωρήθηκαν ανεπαρκή για τη διάγνωση. Οι βιοψίες εκτελέστηκαν είτε στο Α' τμήμα του Πανεπιστημίου Αθηνών (οι περισσότερες από τις περιπτώσεις με ψηλαφητούς όζους) είτε αλλού (κυρίως κάτω από την καθοδήγηση του κέντρου αναφοράς). Τα δείγματα επιστρωμένα σε πλακάκια, στάλθηκαν στο κέντρο αναφοράς από διάφορα νοσοκομεία, με διαφορετικά πρωτόκολλα σχετικά με τα κριτήρια εκτέλεσης βιοψίας FNA σε θυρεοειδή. Μετεγχειρητικές ιστολογικές επαληθεύσεις ήταν διαθέσιμες για 266 ασθενείς (κακοήθειες και μη). Το χαμηλό ποσοστό ιστολογικών επαληθεύσεων οφείλεται στην ετερογενή προέλευση των ασθενών και στην έλλειψη ολοκληρωμένης παρακολούθησης και επανελέγχου των ασθενών. Για την αξιολόγηση των δεδομένων χρησιμοποιήθηκαν περιγραφικά στατιστικά μεγέθη όπως, μέση τιμή, τυπική απόκλιση, ποσοστά, μέγιστο και ελάχιστο. Έγιναν επίσης και χ^2 δοκιμές επιπέδου σημαντικότητας διαφόρων παραμέτρων για να ελεγχθεί η πιθανή συσχέτιση ή η ανεξαρτησία. Για τη συσχέτιση των κυτταρολογικών και των ιστολογικών διαγνώσεων και την αξιολόγηση των εργαστηριακών ευρημάτων, πέραν των περιγραφικών στατιστικών μεγεθών χρησιμοποιήθηκαν και υπολογισμοί της ευαισθησίας, της ειδικότητας, της συνολικής ακρίβειας, της αρνητικής και θετικής αξίας πρόβλεψης (negative and positive predictive value). Προκειμένου να καθοριστεί εάν μια κατηγορία ασθενειών συσχετίζεται ή όχι με συγκεκριμένες κυτταρολογικές παραμέτρους εφαρμόστηκε μέθοδος ελέγχου στατιστικής σημαντικότητας σε επίπεδο 5% ($p < 0,05$). Η διαδικασία ακολουθήθηκε για κάθε κατηγορία ασθενειών ή συνδυασμό τους και για κάθε παράμετρο των

κυτταρολογικών και αρχιτεκτονικών στοιχείων της κυτταρολογικής διάγνωσης. Τα αποτελέσματα της στατιστικής ανάλυσης επέτρεψαν το διαχωρισμό των δεδομένων σε καλοήγη, κακοήγη, νεοπλασματικά, ύποπτα για κακοήθεια και οριακά με χαρακτηριστικά γνωρίσματα μεταξύ ενός καλοήθους και ενός νεοπλασματικού.

Στην συνέχεια αναπτύχθηκε σύστημα υποστήριξης της διάγνωσης χρησιμοποιώντας εξειδικευμένες μεθόδους εξόρυξης δεδομένων. Το σύστημα αποτελείται από τέσσερις βαθμίδες. Η πρώτη βαθμίδα αυτού του συστήματος είναι το περιβάλλον **Σύλλογής Δεδομένων** στην οποία τα δεδομένα αποθηκεύονται στη βάση δεδομένων. Η **Δεύτερη Βαθμίδα** αυτού του συστήματος αφορά στην **Επιλογή Δεδομένων**. Σύμφωνα με την καταγραφή των απαιτήσεων, την εισαγωγή και τη ψηφιοποίηση των στοιχείων, δημιουργήθηκαν 111 χαρακτηριστικά για κάθε ασθενή (record). Τα περισσότερα χαρακτηριστικά είχαν τιμές δυαδικού τύπου, αποτυπώνοντας την ύπαρξη ή μη του κάθε χαρακτηριστικού, ενώ κάποιες άλλες είχαν τιμές τύπων αριθμών ή αλφαριθμητικών χαρακτήρων. Από τα 111 χαρακτηριστικά επιλέχθηκαν 60 χαρακτηριστικά τα οποία περιγράφουν τη δομή των επιχρισμάτων ενώ δημιουργήθηκαν άλλα 7 χαρακτηριστικά τα οποία αφορούσαν στην ομαδοποίηση άλλων χαρακτηριστικών.

Η **Τρίτη Βαθμίδα** του συστήματος αφορά στην εύρεση των **Κατάλληλων Χαρακτηριστικών**. Λόγω του αρχικά υψηλού αριθμού χαρακτηριστικών παραμέτρων (67 ανά περίπτωση), ήταν απαραίτητο να εξαλειφθούν οι χαρακτηριστικές παράμετροι που συσχετίζονταν γραμμικά ή δεν είχαν καμία διαγνωστική πληροφορία. Η μέθοδος επιλογής χαρακτηριστικών εφαρμόστηκε πριν από την ταξινόμηση, με γνώμονα την ανεύρεση ενός υποσυνόλου των χαρακτηριστικών παραμέτρων που βελτιστοποιούν σε ακρίβεια τη διαδικασία ταξινόμησης. Εφαρμόστηκε η τεχνική επιλέουσας πρόσθιας ακολουθιακά μεταβαλλόμενης επιλογής (SFFS). Ο αριθμός των δειγμάτων που χρησιμοποιήθηκαν είναι 2.036 (1.886 καλοήθειες και 150 κακοήθειες). Εξ αυτών, όλες οι κακοήθειες είναι ιστολογικά επιβεβαιωμένες. Επίσης, 140 καλοήθειες είναι ιστολογικά επιβεβαιωμένες με επάρκεια υλικού. Οι υπόλοιπες 1.726 καλοήθειες είναι επιβεβαιωμένες με συνεκτίμηση κλινικών, εργαστηριακών και απεικονιστικών ιατρικών εξετάσεων (υπέρηχοι κ.λπ.). Από τα 2.036 δείγματα, το 25% χρησιμοποιήθηκε για την επιλογή χαρακτηριστικών παραμέτρων, δηλαδή 37 περιπτώσεις κακοήθειας (Malignant) και 472 περιπτώσεις καλοήθειας (Non Malignant). Από την εφαρμογή της τεχνικής (SFFS) επιλέχθηκαν τελικά 12 χαρακτηριστικά ως βέλτιστα για την ταξινόμηση των δεδομένων FNA σε καλοήγη και κακοήγη.

Η **Τέταρτη βαθμίδα** επεξεργασίας είναι η **Εφαρμογής Ταξινόμησης με χρήση Μεθόδων Εξόρυξης Δεδομένων ή Ταξινομητής**. Για το σκοπό αυτό, επιλέχθηκε να εφαρμοστεί μια πληθώρα αξιόπιστων, καλά επιβεβαιωμένων και σύγχρονων μεθόδων εξόρυξης δεδομένων. Το σύστημα εκπαιδεύτηκε και ελέγχθηκε σε σχέση με το δείγμα για το οποίο είχαμε ιστολογικές επιβεβαιώσεις (ground truth). Η ανεξάρτητη εφαρμογή τεσσάρων αξιόπιστων μεθόδων, Δέντρων Αποφάσεων (Decision Trees), Τεχνητών Νευρωνικών Δικτύων (Artificial Neural Network), Μηχανών Στήριξης Διανυσμάτων (Support Vector Machine), και K - κοντινότερου γείτονα (k-NN), έδωσε αποτελέσματα συγκρίσιμα με αυτά της FNA μεθόδου. Περαιτέρω βελτίωση των αποτελεσμάτων επιτεύχθηκε με την εφαρμογή της μεθόδου πλειοψηφικού κανόνα (Majority Vote - CMV) συνδυάζοντας τα αποτελέσματα από την εφαρμογή των τριών καλύτερων αλγορίθμων, ήτοι των Νευρωνικών Δικτύων, Μηχανών Στήριξης Διανυσμάτων και K - κοντινότερου γείτονα. Η τροποποιημένη μέθοδος τεχνητών αυτοάνοσων συστημάτων (Artificial Immune Systems – AIS) χρησιμοποιήθηκε για πρώτη φορά στην ταξινόμηση και παρουσίασε ιδιαίτερα βελτιωμένα αποτελέσματα στην ταξινόμηση των επιχρισμάτων τα οποία χαρακτηρίζονται ύποπτα (suspicious) από τους ειδικούς και αποτελούν το αδύναμο σημείο της μεθόδου FNA. Αυτές οι περιπτώσεις υπόνοιας αποτελούν ένα πολύ δύσκολο κομμάτι για τη διάκριση μεταξύ των καλοηθειών και των κακοηθειών, ακόμα και για τους πλέον ειδικούς. Επειδή όλα τα περιστατικά που χαρακτηρίζονται από την βιοψία FNA ως υπόνοιες αντιμετωπίζονται κλινικά σαν κακοήθειες, η εφαρμογή των αλγοριθμικών μεθόδων βελτιώνει αισθητά τη διαχείριση αυτών των περιπτώσεων μειώνοντας τον αριθμό των άσκοπων χειρουργικών επεμβάσεων θυρεοειδεκτομών.

Λέξεις Κλειδιά

Σύστημα Υποβοήθησης Ιατρικών Αποφάσεων, Εξόρυξη Δεδομένων, Βιοψία Θυρεοειδούς δια Λεπτής Βελόνης, Τεχνητά Νευρωνικά Δίκτυα, Δέντρα Αποφάσεων, K-Πλησιέστερου Γείτονα, Ανοσοποιητικά Συστήματα, Μέθοδος Πλειοψηφικού Κανόνα, Μηχανές Διανυσμάτων Στήριξης, SFFS

Abstract

The Aim of present thesis is the development of an integrated system for supporting diagnosis (Decision Support System - DSS) using for categorizing FNA biopsy smears. Two categories were selected for the FNA smears: malignant and nonmalignant. The system is constituted by the following stages of 1) data collection, 2) data selection 3) choice of suitable clinical and cytological features, 4) application of data mining method for the categorization of FNA biopsy smears. Furthermore a fundamental objective of the doctoral thesis was the improvement of suspect smears (suspicious) categorization, for the latter FNA Biopsy has a known restriction. The system had been trained and checked in relation to the sample that histologic evaluation existed (ground truth). For smears that characterized as nonmalignant by FNA and histological data we're not available, complementary clinical, laboratory and imaging evaluations took into account in order to create the sample.

The smears that were available in this thesis, were collected from FNA biopsies in Pathologoanatomy Laboratory, A' Pathology Department, Medical School of Athens University. Given that the above referred laboratory is a reference center, an important number of FNA smears were sent to it from other laboratories for cross check. The examination files were sorted in chronological order, but there were in paper forms. The requirements for the formation and the design of database system were collected. Based on the material of the diagnosis an improved system was designed and developed for data initialization and coding. The database was developed based on the design and analysis of requirements; in this database data were stored for further investigation. Analysis of the graphical user interface design was performed in parallel to the database design. Taking into account that the system might be used after the completion of thesis, the graphical user interface was designed in order to be user friendly and flexible environment.

According to the methodological approach that was followed, the various cytological characteristic of 9102 FNA smears aspired among 2000-2004 was analyzed statistically. The cytological reports cross correlated with histological diagnoses, aiming to calculate the effect or contribution of each cytological characteristic to a false or true cytological diagnosis and to find the possible sources of erroneous diagnosis. The smears that have blood or a few follicular cells without colloid were characterized as insufficient for further diagnosis. The aspiration was performed either in A' department of Athens University (most of the cases with palpable nodules) or elsewhere (mainly under guidance of the reference center). The acquired smears being send to the reference center from various hospitals with different protocols concerning criteria to perform a thyroid FNA. Histological reports were available for 266 patients. The small number of histological verifications was due to the heterogeneity and the lack of patients files. For evaluating of data, descriptive statistic values were used like mean, standard deviation, percentage, maximum and minimum. In addition to that χ^2 tests of significance were performed in order to check possible correlation or independence. For correlating cytological and histological diagnosis and evaluating laboratory findings, apart from the descriptive statistic parameters also calculated sensitivity, specificity, total accuracy, negative predictive value and positive predictive value. Method of statistical significance in the level of 5% ($p < 0,05$) was applied in order to specify if a disease was correlated to a cytological parameter. Those checks were performed for each disease category in correlation to any cytological parameter. Statistical analysis divided the smears into nonmalignant, malignant, neoplasms, suspicious for malignancy and borderline.

A diagnosis support system was implemented using data mining methods. The system is consisted of four stages. The **First stage** of the system is the **Data Collection** environment, which stores the data to the database. The **Second stage** of this system concerns the **Selection of Data**. User requirements concluded that 111 characteristics are needed to describe each patient (record). Most of them have binary values, presenting existence and not existence, other have alphanumeric and number values. Among them 60 were selected and 7 more are produced from grouping other characteristics. The final analysis reveals that 67 characteristics of the smears are capable for describing the structure of smears in general.

The **Third stage** of system concerns the **Selection of Best Characteristics**. Due to the high number of attributes (67 per case), it was essential to eliminate the characteristics that are connected linearly or do not bring diagnostics information. The choice of characteristics applied before the classification, having the aim of discovering a subset of characteristics that optimizes the process of classification. The technique of Sequential Float Forward Search (SFFS) was applied. The number of patients that used was 2,036 (1886 non malignancies and 150 malignancies). Among them all malignancies were histologically confirmed. In addition to that 140 no malignancies were histologically confirmed in correlation to evaluation of clinics, laboratorial and medical image actions (ultrasounds etc.). Among 2.036 smears the 25% used for characteristics selection, 37 smears of Malignant and smears of Non Malignant. The Sequential Float Forward Search (SFFS) Technique, choose the best 12 elements that they reveal high performance to FNA data categorization.

The **Fourth stage** is the **Application of Classification using Data Mining Methods** or in other words data mining method. For this aim a set of reliable, well confirmed but also modern methods applied. In addition to that the system was trained and was checked using the sample with histological verifications (ground truth). The independent application of four reliable methods, Decision Trees, Artificial Neural Network, Support Vector Machine, and k-NN, resulting to comparable outcomes concerning those of FNA. However, further improvement was achieved with the application of Majority (Majority Vote - CMV) using of previous results of three algorithms Artificial Neural Network, Support Vector Machine, and k-NN. The modified Artificial Immune System (AIS) was applied for first time. AIS presents particularly improved results for the categorization of smears, which are characterised “suspicious” by the experts and is a known weakness of FNA method. These cases constitute a very difficult part for the discrimination among non-malignant and malignant, even for a specialist. Since all these cases are faced clinically using FNA as malignancies, the application of an improved algorithmic method improves accordingly the management of these cases by decreasing the number of useless surgical thyroid operations.

Keywords

Medical Decision Support System, Data Mining, FNA Biopsy, Neural Networks, Decision Trees, k-Nearest Neighborhood, Immune Systems, Majority Vote, Support Vector Systems, Feature Selection.

Ευχαριστίες

Η παρούσα Διδακτορική Διατριβή εκπονήθηκε στη Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου, στο Εργαστήριο Μικροκυμάτων και Οπτικών Ινών στο πλαίσιο του Διατμηματικού Προγράμματος Μεταπτυχιακών Σπουδών στη Βιοϊατρική Τεχνολογία. Η εργασία ήταν η αφορμή να ασχοληθώ εντατικότερα με εφαρμογές της πληροφορικής στον τομέα της υγείας και πιο συγκεκριμένα στο ιδιαίτερα ενδιαφέρον πεδίο της εξόρυξης γνώσης από ιατρικά δεδομένα (Data Mining), κάνοντας χρήση πολλών και διαφορετικών τεχνολογιών όπως προγραμματισμός, βάσεις δεδομένων, εξόρυξη γνώσης, αλγοριθμικές μέθοδοι, στατιστική, καθώς και με ιατρικά θέματα.

Στο σημείο αυτό θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα Επίκουρο Καθηγητή Ε.Μ.Π. κ. Γεώργιο Ματσόπουλο ο οποίος ήταν η αφορμή και η αιτία να ασχοληθώ με την εκπόνηση αυτής της διδακτορικής διατριβής, καθώς επίσης και βοηθός, συμπαραστάτης και οδηγός σε όλη αυτή την προσπάθεια και σε κάθε δυσκολία της. Τον ευχαριστώ πολύ για τον πολύτιμο χρόνο τον οποίο αφιέρωσε αγόγγυστα για το σκοπό αυτό. Επίσης, ευχαριστώ τον καθηγητή Ε.Μ.Π. κ. Νικόλαο Ουζούνoglou για την ευκαιρία που μου προσέφερε να ασχοληθώ με ένα τέτοιο θέμα αλλά και για την πολυτιμότερη βοήθεια, εποπτεία και υποστήριξη του καθόλη τη διάρκεια εκπόνησης της διατριβής. Επιπλέον, θα ήθελα να ευχαριστήσω την Καθηγήτρια της Ιατρικής Σχολής Αθηνών κ. Σοφία Τσελένη - Μπαλαφούτα για την αδιάκοπη και με πολύ υπομονή υποστήριξη της. Ήταν για μένα πραγματική εκπαιδευτρια σε ένα ιατρικό κομμάτι το οποίο δεν γνώριζα. Θέλω όμως να την ευχαριστήσω και για την συμπαράσταση της σε προσωπικό επίπεδο καθώς στάθηκε για μένα πολυτιμότερη φίλη σε δύσκολες στιγμές. Ακόμα, θα ήθελα να ευχαριστήσω θερμά τον καθηγητή εφαρμογών της σχολής Τ.Ι.Ο. του Τ.Ε.Ι. Αθηνών Δρ. Παντελή Ασβεστά, φίλο και συμπαραστάτη, για τις πολύτιμες συμβουλές του και καθοδήγηση του σε δύσκολα σημεία. Ευχαριστίες οφείλω στην εξαιρετική Λέκτορα της Ιατρικής Σχολής Αθηνών κ. Χαρίκλεια Γακοπούλου, για την άποψη συνεργασία μας και για το χρόνο που αφιέρωσε ώστε να προχωρήσει ένα σημαντικό τμήμα της έρευνας. Ακόμα ευχαριστώ τον Γεώργιο Δούκα για τη βοήθεια που μου παρείχε στα αρχικά στάδια της διατριβής, καθώς και το φίλο και συνάδελφο Δρ. Παναγιώτη Κρασόπουλο για τις πολύτιμες διορθώσεις του στο κείμενο της διατριβής.

Τέλος, ευχαριστώ ιδιαίτερα τη σύζυγο μου για την υπομονή της, τη συμπαράσταση και την υποστήριξη που μου παρείχε όλο αυτό το διάστημα καθώς και την κόρη μας Ζωή της οποίας η παρουσία στη ζωή μου υπήρξε πηγή χαράς και άντλησης δύναμης. Να ευχαριστήσω τους γονείς και τους φίλους μου για την κατανόηση και την αγάπη τους. Τέλος, θέλω να ευχαριστήσω θερμά την κ. Κατερίνα Παπαδοπούλου η οποία μου στάθηκε όλη αυτή την περίοδο σε επίπεδο ψυχολογικής υποστήριξης τόσο για την εκπόνηση της διδακτορικής μου διατριβής αλλά και για όλο το φάσμα των προσωπικών θεμάτων που με απασχόλησαν.

Στη σύζυγο μου Αλεξάνδρα

ΚΕΦΑΛΑΙΟ 1	14
1 ΕΙΣΑΓΩΓΗ.....	14
1.1 ΣΚΟΠΟΣ ΔΙΔΑΚΤΟΡΙΚΗΣ ΔΙΑΤΡΙΒΗΣ	14
1.2 ΟΡΓΑΝΩΣΗ ΤΗΣ ΔΙΔΑΚΤΟΡΙΚΗΣ ΔΙΑΤΡΙΒΗΣ	15
ΚΕΦΑΛΑΙΟ 2	17
2 ΙΑΤΡΙΚΑ ΣΤΟΙΧΕΙΑ.....	17
2.1 ΤΙ ΕΙΝΑΙ Ο ΘΥΡΕΟΕΙΔΗΣ.....	17
2.2 ΙΣΤΟΛΟΓΙΑ ΤΟΥ ΘΥΡΕΟΕΙΔΟΥΣ.....	18
2.3 ΕΛΕΓΧΟΣ – ΕΞΕΤΑΣΕΙΣ – ΝΟΣΟΙ ΘΥΡΕΟΕΙΔΟΥΣ.....	19
2.4 Η ΤΕΧΝΙΚΗ ΤΗΣ ΒΙΟΨΙΑΣ ΔΙΑ ΛΕΠΤΗΣ ΒΕΛΟΝΗΣ (FINE NEEDLE ASPIRATION – FNA) - ΑΝΤΙΠΡΟΣΩΠΕΥΤΙΚΟΤΗΤΑ ΤΟΥ ΥΛΙΚΟΥ.....	20
2.5 ΠΑΡΑΜΕΤΡΟΙ ΑΞΙΟΛΟΓΗΣΗΣ ΕΠΙΧΡΙΣΜΑΤΩΝ ΒΙΟΨΙΑΣ FNA.....	21
2.6 ΔΙΑΓΝΩΣΕΙΣ FNA.....	22
2.7 ΑΞΙΟΠΙΣΤΙΑ ΔΙΑΓΝΩΣΕΩΝ.....	22
2.8 ΠΕΡΙΟΡΙΣΜΟΙ ΜΕΘΟΔΟΥ FNA.....	24
2.9 ΒΙΒΛΙΟΓΡΑΦΙΑ.....	24
ΚΕΦΑΛΑΙΟ 3	26
3 ΙΑΤΡΙΚΑ ΣΥΣΤΗΜΑΤΑ ΑΠΟΦΑΣΕΩΝ ΚΑΙ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ.....	26
3.1 ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ.....	26
3.2 ΒΑΣΙΚΕΣ ΚΑΤΗΓΟΡΙΕΣ ΜΕΘΟΔΩΝ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ.....	27
3.3 ΚΥΚΛΟΣ ΖΩΗΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ.....	28
3.4 ΠΡΟΤΥΠΑ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ.....	30
ΔΕΔΟΜΕΝΩΝ OLE DB:.....	30
ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ API ΤΗΣ JAVA:.....	30
CRISP-DM:.....	31
3.5 ΝΕΕΣ ΤΑΣΕΙΣ ΣΤΗΝ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ.....	31
3.6 ΕΦΑΡΜΟΓΗ ΤΗΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ ΣΤΗΝ ΙΑΤΡΙΚΗ.....	31
3.7 ΒΙΒΛΙΟΓΡΑΦΙΑ.....	32
ΚΕΦΑΛΑΙΟ 4	34
4 ΗΛΕΚΤΡΟΝΙΚΟ ΣΥΣΤΗΜΑ ΑΡΧΕΙΟΘΕΤΗΣΗΣ ΔΕΔΟΜΕΝΩΝ ΒΙΟΨΙΑΣ FNA ΘΥΡΕΟΕΙΔΟΥΣ.....	34
4.1 ΕΙΣΑΓΩΓΗ.....	34
4.1.1 ΣΧΕΔΙΑΣΜΟΣ ΚΑΙ ΑΝΑΛΥΣΗ ΑΠΑΙΤΗΣΕΩΝ ΒΑΣΗΣ ΔΕΔΟΜΕΝΩΝ FNA.....	34
4.1.2 ΑΝΑΠΤΥΞΗ – ΥΛΟΠΟΙΗΣΗ ΒΑΣΗΣ ΔΕΔΟΜΕΝΩΝ.....	35
4.1.3 ΣΧΕΔΙΑΣΜΟΣ ΚΑΙ ΑΝΑΛΥΣΗ ΑΠΑΙΤΗΣΕΩΝ ΓΡΑΦΙΚΟΥ ΠΕΡΙΒΑΛΛΟΝΤΟΣ ΕΙΣΑΓΩΓΗΣ ΣΤΟΙΧΕΙΩΝ.....	36
4.1.4 ΑΝΑΠΤΥΞΗ – ΥΛΟΠΟΙΗΣΗ ΓΡΑΦΙΚΟΥ ΠΕΡΙΒΑΛΛΟΝΤΟΣ ΕΙΣΑΓΩΓΗΣ ΔΕΔΟΜΕΝΩΝ.....	36
4.2 ΔΙΑΘΕΣΙΜΑ ΣΤΟΙΧΕΙΑ.....	39
4.3 ΒΙΒΛΙΟΓΡΑΦΙΑ.....	39
ΚΕΦΑΛΑΙΟ 5	41
5 ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΒΙΟΨΙΑΣ FNA	41
5.1 ΕΙΣΑΓΩΓΗ.....	41
5.2 ΥΛΙΚΟ ΜΕΛΕΤΗΣ	41
5.3 ΜΕΘΟΔΟΛΟΓΙΑ.....	41
5.4 ΣΤΑΤΙΣΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ.....	42
5.4.1 ΣΤΑΤΙΣΤΙΚΑ ΣΤΟΙΧΕΙΑ ΣΤΟ ΣΥΝΟΛΟ ΤΟΥ ΠΛΗΘΥΣΜΟΥ.....	42
5.5 ΣΤΑΤΙΣΤΙΚΗ ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ.....	46
5.5.1 ΣΤΑΤΙΣΤΙΚΑ ΣΗΜΑΝΤΙΚΕΣ ΣΥΓΚΡΙΣΕΙΣ ΚΥΤΤΑΡΟΛΟΓΙΚΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΜΕΤΑΞΥ ΟΡΙΑΚΩΝ ΚΑΙ ΝΕΟΠΛΑΣΜΑΤΩΝ.....	46
5.5.2 ΠΕΡΙΣΤΑΤΙΚΑ ΤΑ ΟΠΟΙΑ ΕΧΟΥΝ ΧΑΡΑΚΤΗΡΙΣΤΕΙ ΩΣ ΥΠΟΝΟΙΑ ΚΑΚΟΗΘΕΙΑΣ	46
5.5.3 ΠΕΡΙΣΤΑΤΙΚΑ ΘΕΤΙΚΑ ΣΕ ΚΑΚΟΗΘΕΙΑ	46

5.6 ΣΥΣΧΕΤΙΣΗ ΚΥΤΤΑΡΟΛΟΓΙΚΩΝ ΚΑΙ ΙΣΤΟΛΟΓΙΚΩΝ ΔΙΑΓΝΩΣΕΩΝ.....	47
5.7 ΑΝΑΛΥΣΗ ΤΩΝ ΣΤΑΤΙΣΤΙΚΩΝ ΣΤΟΙΧΕΙΩΝ	52
5.8 ΒΙΒΛΙΟΓΡΑΦΙΑ	53
ΚΕΦΑΛΑΙΟ 6	55
6 ΣΥΣΤΗΜΑΤΑ ΥΠΟΒΟΗΘΗΣΗΣ ΔΙΑΓΝΩΣΗΣ ΒΙΟΨΙΩΝ ΘΥΡΕΟΕΙΔΟΥΣ ΜΕ ΧΡΗΣΗ ΜΕΘΟΔΩΝ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ	55
6.1 ΔΕΥΤΕΡΗ ΒΑΘΜΙΔΑ - ΕΠΙΛΟΓΗ ΔΕΔΟΜΕΝΩΝ	55
6.1.1 ΠΑΡΟΥΣΙΑΣΗ ΔΕΔΟΜΕΝΩΝ	56
6.1.2 ΚΑΘΑΡΙΣΜΟΣ ΔΕΔΟΜΕΝΩΝ.....	56
6.2 ΤΡΙΤΗ ΒΑΘΜΙΔΑ – ΕΥΡΕΣΗ ΚΑΤΑΛΛΗΛΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ (FEATURE SELECTION).....	59
6.2.1 ΜΕΘΟΔΟΣ ΕΠΙΛΟΓΗΣ ΚΑΤΑΛΛΗΛΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΠΑΡΑΜΕΤΡΩΝ	60
6.2.1.1 ΜΕΣΟΣ ΟΡΟΣ ΚΑΙ ΔΙΑΣΠΟΡΑ	60
6.2.1.2 ΜΕΤΡΟ ΑΠΟΣΤΑΣΗΣ	60
6.2.1.3 ΧΡΗΣΗ ΚΥΡΙΩΝ ΣΤΟΙΧΕΙΩΝ.....	61
6.2.2 ΕΠΙΠΛΕΟΥΣΑ ΠΡΟΣΘΙΑ ΑΚΟΛΟΥΘΙΑΚΑ ΜΕΤΑΒΑΛΛΟΜΕΝΗ ΕΠΙΛΟΓΗ (SEQUENTIAL FLOAT FORWARD SEARCH TECHNIQUE - SFFS).....	61
6.2.3 ΕΥΡΕΣΗ ΚΑΤΑΛΛΗΛΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ	63
6.2.4 ΠΑΡΟΥΣΙΑΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ.....	63
6.3 ΤΕΤΑΡΤΗ ΒΑΘΜΙΔΑ - ΣΥΣΤΗΜΑ ΑΠΟΦΑΣΕΩΝ ΚΑΙ ΕΦΑΡΜΟΖΟΜΕΝΕΣ ΤΕΧΝΙΚΕΣ	63
6.4 ΔΕΝΤΡΑ ΑΠΟΦΑΣΕΩΝ - ΘΕΩΡΗΤΙΚΗ ΠΕΡΙΓΡΑΦΗ	63
6.4.1 ΑΛΓΟΡΙΘΜΙΚΟ ΠΛΑΙΣΙΟ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΕΩΝ.....	64
6.4.2 ΠΑΡΟΥΣΙΑΣΗ ΠΕΙΡΑΜΑΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ.....	67
6.4.3 ΑΠΟΤΕΛΕΣΜΑΤΑ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ.....	67
6.5 ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ - ΘΕΩΡΗΤΙΚΗ ΠΕΡΙΓΡΑΦΗ	69
6.5.1 ΑΡΧΙΤΕΚΤΟΝΙΚΗ - ΜΟΝΤΕΛΟ ΝΕΥΡΩΝΑ	70
6.5.2 ΠΡΟΣΘΙΑΣ ΤΡΟΦΟΔΟΣΙΑΣ (FEEDFORWARD) ΔΙΚΤΥΑ.....	70
6.5.2.1 ΕΚΠΑΙΔΕΥΣΗ ΔΙΚΤΥΟΥ	71
6.5.2.2 ΑΛΓΟΡΙΘΜΟΣ ΟΠΙΣΘΟΔΙΑΔΟΣΗΣ (BACKPROPAGATION)	71
6.5.3 ΠΑΡΟΥΣΙΑΣΗ ΠΕΙΡΑΜΑΤΩΝ ΝΕΥΡΩΝΙΚΩΝ ΔΙΚΤΥΩΝ.....	72
6.5.4 ΑΠΟΤΕΛΕΣΜΑΤΑ ΝΕΥΡΩΝΙΚΩΝ ΔΙΚΤΥΩΝ.....	73
6.6 ΜΗΧΑΝΕΣ ΣΤΗΡΙΞΗΣ ΔΙΑΝΥΣΜΑΤΩΝ (SUPPORT VECTOR MACHINES – SVM’S)	75
6.6.1 ΘΕΩΡΗΤΙΚΗ ΠΕΡΙΓΡΑΦΗ	75
6.6.2 ΠΑΡΟΥΣΙΑΣΗ ΠΕΙΡΑΜΑΤΩΝ ΜΣΔ.....	77
6.6.3 ΑΠΟΤΕΛΕΣΜΑΤΑ ΜΣΔ.....	78
6.7 Αλγόριθμος k- ΚΟΝΤΙΝΟΤΕΡΟΥ ΓΕΙΤΟΝΑ (k NEAREST NEIGHBOR ALGORITHM)	79
6.7.1 ΠΑΡΟΥΣΙΑΣΗ ΠΕΙΡΑΜΑΤΩΝ k-ΚΟΝΤΙΝΟΤΕΡΟΥ ΓΕΙΤΟΝΑ	82
6.7.2 ΑΠΟΤΕΛΕΣΜΑΤΑ k-ΚΟΝΤΙΝΟΤΕΡΟΥ ΓΕΙΤΟΝΑ.....	82
6.8 ΑΛΓΟΡΙΘΜΟΣ ΠΛΕΙΟΨΗΦΙΚΟΥ ΚΑΝΟΝΑ.....	84
6.8.1 ΠΑΡΟΥΣΙΑΣΗ ΠΕΙΡΑΜΑΤΩΝ – ΑΠΟΤΕΛΕΣΜΑΤΩΝ	85
6.8.2 ΑΠΟΤΕΛΕΣΜΑΤΑ	85
6.9 ΒΙΒΛΙΟΓΡΑΦΙΑ.....	86
ΚΕΦΑΛΑΙΟ 7	90
7 ΤΕΧΝΗΤΑ ΑΝΟΣΟΠΟΙΗΤΙΚΑ ΣΥΣΤΗΜΑΤΑ (ARTIFICIAL IMMUNE SYSTEMS)	90
7.1 ΒΑΣΙΚΕΣ ΜΕΘΟΔΟΙ ΤΕΧΝΗΤΩΝ ΑΝΟΣΟΠΟΙΗΤΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ (ARTIFICIAL IMMUNE SYSTEMS - AIS)	90
7.1.1 ΑΡΝΗΤΙΚΗΣ ΕΠΙΛΟΓΗΣ	90
7.1.2 ΕΠΙΛΟΓΗΣ ΚΛΩΝΟΥ.....	91
7.1.3 ΑΝΟΣΟΠΟΙΗΤΙΚΑ ΔΙΚΤΥΑ.....	92
7.1.4 ΠΡΟΤΕΙΝΟΜΕΝΗ ΤΕΧΝΙΚΗ.....	93
7.2 ΤΑΞΙΝΟΜΗΣΗ ΜΕ ΤΗΝ ΧΡΗΣΗ ΤΕΧΝΗΤΩΝ ΑΝΟΣΟΠΟΙΗΤΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ (ARTIFICIAL IMMUNE SYSTEMS - AIS).....	94
7.3 ΠΕΡΙΓΡΑΦΗ ΠΡΟΤΕΙΝΟΜΕΝΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΑΞΙΝΟΜΗΣΗΣ AIS	95

7.3.1	ΑΡΧΙΚΟΠΟΙΗΣΗ ΠΛΗΘΥΣΜΟΥ BOXCELLS	95
7.3.2	ΚΛΩΝΟΠΟΙΗΣΗ ΤΩΝ BOXCELLS	96
7.3.3	ΔΙΑΣΤΟΛΗ - ΣΥΡΡΙΚΝΩΣΗ ΤΩΝ BOXCELLS.....	96
7.3.4	ΑΦΑΙΡΕΣΗ BOXCELL	97
7.3.5	ΣΥΓΧΩΝΕΥΣΗ BOXCELLS	97
7.3.6	ΠΡΟΣΘΗΚΗ ΝΕΟΥ BOXCELL	97
7.3.7	ΔΙΑΔΙΚΑΣΙΑ ΤΑΞΙΝΟΜΗΣΗΣ ΒΑΣΙΣΜΕΝΗ ΣΤΟ AIS.....	98
7.3.8	ΠΑΡΟΥΣΙΑΣΗ ΠΕΙΡΑΜΑΤΩΝ ΤΕΧΝΗΤΩΝ ΑΝΟΣΟΠΟΙΗΤΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ (ARTIFICIAL IMMUNE SYSTEMS)	99
7.3.9	ΑΠΟΤΕΛΕΣΜΑΤΑ	99
7.3.10	ΣΥΝΟΛΙΚΗ ΠΑΡΟΥΣΙΑΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΣΥΓΚΡΙΣΗ ΜΕ ΤΗΝ FNA ΚΑΙ ΜΕΤΑΞΥ ΑΛΓΟΡΙΘΜΙΚΩΝ ΜΕΘΟΔΩΝ	101
7.4	ΒΙΒΛΙΟΓΡΑΦΙΑ	104
ΚΕΦΑΛΑΙΟ 8		106
8 ΣΤΟΧΟΙ ΠΟΥ ΕΠΕΤΕΥΧΘΗΣΑΝ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΕΠΕΚΤΑΣΕΙΣ.....		106
8.1	ΣΤΟΧΟΙ ΠΟΥ ΕΠΕΤΕΥΧΘΗΣΑΝ	106
8.2	ΜΕΛΛΟΝΤΙΚΕΣ ΕΠΕΚΤΑΣΕΙΣ	106
ΥΠΟΤΙΤΛΟΙ ΕΙΚΟΝΩΝ.....		109
ΥΠΕΡΤΙΤΛΟΙ ΠΙΝΑΚΩΝ.....		110
ΠΑΡΑΡΤΗΜΑ Α.....		112
ΠΑΡΑΡΤΗΜΑ Β.....		119

1 ΕΙΣΑΓΩΓΗ

Τις τελευταίες δεκαετίες η χρήση πληροφοριακών συστημάτων έχει εξαπλωθεί σε όλους τους εργασιακούς χώρους και τομείς. Αποτελούν το κύριο εργαλείο στην καθημερινή εργασία για την επίλυση προβλημάτων, την εξαγωγή συμπερασμάτων και τη λήψη αποφάσεων. Όλα αυτά τα χρόνια έχουν συσσωρευτεί μεγάλοι όγκοι ηλεκτρονικών δεδομένων οι οποίοι μέχρι πριν λίγα χρόνια παρέμεναν αναξιοποίητοι. Η αξιοποίηση αυτών των δεδομένων και η κρυφή πληροφορία που αυτά περιέχουν αποτελεί την νέα πρόκληση για όσους απασχολούνται στο χώρο της πληροφορικής. Η ανάπτυξη Συστημάτων Υποστήριξης Αποφάσεων (Decision Support Systems) είναι από τις πιο αξιόλογες και σημαντικές προσπάθειες στο χώρο της πληροφορικής και η χρήση τους βοηθά στην επίλυση καθημερινών προβλημάτων σε διάφορους τομείς. Παραδείγματα εφαρμογής τους υπάρχουν στον τομέα της υγείας, στον τραπεζικό τομέα, στην ασφάλεια και στις τηλεπικοινωνίες. Επίσης, αυξάνεται ολοένα και περισσότερο η αναγκαιότητα για τη χρήση εξειδικευμένων συστημάτων ώστε να επιλυθούν επιχειρησιακά προβλήματα. Ο τομέας της υγείας είναι από αυτούς με το πλέον ιδιαίτερο ενδιαφέρον εφαρμογής αυτών των τεχνικών. Η βοήθεια την οποία μπορεί να προσφέρουν τέτοια συστήματα μπορεί να συμβάλει στη λήψη αποφάσεων για τη διάγνωση των ασθενειών. Πολλά τέτοια συστήματα έχουν αναπτυχθεί και εφαρμοστεί για την υποβοήθηση της διάγνωσης σε παθήσεις καρκίνου ύπατος, σε προβλήματα ρύθμισης γλυκόζης, ταξινόμησης χρωμοσωμάτων και αλλού [Error! Reference source not found.]. Στα πλαίσια αυτής της διδακτορικής διατριβής γίνεται μια προσπάθεια να προσεγγίσουμε το πρόβλημα της διάκρισης των καλοήθων από τα κακοήθη επιχρίσματα θυρεοειδούς τα οποία έχουν ληφθεί με τη μέθοδο της βιοψίας δια λεπτής βελόνης (Fine Needle Aspiration - FNA). Από τη μελέτη εξήχθησαν σημαντικά συμπεράσματα σχετικά με την απόδοση των μεθόδων εξόρυξης γνώσης έναντι της μεθόδου διάκρισης από τον ειδικό παθολογοανατόμο.

1.1 ΣΚΟΠΟΣ ΔΙΔΑΚΤΟΡΙΚΗΣ ΔΙΑΤΡΙΒΗΣ

Σκοπός της παρούσας διδακτορικής διατριβής είναι η ανάπτυξη ενός ολοκληρωμένου συστήματος υποστήριξης της διάγνωσης (Decision Support System - DSS) με χρήση μεθόδων εξόρυξης δεδομένων για την ταξινόμηση επιχρισμάτων βιοψίας FNA ως καλοήθη ή κακοήθη. Το σύστημα αυτό αποτελείται από τις βαθμίδες συλλογής δεδομένων, επιλογής κατάλληλων δεδομένων και αξιολόγησης των στοιχείων τους και από μια ολοκληρωμένη βαθμίδα υποστήριξης απόφασης (Decision Support System - DSS) για την ταξινόμηση των επιχρισμάτων βιοψίας σε καλοήθη και κακοήθη. Επίσης βασικός στόχος είναι η βελτίωση της ορθής ταξινόμησης των ύποπτων επιχρισμάτων (suspicious), για τα οποία είναι γνωστός ο περιορισμός της μεθόδου FNA στο να τα κατηγοριοποιήσει ορθά.

Στο πλαίσιο της παρούσας διδακτορικής διατριβής γίνεται μια σημαντική προσπάθεια προκειμένου:

- α) Να συλλεχθούν και να ομαδοποιηθούν κυτταρολογικές διαγνώσεις βιοψίας FNA που διενεργήθηκαν από ειδικούς παθολογοανατόμους αριθμού άνω των 9.000 δειγμάτων.
- β) Να σχεδιαστεί και να αναπτυχθεί ηλεκτρονικό σύστημα συλλογής δεδομένων προκειμένου να αξιοποιηθεί η πληροφορία που περιέχουν τα παραπάνω δεδομένα.
- γ) Να αναπτυχθούν και να εφαρμοστούν προηγμένοι μέθοδοι εξόρυξης για την ταξινόμηση των δεδομένων σε δύο κατηγορίες: καλοήθεια και κακοήθεια.

1.2 ΟΡΓΑΝΩΣΗ ΤΗΣ ΔΙΔΑΚΤΟΡΙΚΗΣ ΔΙΑΤΡΙΒΗΣ

Η διδακτορική διατριβή αποτελείται από τα παρακάτω κεφάλαια:

Στο **Κεφάλαιο 2**, περιγράφονται αρχικά τα στοιχεία που σχετίζονται με την ανατομία, τη φυσιολογία και το ρόλο του θυρεοειδούς ως αδένων του ανθρωπίνου σώματος. Ακολουθεί μια παρουσίαση της ιστολογίας του οργάνου. Στη συνέχεια, αναφέρονται οι σχετικές εξετάσεις ελέγχου του θυρεοειδούς, μία από τις οποίες είναι και η Βιοψία δια Λεπτής Βελόνης (Fine Needle Aspiration - FNA). Ακολουθεί αναφορά στη μέθοδο Βιοψίας δια Λεπτής Βελόνης (FNA), όσον αφορά στα στοιχεία αξιολόγησης, στις πιθανές διαγνώσεις, στην αξιοπιστία της μεθόδου και στους περιορισμούς της. Τέλος παρουσιάζεται αναλυτική λίστα με τις ασθένειες που αφορούν στην παθολογία του θυρεοειδούς και τη συχνότητα εμφάνισής τους.

Στο **Κεφάλαιο 3**, γίνεται μια παρουσίαση των ιατρικών συστημάτων υποστήριξης απόφασης και εξόρυξης δεδομένων. Αναλύεται μέσα από παραδείγματα ο όρος εξόρυξης δεδομένων, οι ανάγκες που δημιούργησαν αυτό το νέο τομέα και παρουσιάζονται οι διαφορές από τις απλές αναζητήσεις σε βάσεις δεδομένων. Στη συνέχεια αναφέρονται οι βασικές κατηγορίες εξόρυξης δεδομένων, καθώς και παραδείγματα από την εφαρμογή τους. Επίσης, παρουσιάζεται ο κύκλος ζωής της εφαρμογής εξόρυξης δεδομένων. Ακολουθεί αναφορά σε ορισμένα διεθνή εμπορικά πρότυπα εξόρυξης δεδομένων όπως DB OLE for DM, το API της JAVA, CRISP-DM και PMML. Τέλος, παρουσιάζονται οι νέες τάσεις στην εξόρυξη δεδομένων και η εφαρμογή της εξόρυξης δεδομένων στην ιατρική.

Στο **Κεφάλαιο 4** αρχικά, περιγράφεται η ανάπτυξη του συστήματος αξιοποίησης δεδομένων βιοψίας FNA θυρεοειδούς. Γίνεται αναλυτική παρουσίαση του σχεδιασμού, της ανάλυσης απαιτήσεων και της ανάπτυξης - υλοποίησης της βάσης δεδομένων για την αποθήκευση των δεδομένων. Ακολούθως, παρουσιάζεται ο σχεδιασμός, η ανάλυση και η ανάπτυξη - υλοποίηση του γραφικού περιβάλλοντος εισαγωγής δεδομένων με οθόνες του συστήματος. Τέλος, γίνεται παρουσίαση των διαθέσιμων στοιχείων και των ιδιαιτεροτήτων τους.

Στο **Κεφάλαιο 5** αρχικά, παρουσιάζονται περιγραφικά στατιστικά στοιχεία για το υλικό της μελέτης και η μεθοδολογία η οποία ακολουθήθηκε. Επιπλέον, γίνεται μια ανάλυση των στατιστικά σημαντικών κυτταρολογικών χαρακτηριστικών για τις κατηγορίες των καλοηθειών – κακοηθειών και τον περιπτώσεων υπόνοιας. Επίσης, παρουσιάζεται η ανάλυση στατιστικά σημαντικής συσχέτισης κυτταρολογικών και ιστολογικών διαγνώσεων, για το τμήμα εκείνο των περιπτώσεων όπου υπήρχαν ιστολογικές επιβεβαιώσεις. Τέλος, παρουσιάζεται μια συνολική ανάλυση των στατιστικών στοιχείων.

Στο **Κεφάλαιο 6** παρουσιάζεται το σύστημα υποβοήθησης ιατρικής απόφασης με σκοπό τη διάγνωση για βιοψίες θυρεοειδούς με χρήση μεθόδων εξόρυξης δεδομένων και σχετική ανάλυση για κάθε μια βαθμίδα του. Στην πρώτη βαθμίδα συλλέγονται τα δεδομένα και αποθηκεύονται στη βάση δεδομένων. Στη δεύτερη βαθμίδα πραγματοποιείται η επιλογή δεδομένων. Στην τρίτη βαθμίδα παρουσιάζεται η επιλογή των κατάλληλων χαρακτηριστικών προκειμένου να χρησιμοποιηθούν τα βέλτιστα χαρακτηριστικά για την ταξινόμηση των δεδομένων. Στη συνέχεια παρουσιάζεται η τέταρτη βαθμίδα που αφορά στις εφαρμοζόμενες τεχνικές εξόρυξης δεδομένων οι οποίες είναι τα Δέντρα Απόφασης (Decision Trees), τα Τεχνητά Νευρωνικά Δίκτυα (Neural Networks), Μηχανές Στήριξης Διανυσμάτων (Support Vector Machines - SVM), κ-Κοντινότερου Γείτονα (k-NN), Πλειοψηφικού Κανόνα (Majority Vote). Για κάθε μια από τις μεθόδους οι οποίες εφαρμόστηκαν στα πλαίσια της παρούσας διδακτορικής διατριβής δίνεται αρχικά θεωρητική προσέγγιση της μεθόδου και στη συνέχεια παρουσιάζονται τα αποτελέσματα της εφαρμογής της και η γραφική αποτίμηση με τη μέθοδο ROC.

Στο **Κεφάλαιο 7** παρουσιάζεται μια νέα μέθοδος εξόρυξης δεδομένων βασισμένη στα Τεχνητά Ανοσοποιητικά Συστήματα (Artificial Immune Systems - AIS). Η συγκεκριμένη μέθοδος βασίζεται στη δομή και λειτουργία του βιολογικού ανοσοποιητικού συστήματος. Το βιολογικό ανοσοποιητικό σύστημα αναγνωρίζει και διακρίνει ανάμεσα στα «οικεία» και «μη οικεία» κύτταρα (αντιγόνα). Αναγνωρίζοντας

τα «μη οικεία», αντιμετωπίζει με επιτυχία κάθε παθογόνο μικροοργανισμό που εισβάλλει στο σώμα. Έτσι ο AIS δρα ως ταξινομητής αναγνωρίζοντας συνεχώς «οικεία» από τα «μη οικεία» δείγματα στο επίπεδο του χώρου των χαρακτηριστικών παραμέτρων. Έπειτα από την εφαρμογή της, στα ίδια δεδομένα στα οποία εφαρμόστηκαν και οι προηγούμενες μέθοδοι, δίνονται τα αποτελέσματα της και παρουσιάζεται η γραφική της αποτίμηση με τη μέθοδο ROC. Τέλος επιχειρείται να παρουσιαστεί μια συνολική συγκριτική μελέτη των στατιστικών αποτελεσμάτων αποδόσεως της FNA και των αποτελεσμάτων των αλγορίθμων εξόρυξης γνώσης για το πρόβλημα της κατάταξης επιχρισμάτων σαν κακοήθη ή καλοήθη και αξιολογείται η AIS ως η αποδοτικότερη.

Τέλος στο **Κεφάλαιο 8** παρουσιάζονται τα γενικά συμπεράσματα της διδακτορικής διατριβής και προτείνονται συγκεκριμένα πεδία για μελλοντικές επεκτάσεις της παρούσας διδακτορικής διατριβής τόσο σε τεχνολογικό όσο και σε ιατρικό επίπεδο.

2 ΙΑΤΡΙΚΑ ΣΤΟΙΧΕΙΑ

2.1 ΤΙ ΕΙΝΑΙ Ο ΘΥΡΕΟΕΙΔΗΣ

Ο θυρεοειδής είναι ένας αδένας του Ενδοκρινικού Συστήματος, ο οποίος παράγει δύο δραστικές ορμόνες τη θυροξίνη (T4) και την τριωδοθυρονίνη (T3) καθώς επίσης και μικρές ποσότητες ανάστροφης τριωδοθυρονίνης (rT3). Επίσης, από τα παραθυλακιοειδή κύτταρα παράγεται η καλσιτονίνη. Βρίσκεται στο κάτω τμήμα του τραχήλου, κάτω από το υοειδές οστό και μπροστά από τους πρώτους κρικοειδείς χόνδρους της τραχείας (Εικόνα 1), δηλαδή στο πρόσθιο μέρος του λαιμού, λίγο πιο κάτω από το μήλο του Αδάμ.

Εικόνα 1. Θέση Θυρεοειδούς

Ο θυρεοειδής αδένας είναι ένα συμπαγές όργανο σχήματος πεταλούδας, το οποίο αποτελείται από δύο λοβούς (Εικόνα 2), το δεξιό και τον αριστερό, το σχήμα των οποίων έχει μορφή πυραμίδας ή ωοειδούς. Οι λοβοί ενώνονται στο κάτω μέρος με ένα εγκάρσιο πέταλο θυρεοειδικού ιστού, το οποίο ονομάζεται ισθμός. Από τον ισθμό, αναπτύσσεται μερικές φορές ο πυραμοειδής λοβός. Το βάρος του θυρεοειδή ανέρχεται σε 15-20gr. Οι διαστάσεις του είναι 3~4 cm στο μήκος (ύψος) λοβού, 1,5~2 cm (πλάτος) λοβού, 1,5~2 cm πάχος [1].

Εικόνα 2. Δομή και Σχήμα Θυρεοειδούς

Στην οπίσθια επιφάνεια του θυρεοειδούς βρίσκονται οι τέσσερις παραθυρεοειδείς αδένες (Εικόνα 3), δύο σε κάθε λοβό, άνω και κάτω. Οι παραθυρεοειδείς αδένες εκκρίνουν την παραθυρεοειδορμόνη, ορμόνη που ρυθμίζει το μεταβολισμό του ασβεστίου.

Εικόνα 3. Παραθυρεοειδείς Αδένες

Ο θυρεοειδής, σε σπάνιες περιπτώσεις, εντοπίζεται και σε άλλες ανατομικές περιοχές, εκτός από την φυσιολογική θέση την οποία περιγράψαμε παραπάνω. Οι έκτοπες θέσεις του θυρεοειδούς σχετίζονται με ανωμαλίες κατά τη διάρκεια της διάπλασης του εμβρύου. Για τις περιπτώσεις αυτές, υπάρχουν πολλές διαφορετικές κλινικές εικόνες οι οποίες εξαρτώνται κάθε φορά από τη μη φυσιολογική θέση στην οποία εντοπίζεται ο θυρεοειδής (Εικόνα 4). Έτσι έχουμε π.χ. το Γλωσσικό Θυρεοειδή, που εντοπίζεται στο ύψος της γλώσσας. Όλες οι πιθανές περιπτώσεις αναφέρονται στην Εικόνα 4 με διαφορετικά χρώματα [1, 2].

Εικόνα 4. Έκτοπες θέσεις Θυρεοειδούς

Η αξία του θυρεοειδούς οφείλεται στην έκκριση των προαναφερόμενων ορμονών θυροξίνη (T4), τριωδοθυρονίνη (T3), καλσιτονίνη καθώς αυτές έχουν επίδραση στην αύξηση της κατανάλωσης του οξυγόνου από τους ιστούς, στη σωματική αύξηση, στην ωρίμανση ορισμένων ιστών, στην ανάπτυξη του κεντρικού νευρικού συστήματος, στο συμπαθητικό σύστημα, στο καρδιαγγειακό, στο γαστρεντερικό καθώς και στο μεταβολισμό των λευκωμάτων, των λιπών και των υδατανθράκων.

2.2 ΙΣΤΟΛΟΓΙΑ ΤΟΥ ΘΥΡΕΟΕΙΔΟΥΣ

Ο θυρεοειδής αποτελείται από τους δύο λοβούς και τον ισθμό. Ο κάθε λοβός αποτελείται από ένα άθροισμα μικρότερων μονάδων που καλούνται λοβία. Το κάθε λοβίο είναι ένα άθροισμα από 20-40 σφαιρικά κυστίδια που είναι η βασική ανατομική μονάδα του θυρεοειδικού ιστού. Στην Εικόνα 5 παρουσιάζεται ιστολογικό δείγμα του θυρεοειδικού κυστιδίου χρησιμοποιώντας το μικροσκόπιο.

Εικόνα 5. Ιστολογία Θυρεοειδούς

Κάθε σφαιρικό κυστίδιο (Εικόνα 6), έχει διάμετρο 200 μm και αποτελείται από:

1. Τη βασική μεμβράνη
2. Ένα στοίχο κυβοειδών - εκκριτικών κυττάρων
3. Το κολλοειδές, μια ουσία λευκωματοειδή στην κοιλότητα του κυστιδίου, η οποία έχει σαν κύριο συστατικό της τη Θυρεοσφαιρίνη, μια γλυκοπρωτεΐνη πλούσια σε ιώδιο.

Εικόνα 6. Θυρεοειδικό κυστίδιο

Στο θυρεοειδή, ανάμεσα στα θυρεοειδικά θυλάκια, βρίσκονται μικρά απομονωμένα αθροίσματα από διανυγή κύτταρα. Αυτά είναι τα παραθυλακιώδη κύτταρα ή κύτταρα C (Εικόνα 7) [1, 2].

Εικόνα 7. Κύτταρα C

2.3 ΕΛΕΓΧΟΣ - ΕΞΕΤΑΣΕΙΣ - ΝΟΣΟΙ ΘΥΡΕΟΕΙΔΟΥΣ

Όταν ο ασθενής παραπονεθεί για συμπτώματα που παραπέμπουν σε ασθένειες του θυρεοειδούς, ο ειδικός θεράπων παθολόγος ή ενδοκρινολόγος θα τον παραπέμψει για μια σειρά χρήσιμων στη διερεύνηση του θυρεοειδούς εξετάσεων, πάντα ανάλογα με τη σοβαρότητα της κατάστασης.

Οι εξετάσεις ελέγχου του θυρεοειδούς είναι:

1. Ανάλυση αίματος για μέτρηση των ορμονών που ελέγχει ή παράγονται από αυτόν (TSH, T4, T3)
2. Υπερηχογράφημα του θυρεοειδούς και του τραχήλου γενικότερα
3. **Βιοψία δια Λεπτής Βελόνης (Fine Needle Aspiration - FNA)**
4. Σπινθηρογράφημα με ιώδιο το οποίο χρησιμοποιείται για το διαχωρισμό «θερμού - ψυχρού» όζου
5. Αξονική ή μαγνητική τομογραφία (σπανιότερα)
6. **Εγχείριση αφαίρεσως του θυρεοειδούς αδένος και βιοψία**

Η εξέταση βιοψίας δια λεπτής Βελόνης (FNA) πραγματοποιείται από ιατρό ειδικότητας παθολογοανατομίας, ο οποίος διαθέτει τις γνώσεις τόσο για τη λήψη του υλικού όσο και για την επεξεργασία του και την τελική διάγνωση.

Οι νόσοι του θυρεοειδούς μπορούν να καταταχθούν στις ακόλουθες δύο κύριες κατηγορίες:

1. Καλοήθειες (οζώδης υπερπλασία, θυρεοειδίτιδα hashimoto, υποξεία θυρεοειδίτιδα de Quervain, αυτοάνοση θυρεοειδοπάθεια κ.α.)
2. Κακοήθειες (θηλώδες καρκίνωμα Ca, αναπλαστικό καρκίνωμα Ca, μυελοειδές καρκίνωμα Ca, λέμφωμα κ.α.)

Η ανεύρεση όζων στο θυρεοειδή με απλή ψηλάφηση είναι συχνή. Υπολογίζεται ότι σε πληθυσμούς της Δυτικής Ευρώπης και της Αμερικής διαπιστώνεται η ύπαρξη ενός ή περισσότερων όζων σε ποσοστά της τάξεως του 1,5% για τους άνδρες και 6,4% για τις γυναίκες. Οι μη ψηλαφητοί όζοι αποτελούν το 90% του συνόλου των θυρεοειδικών όζων και η συχνότητα τους αυξάνεται γραμμικά με την ηλικία. Με τη χρήση των υπερηχογράφων έχει ανέβει θεαματικά το ποσοστό των όζων που μπορούν να εντοπιστούν. Με την υπερηχογραφική υποστήριξη αποδεικνύεται ότι οι ψηλαφητοί όζοι που αρχικά φαίνονται ως μονήρης μπορεί να συνυπάρχουν και με άλλους όζους σε ποσοστό 20%-40%. Ο κίνδυνος κακοήθειας των ψηλαφητών όζων κυμαίνεται σε ποσοστό 2%-3% σε περιπτώσεις ιωδοπενικές και σε ποσοστό 4%-8% σε περιπτώσεις μη ιωδοπενικές. Έρευνες έχουν δείξει ότι ο κίνδυνος κακοήθειας δεν επηρεάζεται από τον αριθμό των όζων [3][4]. Η μέθοδος Βιοψίας δια Λεπτής Βελόνης (Fine Needle Aspiration – FNA) είναι ένα εργαλείο για τον κλινικό ιατρό ώστε να αποφασίσει αν θα υποβληθεί ο ασθενής σε χειρουργική επέμβαση ή όχι [4].

2.4 Η ΤΕΧΝΙΚΗ ΤΗΣ ΒΙΟΨΙΑΣ ΔΙΑ ΛΕΠΤΗΣ ΒΕΛΟΝΗΣ (FINE NEEDLE ASPIRATION – FNA) - ΑΝΤΙΠΡΟΣΩΠΕΥΤΙΚΟΤΗΤΑ ΤΟΥ ΥΛΙΚΟΥ

Θυρεοειδικές βιοψίες έχουν αναφερθεί ήδη από τα τέλη του 19ου αιώνα. Τη δεκαετία του 1920, στις Η.Π.Α., ο Δρ. H. Martin πρότεινε και εφάρμοσε την παρακέντηση θυρεοειδούς ως διαγνωστική μέθοδο. Όμως ο Martin χρησιμοποιούσε χοντρές βελόνες, 18 gr, με τοπική αναισθησία και η μέθοδος δεν έτυχε μεγάλης αποδοχής. Το 1950, η μέθοδος ήρθε πάλι στο προσκήνιο από Σουηδούς ερευνητές, που χρησιμοποίησαν λεπτή βελόνα [4]. Η Βιοψία δια Λεπτής Βελόνης (Fine Needle Aspiration – FNA) είναι σήμερα μια τυποποιημένη, καθιερωμένη και παγκόσμια αποδεκτή διαδικασία στην κλινική διερεύνηση για τους όζους θυρεοειδή. Βοηθά στον καλύτερο διαχωρισμό των ασθενών σε αυτούς που πρέπει να χειρουργηθούν και σε αυτούς που δεν πρέπει με τελικό αποτέλεσμα τη μείωση των εγχειρίσεων θυρεοειδούς. Η FNA είναι μια ανέξοδη μέθοδος που μπορεί να εκτελεσθεί σε ένα τμήμα εξωτερικών ιατρείων και έχει λίγες επιπλοκές. Η FNA συνιστά γενικά μια καλή διαγνωστική μέθοδο στην περίπτωση όζων μεγαλύτερων του 1 cm ανεξάρτητα αν είναι ψηλαφητοί ή μη, μονήρεις ή πολυοζώδεις.[4]

Με τον όρο Βιοψία δια Λεπτής Βελόνης (Fine Needle Aspiration – FNA) ή παρακέντηση θυρεοειδούς, εννοούμε την ιατρική πράξη λήψης ιστολογικού δείγματος από το θυρεοειδή με τη χρήση λεπτής βελόνας και χωρίς χειρουργική επέμβαση. Η τεχνική της FNA είναι γενικά απλή στην εφαρμογή αν και απαιτεί ειδική εκπαίδευση και πραγματοποιείται όπως προαναφέρθηκε σε εξωτερικό ιατρείο. Οι βελόνες που χρησιμοποιούνται είναι 22-25 gr και σύριγγες των 10 ή 20 ml. Η FNA εάν πραγματοποιηθεί σωστά δεν προκαλεί τραυματισμό. Μπορεί να τοποθετηθεί μόνο τοπικό αναισθητικό χωρίς να είναι απαραίτητο σε πολλές περιπτώσεις. Η βελόνα τοποθετείται εντός του όζου και γίνεται αναρρόφηση μετακινώντας τη βελόνα με πολύ μικρές και γρήγορες κινήσεις μπρος και πίσω. Η αναρρόφηση διακόπτεται μόλις το υλικό εμφανιστεί στο στόμιο της σύριγγας. Συνήθως γίνονται περισσότερες της μιας παρακεντήσεις σε

κάθε όζο, για λήψη υλικού από διαφορετικές περιοχές του όζου. Το υλικό επιστρώνεται σε ειδικά υάλινα πλακάκια και εξετάζεται στο μικροσκόπιο από ειδικό. Υπάρχει δυνατότητα να πραγματοποιηθεί FNA και με υπερηχογραφική υποστήριξη, προκειμένου να γίνει παρακέντηση σε όζους μικρότερους των 5mm. Γενικά, δεν εμφανίζονται επιπλοκές και σε σπάνιες περιπτώσεις μπορεί να εμφανιστεί οξεία φλεγμονή [4],[5].

Η επιτυχής διάγνωση με τη βιοψία με λεπτή βελόνη (FNA) βασίζεται στους ακόλουθους βασικούς παράγοντες: την επιτυχή λήψη του υλικού και τη σωστή κυτταρολογική αξιολόγηση των επιχρισμάτων στο μικροσκόπιο [6],[7]. Ωστόσο, η εισροή διαγνωστικών μηνυμάτων αρχίζει ήδη κατά τη διενέργεια της βιοψίας (αίσθηση σύστασης αλλοίωσης, αποστιτανώσεων, ινώσεως, παρουσία υγρού), πληθαίνει με την μακροσκοπική παρατήρηση του υλικού (λ.χ. κολλοειδές, στεγνό-ινώδες, αιματηρό, νεκρωτικό, πύον, βλέννα) και κορυφώνεται κατόπιν σωστής χρώσης των επιχρισμάτων με τη μικροσκοπική παρατήρηση. Επομένως, είναι μεγάλης σημασίας ο ειδικός παθολογοανατόμος να διενεργεί προσωπικά τη βιοψία. Έχει διαπιστωθεί και βιβλιογραφικά πως ο ειδικός που διενεργεί την FNA βρίσκεται σε πλεονεκτική θέση καθώς έχει και τη δια ζώσης επαφή κατά τη διενέργεια της [8].

Για να αξιολογηθεί η ευαισθησία και η ειδικότητα της μεθόδου θα πρέπει στους ασθενείς που θα χειρουργηθούν να συγκριθούν τα κυτταρολογικά ευρήματα της FNA με τη μετεγχειρητική ιστολογική διάγνωση. Οι περιπτώσεις στις οποίες έχουμε τόσο την κυτταρολογική έκθεση της FNA όσο και τη μετεγχειρητική ιστολογική διάγνωση αποτελούν το «χρυσό» δείγμα (golden standard). Στις περισσότερες μελέτες της βιβλιογραφίας, η ειδικότητα (specificity) και η ευαισθησία (sensitivity) FNA είναι υψηλές αλλά παρουσιάζουν μεγάλο εύρος τιμών με όρια που κυμαίνονται μεταξύ 65%-98% και 72%-100%, αντίστοιχα [5]. Το εύρος των τιμών δικαιολογεί ότι η εφαρμογή της μεθόδου FNA παρουσιάζει διαφορές μεταξύ των κέντρων λόγω των διαφορετικών μέσων και του επιπέδου εμπειρίας.

Απαραίτητη προϋπόθεση για την αξιόπιστη διάγνωση με την FNA είναι η αντιπροσωπευτικότητα του ληφθέντος υλικού. Αυτή έχει να κάνει τόσο με την ποσότητα του υλικού όσο και με την ποιότητά του, δηλαδή την παρουσία επαρκών παθολογικών στοιχείων για την ορθή διάγνωση. Για το λόγο αυτό, διενεργούνται περισσότερες της μιας παρακεντήσεις σε κάθε όζο, ώστε να υπάρχει υλικό αξιολόγησης από διαφορετικά σημεία του. Η ποιότητα του ληφθέντος υλικού έχει να κάνει και με την αναγνώριση της θυρεοειδικής φύσης της αλλοίωσης δεδομένου ότι σπανίως μη θυρεοειδικές διογκώσεις της περιοχής του τραχήλου παρακεντούνται ως θυρεοειδικοί όζοι (λεμφαδένες, σιελογόνοι αδένες, μορφώματα μαλακών μορίων, μετεγχειρητικά κοκκιώματα, παραθυρεοειδικές βραγχιακές κύστες). Γενικά, οποιαδήποτε αμφισβήτηση της αντιπροσωπευτικότητας του υλικού αναφέρεται απαραίτητα στην έκθεση κυτταρολογικής αξιολόγησης [3].

2.5 ΠΑΡΑΜΕΤΡΟΙ ΑΞΙΟΛΟΓΗΣΗΣ ΕΠΙΧΡΙΣΜΑΤΩΝ ΒΙΟΨΙΑΣ FNA

Σε επάρκεια υλικού, η κυτταρολογική αξιολόγηση ακολουθεί την πορεία σύνθεσης ενός πάζλ. Ο ειδικός καλείται από τα επί μέρους διατηρημένα στοιχεία ενός ιστού που αναρροφήθηκε ("ξεριζώθηκε") και επιστρώθηκε να ανασυνθέσει νοητά την αρχική ιστολογική εικόνα. Είναι λοιπόν προφανές ότι για τη διάγνωση με FNA απαιτούνται άριστη γνώση ερμηνείας των ιστολογικών ευρημάτων στο μικροσκόπιο, εμπειρία στην αντιστοιχία δομών επιχρισμάτων-ιστολογικών τομών και συνδυαστική ικανότητα [3]. Οι παράμετροι που αξιολογούνται από τον ειδικό στα επιχρίσματα θυρεοειδούς είναι οι εξής (από τη μικρότερη προς τη μεγαλύτερη μικροσκοπική μεγέθυνση):

- Υπόστρωμα (κολλοειδές λεπτό/παχύ, ίζημα ή σε σωματίδια, αίμα, φαγοκύτταρα, νεκρωτικό υλικό, λεμφοκυτταρικά στοιχεία, πύον, βλέννα κ.λπ.)
- Κυτταροβρίθεια
- Διάταξη κυττάρων (σκαϊώδης αρχιτεκτονική)
- Κυτταρική μορφολογία (μέγεθος, σχήμα, όρια, χρωστικότητα-κοκκίωση κυτταροπλάσματος, μέγεθος, σχήμα, θέση πυρήνα, χρωματίνη, έγκλειστα, μιτώσεις, ανισοκαρύωση, πολυμορφία, πολυπύρηνες μορφές, κ.ά.)

Επομένως, είναι εύλογο ότι η διαγνωστική ακρίβεια, ευαισθησία και ειδικότητα της μεθόδου FNA εξαρτώνται εκτός από την εμπειρία των ειδικών ιατρών και από το είδος της αλλοίωσης.

2.6 ΔΙΑΓΝΩΣΕΙΣ FNA

Υπάρχουν τέσσερις κύριες κατηγορίες κατάταξης των κυτταρολογικών θυρεοειδικών επιχρισμάτων με FNA:

1. Καλοήθειες
2. Κακοήθειες
3. Υπόνοιες κακοήθειας
4. Οριακά

Στην Εικόνα 8, δίνονται παραστατικά οι πιθανές γνωματεύσεις της κυτταρολογικής εξέτασης βιοψίας FNA. Κάθε σχήμα έλλειψης περικλείει μια κατηγορία με όλες τις διαφορετικές περιπτώσεις που περιλαμβάνει αυτή. Σε κάθε περίπτωση κυτταρολογική εξέταση (βιοψίας) μπορεί το ίδιο επίχρισμα ή σειρά πολλαπλών επιχρισμάτων από ένα ασθενή, να κατατάσσονται σε περισσότερες από μία από τις παραπάνω κατηγορίες. Αυτό συμβαίνει γιατί η παρακέντηση σε ένα σημείο του θυρεοειδούς δίνει τα αυστηρά τοπικά χαρακτηριστικά του συγκεκριμένου σημείου. Έτσι, για παράδειγμα, σε περιπτώσεις στις οποίες δεν έχει εξαπλωθεί κάποιας μορφής καρκίνωμα μπορεί στο σημείο στο οποίο πραγματοποιείται η βιοψία να εμφανίζεται κυτταρολογική διάγνωση καλοήθειας ενώ σε άλλο να εμφανίζεται κυτταρολογική διάγνωση κακοήθειας. Σε αυτή την περίπτωση, στη διάγνωση λαμβάνεται υπόψη το χειρότερο σενάριο, αυτό της κακοήθειας. Τελικά, η πραγματική κατάσταση του ασθενούς διαπιστώνεται με την ιστολογική εξέταση έπειτα από χειρουργική επέμβαση θυρεοειδούς [3].

Εικόνα 8 Πίνακας Διαγνώσεων FNA

2.7 ΑΞΙΟΠΙΣΤΙΑ ΔΙΑΓΝΩΣΕΩΝ

Σύμφωνα με τη βιβλιογραφία, δεν διαγιγνώσκονται όλες οι παθήσεις του θυρεοειδούς το ίδιο εύκολα.

Με καλή αξιοπιστία διαγιγνώσκονται:

Καλοήθειες παθήσεις

Οζώδης υπερπλασία (μεγαλοθυλακιώδης "κολλοειδής" τύπος), λεμφοκυτταρικός τύπος θυρεοειδίτιδας Hashimoto (χωρίς μεγάλη διέγερση) ή οξύφυλλος τύπος (χωρίς σχηματισμό ευμεγέθους οξύφυλλου όζου), υποξεία θυρεοειδίτιδα de Quervain (στο οξύ στάδιο), μετεγχειρητικό κοκκίωμα κ.ά.

Κακοήθεις παθήσεις

Θηλώδες Ca (πλην λιγιστών περιπτώσεων πολύ υψηλής διαφοροποίησης ή εκτεταμένης κυστικής εκφύλισης), Μυελοειδές Ca (κλασσικός τύπος), χαμηλής διαφοροποίησης Ca, αναπλαστικό ή Επιδερμοειδές Ca (πλην δυσχέρειας στη λήψη λόγω εκτεταμένης νέκρωσης) [3].

Μέτρια έως χαμηλή διαγνωστική ακρίβεια:

- Όζοι με εκτεταμένη κυστική εκφύλιση, ίνωση, αποτιτάνωση (δυσχέρεια λήψης)
- Νόσος Graves με λήψη αντιθυρεοειδικών
- Νόσος Hashimoto με επιθηλιακή ατυπία, εκτεταμένη μετάπλαση (οξύφιλη ή πλακώδη) ή έντονη ανοσολογική διέγερση (συχνά ψευδώς θετικά επιχρίσματα)
- Νόσος hashimoto με ανάπτυξη λεμφώματος (απαραίτητη και συχνά επίπονη η διαπίστωση μονοκλωνικού πληθυσμού στις ιστολογικές τομές)
- Εστία θηλώδους Ca σε έδαφος νόσου Hashimoto (συχνά ψευδώς αρνητικά επιχρίσματα).

Η νόσος Hashimoto αποτελεί συχνά συγχυτικό διαγνωστικό παράγοντα προς αμφότερες τις κατευθύνσεις.

"Γκρίζα ζώνη" της βιοψίας του θυρεοειδούς

Οι περιπτώσεις που αποτελούν τη "γκρίζα ζώνη" από διαγνωστική άποψη στο θυρεοειδή είναι:

- α) "Αδενωματώδεις" υπερπλαστικοί όζοι
- β) Θυλακιώδη αδενώματα (ή οξύφιλα από κύτταρα Huerthle), πιθανόν να κατηγοριοποιηθούν ως οριακά και νεοπλάσματα
- γ) Θυλακιώδη καρκινώματα (ή οξύφιλα από κύτταρα Huerthle), πιθανόν να κατηγοριοποιηθούν ως οριακά και νεοπλάσματα

Παρά την προφανή τους καλοήθεια, οι περιπτώσεις των ομάδων (α) και (β) εγείρουν υπόνοια στα επιχρίσματα βιοψίας με βελόνη (αναγκαστικά ψευδώς ύποπτες διαγνώσεις), αφού το κριτήριο που τις ταξινομεί τελικά ιστολογικά σαν καλοήθειες, δηλαδή η παρουσία πλήρους κάψας που δεν διηθείται από τα επιθηλιακά στοιχεία, μοιραία απουσιάζει στο υλικό της FNA. Όταν κυτταρολογικά το αποτέλεσμα κυμαίνεται περισσότερο μεταξύ (α) και (β) χαρακτηρίζεται το επίχρισμα σαν θυλακιώδης αλλοίωση (follicular lesion), ενώ σαν θυλακιώδης νεοπλασία (follicular neoplasia) χαρακτηρίζονται κυρίως πολύ κυτταροβριθείς και πιθανών άτυποι όζοι, οπότε θεωρείται σκόπιμη η ιστολογική διερεύνηση. Αντίστοιχα, στην παρουσία οξύφυλων κυττάρων αναφέρεται ως οξύφιλος πιθανόν μεταπλαστικός όζος (oxyphilic nodule) ή οξύφιλη νεοπλασία (από κύτταρα Huerthle). Η γκρίζα αυτή ζώνη αξιολογείται με βάση κάποια πενιχρά μορφολογικά κριτήρια [9] και τη γενικότερη εμπειρία του ειδικού. Πολύ χρήσιμη για τη διαχείριση των ασθενών που υπάγονται σ' αυτές τις ομάδες είναι και η προφορική επικοινωνία με τον ενδοκρινολόγο, ώστε να συνεκτιμηθούν και λοιπές διαγνωστικές παράμετροι (ενδεικτικά αναφέρονται μέγεθος όζου, ηλικία ασθενούς, υπερηχογραφική παρουσία κάψας, συνύπαρξη άλλων όζων, αντιθυρεοειδικά αντισώματα, κύηση κ.λπ.).

Στο διαφοροδιαγνωστικό εύρος της "γκρίζας ζώνης" υπεισέρχονται και τα παραθυρεοειδικά αδενώματα, με αυξανόμενη μάλιστα συχνότητα, λόγω της ευρείας εφαρμογής της FNA σε μικρούς μη ψηλαφητούς όζους που παρακεντούνται ως θυρεοειδικοί με υπερηχογραφική καθοδήγηση. Παθογνωμονικά κριτήρια δεν υφίστανται και συχνά λόγω της κυτταροβρίθειας τους τέτοια αδενώματα οδηγούν σε υπόνοια νεοπλασίας. Για το λόγο αυτό, είναι σκόπιμο όταν υπάρχει διάγνωση θυλακιώδους ή οξύφυλης αλλοίωσης να διενεργείται επίσης ο σχετικός βιοχημικός έλεγχος των παραθυρεοειδών [3].

2.8 ΠΕΡΙΟΡΙΣΜΟΙ ΜΕΘΟΔΟΥ FNA

Είναι γενικά αποδεκτό ότι δύο είναι οι σημαντικοί περιορισμοί της μεθόδου FNA: τα ανεπαρκή δείγματα και τα ύποπτα ή απροσδιόριστα δείγματα. Η συχνότητα των ανεπαρκών δειγμάτων ποικίλλει μεταξύ 1,64% και 25% σε δείγματα διαφορετικών μελετών [10],[11],[12], ανάλογα με την εμπειρία του παθολογοανατόμου που εκτελεί τη βιοψία, τη φύση του όζου και την αυστηρότητα των κυτταροπαθολογικών κριτηρίων που εφαρμόζονται, καθώς δεν υπάρχει σαφής ορισμός της επάρκειας. Η διάγνωση των ανεπαρκών δειγμάτων παραμένει θέμα εμπειρίας του ειδικού, το οποίο είναι δύσκολο να τυποποιηθεί.

Ο δεύτερος σημαντικός περιορισμός της μεθόδου FNA είναι ο καθορισμός των κριτηρίων κατάταξης των επιχρισμάτων στην "ύποπτη ή απροσδιόριστη" (grey zone) περιοχή (υπόνοια και οριακά). Σε μελέτες με μεγάλο αριθμό δειγμάτων, τα ύποπτα δείγματα αποτελούν περίπου το 11% [10],[13] και συνδέονται με κακοήθεια σε ποσοστό 20%-25% [14],[15],[16]. Η χειρουργική επέμβαση είναι απαραίτητη για τον προσδιορισμό της κατάστασης. Στις περισσότερες μελέτες μέχρι σήμερα, αυτή η κατηγορία περιλαμβάνει μια ετερογενή ομάδα επιχρισμάτων, η μεγάλη πλειοψηφία της οποίας αντιπροσωπεύει επιχρίσματα, θυλακιώδη νεοπλασματικά και οριακά που δεν εκπληρώνουν τα κυτταρολογικά κριτήρια της κακοήθειας. Ως αποτέλεσμα, η κατηγορία των υπόπτων περιλαμβάνει όζους με διαφορετικό βαθμό υπόνοιας και διαφορετικά δεδομένα για την περαιτέρω διαχείριση του ασθενούς.

Όσον αφορά επιχρίσματα με κύτταρα θυλακιώδη και Huerthle, οι προηγούμενες μελέτες έχουν δείξει ότι υποδιαιρούνται σε τρεις τύπους: στον τύπο I (πιθανό καρκίνωμα), τον τύπο II (ιδιαίτερα ύποπτα όζος) και τον τύπο III (πιθανό αδένωμα) [17]. Δεδομένου ότι η ιστολογική διάγνωση σε αυτές τις περιπτώσεις εξαρτάται από την εικόνα διήθησης της κάψας ή/και αγγειακής διήθησης, χαρακτηριστικά τα οποία είναι μη αναγνωρίσιμα στην κυτταρολογική διάγνωση, η κυτταρολογική διάγνωση ενός θυλακιώδους ή Huerthle κυτταρικού νεοπλάσματος αποτελεί μια ένδειξη για χειρουργική επέμβαση και δεν πρέπει να περιληφθούν αυτά τα δείγματα στην "απροσδιόριστη κατηγορία".

2.9 ΒΙΒΛΙΟΓΡΑΦΙΑ

1. www.thyro.gr.
2. **Patton, Thibodeau.** *Anatomy & Physiology*. 3rd.
3. **Τσελένη – Μπαλαφούτα, Σ.,** *Παρακέντηση Θυρεοειδούς: Κυτταρολογική Αξιολόγηση. Σημειώσεις μαθημάτων Ιατρικής Σχολής Αθηνών*, 2002.
4. **Βασιλόπουλος, Χάρης,** *Παρακέντηση Θυρεοειδούς: Κλινική προσέγγιση. Εντατική εκπαίδευση στην Ενδοκρινολογία, 7ος κύκλος, Θυρεοειδής*, Έκδοση Ελληνικής Ενδοκρινολογικής Εταιρείας, 2004
5. **Suen, K.C.,** *Fine-needle aspiration biopsy of the thyroid*. 5, 2002, CMAJ, Τόμ. 167, σσ. 491-495.
6. **Droese M.,** *Aspiration cytology of thyroid*. s.l. : Schattauer Verlag, 1995.
7. **Groellner, H. Gharib and J.R.,** *Fine needle aspiration of thyroid nodules: an appraisal*. 1993, Ann. Intern. Med., Τόμ. 118, σσ. 282-289.
8. **De May R.,** *The art and science of cytopathology. II. Fine needle aspiration*. s.l. : ASCP Press.
9. **Baloch et al.,** *Diagnosis of follicular neoplasm, a gray zone in thyroid fine needle aspiration cytology*. 2001, Diagn Cytopathol 26: 41.
10. **Gharib, H., Goellner, J.R., Johnson, D.A.** *Fine-needle aspiration cytology of the thyroid. A 12-year experience with 11,000 biopsies*. Clin Lab Med, 13, 699-709, (1993).
11. **Mandreker, S.R., Nadkarni, N.S., Pinto, R.G., et al** *Role of fine needle aspiration cytology as the initial modality in the investigation of thyroid lesions*. Acta Cytol, 39, 898-904, (1995).
12. **Ravetto, C., Colombo, L.& Dottorini, M.E.** *Usefulness of fine-needle aspiration in the diagnosis of thyroid carcinoma: a retrospective study in 37,895 patients*. Cancer, 90, 357-3, (2000)
13. **Cersosimo, E., Gharib, H., Suman, V.J., et al** *"Suspicious" thyroid cytologic findings: outcome in patients without immediate surgical treatment*. Mayo Clin Proc, 68, 343-348, (1993).
14. **La Rosa, G.L., Belfiore, A., Giuffrida, D., et al** *Evaluation of the fine needle aspiration biopsy in the preoperative selection of cold thyroid nodules*. Cancer, 67, 2137-2141, (1991).

15. **Gharib, H.** *Fine-needle aspiration biopsy of thyroid nodules: advantages, limitations, and effect.* Mayo Clin Proc, 69, 44-49, (1994).
16. **Schlinkert, R.T., van Heerden, J.A., Goellner, J.R., et al** *Factors that predict malignant thyroid lesions when fine-needle aspiration is "suspicious for follicular neoplasm".* Mayo Clin Proc 72, 913-916, (1997).
17. **Willems, J.S. & Löwhagen, T.** *The role of fine-needle aspiration cytology in the management of thyroid disease.* Clin Endocrinol Metab, 10, 267-273, (1981).

3 ΙΑΤΡΙΚΑ ΣΥΣΤΗΜΑΤΑ ΑΠΟΦΑΣΕΩΝ ΚΑΙ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Τα τελευταία χρόνια, η εφαρμογή μεθόδων εξόρυξης δεδομένων (Data Mining) ή ορθότερα εξόρυξης γνώσης από δεδομένα είναι ιδιαίτερα διαδεδομένη. Στα σύγχρονα συστήματα διαχείρισης πελατών, η προσθήκη μεθόδων και εργαλείων εξόρυξης δεδομένων, βοηθά στην επίλυση επιχειρησιακών προβλημάτων όπως την κατανομή των πελατών, τη βελτίωση εξυπηρέτησης των πελατών και άλλων εφαρμογών [1].

3.1 ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

Η εξόρυξη δεδομένων είναι βασικό μέλος της οικογένειας προϊόντων επιχειρησιακής νοημοσύνης (Business Intelligence - BI). Στη βιβλιογραφία, ο ορισμός της αναφέρεται ως η σύνθετη διαδικασία εξαγωγής συγκεκριμένης, προηγουμένως άγνωστης και δυνητικά ωφέλιμης γνώσης από δεδομένα. Κατά τη διάρκεια της προηγούμενης δεκαετίας, μεγάλοι όγκοι δεδομένων έχουν συσσωρευτεί και αποθηκευτεί στις βάσεις δεδομένων. Ένα μεγάλο μέρος αυτών προέρχεται από επιχειρησιακό λογισμικό, όπως οι οικονομικές εφαρμογές, η διαχείριση των επιχειρηματικών πόρων (Enterprise Resource Planning - ERP), η διαχείριση σχέσης πελατών (Customer Relation Management - CRM), και από τον παγκόσμιο ιστό. Το αποτέλεσμα αυτής της συλλογής είναι ότι οι οργανισμοί έχουν γίνει πλούσιοι σε δεδομένα και φτωχοί σε γνώση. Οι βάσεις δεδομένων έχουν γίνει τόσο μεγάλες και αυξάνονται τόσο γρήγορα σε μέγεθος ώστε η πρακτική χρήση τους να είναι περιορισμένη. Ο κύριος σκοπός της εξόρυξης δεδομένων είναι να εξαχθούν τα κρυμμένα πρότυπα από τα δεδομένα και να αυξηθεί η εγγενής αξία τους μετατρέποντας τα δεδομένα σε γνώση. Μπορεί να τεθεί το ερώτημα, γιατί δεν γίνεται η αναζήτηση αυτής της γνώσης με τη χρήση των κλασικών ερωτήσεων SQL, ή αλλιώς ποιες είναι οι θεμελιώδεις διαφορές μεταξύ εξόρυξης δεδομένων και τεχνολογιών σχεσιακών βάσεων δεδομένων. Το ακόλουθο παράδειγμα μπορεί να δώσει μια ικανοποιητική προσέγγιση. Ας υποθέσουμε ότι έχουμε ένα πίνακα με στοιχεία ασθενών όπως το γένος, την ηλικία, μετρήσεις αίματος, DNA, ιστορικό ασθενειών και ιστορικό καρκίνων σε συγγενείς. Θέλουμε να δούμε τι είναι αυτό που οδηγεί κάποιον στο να εμφανίσει καρκίνο. Μπορεί να γραφεί μια SQL ερώτηση για να καταδείξει πόσοι άνδρες σε σχέση με το πόσες γυναίκες έχουν καρκίνο. Μπορεί επίσης να γραφεί μια SQL ερώτηση που θα δείξει ποιά είναι η επιρροή της παραμέτρου «ιστορικό καρκίνου σε συγγενικά πρόσωπα». Αλλά τι γίνεται με τους άνδρες πάνω από 30 που έχουν συγγενείς με ιστορικό ασθενών ή με τις γυναίκες που δεν έχουν; Θα πρέπει να γραφούν εκατοντάδες αυτών των ερωτήσεων για να καλυφθούν όλοι οι πιθανοί συνδυασμοί. [1][2].

Ακόμα δυσκολότερη καθίσταται η επεξεργασία στην περίπτωση που τα δεδομένα είναι σε αριθμητικές τιμές. Τότε θα πρέπει να επιλεγούν αυθαίρετες περιοχές αριθμητικών τιμών, να υπάρχουν εκατοντάδες στήλες στον πίνακα και θα καταλήγαμε γρήγορα σε έναν αριθμό SQL ερωτήσεων δύσκολα διαχειρίσιμων για να απαντηθεί μια βασική ερώτηση. Αντίθετα, η προσέγγιση του προβλήματος με μεθόδους εξόρυξης δεδομένων είναι μάλλον απλή. Γίνεται επιλογή του σωστού αλγόριθμου εξόρυξης δεδομένων και επιλέγεται η χρήση κάθε στήλης δεδομένων, σαν στήλης εισόδου ή στήλης πρόβλεψης. Σύμφωνα με το προηγούμενο παράδειγμα, ένας αλγόριθμος δέντρων απόφασης θα μπορούσε να αναδείξει την επιρροή του ιστορικού καρκίνου σε συγγενικά πρόσωπα στην εμφάνιση καρκίνου στο προς εξέταση άτομο. Στην προκειμένη περίπτωση, θα επιλέγαμε όλες τις στήλες σαν στήλες εισόδου εκτός από την στήλη «εμφάνιση καρκίνου» η οποία θα ήταν η στήλη πρόβλεψης. Ο αλγόριθμος θα επεξεργαστεί τα δεδομένα, θα αναλύσει την επίδραση κάθε ιδιότητας εισόδου και θα επιλέξει την πιο σημαντική ιδιότητα ώστε να διαχωρίσει τα δεδομένα σε δύο υποσύνολα, έτσι ώστε οι τιμές της παραμέτρου εξόδου (ύπαρξη καρκίνου) να είναι όσο το δυνατόν πιο διαφορετικές στα δύο υποσύνολα. Αυτή η διαδικασία

επαναλαμβάνεται σε κάθε υποσύνολο έως ότου έχει ολοκληρωθεί το δέντρο. Μόλις η διαδικασία εκπαίδευσης του δέντρου ολοκληρωθεί, μπορεί κάποιος να ανακαλύψει τα πρότυπα διατρέχοντας το δέντρο. Κάθε πορεία από τη ρίζα σε έναν κόμβο φύλλων διαμορφώνει έναν κανόνα με αποτέλεσμα να μπορούν να εξαχθούν συμπεράσματα όπως για παράδειγμα ότι ασθενείς με αυξημένους καρκινικούς δείκτες στις αιματολογικές εξετάσεις και με ιστορικό καρκίνου σε συγγενείς τους έχουν συγκεκριμένο ποσοστό πιθανοτήτων να αναπτύξουν καρκίνο. Με αυτόν τον τρόπο, εξάγεται γνώση από τα υπάρχοντα δεδομένα.

3.2 ΒΑΣΙΚΕΣ ΚΑΤΗΓΟΡΙΕΣ ΜΕΘΟΔΩΝ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Οι δύο κύριοι στόχοι της εξόρυξης δεδομένων είναι η περιγραφή και η πρόβλεψη. Η περιγραφή αφορά στην αναπαράσταση των δεδομένων μιας πολύπλοκης βάσης δεδομένων με ένα κατανοητό και αξιοποιήσιμο τρόπο και η πρόβλεψη στην ανεύρεση κρυμμένων προτύπων, αποκάλυψη μη αναμενόμενων σχέσεων και πρόβλεψη μελλοντικών συνθηκών, συμπεριφορών και τάσεων. [1]

Ταξινόμηση (classification): Η ταξινόμηση είναι μία από τις δημοφιλέστερες κατηγορίες εξόρυξης δεδομένων. Επιχειρησιακά προβλήματα όπως η διαχείριση κινδύνου επιλύονται συνήθως με την ταξινόμηση. Η ταξινόμηση αφορά στην κατάταξη των περιπτώσεων σε κατηγορίες με βάση μια προβλέψιμη ιδιότητα. Κάθε περίπτωση έχει ένα σύνολο ιδιοτήτων, μια από τις οποίες χαρακτηρίζεται ως ιδιότητα κατηγορίας (προβλέψιμη ιδιότητα). Σκοπός είναι να παραχθεί ένα μοντέλο το οποίο θα περιγράφει την ιδιότητα κατηγορίας σαν συνάρτηση των ιδιοτήτων εισόδου. Στην περίπτωση του συνόλου δεδομένων ασθενών που χρησιμοποιήθηκαν στην παρούσα διδακτορική διατριβή δύο κατηγορίες εξετάστηκαν η ύπαρξη ή μη καρκίνου. Για να εκπαιδευτεί το μοντέλο (supervised), πρέπει για τα δεδομένα εκπαίδευσης να είναι γνωστή η κατάσταση της ιδιότητας «ύπαρξη καρκίνου», κάτι το οποίο προκύπτει από το ιστορικό των ασθενών. Αντίστοιχα και για την περίπτωση «μη ύπαρξη καρκίνου» Χαρακτηριστικοί αλγόριθμοι ταξινόμησης είναι τα δέντρα απόφασης, τα τεχνητά νευρωνικά δίκτυα, τα Naïve Bayes κ.α. [2][3][4][5][6].

Ομαδοποίηση (clustering): Η ομαδοποίηση ή αλλιώς κατάτμηση χρησιμοποιείται για να προσδιορίσει τους φυσικούς σχηματισμούς ομάδων από τα δεδομένα με βάση ένα σύνολο κοινών ιδιοτήτων. Οι περιπτώσεις μέσα στην ίδια ομάδα έχουν περισσότερες ή λιγότερες παρόμοιες τιμές ιδιοτήτων. Συγκεκριμένα, έστω ένα σύνολο δεδομένων πελατών που περιέχει δύο ιδιότητες: ηλικία και εισόδημα. Ένας αλγόριθμος ομαδοποίησης συγκεντρώνει το σύνολο δεδομένων με βάση αυτές τις ιδιότητες. Η ομάδα 1 περιέχει το νεώτερο πληθυσμό με χαμηλό εισόδημα. Η ομάδα 2 περιέχει τους μέσης ηλικίας πελάτες με υψηλό εισόδημα. Η ομάδα 3 τους μεγαλύτερης ηλικίας με χαμηλό εισόδημα και ούτω καθεξής. Οι αλγόριθμοι ομαδοποίησης χαρακτηρίζονται ως μη επιβλέψιμοι (non supervised). Καμία ιδιότητα δεν χρησιμοποιείται για να καθοδηγήσει τη διαδικασία κατάρτισης, όλες οι ιδιότητες εισόδου αντιμετωπίζονται εξίσου [2][3][4][5][6][7].

Ένωση (association): Η ένωση είναι επίσης μια από τις δημοφιλείς μεθόδους εξόρυξης δεδομένων. Μια χαρακτηριστική εφαρμογή της ένωσης είναι η ανάλυση του πίνακα πωλήσεων μιας επιχείρησης ώστε προσδιορισθούν εκείνα τα προϊόντα τα οποία πωλούνται συχνά μαζί στο ίδιο καλάθι αγορών. Η συνήθης χρήση της ένωσης είναι να προσδιορίσει τα κοινά σύνολα στοιχείων (συχνά σέτ στοιχείων) και τους κανόνες δημιουργίας αυτών με σκοπό την στοχευόμενη πώληση.

Από την άποψη της ένωσης, κάθε προϊόν, ή γενικότερα κάθε ιδιότητα/αξία θεωρείται στοιχείο. Οι περισσότεροι αλγόριθμοι τύπου ένωσης βρίσκουν τα συχνά σέτ στοιχείων με την ανίχνευση του συνόλου δεδομένων πολλές φορές. Το κατώτατο όριο συχνότητας (υποστήριξης) καθορίζεται από το χρήστη πριν την επεξεργασία. Παραδείγματα τέτοιων αλγορίθμων είναι οι αλγόριθμοι που βασίζονται στην υποστήριξη κανόνα (support rule) και στην εμπιστοσύνη κανόνα (confidence rule) [2][3][4][5][6][7].

Παλινδρόμηση (regression): Η μέθοδος της παλινδρόμησης είναι παρόμοια με την ταξινόμηση. Η κύρια διαφορά είναι ότι η προβλέψιμη ιδιότητα είναι συνεχής αριθμός. Οι τεχνικές παλινδρόμησης

χρησιμοποιούνται ευρέως στον τομέα της στατιστικής. Η γραμμική συμμεταβολή και η λογιστική παλινδρόμηση αποτελούν δημοφιλέστερες μεθόδους παλινδρόμησης. Άλλες μέθοδοι παλινδρόμησης περιλαμβάνουν τα δέντρα παλινδρόμησης και τα τεχνητά νευρωνικά δίκτυα. Η παλινδρόμηση μπορεί να εφαρμοστεί σε διάφορους τομείς, όπως στη μετεωρολογία για να προβλεφθούν οι ταχύτητες ανέμου με βάση τη θερμοκρασία, την πίεση αέρα, και την υγρασία [2][3][4][5][6][7].

Πρόβλεψη (forecasting): Η πρόβλεψη είναι μια ακόμα σημαντική μέθοδος εξόρυξης δεδομένων. Μπορεί να βοηθήσει στην απάντηση ερωτημάτων όπως ποια θα είναι η αξία ενός αποθεματικού αύριο; Ποιό θα είναι το ποσοστό πωλήσεων αναψυκτικών για τον επόμενο μήνα; Τα δεδομένα εισόδου είναι τύπου χρονικής σειράς. Οι μέθοδοι πρόβλεψης εξετάζουν γενικές τάσεις και περιοδικότητα. Ως δημοφιλέστερη μέθοδος πρόβλεψης θεωρείται η ARIMA, η οποία υλοποιεί τη μεθοδολογία Auto Regressive Integrated Moving Average Model [2][3][4][5][6][7].

Ανάλυση Ακολουθίας (sequence analysis): Η ανάλυση ακολουθίας χρησιμοποιείται για τη δημιουργία μοντέλων σε μια διακριτή σειρά. Μια ακολουθία αποτελείται από μια σειρά διακριτών τιμών (ή καταστάσεων). Μια ακολουθία DNA είναι μια μακρά ακολουθία από τέσσερα διαφορετικά μέρη: Αδενίνη (Adenine), Θυμίνη (Thymine), Κυτοσίνη (Cytosine) και Γουανίνη (Guanine). Μια επιλογή στον Παγκόσμιο Ιστό είναι μια ακολουθία από ιστοσελίδες. Οι αγορές πελατών μπορούν επίσης να διαμορφωθούν ως στοιχεία ακολουθίας. Χαρακτηριστικό παράδειγμα αποτελεί ένας πελάτης που αγοράζει αρχικά έναν υπολογιστή, έπειτα ένα μικρόφωνο και τελικά μια Web κάμερα. Το κοινό των μεθόδων ανάλυσης ακολουθίας και ένωσης (association) είναι ότι κάθε μεμονωμένη περίπτωση περιέχει ένα σύνολο στοιχείων ή καταστάσεων. Η διαφορά τους είναι ότι οι μέθοδοι ανάλυσης ακολουθίας αναλύουν τις μεταβάσεις καταστάσεων ενώ η μέθοδος ένωσης θεωρεί κάθε στοιχείο ίσο και ανεξάρτητο. Σύμφωνα με τη μέθοδο ανάλυσης ακολουθίας, το να αγοράσει κάποιος έναν υπολογιστή προτού αγοράσει μικρόφωνο είναι μια διαφορετική ακολουθία από το να αγοράσει μικρόφωνο πριν από έναν υπολογιστή. Για έναν αλγόριθμο ένωσης, αυτά θεωρούνται όμοια. Η ανάλυση ακολουθίας είναι ένας σχετικά νέος τρόπος εξόρυξης δεδομένων. Είναι αρκετά σημαντική σε δύο κυρίως τύπους εφαρμογών: Ανάλυση Παγκόσμιου Ιστού και ανάλυση DNA. Υπάρχουν διάφορες τεχνικές ανάλυσης ακολουθίας διαθέσιμες όπως οι αλυσίδες Markov και άλλες [2][3][4][5][6][7].

3.3 ΚΥΚΛΟΣ ΖΩΗΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Ο κύκλος ζωής εξόρυξης δεδομένων αποτελείται από τα ακόλουθα βήματα [1]:

ΒΗΜΑ 1: ΚΑΤΑΝΟΗΣΗ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ: Το πρώτο βήμα είναι να γίνει κατανοητό το πρόβλημα και να λάβει μια μορφή. Αυτό είναι αναγκαίο για να μπορούν να παραχθούν χρήσιμα συμπεράσματα και να επιλεγθούν οι κατάλληλες μέθοδοι.

ΒΗΜΑ 2: ΣΥΛΛΟΓΗ ΔΕΔΟΜΕΝΩΝ: Το δεύτερο βήμα της εξόρυξης δεδομένων είναι η συλλογή δεδομένων. Τα δεδομένα αποθηκεύονται σε ένα σύστημα βάσης δεδομένων. Τα δεδομένα κάποιες φορές δεν είναι πολλά και πιθανόν να πρέπει να συλλεχθούν περισσότερα καθώς επίσης μπορεί να μην περιέχουν όλες τις απαιτούμενες πληροφορίες και να πρέπει να εμπλουτιστούν.

ΒΗΜΑ 3: ΕΠΙΛΟΓΗ ΔΕΔΟΜΕΝΩΝ: Είναι ένα βήμα το οποίο απαιτεί συνήθως αρκετό χρόνο και βαθιά γνώση των δεδομένων. Ο σκοπός του είναι ο «καθαρισμός» των δεδομένων και η αφαίρεση του “θορύβου” των άσχετων πληροφοριών για παράδειγμα, η αφαίρεση στοιχείων που δεν έχουν αξία για την έρευνα, καθώς και ο μετασχηματισμός των δεδομένων, όπως μετατροπή τύπου δεδομένων (αριθμητικά σε κατηγορικά), ομαδοποίηση, συνάθροιση, προσθήκη τιμών σε περιπτώσεις έλλειψης τους, αφαίρεση ανώμαλων τιμών (outliers), σύμπτυξη δεδομένων.

ΒΗΜΑ 4: ΕΥΡΕΣΗ ΚΑΤΑΛΛΗΛΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ: Το κύριο ζητούμενο κατά τη διαδικασία ευρέσεως των κατάλληλων χαρακτηριστικών είναι να ανακαλυφθεί το κατάλληλο υποσύνολο

των χαρακτηριστικών παραμέτρων, τέτοιο ώστε να αντιπροσωπεύει με τον πληρέστερο τρόπο όλες τις περιπτώσεις του δείγματος. Αυτό έχει σαν αποτέλεσμα την περαιτέρω απλούστευση των προς επεξεργασία στοιχείων με τη μείωση της διάστασης τους. Οι μέθοδοι ευρέσεως κατάλληλων χαρακτηριστικών είναι ανεξάρτητες από τις μεθόδους εξόρυξης δεδομένων, αν και κάποιες από αυτές έχουν ενσωματωθεί σε κάποιες μεθόδους εξόρυξης δεδομένων.

ΒΗΜΑ 5: ΑΝΑΠΤΥΞΗ ΜΟΝΤΕΛΟΥ ΑΠΟΦΑΣΗΣ: Πριν δημιουργηθεί οποιοδήποτε μοντέλο, πρέπει να καθοριστεί η μέθοδος εξόρυξης δεδομένων που θα χρησιμοποιηθεί (ταξινόμηση, ένωση, κατάτμηση ή άλλο). Σε αυτό το στάδιο, χρειάζεται συνεργασία με αυτούς που κατέχουν τη γνώση του συγκεκριμένου αντικειμένου. Η ανάπτυξη του μοντέλου είναι ο πυρήνας της διαδικασίας εξόρυξης δεδομένων, αν και δεν είναι τόσο απαιτητικό στάδιο σε πόρους και χρόνο όσο ο καθαρισμός των δεδομένων.

ΒΗΜΑ 6: ΥΠΟΒΟΛΗ ΕΚΘΕΣΗΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ: Σε αυτό το στάδιο παραδίδονται οι εκθέσεις των συμπερασμάτων ώστε να παρθούν αποφάσεις όπου χρειάζεται.

ΒΗΜΑ 7: ΠΡΟΒΛΕΨΗ: Τα αποκαλυπτόμενα πρότυπα χρησιμοποιούνται ώστε να προβλεφθούν μελλοντικές συμπεριφορές, τάσεις και μοτίβα. Για να δώσουμε τις προβλέψεις, πρέπει να έχουμε εκτός από το εκπαιδευμένο πρότυπο και ένα σύνολο νέων υποθέσεων.

ΒΗΜΑ 8: ΑΠΟΤΙΜΗΣΗ: Πρόκειται για το τελευταίο στάδιο της δημιουργία του πλήρους προβλεπτικού συστήματος. Σε αυτό χρησιμοποιούνται κατάλληλες μετρικές μέθοδοι για την αποτίμηση της απόδοσης των αλγορίθμων εξόρυξης δεδομένων. Έχει αναπτυχθεί ένα πλήθος από αυτές τις μεθόδους με αντιπροσωπευτικότερη και πλέον αξιόπιστη τη γραφική μέθοδο αποτίμησης των καμπυλών Receiving Operating Characteristics (ROC). Η αποτίμηση με τις καμπύλες ROC βασίζεται στον πίνακα συσχέτισης (confusion matrix).

		Αποτέλεσμα βάση γνώσης (Golden Standard)		
		Θετικά	Αρνητικά	
Αποτέλεσμα Ελέγχου (Test)	Θετικά (Positive)	Αληθώς Θετικά (True Positive)	Ψευδώς Θετικά (False Positive)	Θετική προφητική αξία
	Αρνητικά (Negative)	Ψευδώς Αρνητικά (False Negative)	Αληθώς Αρνητικά (True Negative)	Αρνητική προφητική αξία
		Ευαισθησία (Sensitivity)	Ειδικότητα (Specificity)	

Πίνακας 3-1 Confusion Matrix

Συγκεκριμένα ισχύουν τα ακόλουθα:

- ευαισθησία = $\text{Αληθώς Θετικά} / (\text{Αληθώς Θετικά} + \text{Ψευδώς Αρνητικά})$
- ειδικότητα = $\text{Αληθώς Αρνητικά} / (\text{Ψευδώς Θετικά} + \text{Αληθώς Αρνητικά})$
- αρνητική προφητική αξία = $\text{Αληθώς Αρνητικά} / (\text{Ψευδώς Αρνητικά} + \text{Αληθώς Αρνητικά})$
- θετική προφητική αξία = $\text{Αληθώς Θετικά} / (\text{Αληθώς Θετικά} + \text{Ψευδώς Θετικά})$

Η καμπύλη ROC προκύπτει ως γραφική παράσταση των ψευδώς θετικών προς την ευαισθησία. Στην Εικόνα 9 εμφανίζεται μια τυπική καμπύλη ROC. Όσο πιο κοντά στον άξονα της ευαισθησίας βρίσκεται η

καμπύλη τόσο πιο αποδοτικό είναι το μοντέλο. Επίσης, η έννοια του εμβαδού κάτω από την καμπύλη προσδιορίζει την ακρίβεια του μοντέλου. Έτσι μοντέλο με εμβαδό 1 θεωρείται ιδανικό μοντέλο [9].

Εικόνα 9 Παράδειγμα καμπύλης ROC

3.4 ΠΡΟΤΥΠΑ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

ΔΕΔΟΜΕΝΩΝ OLE DB:

Το πρότυπο Object Linking Embedding (OLE) DB για Data Mining παρουσιάστηκε από την εταιρεία Microsoft το 1999. Ο πυρήνας του προτύπου είναι η DMX (Data Mining eXtensions) μια γλώσσα 4^{ης} γενιάς για τη δημιουργία των μοντέλων εξόρυξης δεδομένων. Το API αυτό μπορεί να χρησιμοποιηθεί για τη δημιουργία αυτόνομων εφαρμογών εξόρυξης δεδομένων χρησιμοποιώντας οποιαδήποτε πλατφόρμα ανάπτυξης. Υποστηρίχθηκε από διάφορους προμηθευτές συμπεριλαμβανομένων των Angoss, KXEN, και Megarputer. Ουσιαστικά ενοποιεί δύο τεχνολογίες, της OLE DB και της SQL [7].

ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ API ΤΗΣ JAVA:

Το Application Interface (API) της Java για την εξόρυξη δεδομένων ανέπτυξε η εταιρία Oracle με την εμφάνιση της έκδοσης Oracle 10g. Ο σκοπός είναι να βοηθήσει στην ανάπτυξη αυτόνομων εφαρμογών με χρήση της γλώσσας Java. Το πρότυπο χωρίζει τα στοιχεία πηγής (φυσικά στοιχεία) και τα λογικά στοιχεία έννοιες. Τα στοιχεία πηγής μπορούν να είναι οποιαδήποτε αρχεία δεδομένων σχεσιακών βάσεων ή κειμένων [7].

Πρότυπα PMML: Η Predictive Model Markup Language (PMML) είναι μια γλώσσα σήμανσης βασισμένη στην XML. Αναπτύχθηκε από το Data Mining Group (DMG) για να επιτρέψει το σχεδιασμό μοντέλων για χρήση τους σε προβλεπτική ανάλυση και εξόρυξη δεδομένων που να μπορούν να χρησιμοποιούνται και σε άλλες εφαρμογές λογισμικού. Η PMML παρέχει μια μέθοδο ορισμού προτύπων ανεξάρτητη από τον κατασκευαστή της εφαρμογής, έτσι δεν υπάρχουν περιορισμοί ασυμβατότητας λογισμικού για την ανταλλαγή των μοντέλων μεταξύ των εφαρμογών. Επιτρέπει στους χρήστες να κάνουν χρήση ενός λογισμικού, να παράγουν ένα μοντέλο και να το χρησιμοποιήσουν σε ένα άλλο λογισμικό για να κάνουν την απεικόνιση, την ανάλυση και εκτίμηση του μοντέλου. Η PMML υποστηρίζει τον καθορισμό του περιεχομένου για τους ακόλουθους αλγόριθμους εξόρυξης δεδομένων, Polynomial regression, Logistic regression, General regression, Center-based clusters, Density-based clusters, Trees, Associations, Neural nets, Naïve Bayes, Sequences, Text model, Vector machine. Όλοι οι σημαντικοί προμηθευτές μοντέλων εξόρυξης δεδομένων υποστηρίζουν τα πρότυπα PMML [7].

CRISP-DM:

Η μεθοδολογία εξόρυξης δεδομένων Crisp-DM παρουσιάστηκε το 1996 από τρεις εταιρείες, τις SPSS (ISL τότε), NCR, και Daimler Chrysler. Η μεθοδολογία Crisp-DM δεν περιγράφει μια ιδιαίτερη τεχνική εξόρυξης δεδομένων αλλά εστιάζει στη διαδικασία του κύκλου ζωής ενός έργου εξόρυξης δεδομένων. Αποτελείται από ένα σύνολο εργασιών τεσσάρων επιπέδων. Τα τέσσερα επίπεδα είναι: η φάση, η γενική εργασία, η εξειδικευμένη εργασία και η περίπτωση διαδικασίας. Στο πρώτο επίπεδο, η διαδικασία εξόρυξης δεδομένων είναι οργανωμένη σε φάσεις. Κάθε φάση είναι διαιρεμένη σε αρκετές δευτέρου επιπέδου γενικές εργασίες. Το τρίτο επίπεδο, της εξειδικευμένης εργασίας είναι το σημείο στο οποίο περιγράφεται πως οι δράσεις που ορίζονται στο προηγούμενο, γενικό επίπεδο, θα εφαρμοστούν σε συγκεκριμένες ειδικές περιπτώσεις. Για παράδειγμα, αν στο δεύτερο επίπεδο υπάρχει μια γενική εργασία του τύπου «καθαρισμός δεδομένων», στο τρίτο επίπεδο θα περιγράφεται που είναι διαφορετική αυτή η εργασία στην περίπτωση των αριθμητικών δεδομένων σε σχέση με τα κατηγορικά δεδομένα. Το τέταρτο επίπεδο της διαδικασίας είναι μια καταγραφή των δράσεων, αποφάσεων και αποτελεσμάτων μιας πραγματικής εφαρμογής εξόρυξης δεδομένων. Περιγράφει δηλαδή, τι θα συμβεί σε όλα τα βήματα, σε ένα πραγματικό πρόβλημα και όχι σε γενικό επίπεδο [7].

3.5 ΝΕΕΣ ΤΑΣΕΙΣ ΣΤΗΝ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

Η εξόρυξη δεδομένων είναι σχετικά νέα τεχνολογία έναντι της τεχνολογίας βάσεων δεδομένων. Θεωρείται ακόμα μια αναδυόμενη τεχνολογία. Μέχρι σήμερα, τα περισσότερα λογισμικά πακέτα επεξεργάζονται τα δεδομένα με στόχο τη στατιστική ανάλυση παρά την εξόρυξη γνώσης. Πρόσφατα διάφοροι προμηθευτές μοντέλων εξόρυξης δεδομένων, συμπεριλαμβανομένης της Microsoft, πραγματοποίησαν συμφωνία για τη δημιουργία των API's στο χώρο της εξόρυξης δεδομένων. Στα επόμενα χρόνια, θα υπάρξουν όλο και περισσότεροι προγραμματιστές που θα είναι σε θέση να χτίσουν πρότυπα εξόρυξης δεδομένων. Κατά τη διάρκεια των επόμενων ετών προβλέπεται να κυριαρχήσουν οι ακόλουθες τάσεις [2]:

A) Ενσωματωμένη εξόρυξη δεδομένων: Όλο και περισσότερες επιχειρησιακές εφαρμογές θα περιλάβουν τα χαρακτηριστικά γνωρίσματα της εξόρυξης δεδομένων με στόχο την πρόβλεψη. Παραδείγματος χάριν, μια εφαρμογή CRM θα επιτρέπει στους χρήστες να προβλέψουν τις πωλήσεις προϊόντων, στους λιανοπωλητές να συστήνουν τα κατάλληλα προϊόντα στους πελάτες. Η δυνατότητα αυτή δίνεται γιατί τα διαθέσιμα στο εμπόριο APIs εξόρυξης δεδομένων, όπως το OLE DB για Data Mining, επιτρέπουν στους υπεύθυνους βάσεων δεδομένων και ανάπτυξης εφαρμογών να χρησιμοποιήσουν τα χαρακτηριστικά γνωρίσματα της εξόρυξης δεδομένων και να τα ενσωματώσουν στις επιχειρησιακές εφαρμογές.

B) Πακέτα εξόρυξης δεδομένων για κάθετες εφαρμογές: Η εξόρυξη δεδομένων μπορεί να εφαρμοστεί σχεδόν σε κάθε τομέα. Σήμερα, σημαντικές εφαρμογές της εξόρυξης δεδομένων υπάρχουν στον χρηματοοικονομικό τομέα, στον τομέα της ασφάλειας, και στις τηλεπικοινωνίες. Υπάρχει μια αυξανόμενη ανάγκη για χρήση εξειδικευμένων τεχνικών εξόρυξης δεδομένων ώστε να λυθούν επιχειρησιακά προβλήματα σε πολλούς κάθετους τομείς. Παραδείγματος χάριν, στον τομέα της υγειονομικής περίθαλψης, εφαρμόζονται οι ειδικές τεχνικές εξόρυξης δεδομένων για να αναλυθούν οι ακολουθίες DNA. Στις εφαρμογές ασφάλειας δικτύων, χρησιμοποιούνται αλγόριθμοι κατάρτισης πραγματικού χρόνου για να ανιχνευθούν παρειασφρήσεις στα δίκτυα. Επίσης, μη παραδοσιακές τεχνικές εξόρυξης δεδομένων χρησιμοποιούνται για να αναλυθούν τα μη δομημένα στοιχεία στον Παγκόσμιο ιστό. Ένας ακόμα κάθετος τομέας εφαρμογής είναι η εξόρυξη δεδομένων από κείμενο.

3.6 ΕΦΑΡΜΟΓΗ ΤΗΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ ΣΤΗΝ ΙΑΤΡΙΚΗ

Ο χώρος της ιατρικής παράγει έναν συνεχώς αυξανόμενο όγκο δεδομένων και συνεπώς όλο και περισσότερη κρυμμένη πληροφορία υπάρχει σε αυτά. Είναι κοινά αποδεκτό από όλους τους φορείς υγείας ότι η αποτελεσματική διαχείριση των ιατρικών δεδομένων παίζει καθοριστικό ρόλο στην παροχή υγείας υψηλού επιπέδου [10]. Η ανάλυση των ιατρικών δεδομένων μπορεί να συνεισφέρει στη βελτίωση της αποτελεσματικότητας με ταυτόχρονη μείωση του χρόνου και του κόστους. Η αποκτώμενη γνώση με

τις μεθόδους εξόρυξης δεδομένων μπορεί να αξιοποιηθεί από την ιατρική έρευνα, τόσο στο επίπεδο της διάγνωσης όσο και της θεραπείας [11].

Τα ιατρικά δεδομένα είναι από τη φύση τους ετερογενή, περιλαμβάνουν αποτελέσματα απεικονιστικά, γραφήματα, κείμενο από την κλινική εξέταση, εργαστηριακές μετρήσεις σε αριθμούς καθώς και δεδομένα σε άλλες μορφές. Έτσι, η εξόρυξη της κρυμμένης γνώσης από αυτά πρέπει να γίνει από συνδυασμό εικόνων, σχημάτων, κειμένων, αριθμών, το οποίο είναι δυσκολότερο από τις κλασσικές περιπτώσεις επεξεργασίας δεδομένων σε αριθμούς και κατηγορίες. Οι σύγχρονες τεχνολογίες εξόρυξης δεδομένων επιτρέπουν πλέον τη διαχείριση της ετερογενούς φύσης των ιατρικών δεδομένων. Πρώτιστα η δυνατότητα επεξεργασίας της φυσικής γλώσσας και οι τεχνικές εξόρυξης δεδομένων από κείμενα επιτρέπουν την εξαγωγή πληροφορίας και γνώσης από τις ιατρικές σημειώσεις και τις κλινικές εξετάσεις [12],[13]. Ιατρική οντολογία και ορολογία μπορούν να ανιχνευθούν με τη χρήση μεθόδων εξόρυξης δεδομένων του παγκόσμιου ιστού και με τεχνικές εκμάθησης οντολογίας [13], [14].

Η έμφαση που δίνεται σήμερα σε ιατρικές πράξεις στηριζόμενες σε τεκμήρια (Evidence-based Medicine - EBM) είναι ένας από τους κύριους λόγους που ενισχύουν την εφαρμογή μεθόδων εξόρυξης δεδομένων στην ιατρική πρακτική [11]. Αυτές οι ιατρικές πράξεις συνήθως αποτελούν οδηγίες κλινικής πρακτικής ή κανόνες κλινικών αποφάσεων [15]. Η διασύνδεση της εξόρυξης δεδομένων με την ιατρική πρακτική υλοποιείται με την υπάρχουσα προσπάθεια ανάπτυξης αυτόματων συστημάτων υποστήριξης απόφασης. Η ύπαρξη συγκεκριμένης οντολογίας επιτρέπει να ξεκαθαριστεί ποιοί κανόνες αποφάσεων επιβεβαιώνουν την ιατρική γνώση [16]. Τα ιατρικά συστήματα υποστήριξης απόφασης υποστηρίζουν τους ιατρούς στη λήψη ιατρικών αποφάσεων [16].

Ένα πρόβλημα στο οποίο προσκρούει η εφαρμογή της εξόρυξης δεδομένων στην ιατρική είναι η έλλειψη στοιχειώδους οργάνωσης και αρχειοθέτησης των ιατρικών δεδομένων. Ο όγκος, η πολυπλοκότητα, η πολυσύνθετη δομή της ιατρικής πληροφορίας δυσχεραίνουν την προσπάθεια σε αυτό το επίπεδο. Επίσης, η απουσία ενιαίου ιατρικού φακέλου, η έλλειψη και η ετερογένεια των πληροφοριακών συστημάτων και η έλλειψη διαλειτουργικότητας, η ανομοιομορφία στην κωδικοποίηση των δεδομένων είναι στοιχεία που πρέπει να ξεπεραστούν ώστε να ανοίξει ο δρόμος προς την αποτελεσματική εφαρμογή των μεθόδων εξόρυξης δεδομένων στο χώρο της υγείας. Διάφορες σημαντικές προσπάθειες έχουν γίνει διεθνώς για να συνδέσουν τους κλινικούς και ερευνητικούς στόχους με τους διοικητικούς, όπως το πρόγραμμα I2B2 στο Χάρβαρντ [18], ή, σε μικρότερη κλίμακα, το Hemostat [15] και το σύστημα Rhene στην Ιταλία [14]. Επιπλέον, διάφορες εμπορικές λύσεις για την κοινή διαχείριση των πληροφοριών, των στοιχείων, και της γνώσης είναι διαθέσιμες στην αγορά.

Τέλος, ένα ακόμα σημαντικό στοιχείο της ιδιαιτερότητας των ιατρικών δεδομένων έχει να κάνει με το ιατρικό απόρρητο και την ασφαλή διαχείριση τους. Τα δεδομένα συλλέγονται από ασθενείς και πρέπει να αποτραπεί κάθε πιθανότητα κακής χρήσης τους. Υπάρχει λοιπόν ηθική, νομική αλλά και κοινωνική υποχρέωση να προστατευθεί το ιατρικό απόρρητο και εγείρονται σοβαρά ζητήματα όπως το ζήτημα της ιδιοκτησίας των ιατρικών δεδομένων, η πιθανότητα νομικής εμπλοκής κατά τη χρήση τους καθώς και οι πιθανές συνέπειες από αυτήν [19].

3.7 ΒΙΒΛΙΟΓΡΑΦΙΑ

1. **Bao, HO Tu.** *Introduction to knowledge discovery and data mining*, course, Institute of Information Technology National Center for Natural Science and Technology, <http://www.ebook.edu.vn/?page=1.9&view=1694>, [Web Accessed 22th January 2008].
2. **MacLennan, ZhaoHui Tang and Jamie.** *Data Mining with SQL Server 2005*. s.l.: Wiley Publishing, Inc., 2005.
3. **Mehmed Kantardzic.** *Data Mining: Concepts, Models, Methods, and Algorithms*, John Wiley & Sons © 2003.
4. **Oded Maimon and Lior Rokach.** *Data mining and knowledge discovery handbook*. Tel-Aviv University, Israel, Springer 2005.

5. **Nada Lavrac.** *Selected techniques for data mining in medicine.* Department of Intelligent Systems, J. Stefan Institute, 1000 Ljubljana, Slovenia.
6. **Βαζιργιάννης Μ. – Χαλκίδη Μ.,** *Εξόρυξη γνώσης από βάσεις δεδομένων,* Τυπωθήτω, Αθήνα 2003.
7. **Jiawei Han and Kamber Micheline,** *Data mining concepts and techniques,* Simon Fraser University, Academic Press, USA 2001
8. **Dunham Margaret,** *Data Mining Introductory and Advanced Topics,* Prentice Hall 2004.
9. The area under an ROC Curve, <http://gim.unmc.edu/dxtests/roc3.htm>.
10. **Krzysztof j. Giosa, G. William Mooree,** *Uniqueness of medical data mining,* Joornal of Artificial intelligence in medicine, 2002.
11. **Sackett, D. L., Rosenberg, W. M., Gray, J. A., Haynes, R B., Richardson, W. S.,** *Evidence based medicine: what it is and what it isn't.* BMJ, 312 (7023), 71-2, 2004.
12. **Savova, G. K., Ogren, P. V., Duffy, P. H., Buntrock, J. D., Chute, C. G.,** *Mayo clinic NLP system for patient smoking status identification.* J Am Med Inform Assoc, 15(1), 25-8, 2008.
13. **Cimiano, A., Hoto, A., Staab, S.,** *Learning concept hierarchies from text corpora using formal concept analysis.* Journal of Artificial Intelligence Research, 24, 305-339, 2005.
14. **Montani, S., Portinale, L., Leonardi, G., Bellazzi, R.,** *Case-based retrieval to support the treatment of end stage renal failure patients.* Artif Intell Med, 37(1), 31-42, 2006.
15. **Bellazzi, R., Larizza, C., Magni, P., Bellazzi, R.,** *Temporal data mining for the quality assessment of hemodialysis services.* Artif Intell Med, 34(1), 25-39, 2005.
16. **Raj, R., O'Connor, M. J., Das, A. K.,** *An Ontology-Driven Method for Hierarchical Mining of Temporal Patterns: Application to HIV Drug Resistance Research.* AMIA Symp., 2008.
17. **Schmidt, R., Montani, S., Bellazzi, R., Portinale, L., Gierl, L.** *Case-based reasoning for medical knowledge-based systems.* Int J Med Inform, 64(2-3), 355-367, 2001.
18. **Heinze, D. T., Morsch, M. L., Potter, B. C., Sheffer, R.E Jr.,** *Medical i2b2 NLP smoking challenge: the A-Life system architecture and methodology.* J Am Med Inform Assoc, 15(1), 40-3, 2008.
19. **Petr Berka, Jan Rauch, Djamel Abdelkader Zighed,** *Data Mining and Medical Knowledge Management: Cases and Applications,* Information Science Reference, Hershey, USA, 2009.

4 ΗΛΕΚΤΡΟΝΙΚΟ ΣΥΣΤΗΜΑ ΑΡΧΕΙΟΘΕΤΗΣΗΣ ΔΕΔΟΜΕΝΩΝ ΒΙΟΨΙΑΣ FNA ΘΥΡΕΟΕΙΔΟΥΣ

4.1 ΕΙΣΑΓΩΓΗ

Στα πλαίσια της παρούσας διδακτορικής διατριβής συλλέχθηκαν εξετάσεις βιοψίας FNA θυρεοειδούς από το Εργαστήριο Παθολογοανατομίας του Α' Τμήματος Παθολογίας της Ιατρικής Σχολής του Πανεπιστημίου Αθηνών. Δεδομένου ότι το εν λόγω Εργαστήριο λειτουργεί και σαν κέντρο αναφοράς, σημαντικός αριθμός των δειγμάτων εστάλησαν εκεί από άλλα Εργαστήρια για επανέλεγχο. Το αρχειακό υλικό ήταν πολύ καλά ταξινομημένο σε χρονολογική σειρά αλλά ήταν σε έντυπη μορφή.

4.1.1 ΣΧΕΔΙΑΣΜΟΣ ΚΑΙ ΑΝΑΛΥΣΗ ΑΠΑΙΤΗΣΕΩΝ ΒΑΣΗΣ ΔΕΔΟΜΕΝΩΝ FNA

Αρχικά, πραγματοποιήθηκε ο σχεδιασμός απαιτήσεων για τη δομή και το σχεδιασμό της βάσης δεδομένων. Με βάση τα στοιχεία από την τεκμηριωμένη διάγνωση σχεδιάστηκε και αναπτύχθηκε προηγμένο σύστημα για την κωδικοποίηση και αρχικοποίηση των δεδομένων. Διαμορφώθηκαν οι ακόλουθες πέντε κατηγορίες δεδομένων:

1. Δημογραφικά Στοιχεία, τα οποία αφορούν στα στοιχεία ταυτοποίησης του ασθενούς
2. Γενικά στοιχεία, τα οποία αφορούν στοιχεία για την εξέταση όπως π.χ. ημερομηνία παρακέντησης, επάρκεια υλικού κ.λ.π.
3. Αποτελέσματα Κυτταρολογικής Εξέτασης, τα οποία αφορούν στις τελικές διαγνώσεις της κυτταρολογικής εξέτασης.
4. Επιθηλιακά Στοιχεία, τα οποία αφορούν παραμέτρους/χαρακτηριστικά που σχετίζονται με την επιθηλιακή σύσταση του αναρροφώμενου υλικού.
5. Κυτταρολογικά Στοιχεία, τα οποία αφορούν παραμέτρους/χαρακτηριστικά τα οποία σχετίζονται με την κυτταρολογική σύσταση του υλικού.

Σύμφωνα με την καταγραφή των απαιτήσεων, την εισαγωγή και τη ψηφιοποίηση των στοιχείων, δημιουργήθηκαν 111 χαρακτηριστικά για κάθε ασθενή (record). Τα περισσότερα χαρακτηριστικά είχαν τιμές δυαδικού τύπου, αποτυπώνοντας την ύπαρξη ή μη του κάθε χαρακτηριστικού, ενώ κάποιες άλλες είχαν τιμές τύπων αριθμών ή αλφαριθμητικών χαρακτήρων. Από τα 111 χαρακτηριστικά επιλέχθηκαν 60 χαρακτηριστικά τα οποία περιγράφουν τη δομή των επιχρισμάτων ενώ δημιουργήθηκαν άλλα 7 χαρακτηριστικά τα οποία αφορούσαν στην ομαδοποίηση άλλων χαρακτηριστικών.

Στο παράρτημα Α παρατίθεται ο πίνακας όλων των στοιχείων τα οποία καταγράφηκαν. Συγκεκριμένα, στον πίνακα αυτό παρατίθεται το όνομα της κάθε παραμέτρου (ελληνικά και αγγλικά), το όνομα του πεδίου της βάσης το οποίο αντιστοιχεί σε αυτή και οι αποδεκτές τιμές του πεδίου της βάσης. Επίσης, στην τελευταία δεξιά στήλη υπάρχει διαχωρισμός των στοιχείων κατά κατηγορία (π.χ. δημογραφικά, ασθένειες κλπ.). Στην Εικόνα 10 παρουσιάζεται το μπλοκ διάγραμμα της βάσης δεδομένων. Υπάρχουν δύο πίνακες, DEMOGRAFICS και VISIT. Στον πίνακα DEMOGRAFICS πρωτεύων κλειδί είναι το πεδίο CODE και στον VISIT πρωτεύων κλειδί είναι ο συνδυασμός των πεδίων CODE (ξένο κλειδί – foreign key) και CODE_VISIT.

Εικόνα 10 Μπλοκ διάγραμμα βάσης εφαρμογής συλλογής δεδομένων FNA

4.1.2 ΑΝΑΠΤΥΞΗ – ΥΛΟΠΟΙΗΣΗ ΒΑΣΗΣ ΔΕΔΟΜΕΝΩΝ

Η βάση καθώς και το περιβάλλον εισαγωγής δεδομένων που περιγράφεται παρακάτω εγκαταστάθηκαν σε ένα φορητό υπολογιστή με χαρακτηριστικά CPU Intel στα 2 Ghz, 2GB RAM και λειτουργικό σύστημα Windows XP Professional Έκδοση 2002 και Service Pack 3.

Με τη βοήθεια του σχεδιασμού και ανάλυσης απαιτήσεων αναπτύχθηκε και υλοποιήθηκε η βάση δεδομένων στην οποία αποθηκεύτηκαν τα δεδομένα προς επεξεργασία. Για αυτό το σκοπό επιλέχθηκε η λύση Microsoft SQL Server 2005 [1] γιατί είναι ανοικτή και ευέλικτη σε πολλά εργαλεία, δοκιμασμένη σε πολλές εφαρμογές, διαθέτει καλή υποστήριξη, είναι γενικά ελαφριά εφαρμογή, η ανάπτυξη εφαρμογών είναι προσιτή χωρίς ιδιαιτερότητες. Επίσης, διαθέτει ευρεία γκάμα προϊόντων, ώστε να υποστηρίζει εξυπηρετητές, προσωπικούς υπολογιστές και φορητές συσκευές. Μπορεί να υποστηρίξει χρήστες από διάφορα λειτουργικά συστήματα. Η έκδοση SQL 2005 κατέχει μεγάλο μερίδιο στην αγορά.

Στην Εικόνα 11, παρουσιάζεται μια τυπική οθόνη του περιβάλλοντος σχεδιασμού της βάσης. Μεταξύ του πίνακα δημογραφικών και εξετάσεων υπάρχει σύνδεση ένα προς πολλά. Σε ένα ασθενή μπορεί να αντιστοιχούν πολλές εξετάσεις. Το πρωτεύον κλειδί για τον πίνακα δημογραφικών στοιχείων ασθενών είναι το πεδίο CODE ενώ για τον πίνακα επισκέψεων (VISIT), το πρωτεύον κλειδί είναι ο συνδυασμός των πεδίων CODE (σαν ξένο κλειδί) και του CODE_VISIT (σαν κλειδί του πίνακα επισκέψεων (VISIT)).

Εικόνα 11 Σχεδιασμός Βάσης Δεδομένων με εργαλεία SQL Server 2005

4.1.3 ΣΧΕΔΙΑΣΜΟΣ ΚΑΙ ΑΝΑΛΥΣΗ ΑΠΑΙΤΗΣΕΩΝ ΓΡΑΦΙΚΟΥ ΠΕΡΙΒΑΛΛΟΝΤΟΣ ΕΙΣΑΓΩΓΗΣ ΣΤΟΙΧΕΙΩΝ

Παράλληλα με το σχεδιασμό της βάσης, έγινε και η προεργασία για το σχεδιασμό και την ανάλυση απαιτήσεων του γραφικού περιβάλλοντος εισαγωγής δεδομένων. Λαμβάνοντας υπόψη ότι το σύστημα θα μπορούσε να χρησιμοποιηθεί και πέρα από τα πλαίσια της παρούσας διδακτορικής διατριβής λήφθηκε μέριμνα ώστε να παρέχεται ένα φιλικό και ευέλικτο περιβάλλον προς το χρήστη [2].

Για το σχεδιασμό του γραφικού περιβάλλοντος του συστήματος λήφθηκαν υπό όψιν οι κυτταρολογικές εκθέσεις και η δομή τους, η λογική σειρά με την οποία διαβάζονται οι κυτταρολογικές εκθέσεις, η ευκολία επιλογής κάθε παραμέτρου και οι διάφοροι περιορισμοί σε επιλογές παραμέτρων, όπως αποκλεισμός επιλογής αλληλοσυγκρουόμενων χαρακτηριστικών.

4.1.4 ΑΝΑΠΤΥΞΗ - ΥΛΟΠΟΙΗΣΗ ΓΡΑΦΙΚΟΥ ΠΕΡΙΒΑΛΛΟΝΤΟΣ ΕΙΣΑΓΩΓΗΣ ΔΕΔΟΜΕΝΩΝ

Για την υλοποίηση του περιβάλλοντος εισαγωγής δεδομένων χρησιμοποιήθηκε το εργαλείο Microsoft Visual Studio 2005 και η γλώσσα αντικειμενοστραφούς προγραμματισμού VB.NET [3].

Στο κομμάτι εισαγωγής δημογραφικών στοιχείων, οι ασθενείς εμφανίζονται με αλφαβητική σειρά σε ένα πλαίσιο Grid (Εικόνα 12). Η ταξινόμηση (sorting) μπορεί να γίνει με οποιαδήποτε στήλη επιλεγεί. Επίσης, λόγω του μεγάλου όγκου των δεδομένων, αναπτύχθηκε λειτουργικότητα για την εύκολη αναζήτηση ασθενών με δυνατότητα εμφάνισης των ασθενών που το επώνυμο τους αρχίζει από το ίδιο γράμμα. Επιπλέον, δημιουργήθηκε λίστα προεπιλεχθέντων στοιχείων για την αύξηση ταχύτητας εισαγωγής δεδομένων. Στις επιλογές επεξεργασίας που αναπτύχθηκαν περιλαμβάνονται η εισαγωγή νέου ασθενούς, η διόρθωση, η διαγραφή, η αποθήκευση στοιχείων ή αλλαγών, η επισκόπηση της καρτέλας εξετάσεων του επιλεγμένου ασθενούς. Στην Εικόνα 12, παρουσιάζεται η οθόνη της καρτέλας των δημογραφικών στοιχείων ασθενούς [3].

Λογισμικό Διαχείρισης Βάσης Δεδομένων Ασθενών Θυρεοειδούς

Κωδικός	Επώνυμο	Όνομα	Φύλο	Ηλικία
5566			Γυναίκα	30
5896			Γυναίκα	41
4097			Γυναίκα	0
5409			Άνδρας	44
188			Γυναίκα	48
2741			Γυναίκα	37
6776			Άνδρας	25
3735			Γυναίκα	0
2494			Γυναίκα	50
8873			Γυναίκα	36
776			Γυναίκα	39

Α Β Γ Δ Ε Ζ Η Θ
Ι Κ Λ Μ Ν Ξ Ο Π
Ρ Σ Τ Υ Φ Χ Ψ Ω

ΟΛΟΙ

Εισαγωγή Στοιχείων Νέου Ασθενούς
Διόρθωση Στοιχείων Επιλεγμένου Ασθενούς
Διαγραφή Επιλεγμένου Ασθενούς
Αποθήκευση Αλλαγών
Επισκόπηση Καρτέλας Εξετάσεων Ασθενούς

Κωδικός Ασθενούς: 5566
Επώνυμο Ασθενούς:
Όνομα Ασθενούς:
Φύλο: Γυναίκα
Ημερομηνία Γέννησης: 15/5/1972
Ηλικία: 30
Ιστορικό Ασθενούς:
Εξοδος

Εικόνα 12 Οθόνη Δημογραφικών Στοιχείων Ασθενή

Για την καρτέλα των εξετάσεων λάβαμε υπόψη ότι υπάρχει η διασύνδεση ενός ασθενούς με πολλές εξετάσεις (παρακεντήσεις) οπότε το περιβάλλον εμφανίζει όλες τις εξετάσεις με χρονική σειρά. Ο κωδικός επίσκεψης, σε συνδυασμό με τον κωδικό ασθενούς, είναι το πρωτεύον κλειδί του πίνακα και αυξάνεται αυτόματα [5],[6] (Εικόνα 13). Για τη διευκόλυνση της εισαγωγής στοιχείων χρησιμοποιήθηκε

η λογική των καρτελών (tabs) ώστε να διαχωριστούν οι διακριτές οντότητες της εκθέσεως (Εικόνα 13). Έτσι προέκυψαν οι ακόλουθες τυπικές οθόνες:

- Οθόνη γενικών στοιχείων εξέτασης, η οποία αφορά στοιχεία της εκθέσεως που δεν κατατάσσονται στις άλλες κατηγορίες (Εικόνα 13)
- Οθόνη αποτελεσμάτων κυτταρολογικής εξέτασης, σε συνδυασμό με την τελική μετεγχειρητική ιστολογική διάγνωση (Εικόνα 14)
- Οθόνη επιθηλιακών στοιχείων (Εικόνα 15)
- Οθόνη κυτταρολογικών στοιχείων (Εικόνα 16)

Ακολουθήθηκε η λογική συμπλήρωσης των στοιχείων από αριστερά προς τα δεξιά της κάθε οθόνης. Το μεγαλύτερο πλήθος των στοιχείων ήταν δυαδικής λογικής 0 ή 1 (υπάρχει ή δεν υπάρχει), γι' αυτό επιλέχθηκε η λύση των check boxes για την συμπλήρωση των περισσότερων δεδομένων. Επίσης, προστέθηκαν τεχνικές και έλεγχοι ώστε να αποκλείεται η επιλογή αντικρουόμενων στοιχείων.

Εικόνα 13 Οθόνη Καρτέλας Εξετάσεων Ασθενούς - Γενικά στοιχεία εξέτασεων

Λογισμικό Διαχείρισης Βάσης Δεδομένων Ασθενών Θυρεοειδούς

Καρτέλα Εξετάσεων Ασθενούς

Εξετάσεις Ασθενούς

Κωδικός Ασθενούς	Κωδικός Επίσκεψης	Ημερομηνία Ε
5566	1	15/5/2002 10:0

Εισαγωγή Νέας Εξέτασης Ασθενούς

Διαγραφή Επιλεγμένης Εξέτασης Ασθενούς

Διόρθωση Στοιχείων Επιλεγμένης Εξέτασης Ασθενούς

Αποθήκευση Αλλαγών

Κλείσιμο

Γενικά Στοιχεία Εξετάσεων

Αποτέλεσμα Κυτταρολογικής Εξέτασης

Επιθηλιακά Στοιχεία

Κυτταρολογικά Στοιχεία

Καλοήθειες

- ☐ Οξώδης Υπερπλάσια
- ☐ Θυρεοειδίτιδα HASHIMOTO
- ☐ Οξυφιλος Οζος
- ☐ Στοιχεία Θυρεοειδίτιδας
- ☐ Υποξεία Θυρεοειδίτις de Quervain
- ☐ Μη Ειδική Κυστική
- ☐ Μη Ταξινομούμενη Θυρεοειδίτις
- ☐ Αυτοάνοση Θυρεοειδοπάθεια

Κακοήθειες

- ☒ Θηλώδης Καρκίνωμα Ca
- ☐ Θηλώδης Καρκίνωμα Ποικιλία
- ☐ Αναπλαστικό Καρκίνωμα Ca
- ☐ Μυελοειδές Καρκίνωμα Ca
- ☐ Αταξινόμητα
- ☐ Λέμφωμα
- ☐ Λοιπά
- ☒ Νεόπλασμα Θυλακώδεις
- ☐ Νεόπλασμα Huerthle
- ☐ Καρκίνωμα Χαμηλής Διαφοροποίησης
- ☐ Μετάσταση

Άλλο

- ☐ Παραθυρεοειδής
- ☐ Λεμφαδένας
- ☐ Βραγχιακή Κύστη
- ☐ Θυρεογλωσσικού Πόρο
- ☐ Φαγοκοκκύωμα
- ☐ Μεσεγχυματικό
- ☐ Ογκος Καροτιδικού
- ☐ Σιειλογόνος

Οριακά

- ☐ Θυλακώδη
- ☐ Huerthle

☐ Υπόνοια Κακοήθειας

Εικόνα 14 Οθόνη Καρτέλας Εξετάσεων Ασθενούς – Αποτελέσματα κυτταρολογικής εξέτασης

Λογισμικό Διαχείρισης Βάσης Δεδομένων Ασθενών Θυρεοειδούς

Καρτέλα Εξετάσεων Ασθενούς

Εξετάσεις Ασθενούς

Κωδικός Ασθενούς	Κωδικός Επίσκεψης	Ημερομηνία Ε
5566	1	15/5/2002 10:0

Εισαγωγή Νέας Εξέτασης Ασθενούς

Διαγραφή Επιλεγμένης Εξέτασης Ασθενούς

Διόρθωση Στοιχείων Επιλεγμένης Εξέτασης Ασθενούς

Αποθήκευση Αλλαγών

Κλείσιμο

Γενικά Στοιχεία Εξετάσεων

Αποτέλεσμα Κυτταρολογικής Εξέτασης

Επιθηλιακά Στοιχεία

Κυτταρολογικά Στοιχεία

Υπόστρωμα

- ☐ Πολύ Καλοήθειδες
- ☐ Λίγα Καλοήθειδες
- ☐ Αίμα
- ☐ Αμυλοειδές

Κυτταροβρίθεια

- ☐ Μικρή
- ☒ Μεγάλη

Συνολική Κυττάρων

- ☐ Καλή
- ☒ Χαλαρή

Πρόσθετα Στοιχεία Υποστρώματος

- ☐ Πολλά Λεμφοκύτταρα/Λεμφοτικό
- ☒ Λίγα Λεμφοκύτταρα/Λεμφοτικό
- ☐ Νεκρωτικό
- ☐ Ασβέστιο
- ☐ Λιποειδή
- ☐ Πολλά Φαγοκύτταρα/Μακροφάγα
- ☐ Λίγα Φαγοκύτταρα/Μακροφάγα
- ☐ Ποιμορροφύρηνα
- ☐ Γιγαντοκύτταρα
- ☐ Άλλο
- ☐ Detritus
- ☐ Συνδετικό Υπόστρωμα
- ☐ Ψαμμώδη
- ☐ Ινοβλάστες
- ☐ Ιστιοκύτταρα

Αρχιτεκτονικά Στοιχεία

- ☐ Μονοστιβάδωση
- ☐ Μικροβυθία
- ☒ Πολλοί Θυλακώδεις Σχηματισμοί
- ☐ Λίγοι Θυλακώδεις Σχηματισμοί
- ☒ Συμπαγείς Σχηματισμοί
- ☐ Πολλοί Θηλώδεις Σχηματισμοί
- ☐ Λίγοι Θηλώδεις Σχηματισμοί
- ☐ Ελεύθερα (Διάχυτα)

Εικόνα 15 Οθόνη Καρτέλας Εξετάσεων Ασθενούς – Επιθηλιακά στοιχεία

Λογισμικό Διαχείρισης Βάσης Δεδομένων Ασθενών Θυρεοειδούς

Καρτέλα Εξετάσεων Ασθενούς

Εξετάσεις Ασθενούς

Κωδικός Ασθενούς	Κωδικός Επίσκεψης	Ημερομηνία Ε
5566	1	15/5/2002 10:0

Εισαγωγή Νέας Εξέτασης Ασθενούς

Διαγραφή Επιλεγμένης Εξέτασης Ασθενούς

Διόρθωση Στοιχείων Επιλεγμένης Εξέτασης Ασθενούς

Αποθήκευση Αλλαγών

Κλείσιμο

Γενικά Στοιχεία Εξετάσεων | Αποτέλεσμα Κυτταρολογικής Εξέτασης | Επιθηλιακά Στοιχεία | Κυτταρολογικά Στοιχεία

Στοιχεία Κυττάρων

☒ Μονοκύρην

☐ Δικύρην

☐ Πολυκύρην

Μορφή Κυττάρων

☐ Ανισοκυττάρωση

☒ Ομολή

☐ Πολυμορφία Κυττάρων

Κυτταρόπλασμα

☐ Ατρακτοειδές

☒ Κλασσικό

☐ Πλασμοκυττοειδές

☐ Διαυγές

☐ Πολυγωνικά

☐ Κοκκώση

☐ Καινοτοπίωση

Μορφή Πυρήνα

☒ Ανισοκαρύωση

☐ Έγκλειστα

☐ Πυκνωτικός Πυρήνα

☐ Ισοπυρήνωση/Ομοιόμορφος

☒ Πυρηνικός Συνωστισμός

☐ Πυρήνια

☐ Πολυμορφία Πυρήνα

Τύπος Κυττάρων

☒ Βυλακικά

☐ Πολύα Άτυπα

☐ Λίγα Άτυπα

☐ Επιδερμοειδή

☐ Οξυφιλά

Όρια Κυττάρων

☒ Σαφή

☐ Μη Σαφή

Θέση Πυρήνα στο Κύτταρο

☒ Έκκεντρος

☐ Κεντρικός

Εικόνα 16 Οθόνη Καρτέλας Εξετάσεων Ασθενούς - Κυτταρολογικά στοιχεία

4.2 ΔΙΑΘΕΣΙΜΑ ΣΤΟΙΧΕΙΑ

Η εισαγωγή των δεδομένων ήταν μια μακροσκελής και επίπονη διαδικασία η οποία όμως ανέδειξε νέες απαιτήσεις στο σχεδιασμό της βάσης και στην τήρηση των στοιχείων, με γνώμονα να μην καταστρατηγηθούν οι βασικές αρχές της ακεραιότητας και πληρότητας των δεδομένων. Επίσης, νέες απαιτήσεις αναδείχθηκαν στο σχεδιασμό του περιβάλλοντος δοκιμών καθώς και στην απλοποίηση του συστήματος εισαγωγής νέων δεδομένων.

Τα δεδομένα τα οποία εισήχθησαν τελικά στο σύστημα ήταν συνολικά FNA βιοψίες 9.102 ασθενών ηλικιών μεταξύ 6 και 94 ετών (με αναλογία ανδρών/γυναικών: 1/5.78). Οι βιοψίες FNA αφορούσαν στη περίοδο 2000-2004. Η μεγάλη πλειοψηφία των ασθενών ήταν Έλληνες, από μια ετερογενή γεωγραφική περιοχή. Το γεγονός ότι το Α' Τμήμα Παθολογοανατομίας του Πανεπιστημίου Αθηνών αποτελεί κέντρο αναφοράς για βιοψίες FNA του θυρεοειδούς αποτελεί παράγοντα αύξησης της αξιοπιστίας και καταλληλότητας του δείγματος. Οι βιοψίες εκτελέστηκαν είτε στο Α' Τμήμα του Πανεπιστημίου Αθηνών (οι περισσότερες από τις περιπτώσεις με ψηλαφητούς όζους) είτε αλλού (κυρίως κάτω από την καθοδήγηση του κέντρου αναφοράς). Στη δεύτερη περίπτωση τα δείγματα εστάλησαν σε πλακάκια στο κέντρο αναφοράς, προερχόμενα από διάφορα νοσοκομεία, με διαφορετικά πρωτόκολλα εκτέλεσης της βιοψίας FNA σε θυρεοειδή και συνοδεύτηκαν από τα εξής στοιχεία: φύλλο, ηλικία ασθενούς και κλινική εικόνα.

Οι κυτταρολογικές εκθέσεις ανακτήθηκαν από τα φυσικά αρχεία τα οποία περιλαμβάνουν σύγκριση των στοιχείων των ασθενών καθώς επίσης και ένα σημαντικό αριθμό αναφερόμενων κυτταρολογικών χαρακτηριστικών από την τελική κυτταρολογική διάγνωση. Ιστολογικές διαγνώσεις κατόπιν αφαίρεσης του θυρεοειδούς ήταν διαθέσιμες για 266 ασθενείς.

4.3 ΒΙΒΛΙΟΓΡΑΦΙΑ

1. **Dumler Michelle**, *Microsoft SQL Server 2005 Product Guide*, Microsoft, September 2005.
2. **Shari Lawrence Pfleeger**, *Software Engineering Theory and Practice*. Prentice Hall, February 27, 2009
3. **Foxall James**, *Sams Teach Yourself Visual Basic .NET in 24 Hours*, Sams publishing, 2002.

4. **Grimes Fergal**, *Microsoft .NET for Programmers*, Manning Publications Co., 2002.
5. **Stephens Rod**, *Visual Basic .NET Database Programming*, Que Publishing, 2002.
6. **Dobson Rick**, *Programming Microsoft SQL Server 2000 with Visual Basic .NET*, Microsoft Press 2002.

5 ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΒΙΟΨΙΑΣ FNA

5.1 ΕΙΣΑΓΩΓΗ

Στην παρούσα ενότητα της διδακτορικής διατριβής περιγράφεται η στατιστική ανάλυση των 9.102 συλλεχθέντων δειγμάτων FNA ως προς τα κυτταρολογικά χαρακτηριστικά τους και τις διαγνώσεις. Οι κυτταρολογικές διαγνώσεις των συγκεκριμένων δειγμάτων συσχετίστηκαν με τις ιστολογικές διαγνώσεις, στοχεύοντας στον υπολογισμό της πιθανής επίδρασης και συμβολής κάθε κυτταρολογικού χαρακτηριστικού σε μια ορθή ή ψευδή κυτταρολογική διάγνωση, έτσι ώστε να προσδιοριστούν οι πιθανές πηγές λανθασμένης διάγνωσης.

Τα δείγματα τα οποία περιείχαν μόνο αίμα ή πολύ λίγα θυλακειώδη κύτταρα χωρίς κολλοειδές θεωρήθηκαν ανεπαρκή για τη διάγνωση. Τα αποτελέσματα της στατιστικής ανάλυσης επέτρεψαν το διαχωρισμό των δεδομένων σε καλοήγη, κακοήγη, νεοπλασματικά, ύποπτα για κακοήθεια και οριακά με χαρακτηριστικά γνωρίσματα μεταξύ ενός καλοήθους και ενός νεοπλασματικού.

5.2 ΥΛΙΚΟ ΜΕΛΕΤΗΣ

Το υλικό της μελέτης ήταν ένα σύνολο 9.102 βιοψιών FNA του θυρεοειδούς που ελήφθησαν κατά την διάρκεια της περιόδου 2000 - 2004 από ασθενείς ηλικίας μεταξύ 6 και 94 ετών (με αναλογία πλήθους ανδρών/γυναικών: 1/5.78). Η μεγάλη πλειοψηφία των ασθενών ήταν Έλληνες, από μια ετερογενή γεωγραφική περιοχή. Το γεγονός ότι το Α' Τμήμα Παθολογοανατομίας του Πανεπιστημίου Αθηνών αποτελεί κέντρο αναφοράς για βιοψίες FNA του θυρεοειδή αποτελεί παράγοντα αύξησης της αξιοπιστίας και καταλληλότητας του δείγματος. Οι βιοψίες εκτελέστηκαν είτε στο Α' τμήμα του Πανεπιστημίου Αθηνών (οι περισσότερες από τις περιπτώσεις με ψηλαφητούς όζους) είτε αλλού (κυρίως κάτω από την καθοδήγηση του κέντρου αναφοράς). Τα δείγματα επιστρωμένα σε πλακάκια, στάλθηκαν στο κέντρο αναφοράς από διάφορα νοσοκομεία, με διαφορετικά πρωτόκολλα σχετικά με τα κριτήρια εκτέλεσης βιοψίας FNA σε θυρεοειδή. Μετεγχειρητικές ιστολογικές επαληθεύσεις ήταν διαθέσιμες για 266 ασθενείς (κακοήθειες και μη).

Όλα τα δείγματα που είχαν ιστολογική επιβεβαίωση διερευνήθηκαν προκειμένου να καθοριστούν τα κυριότερα κυτταρολογικά γνωρίσματα που θα εμφανιστούν σε μια ορθή διάγνωση μειώνοντας την πιθανότητα λανθασμένης διάγνωσης.

5.3 ΜΕΘΟΔΟΛΟΓΙΑ

Για την αξιολόγηση των στοιχείων εφαρμόστηκαν στατιστικές μέθοδοι με χρήση του πακέτου λογισμικού. Χρησιμοποιήθηκαν περιγραφικά στατιστικά μεγέθη για την ανάλυση των δεδομένων όπως, μέση τιμή, τυπική απόκλιση, ποσοστά, μέγιστο και ελάχιστο. Έγιναν επίσης χ^2 δοκιμές επιπέδου σημαντικότητας διαφόρων παραμέτρων για να ελεγχθεί η πιθανή συσχέτιση ή ανεξαρτησία τους.

Για τη συσχέτιση των κυτταρολογικών και των ιστολογικών διαγνώσεων και την αξιολόγηση των εργαστηριακών ευρημάτων, πέραν των περιγραφικών στατιστικών μεγεθών χρησιμοποιήθηκαν και υπολογισμοί της ευαισθησίας, της ειδικότητας, της συνολικής ακρίβειας, της αρνητικής και θετικής αξίας πρόβλεψης (negative and positive predictive value). Η ευαισθησία είναι το μέτρο της συχνότητας των αληθώς θετικών ευρημάτων ανάμεσα στο σύνολο εκείνων που πάσχουν από τη νόσο. Η ειδικότητα είναι μέτρο της συχνότητας των αληθώς αρνητικών ευρημάτων ανάμεσα στο σύνολο εκείνων που δεν πάσχουν από τη νόσο. Αντίστοιχα, η αρνητική προβλεπτική τιμή είναι το μέτρο της μετρούμενης από το τεστ

συχνότητας όσων δεν πάσχουν σε σχέση με εκείνους που βρέθηκαν από το τεστ να πάσχουν και θετική προβλεπτική τιμή είναι το μέτρο της συχνότητας όσων πάσχουν σε σχέση με εκείνους που βρέθηκαν από τη μέθοδο FNA να πάσχουν. Ο Πίνακας 5-1 δίνει παραστατικά τους ορισμούς αυτούς.

Πίνακας 5-1 Ορισμοί στατιστικών όρων ευαισθησίας, ειδικότητας, αρνητικής αξίας πρόβλεψης, θετικής προφητικής αξίας

		Αποτέλεσμα βάση γνώσης (Golden Standard)		
		Θετικά	Αρνητικά	
Αποτέλεσμα Ελέγχου (Test)	Θετικά (Positive)	Αληθώς Θετικά (True Positive)	Ψευδώς Θετικά (False Positive)	Θετική προφητική αξία
	Αρνητικά (Negative)	Ψευδώς Αρνητικά (False Negative)	Αληθώς Αρνητικά (True Negative)	Αρνητική προφητική αξία
		Ευαισθησία (Sensitivity)	Ειδικότητα (Specificity)	

Προκειμένου να καθοριστεί εάν μια κατηγορία ασθενειών συσχετίζεται ή όχι με συγκεκριμένες κυτταρολογικές παραμέτρους εφαρμόστηκε μέθοδος ελέγχου στατιστικής σημαντικότητας σε επίπεδο 5% ($p < 0,05$). Η διαδικασία ακολουθήθηκε για κάθε κατηγορία ασθενειών ή συνδυασμό τους και για κάθε παράμετρο των κυτταρολογικών και αρχιτεκτονικών στοιχείων της κυτταρολογικής διάγνωσης.

5.4 ΣΤΑΤΙΣΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

5.4.1 ΣΤΑΤΙΣΤΙΚΑ ΣΤΟΙΧΕΙΑ ΣΤΟ ΣΥΝΟΛΟ ΤΟΥ ΠΛΗΘΥΣΜΟΥ

Από το σύνολο των 9.102 περιπτώσεων, ο αριθμός των δειγμάτων που αντιστοιχούν σε θυρεοειδικά δείγματα ήταν 8.961, συμπεριλαμβανομένων 723 χωρίς επαρκές υλικό (146 με κάποια πιθανή διάγνωση και 577 χωρίς διάγνωση) 141 δείγματα αποδείχθηκαν μη θυρεοειδικά (72 λεμφαδένες (51,06%), 27 σιελογόνοι αδένες (19,15%), 16 ξένα σώματα (11,35%), 11 μεσεγχυματικά (7,8%), 8 θυρεογλωσσικές κύστες (5,67%), 4 βραγχιακές κύστες (2,84%) και 3 παραθυρεοειδείς αδένες (2,13%)).

Μεταξύ του πληθυσμού μελέτης υπήρξαν 1.322 άνδρες (14,8%) και 7.639 γυναίκες (85,25%), με μια αναλογία ανδρών προς γυναίκες 1:5,78. Η ηλικία των ασθενών κυμάνθηκε μεταξύ 6 και 94 ετών (μέση ηλικία: 45,84 έτη, τυπική απόκλιση 17,7). Καμία στατιστικά σημαντική διαφορά δεν καταγράφηκε μεταξύ της μέσης ηλικίας των ανδρών (46,51 έτη, τυπική απόκλιση 18,64) και των γυναικών (45,73 έτη, τυπική απόκλιση 17,53).

Μονήρης όζος θυρεοειδή παρατηρήθηκε σε 6.837 (76,3%) ασθενείς ενώ τα υπόλοιπα 2.124 (23,7%) περιστατικά κατηγοριοποιήθηκαν ως πολυοζώδη. Οι μονήρεις όζοι ήταν εξίσου κατανομημένοι μεταξύ δεξιού λοβού 3.398 (37,93%) και αριστερού λοβού 3.219 (38,25%), ενώ 244 (2,9%) μονήρεις όζοι εντοπίστηκαν στον ισθμό και 1.760 (20,92%) και στους δύο λοβούς.

Καμία στατιστική σχέση δεν παρατηρήθηκε μεταξύ του φύλου των ασθενών ή της ηλικίας και της θέσης των όζων. Η κατανομή των κύριων κυτταρολογικών διαγνώσεων στους ασθενείς με μονήρεις όζους ή πολυοζώδεις παρουσιάζεται στον Πίνακα 5-2 όπου οι μονήρεις όζοι συνδέθηκαν με ένα υψηλότερο ποσοστό κακοήθων κυτταρολογικών διαγνώσεων έναντι των πολυοζωδών (Πίνακας 5-2).

Πίνακας 5-2 Κατανομή κύριων κυτταρολογικών διαγνώσεων σε ασθενείς με μονήρεις ή πολυοζώδεις όζους

ΚΥΤΤΑΡΟΛΟΓΙΚΗ ΔΙΑΓΝΩΣΗ	ΠΛΗΘΟΣ	ΜΟΝΗΡΕΙΣ ΟΖΟΙ			ΠΟΛΥΟΖΩΔΕΙΣ		
		Πλήθος (N)	*	**	Πλήθος (N)	*	**
Ανεπαρκές Υλικό	723	624	9,13%	86,31%	99	4,66%	13,69%
Καλοήθεια	7.481	5.566	81,41%	74,40%	1.915	90,16%	25,60%
Απροσδιόριστα							
Οριακά	108	90	1,32%	83,33%	18	0,85%	16,67%
Υπόνοια	300	255	3,73%	85,00%	45	2,12%	15,00%
Νεόπλασμα	64	55	0,80%	85,94%	9	0,42%	14,06%
Κακοήθεια	285	247	3,61%	86,67%	38	1,79%	13,33%
Σύνολο	8.961	6.837	100,00%	76,30%	2.124	100,00%	23,70%

* Ποσοστό σε σχέση με το σύνολο μονήρων/πολυοζωδών

** Ποσοστό σε σχέση με τον συνολικό αριθμό των περιπτώσεων σε κάθε κατηγορία

Τα 8.238 επαρκή δείγματα ταξινομήθηκαν ως καλοήθης, κακοήθης, νεοπλασματικά (θυλακοειδή ή ογκολογικά ή αβέβαιης βιολογικής συμπεριφοράς) και απροσδιόριστα. Η τελευταία ομάδα υποδιαιρέθηκε σε ύποπτα και οριακά.

Από την πλευρά των καλοήθων, η κατηγορία της οζώδους υπερπλασίας είναι συντριπτικά η πολυπληθέστερη διάγνωση καλοήθειας (6.160 περιπτώσεις - 82,34%) και ακολουθεί η Hashimoto/Θυρεοειδίτιδα. Η κατανομή των καλοήθων κυτταρολογικών διαγνώσεων παρουσιάζεται στον Πίνακα 5-3.

Πίνακας 5-3 Κατανομή Καλοήθων Κυτταρολογικών Διαγνώσεων

	Άνδρες		Γυναίκες		Σύνολο	Ποσοστό %
Οζώδης Υπερπλασία	903	88,88%	5.257	81,31%	6.160	82,34%
Hashimoto/Θυρεοειδίτιδα	40	3,94%	568	8,79%	608	8,13%
Άλλο	73	7,19%	640	9,9%	713	9,53%
Σύνολο	1.016	100,00%	6.465	100,00%	7.481	100,00%

Από τις κακοήθειες η πιο κοινή κυτταρολογική διάγνωση είναι του θηλώδους καρκινώματος (257 περιπτώσεις/90,18%), ακολουθεί το μυελωειδές, το αναπλαστικό καρκίνωμα, το μεταστατικό καρκίνωμα, το λέμφωμα, και το χαμηλής διαφοροποίησης (Πίνακας 5-4).

Πίνακας 5-4 Κατανομή κακώθων κυτταρολογικών διαγνώσεων

Κυτταρολογική Διάγνωση	n	%
Θηλώδες Καρκίνωμα	257	90,18%
Μυελωειδές Καρκίνωμα	14	4,91%
Μη ταξινομούμενο Καρκίνωμα	4	1,40%
Αναπλαστικό Καρκίνωμα	3	1,05%
Μεταστατικό Καρκίνωμα	3	1,05%
Καρκίνωμα Χαμηλής Διαφοροποίησης	2	0,70%
Λέμφωμα	2	0,70%
Σύνολο	285	100,00%

Ο Πίνακας 5-5 παρουσιάζει αναλυτική πληθυσμιακή μελέτη για τα κυτταρολογικά χαρακτηριστικά. Με βάση τα στοιχεία του πίνακα προκύπτει ότι η μεγάλη κυτταροβρίθεια είναι κυρίαρχο χαρακτηριστικό σε όλες τις κακοήθειες (στήλες 5 έως 12) ενώ αντίθετα η μικρή κυτταροβρίθεια κυριαρχεί σε όλες τις καλοήθειες (στήλες 1 έως 4) και στο οριακό οξύφιλο (στήλες 15 έως 16). Αντίστοιχη παρατήρηση μπορούμε να κάνουμε και για τη συνοχή του κυττάρου, καλή συνοχή παρατηρείται σε συντριπτικό ποσοστό στις καλοήθειες (στήλες 1 έως 4) ενώ χαλαρή συνοχή σε μεγάλο ποσοστό στις κακοήθειες (στήλες 5 έως 12). Αντίθετα, η παρουσία κολλοειδούς δεν παρουσιάζει αξιοσημείωτες διαφορές. Επίσης, σχετικά με τη μονοστοιβάδωση παρατηρούνται μεγάλα ποσοστά παρουσίας στις καλοήθειες (στήλες 1 έως 4) και τα οριακά (στήλες 13 έως 16) και αντίθετα μικρά ποσοστά στις κακοήθειες (στήλες 5 έως 12). Οι θηλώδεις σχηματισμοί φαίνεται να είναι χαρακτηριστικό γνώρισμα των θηλωδών περιπτώσεων, καρκίνωμα, νεόπλασμα και οριακό (στήλες 5-6, 9-10, 13-14). Αντίστοιχα, τα οξύφιλα κύτταρα φαίνονται να είναι χαρακτηριστικό γνώρισμα των οξύφιλων περιπτώσεων, hashimoto, νεόπλασμα και οριακό (στήλες 3-4, 11-12, 15-16). Τα πολυπύρρηνα γιγαντοκύτταρα, τα λεμφοκύτταρα και τα έγκλειστα παρουσιάζονται σε πολύ υψηλό ποσοστό μόνο στο θηλώδες καρκίνωμα. Τέλος, τα λεμφοκύτταρα είναι συντριπτικά πολλά στη νόσο Hashimoto.

Πίνακας 5-5 Κυτταρολογική Ανάλυση διαφόρων κατηγοριών Ασθενειών

Κυτταρολογικά Χαρακτηριστικά	Οξώδης Υπερπλασία 6160		Hashimoto/ Θυρεοειδίτης 608		Θηλώδεις Καρκίνωμα ³ 261		Μυελωειδές Καρκίνωμα 14		Νεόπλασμα Θηλώδεις 20		Νεόπλασμα Οξύφιλο 44		Οριακό Θηλώδεις 56		Οριακό Οξύφιλο 52	
	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Κυτταροβρίθεια																
Μικρή	4887	79,33%	405	66,61%	23	8,81%	0	0,00%	0	0,00%	3	6,82%	5	8,93%	29	55,77%
Μεγάλη	1270	20,62%	203	33,39%	217	83,14%	14	100,00%	20	100,00%	32	72,73%	51	91,07%	15	28,85%
Συνοχή Κυττάρου																
Καλή	6155	99,92%	608	100,00%	24	9,20%	0	0,00%	1	5,00%	5	11,36%	24	42,86%	26	50,00%
Χαλαρή	0	0,00%	0	0,00%	211	80,84%	11	78,57%	18	90,00%	35	79,55%	21	37,50%	21	40,38%
Παρουσία Κολλοειδούς																
Λίγο	3647	59,20%	602	99,01%	111	42,53%	1	7,14%	8	40,00%	16	36,36%	41	73,21%	38	73,08%
Αφθονία	2507	40,70%	6	0,99%	15	5,75%	2	14,29%	0	0,00%	4	9,09%	2	3,57%	9	17,31%
Μονοστοιβάδωση	6155	99,92%	607	99,84%	11	4,21%	2	14,29%	0	0,00%	2	4,55%	28	50,00%	20	38,46%
Θηλακιώδεις Σχηματισμοί¹	1015	16,48%	5	0,82%	25	9,58%	0	0,00%	18	90,00%	1	2,27%	43	76,79%	0	0,00%
Μικροθυλάκια²	1675	27,19%	1	0,16%	1	0,38%	0	0,00%	0	0,00%	0	0,00%	3	5,36%	0	0,00%
Φαγοκύτταρα	4852	78,77%	9	1,48%	75	28,74%	2	14,29%	0	0,00%	11	25,00%	6	10,71%	17	32,69%
Πολυπύρρηνα Γιγαντοκύτταρα	15	0,24%	8	1,32%	95	36,40%	0	0,00%	0	0,00%	2	4,55%	0	0,00%	1	1,92%
Οξύφιλα κύτταρα	52	0,84%	278	45,72%	25	9,58%	2	14,29%	3	15,00%	43	97,73%	11	19,64%	51	98,08%
Λεμφοκύτταρα	1074	17,44%	607	99,84%	73	27,97%	1	7,14%	2	10,00%	4	9,09%	16	28,57%	14	26,92%
Έγκλειστα	0	0,00%	0	0,00%	86	32,95%	2	14,29%	0	0,00%	0	0,00%	0	0,00%	0	0,00%

1. Κύτταρα σε θηλώδεις σχηματισμούς
2. Τρισδιάστατη διάταξη κυττάρου με βάση μεμβράνη
3. Συμπεριλαμβάνει θηλώδεις, αναπλαστικό και χαμηλής διαφοροποίησης

5.5 ΣΤΑΤΙΣΤΙΚΗ ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

5.5.1 ΣΤΑΤΙΣΤΙΚΑ ΣΗΜΑΝΤΙΚΕΣ ΣΥΓΚΡΙΣΕΙΣ ΚΥΤΤΑΡΟΛΟΓΙΚΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΜΕΤΑΞΥ ΟΡΙΑΚΩΝ ΚΑΙ ΝΕΟΠΛΑΣΜΑΤΩΝ

Στατιστικές αναλύσεις πραγματοποιήθηκαν σε μια προσπάθεια να αποφασιστούν στατιστικά σημαντικές σχέσεις σε επίπεδο $p < 0,05$ ή $p < 0,01$ στα κυτταρολογικά στοιχεία μεταξύ των επιχρισμάτων τα οποία χαρακτηρίστηκαν ως οριακά και αυτών που χαρακτηρίστηκαν ξεκάθαρα νεοπλασματικά.

1. Στοιχεία υποστρώματος: παρατηρήθηκαν οι εξής διαφορές για τα οριακά σε σχέση με τα νεοπλάσματα: Για τα νεοπλάσματα η δοκιμασία χ^2 σε σχέση με την παρουσία λεμφοκυττάρων έδωσε $p=0,000$ άρα η σχέση νεοπλασμάτων και παρουσίας λεμφοκυττάρων είναι στατιστικά σημαντική σε επίπεδο 0,1%.

2. Αρχιτεκτονικά χαρακτηριστικά: Η χαλαρή συνοχή και η απουσία μονοστοιβάδωσης χαρακτηρίζουν κυρίως τα νεοπλάσματα σε επίπεδο στατιστικής σημαντικότητας 0,1% και για τα δύο.

3. Στοιχεία κυττάρου: Τα άτυπα κύτταρα και πολυγωνικά κύτταρα συνδέονται με ιδιαίτερη σημαντικότητα στη διάγνωση ενός νεοπλάσματος (15,6%) ($p=0,028\%$ και $p=0,001$ αντίστοιχα), όπως επίσης και η παρουσία διπύρηνων κυττάρων ($p=0,001$) και πολυπύρηνων κυττάρων ($p=0,023$).

4. Μορφολογία πυρήνα:

Στα μορφολογικά χαρακτηριστικά τα οποία έχουν στατιστικά σημαντική σχέση με τα νεοπλάσματα βρέθηκαν η μεταβολή του μεγέθους ($p=0,001$), η απουσία ομαλών παρυφών ($p=0,023$), τα σαφή όρια ($p=0,0001$), οι έκκεντροι πυρήνες ($p=0,0001$), η ανισοκαρύωση ($p=0,0001$) και η παρουσία πυρήνιων ($p=0,0001$).

5.5.2 ΠΕΡΙΣΤΑΤΙΚΑ ΤΑ ΟΠΟΙΑ ΕΧΟΥΝ ΧΑΡΑΚΤΗΡΙΣΤΕΙ ΩΣ ΥΠΟΝΟΙΑ ΚΑΚΟΗΘΕΙΑΣ

Από τους 300 ασθενείς (3,34%) με υπόνοια κακοήθειας, ιστολογική επιβεβαίωση είχαμε για 55 περιπτώσεις (18,3% του συνολικού αριθμού των περιπτώσεων υπόνοιας). Οι 38 (69%) από αυτές αποδείχτηκαν κακοήθεις (αληθώς ύποπτα - true suspicious). Από τη σύγκριση των κυτταρολογικών χαρακτηριστικών των αληθώς υπόπτων (true suspicious) σε σχέση με τις άλλες υπόνοιες προκύπτουν οι εξής στατιστικώς σημαντικές σχέσεις:

1. Σε σχέση με τα στοιχεία υποστρώματος, η απουσία ή αφθονία κολλοειδούς ($p=0,043$), η υψηλή κυτταροβρίθεια ($p=0,0001$) και η απουσία φαγοκυττάρων ($p=0,005$).
2. Σε σχέση με τα αρχιτεκτονικά χαρακτηριστικά, η χαλαρή συνοχή ($p=0,0001$), η απουσία μονοστοιβάδωσης ($p=0,0001$), η απουσία μικροθυλακίων ($p=0,003$), η παρουσία συμπαγών σχηματισμών ($p=0,0001$), η παρουσία θηλωδών σχηματισμών ($p=0,0001$).
3. Σε σχέση με τα κυτταρολογικά στοιχεία, η παρουσία γιγαντοκυττάρων ($p=0,0001$), η παρουσία διπύρηνων και πολυπύρηνων κυττάρων ($p=0,0001$) και η υψηλή εκπροσώπηση οξύφιλων κυττάρων ($p=0,0001$).
4. Σε σχέση με τη μορφολογία κυττάρων, η ανισοκύτωση ($p=0,0001$), τα μη φυσιολογικά όρια ($p=0,0001$), το καθαρό κυτταρόπλασμα ($p=0,0001$), τα πολυγωνικά σχήματα ($p=0,0001$), οι έκκεντρες τοποθετήσεις πυρήνων ($p=0,0001$), η ανισοκαρύωση ($p=0,0001$).

Η παρουσία ή απουσία λεμφοκυττάρων δεν είναι στατιστικώς σημαντικά συσχετισμένη με τα αληθώς ύποπτα περιστατικά ενώ χαρακτηριστικά όπως νεκρωτικό, αμυλοϋδές, ασβεστώσεις και ψαμμώματα αποκλείονται από τη στατιστική ανάλυση εξαιτίας του μικρού αριθμού.

5.5.3 ΠΕΡΙΣΤΑΤΙΚΑ ΘΕΤΙΚΑ ΣΕ ΚΑΚΟΗΘΕΙΑ

Συνολικά 285 περιστατικά χαρακτηρίστηκαν σαν θετικά για κακοήθεια. Όταν εξετάστηκαν τα κυτταρολογικά χαρακτηριστικά των κακοηθειών ένα προς ένα σε σύγκριση με τα άλλα περιστατικά βρέθηκαν τα ακόλουθα στοιχεία στατιστικής σημαντικότητας:

1. Χαρακτηριστικά Υποστρώματος: τα κακοήθη περιστατικά σε σύγκριση με όλα τα άλλα παρουσιάζουν τις εξής στατιστικά σημαντικές σχέσεις: α) την απουσία και αφθονία κολλοειδούς 94,4% με επίπεδο στατιστικής σημαντικότητας $p=0,0001$, β) την ανεπάρκεια κολλοειδούς 59,8% με επίπεδο στατιστικής σημαντικότητας $p=0,0001$, γ) την απουσία φαγοκυττάρων με επίπεδο στατιστικής σημαντικότητας $p=0,005$, δ) την υψηλή κυτταροβρίθεια 83,3% με επίπεδο στατιστικής σημαντικότητας $p=0,0001$ και ε) την παρουσία λεμφοκυττάρων με επίπεδο στατιστικής σημαντικότητας $p=0,013$.
2. Αρχιτεκτονικά Στοιχεία: τα κακοήθη περιστατικά σε σύγκριση με όλα τα άλλα παρουσιάζουν τις εξής στατιστικά σημαντικές σχέσεις: α) τη χαλαρή συνοχή 80,9% με επίπεδο στατιστικής σημαντικότητας $p=0,0001$, β) την απουσία μονοστοιβάδωσης 93,2% με επίπεδο στατιστικής σημαντικότητας $p=0,0001$, γ) την απουσία μικροθυλακίων 99,2% με επίπεδο στατιστικής σημαντικότητας $p=0,0001$, δ) την παρουσία συμπαγών σχηματισμών 12,4% με επίπεδο στατιστικής σημαντικότητας $p=0,0001$, ε) την παρουσία θηλωδών σχηματισμών 81,3% με επίπεδο στατιστικής σημαντικότητας $p=0,0001$.
3. Τύπους Κυττάρων: στατιστικά σημαντική σχέση εμφανίζεται για την απουσία θυλακικών κυττάρων 84.5% με επίπεδο στατιστικής σημαντικότητας $p=0.0001$.

5.6 ΣΥΣΧΕΤΙΣΗ ΚΥΤΤΑΡΟΛΟΓΙΚΩΝ ΚΑΙ ΙΣΤΟΛΟΓΙΚΩΝ ΔΙΑΓΝΩΣΕΩΝ

Ο συσχετισμός κυτταρολογικών FNA και ιστολογικών διαγνώσεων έχει μεγάλη σημασία γιατί με αυτό τον τρόπο αναδεικνύεται η διαγνωστική αξία της μεθόδου. Η τυπική διαδικασία προβλέπει αυτή τη σύγκριση εφόσον υπάρχουν και οι ιστολογικές διαγνώσεις. Επίσης, κάποιες παράλληλες εργαστηριακές εξετάσεις, όπως αιματολογικές και υπερηχογραφήματα μπορεί να βοηθήσουν τον ειδικό στη διάγνωση κακοηθειών. Σε αυτή την ενότητα παρουσιάζονται αναλυτικά τα στοιχεία της συσχέτισης κυτταρολογικών και ιστολογικών διαγνώσεων οι οποίες είναι πλήρως επιβεβαιωμένες αλλά και όσων θεωρήθηκαν ως επιβεβαιωμένες με τη συνεκτίμηση επιπλέον εργαστηριακών εξετάσεων.

Στην περίπτωση που λάβουμε υπόψη αποκλειστικά και μόνο τις 266 ιστολογικές επιβεβαιώσεις, ο συσχετισμός των κυτταρολογικών με τις ιστολογικές διαγνώσεις αναλύθηκε και αναδείχθηκε ότι 137 δείγματα ήταν αληθώς αρνητικά (51,5%) και 89 αληθώς θετικά (33,5%). Υπήρξαν 15 ψευδώς-αρνητικά (5,6%), 1 ψευδώς θετικό (0,4%) και 24 ψευδώς ύποπτα δείγματα (9,0%). Συνεπώς, η ευαισθησία (sensitivity) της μεθόδου FNA ήταν 85,58% και η ειδικότητα (specificity) 84,57%, όταν τα ψευδώς θετικά και τα ψευδώς ύποπτα χαρακτηρίστούν σαν ψευδώς θετικά, και ευαισθησία 99,28 όταν τα ψευδώς θετικά και τα ψευδώς ύποπτα δεν συνυπολογιστούν. Η συνολική ακρίβεια κυμάνθηκε από 84,96% έως 93,39%. Η αρνητική προβλεπτική τιμή της FNA ήταν 90,13% και η θετική προβλεπτική τιμή ήταν 78,07% όταν τα ψευδώς θετικά και τα ψευδώς ύποπτα συγκεντρώθηκαν σε μια ομάδα και 98,89% όταν τα ψευδώς ύποπτα δεν εξετάστηκαν.

Στην περίπτωση που λάβουμε υπόψη και τις διαγνώσεις με επιβεβαιώσεις βοηθούμενες από άλλες παράλληλες εργαστηριακές εξετάσεις, ο συσχετισμός των κυτταρολογικών αποτελεσμάτων με τις ιστολογικές, συμπεριλαμβανομένων αυτών από τις συμπληρωματικές μεθόδους, έχει ως εξής. Από ιστολογική ανάλυση, για τα 145 καλοήθη και 121 κακοήθη αλλά και για 1.378 που διαγνώστηκαν ως καλοήθη από συμπληρωματικές εργαστηριακές εξετάσεις (υπέρηχο, αιματολογικές εξετάσεις), αναδείχθηκε ότι υπήρχαν 1.502 αληθώς αρνητικά (91,7%), 108 αληθώς θετικά (6,6%), 13 ψευδώς-αρνητικά (0,07%), 15 ψευδώς θετικά (0,09%). Συνεπώς, η ευαισθησία της μεθόδου FNA ήταν 89,33%, η ειδικότητα 98,98%, συνολική ακρίβεια 98,3%. Η αρνητική προφητική αξία της FNA ήταν 99,14% και η θετική προφητική αξία 87,58%.

Από τα παραπάνω, συμπεραίνεται ότι με το συνυπολογισμό των ιστολογικά επιβεβαιωμένων περιπτώσεων, και αυτών που η διάγνωση τους υποβοηθείται από συμπληρωματικές εργαστηριακές εξετάσεις, βελτιώνονται τα αποτελέσματα της μεθόδου FNA περισσότερο. Άρα, αν συγκρίνουμε αυτό το βελτιωμένο αποτέλεσμα της FNA με αποτελέσματα άλλων μη στατιστικών μεθόδων και τα δεύτερα αποδειχθούν καλύτερα, αυτό στην πραγματικότητα θα έχει μεγαλύτερη αξία αφού το αποτέλεσμα τους θα έχει συγκριθεί με κάτι καλύτερο από την κλασσική FNA.

Στους παρακάτω πίνακες αναλύονται με μεγαλύτερη λεπτομέρεια όσα στοιχεία προέκυψαν για τη σύγκριση κυτταρολογικών με απόλυτα επιβεβαιωμένες ιστολογικές διαγνώσεις και παρέχονται στοιχεία για τα στατιστικά σημαντικά χαρακτηριστικά των επιχρισμάτων. Πιο συγκεκριμένα στον **Πίνακα 5-6** παρουσιάζεται πληθυσμιακή και ποσοστιαία ανάλυση των κατηγοριών, αληθώς αρνητικών, ψευδώς αρνητικών, αληθώς θετικών και ψευδώς υπόπτων με κυτταρολογικά χαρακτηριστικά που εμφάνισαν σημαντικές ποσοστιαίες διαφορές για την διάκριση καλοηθειών και κακοηθειών. Ειδικότερα, αναδεικνύεται ότι η παρουσία υψηλής κυτταροβρίθειας είναι σε μεγαλύτερο ποσοστό στα αληθώς θετικά (74,2% έναντι 37,5%) όπως και το άφθονο κολλοειδές (32,6% έναντι 12,5%). Η μονοστοιβάδωση από την άλλη δεν παρουσιάζεται να είναι ένα σημαντικό στοιχείο αξιολόγησης μια και εμφανίζει πολύ μεγάλα ποσοστά τόσο στα αληθώς αρνητικά (91,2%) όσο και στα ψευδώς αρνητικά (100%) αλλά αντίστοιχα και πολύ μικρό ποσοστό στα Ψευδώς Ύποπτα (20,8%) και αληθώς θετικά (10,1%). Σχετικά με τους θυλακιώδεις σχηματισμούς είναι αξιοσημείωτο ότι στα ψευδώς αρνητικά παρουσιάζονται σε ποσοστό 0%. Επίσης, τα μικροθυλάκια εμφανίζονται σε κάποιο ποσοστό μόνο στα αληθώς αρνητικά. Τα πολυπύρρηνα γιγαντοκύτταρα από την άλλη εμφανίζονται σε αξιοσημείωτο ποσοστό στα αληθώς θετικά (32,6%), ενώ τα οξύφιλα σε αξιοσημείωτο ποσοστό στα ψευδώς ύποπτα (41,7%). Τέλος, τα έγκλειστα εμφανίζονται να είναι το απόλυτο χαρακτηριστικό για τα αληθώς θετικά 19,1% έναντι 0% στις άλλες κατηγορίες.

Στον Πίνακα 5-7, παρουσιάζεται μια ειδικότερη σύγκριση κυτταρολογικών χαρακτηριστικών μεταξύ αληθώς θετικών και ψευδώς θετικών/υπόπτων. Σημαντικές διαφορές παρουσιάζονται στους άφθονους θηλώδεις σχηματισμούς (24,7% στα αληθώς θετικά έναντι 54,2% στα ψευδώς θετικά/ύποπτα). Επίσης, στα οξύφιλα κύτταρα παρουσιάζεται διαφορά 29,2% στα αληθώς θετικά έναντι 62,5% στα ψευδώς θετικά/ύποπτα. Αξιοσημείωτα χαρακτηριστικά είναι η πολυμορφία πυρήνα και τα πυρήνια τα οποία εμφανίζονται μόνο στα ψευδώς θετικά/ύποπτα σε ποσοστό 4,2% έκαστο. Επίσης, τα έγκλειστα εμφανίζονται μόνο στα αληθώς θετικά σε ποσοστό 19,1%.

Τέλος, ο Πίνακας 5-8 παρουσιάζει την πληθυσμιακή και ποσοστιαία κατανομή των κυτταρολογικών διαγνώσεων προς τις ιστολογικές επιβεβαιώσεις. Στον πίνακα επισημαίνονται με ΑΑ τα Αληθώς Αρνητικά, με ΨΑ τα Ψευδώς Αρνητικά, με ΑΘ τα Αληθώς Θετικά και με ΨΥ τα Ψευδώς Ύποπτα. Από την μελέτη του Πίνακα 5-8 προκύπτει ότι σε συντριπτική πλειοψηφία όσα χαρακτηρίστηκαν ως κακοήθειες από την κυτταρολογική ήταν και κακοήθειες με εξαίρεση μία περίπτωση (ΨΘ - στήλη 6). Αυτό συμβαίνει γιατί η μέθοδος FNA είναι γενικά αξιόπιστη και επίσης ο ειδικός παθολογοανατόμος ήταν πολύ έμπειρος. Αντίστοιχα, για τους ίδιους λόγους ΨΑ εμφανίζονται σε μικρό ποσοστό (στήλη 1 - 11,46%, στήλη 2 - 1,04%, στήλη 4 - 3,13%) και έχουν να κάνουν κυρίως με τις περιπτώσεις όπου στην κυτταρολογική εξέταση αφορούσαν οξυφιλία, με την μορφή οξύφιλου όζου, σε περιπτώσεις Hashimoto ή οζώδους υπερπλασίας με στοιχεία οξυφιλίας. Τέλος, οι περιπτώσεις ΨΥ εμφανίζονται κυρίως όταν εμπλέκονται στοιχεία οξυφιλίας, οριακά ή νεοπλάσματα. Έτσι, 3,08% των ιστολογικών επιβεβαιωμένων περιπτώσεων οζώδους υπερπλασίας είχαν χαρακτηριστεί κυτταρολογικά ψευδώς ως νεοπλάσματα και 10,77% ως ύποπτα.

Πίνακας 5-6 Κυτταρολογικές παράμετροι συσχετιζόμενες με Αληθώς Αρνητικά, Ψευδώς Αρνητικά, Αληθώς Θετικά και Ψευδώς Υποπτα

Κυτταρολογικά Χαρακτηριστικά	Ψευδώς Θετικών/Υπόπτων (Καλοήθη)	Αληθώς Αρνητικά (Καλοήθη)	Ψευδώς Αρνητικά (Κακοήθη)	Αληθώς Θετικά (Κακοήθη)
Κυτταροβρίθεια Μικρή Υψηλή	8/24 (33,3%) 9/24 (37,5%)	89/137 (65%) 46/137 (33,6%)	11/15 (73,3%) 4/15 (26,7%)	11/89 (12,4%) 66/89 (74,2%)
Κολλοειδές Λίγο Αφθονο	14/24 (58,3%) 3/24 (12,5%)	83/137 (60,6%) 7/137 (5,1%)	13/15 (86,7%) 0/15 (0%)	49/89 (55,1%) 29/89 (32,6%)
Μονοστοιβάδωση	5/24 (20,8%)	125/137 (91,2%)	15/15 (100%)	9/89 (10,1%)
Θυλακιώδεις Σχηματισμοί Λίγοι Πολλοί Σύνολο %	3/24 (12,5%) 1/24 (4,2%) 4/24 (16,7%)	17/137 (12,4%) 11/137 (8,0%) 28/137 (20,4%)	0/15 (0%) 0/15 (0%) 0/15 (0%)	7/89 (7,9%) 4/89 (4,5%) 11/89 (12,4%)
Μικροθυλάκια	0/24 (0%)	31/137 (22,6%)	0/15 (0%)	0/89 (0%)
Φαγοκύτταρα Λίγα Πολλά Σύνολο %	5/24 (20,8%) 3/24 (12,5%) 8/24 (33,3%)	35/137(25,5%) 37/137 (27%) 72/137 (52,5%)	5/15 (33,3%) 2/15 (13,3%) 7/15 (46,6%)	27/89 (30,3%) 7/89 (7,9%) 34/89 (38,2%)
Πολυπύρηννα Γιγαντοκύτταρα	3/24 (12,5%)	0/137 (0%)	0/15 (0%)	29/89 (32,6%)
Οξύφιλα κύτταρα	10/24 (41,7%)	30/137 (21,9%)	1/15 (7,0%)	18/89 (20,2%)
Συνοχή Καλή Χαλαρή	7/24 (29,2%) 14/24 (58,3%)	126/137 (92%) 8/137 (5,83%)	15/15 (100%) 0/15 (0%)	14/89 (15,7%) 63/89 (70,8%)
Λεμφοκύτταρα Λίγα Πολλά Σύνολο %	7/24 (29,2%) 5/24 (20,8%) 12/24 (50,0%)	8/137 (5,8%) 20/137 (14,6%) 28/137 (20,4%)	5/15 (33,3%) 6/15 (40,0%) 11/15 (73,3%)	15/89 (16,9%) 8/89 (9,0%) 23/89 (25,9%)
Έγκλειστα	0/24 (0%)	0/137 (0%)	0/15 (0%)	17/89 (19,1%)

Πίνακας 5-7 Σύγκριση διαφόρων κυτταρολογικών χαρακτηριστικών μεταξύ Αληθώς Θετικών και Ψευδώς Θετικών/Υπόπτων

Κυτταρολογικά Χαρακτηριστικά	Αληθώς Θετικά	Ψευδώς Θετικά/Υποπτα
1. Υπόστρωμα		
Άφθονο Κολλοειδές	32,6%	12,5%
Χαμηλή Κυτταροβρίθεια	12,4%	33,3%
Ελεύθερα κύτταρα	14,8%	37,5%
2. Αρχιτεκτονικά Στοιχεία		
Θηλώδεις σχηματισμοί		
άφθονοι	24,7%	54,2%
λίγοι	33,7%	37,5%
3. Τύποι κυττάρων		
Θυλακικά κύτταρα		
Οξύφιλα κύτταρα	29,2%	62,5%
Γιγαντοκύτταρα	20,2%	41,7%
Άτυπα κύτταρα	32,6%	12,5%
Πολλά	51,7%	20,8%
Λίγα	30,3%	33,3%
4. Χαρακτηριστικά κυττάρων		
Ανισοκαρύωση	70,8%	50%
Πυκνωτικός Πυρήνας	0%	0%
Πολυμορφία Πυρήνα	0%	4,2%
Πυρήνια	0%	4,2%
Έγκλειστα	19,1%	0%

Πίνακας 5-8 Αποτελέσματα Κυτταρολογικών προς Ιστολογικές και κατανομή τους

		Κυτταρολογική Διάγνωση										
		Καλοήθεια				Νεόπλασμα	Κακοήθεια				Υποπτο	Οριακό
		1	2	3	4	5	6	7	8	9	10	11
Ιστολογική επιβεβαίωση	Πλήθος	Οζώδης Υπερπλασία	Οξύφιλος Όζος	Απόστημα	Hashimoto/Αυτοάνοση Θυρεοειδίτιδα	Θηλώδες/Οξύφιλο Νεόπλασμα	Θηλώδες Καρκίνωμα	Μυελωειδές Καρκίνωμα	Οξύφιλο Καρκίνωμα	Λέμφωμα		
Οζώδης Υπερπλασία	130	97 (ΑΑ)	10(ΑΑ)			4(ΨΥ)	1(ΨΘ)				14(ΨΥ)	4(ΨΥ)
		74,62%	7,69%			3,08%	0,77%				10,77%	3,08%
Οξύφιλος Όζος	4		3(ΑΑ)									1(ΨΥ)
			75,00%									25,00%
Απόστημα	1			1(ΑΑ)								
				100,00%								
Hashimoto/Αυτοάνοση Θυρεοειδίτιδα	13		3(ΑΑ)		9(ΑΑ)						1(ΨΥ)	
			23,08%		69,23%						7,69%	
Θηλώδες/Οξύφιλο Αδένωμα	14					14(ΑΑ)						
						100,00%						
Θηλώδες Καρκίνωμα	96	11(ΨΑ)	1(ΨΑ)		3(ΨΑ)		42(ΑΘ)				39(ΑΘ)	
		11,46%	1,04%		3,13%		43,75%				40,63%	
Μυελωειδές Καρκίνωμα	4							4(ΑΘ)				
								100,00%				
Οξύφιλο Καρκίνωμα	1								1(ΑΘ)			
									100,00%			
Λέμφωμα	3									3(ΑΘ)		
										100,00%		
Σύνολο	266	108	17	1	12	18	43	4	1	3	54	5
Σύνολο (%)		40,60%	6,39%	0,38%	4,51%	6,77%	16,17%	1,50%	0,38%	1,13%	20,30%	1,88%

ΑΑ: Αληθώς Αρνητικά, ΨΑ: Ψευδώς Αρνητικά, ΑΘ: Αληθώς Θετικά, ΨΥ: Ψευδώς Υποπτα

5.7 ΑΝΑΛΥΣΗ ΤΩΝ ΣΤΑΤΙΣΤΙΚΩΝ ΣΤΟΙΧΕΙΩΝ

Ένα από τα σημαντικότερα συμπεράσματα της μελέτης είναι το εξαιρετικά χαμηλό ποσοστό ψευδώς θετικών δειγμάτων χωρίς τη συνακόλουθη αύξηση των απροσδιόριστων δειγμάτων. Στην πραγματικότητα, μόνο ένα ψευδώς θετικό αποτέλεσμα καταγράφηκε επιβεβαιώνοντας την υψηλή θετική προβλεπτική αξία και τη διαγνωστική ακρίβεια της FNA (98,89% και 93,39%, αντίστοιχα). Αυτό οφείλεται στο γεγονός ότι ο ειδικός παθολογοανατόμος διέθετε μεγάλη εμπειρία στο αντικείμενο.

Οι δύο σημαντικοί περιορισμοί της FNA έχουν να κάνουν με τα ανεπαρκή δείγματα και τα ύποπτα ή απροσδιόριστα δείγματα. Σε αυτή τη μελέτη, τα ανεπαρκή παρουσιάστηκαν σε χαμηλό ποσοστό, 1,39% περιείχαν μόνο αίμα αποκλείοντας κάθε διάγνωση και σε ένα πρόσθετο 6,65% των περιπτώσεων η εξέταση κατέληξε σε μια διάγνωση με τον περιορισμό της ανεπάρκειας υλικού. Από τα αποτελέσματα της μελέτης προέκυψε ότι το ποσοστό των ανεπαρκών επιχρίσμάτων ήταν υψηλότερο στους άνδρες έναντι των γυναικών, 11,31% και 7,38% αντίστοιχα. Αυτή η διαφορά οφείλεται πιθανώς στο καλύτερα αναπτυγμένο μυϊκό ιστό στους άνδρες γεγονός που δυσκολεύει τη ψηλάφηση και τη βιοψία.

Ο δεύτερος σημαντικός περιορισμός της μεθόδου FNA είναι ο καθορισμός των κριτηρίων κατάταξης των επιχρίσμάτων στην "ύποπτη ή απροσδιόριστη" (grey zone) περιοχή (υπόνοια και οριακά). Η χειρουργική επέμβαση είναι απαραίτητη για τον προσδιορισμό της κατάστασης των δειγμάτων. Αυτή η κατηγορία περιλαμβάνει μια ετερογενή ομάδα επιχρίσμάτων, η μεγάλη πλειοψηφία της οποίας αντιπροσωπεύει επιχρίσματα, θυλακιώδη νεοπλασματικά και οριακά που δεν εκπληρώνουν τα κυτταρολογικά κριτήρια της κακοήθειας. Ως αποτέλεσμα, η κατηγορία των υπόπτων περιλαμβάνει όζους με διαφορετικό βαθμό υπόνοιας και διαφορετικά δεδομένα για την περαιτέρω διαχείριση του ασθενούς. Όσον αφορά επιχρίσματα με κύτταρα θυλακοειδή και Huerthle, προηγούμενες μελέτες έχουν δείξει ότι υποδιαιρούνται σε τρεις τύπους: στον τύπο I (πιθανό καρκίνωμα), τον τύπο II (ιδιαίτερα ύποπτα τραύματα) και τον τύπο III (πιθανό αδένωμα) [16]. Δεδομένου ότι η ιστολογική διάγνωση σε αυτές τις περιπτώσεις βασίζεται στην εμφάνιση διήθησης της κάψας ή/και αγγειακής διήθησης, χαρακτηριστικά μη αναγνωρίσιμα στο κυτταρολογικό επίπεδο, η κυτταρολογική διάγνωση ενός θυλακοειδούς ή Huerthle κυτταρικού νεοπλασματος αποτελεί μια ένδειξη για χειρουργική επέμβαση και δεν πρέπει να περιληφθούν αυτές οι περιπτώσεις στην "απροσδιόριστη κατηγορία". Η στατιστική ανάλυσή υποστήριξε αυτή τη διάκριση όπου "τα νεοπλάσματα" διέφεραν με ένα στατιστικά σημαντικό βαθμό βάσει συγκεκριμένων κυτταρολογικών παραμέτρων από τα οριακά επιχρίσματα. Τα κυτταρολογικά χαρακτηριστικά όπως η χαλαρή συνοχή, η απουσία μονοκυττάρων και η πυρηνική ατυπία αποτελούν παραμέτρους που συνδέθηκαν σημαντικά με τα νεοπλάσματα παρά με τα οριακά επιχρίσματα. Κατά συνέπεια, το κύριο πρόβλημα του απροσδιόριστου επιχρίσματος είναι η καλύτερη διάκριση μεταξύ των υπερπλαστικών/μεταπλαστικών επιχρίσμάτων (επιχρίσματα οριακά) και των νεοπλασματικών συνθηκών καθώς επίσης και της διάκρισης των άτυπων επιχρίσμάτων (ύποπτα για κακοήθεια).

Στο δείγμα που χρησιμοποιήθηκε σε αυτή τη μελέτη παρουσιάστηκε ένα ποσοστό υπόπτων για κακοήθεια διαγνώσεων 3,31% (300 περιπτώσεις) που είναι σημαντικά χαμηλότερο από τα προηγούμενα εμφανιζόμενα στη βιβλιογραφία (11-17%) [7],[12],[13]. Παράλληλα, το ποσοστό των πραγματικά κακοήθων περιπτώσεων μεταξύ των υπόπτων αυξήθηκε σε 75% περίπου, ποσοστό που είναι σημαντικά υψηλότερο από το κατά μέσο όρο 20% που αναφέρεται στη βιβλιογραφία. Οι ιστολογικές επαληθεύσεις ήταν διαθέσιμες σε ποσοστό περίπου 18% των υπόπτων περιπτώσεων. Έτσι μπορεί να εξαχθεί το συμπέρασμα ότι το ποσοστό των πραγματικά και ιστολογικά επιβεβαιωμένων κακοήθων περιπτώσεων μεταξύ των χαρακτηριζόμενων ως ύποπτων από την κυτταρολογική εξέταση είναι ακόμα υψηλότερο από το αναφερόμενο. Επιπλέον, ακόμη και μέσα στην υποομάδα υπόπτων για κακοήθεια υπήρξε υψηλή ετερογένεια αφού ο βαθμός βεβαιότητας κυμάνθηκε από την σχεδόν απόλυτη βεβαιότητα ύπαρξης κακοήθειας ως την αναφορά μιας εστιακής ατυπίας που χρήζει περαιτέρω αξιολόγησης. Τέλος, σε ποσοστό 12% του συνολικού αριθμού των υπόπτων περιπτώσεων, τα επιχρίσματα χαρακτηρίστηκαν πρόσθετα ως ανεπαρκή ή μη αντιπροσωπευτικά. Σε αυτές τις περιπτώσεις ενδείκνυται η επανάληψη της FNA.

Σχετικά με τα κυτταρολογικά τους χαρακτηριστικά, τα ψευδώς ύποπτα επιχρίσματα χαρακτηρίστηκαν από την υψηλή κυτταροβρίθεια, το αραιό κολλοειδές, την παρουσία οξύφιλων κυττάρων και την παρουσία λεμφοκυττάρων. Εντούτοις, όταν συγκρίθηκαν τα ψευδώς ύποπτα επιχρίσματα με τα αληθώς θετικά, τα ψευδώς ύποπτα επιχρίσματα έδειξαν σημαντικά υψηλότερα ποσοστά περιπτώσεων με άφθονο κολλοειδές ($p=0,021$), χαμηλή κυτταροβρίθεια ($p=0,021$) και ελεύθερα κύτταρα ($p=0,018$). Η παρουσία οξύφυλλων κυττάρων καταγράφηκε ως στατιστικώς σημαντική παράμετρος στις ψευδώς ύποπτες απ' ότι στις αληθώς θετικές για κακοήθεια περιπτώσεις, ενώ οι θυλώδεις σχηματισμοί, τα γιγαντιαία κύτταρα και η πυρηνική ατυπία εμφάνιζαν στατιστικά σημαντική σχέση με τα αληθώς θετικά επιχρίσματα ($p=0,006$ και $p=0,04$). Το μόνο απόλυτα ιδιαίτερο χαρακτηριστικό για τα αληθώς θετικά επιχρίσματα είναι η παρουσία εγκλείστων. Παρόλα αυτά εμφανίστηκε μόνο σε ποσοστό 20% των επιχρισμάτων.

Οι θηλακιώδεις σχηματισμοί ήταν το κυρίαρχο κυτταρολογικό χαρακτηριστικό του θυλώδους νεοπλάσματος αλλά όχι των νεοπλασμάτων Huerthle. Τα τελευταία συσχετίστηκαν συχνότερα με κυστικές αλλαγές όπως απεικονίζονται από το σημαντικά υψηλότερο ποσοστό των φαγοκύτταρων που παρατηρήθηκαν σε αυτά. Περισσότερα λεμφοκύτταρα βρέθηκαν στα θυλώδη νεοπλάσματα από τα νεοπλάσματα Huerthle. Τα οξύφιλα κύτταρα βρέθηκαν σε μεγάλο αριθμό των οξυφίλων νεοπλασμάτων αλλά και σε περίπου 20% των θυλώδων νεοπλασμάτων κάτι που αναδεικνύει τις συνδυασμένες ιστολογικές εικόνες που παρατηρούνται και δημιουργούν διαφοροδιαγνωστικά προβλήματα.

Η τελική ιστολογική διάγνωση ήταν θετική για την κακοήθεια σε 105 περιπτώσεις. 16 από αυτές ήταν ταξινομημένες κυτταρολογικά ως καλοήγη (ψευδώς αρνητικά δείγματα: 5,97%) ενώ η ιστολογική διάγνωση ήταν θηλώδης καρκίνωμα. Η πλειοψηφία αυτών των περιπτώσεων παρερμηνεύθηκε κυρίως ως θυρεοειδίτις σε συνδυασμό με οζώδη υπερπλασία.

5.8 ΒΙΒΛΙΟΓΡΑΦΙΑ

1. **Miller, T.R., Abele, J.S. & Greenspan, F.S** *Fine-needle aspiration biopsy in the management of thyroid nodules*. West J Med, 134, 198-205, (1981).
2. **Kendall, C.H.** *Fine needle aspiration of thyroid nodules: three years' experience*. J Clin Pathol, 42, 23-27, (1989).
3. **Lowhagen, T. & Linsk, J.** *Aspiration biopsy cytology of the thyroid gland*. In Clinical Aspiration Cytology (ed. J.A. Linsk, S. Franzen), Lippincott New York 1989.
4. **Caruso, D., Mazzaferri, E.L.** *Fine needle aspiration biopsy in the management of thyroid nodules*. Endocrinologist, 1, 194-202, (1991).
5. **Droese, M.,** *Aspiration Cytology of the Thyroid*. Schattauer Stuttgart., (1995)
6. **Leonard, N., Melcher, D.H.** *To operate or not to operate? The value of fine needle aspiration cytology in the assessment of thyroid swellings*. J Clin Pathol, 50, 941-943, (1997).
7. **Gharib, H., Goellner, J.R., Johnson, D.A.** *Fine-needle aspiration cytology of the thyroid. A 12-year experience with 11,000 biopsies*. Clin Lab Med, 13, 699-709, (1993).
8. **Galera-Davidson, H.** *Diagnostic problems in thyroid FNAs*. Diagn Cytopathol, 17, 422-428, (1997).
9. **Cáp, J., Ryska, A., Rehorková, P., et al.** *Sensitivity and specificity of the fine needle aspiration biopsy of the thyroid: clinical point of view*. Clin Endocrinol (Oxf), 51, 509-515, (1999).
10. **Mandreker, S.R., Nadkarni, N.S., Pinto, R.G., et al** *Role of fine needle aspiration cytology as the initial modality in the investigation of thyroid lesions*. Acta Cytol, 39, 898-904, (1995).
11. **Ravetto, C., Colombo, L. & Dottorini, M.E.** *Usefulness of fine-needle aspiration in the diagnosis of thyroid carcinoma: a retrospective study in 37,895 patients*. Cancer, 90, 357-3, (2000).
12. **Cersosimo, E., Gharib, H., Suman, V.J., et al** *"Suspicious" thyroid cytologic findings: outcome in patients without immediate surgical treatment*. Mayo Clin Proc, 68, 343-348, (1993).
13. **La Rosa, G.L., Belfiore, A., Giuffrida, D., et al** *Evaluation of the fine needle aspiration biopsy in the preoperative selection of cold thyroid nodules*. Cancer, 67, 2137-2141, (1991).
14. **Gharib, H.** *Fine-needle aspiration biopsy of thyroid nodules: advantages, limitations, and effect*. Mayo Clin Proc, 69, 44-49, (1994).

15. **Schlinkert, R.T., van Heerden, J.A., Goellner, J.R., et al** *Factors that predict malignant thyroid lesions when fine-needle aspiration is "suspicious for follicular neoplasm".* Mayo Clin Proc 72, 913-916, (1997).
16. **Willems, J.S. & Löwhagen, T.** *The role of fine-needle aspiration cytology in the management of thyroid disease.* Clin Endocrinol Metab, 10, 267-273, (1981).
17. **Cox, D. R. and Hinkley, D. V.** *Theoretical Statistics.* s.l. : Chapman and Hall, 1974. σ. 92.
18. **Chernoff, H. and E.L., Lehmann. s.l. :** *The use of maximum likelihood estimates in χ^2 tests for goodness-of-fit.* The Annals of Mathematical Statistics, 1954, Vol. 25, σ. 579–586.
19. **Hill, T. & Lewicki, P.** *Statistics Methods and Applications.* Tulsa : StatSoft, 2007.
20. **Feinstein, Alvan.** *Principles of Medical Statistics.* s.l. : Chapman & Hall, 2002.
21. **Mood, Duane C. Boes.** *Introduction to theory of Statistics.* σ. 410.

6 ΣΥΣΤΗΜΑΤΑ ΥΠΟΒΟΗΘΗΣΗΣ ΔΙΑΓΝΩΣΗΣ ΒΙΟΨΙΩΝ ΘΥΡΕΟΕΙΔΟΥΣ ΜΕ ΧΡΗΣΗ ΜΕΘΟΔΩΝ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Η ανάπτυξη ενός συστήματος υποβοήθησης ιατρικής απόφασης αποτελείται από ένα σύνολο βαθμίδων επεξεργασίας δεδομένων (Εικόνα 17). Στην **Πρώτη Βαθμίδα** πραγματοποιείται η **συλλογή δεδομένων**. Τα δεδομένα αποθηκεύονται σε μια βάση δεδομένων. Στη **Δεύτερη Βαθμίδα** πραγματοποιείται η **επιλογή δεδομένων** Εικόνα 17. Η **Τρίτη Βαθμίδα**, **εύρεσης κατάλληλων χαρακτηριστικών**, αφορά στη συρρίκνωση του μεγέθους των χαρακτηριστικών σε εκείνα που συμβάλλουν στην καλύτερη απόδοση του συστήματος. Η **Τέταρτη βαθμίδα** επεξεργασίας αποτελεί την «καρδιά» του συστήματος και είναι η **εφαρμογής ταξινόμησης με χρήση μεθόδων εξόρυξης δεδομένων** ή αλλιώς της εξόρυξης δεδομένων για την κατηγοριοποίηση των FNA δεδομένων. Στην παρούσα διδακτορική διατριβή παρουσιάστηκε ένα πλήρες σύστημα υποβοήθησης ιατρικής απόφασης (Decision Support System - DSS) με όλες τις προαναφερθείσες βαθμίδες για την ταξινόμηση επιχρισμάτων βιοψίας FNA με χρήση μεθόδων εξόρυξης δεδομένων. Δύο κατηγορίες επιλέχθηκαν για τα δείγματα FNA: καλοήθεια και κακοήθεια. Επίσης, βασικός στόχος ήταν η βελτίωση της ορθής ταξινόμησης των ύποπτων επιχρισμάτων (suspicious), για τα οποία είναι γνωστή η αδυναμία της μεθόδου FNA να τα ταξινομήσει. Το σύστημα εκπαιδεύτηκε και ελέγχθηκε σε σχέση με το δείγμα για το οποίο είχαμε ιστολογικές επιβεβαιώσεις - κατηγοριοποιήσεις (ground truth). Για περιπτώσεις οι οποίες χαρακτηρίστηκαν ως μη κακοήθεις από την FNA, και για τις οποίες δεν είχαμε ιστολογικές επιβεβαιώσεις, το δείγμα προέκυψε από την συνεκτίμηση και άλλων κλινικών, εργαστηριακών και απεικονιστικών εξετάσεων.

Εικόνα 17 Βαθμίδες Συστήματος Ιατρικής Υποβοήθησης

6.1 ΔΕΥΤΕΡΗ ΒΑΘΜΙΔΑ - ΕΠΙΛΟΓΗ ΔΕΔΟΜΕΝΩΝ

Στις εφαρμογές εξόρυξης γνώσης σημαντικό βήμα αποτελεί η προετοιμασία και η επιλογή χαρακτηριστικών μεγεθών. Η επιλογή ή προεπεξεργασία των δεδομένων μπορεί να έχει πολλές διακριτές και διαφορετικές μορφές. Έτσι μπορεί να πρόκειται για καθαρισμό δεδομένων, μετασχηματισμό,

ομαδοποίηση, αφαίρεση ασυνήθιστων τιμών (outliers), μείωση του μεγέθους των δεδομένων εν γένει, προσθήκη ή και αφαίρεση στοιχείων που λείπουν καθώς και συμπλήρωση μη υπαρχόντων τιμών.

Καθαρισμός δεδομένων μπορεί να γίνει στην περίπτωση που έχουμε δεδομένα τα οποία δεν είναι πραγματικά ή δεν περιέχουν όλα τα υποχρεωτικά πεδία για την έρευνα μας. Μετασχηματισμό μπορεί να πραγματοποιήσουμε αν θέλουμε να επεξεργαστούμε σε μια διαφορετική διάσταση ένα στοιχείο, όπως στην περίπτωση των αριθμητικών δεδομένων. Ομαδοποίηση είναι η περίπτωση όπου περισσότερα του ενός χαρακτηριστικά μπορούν να ομαδοποιηθούν και να συνθέσουν ένα νέο χαρακτηριστικό το οποίο θα μας δώσει μια ενιαία πληροφορία. Ασυνήθιστες τιμές έχουμε σε περιπτώσεις όπου οι τιμές κάποιων στοιχείων απέχουν πολύ από τις συνηθισμένες τιμές ή έχει γίνει εσφαλμένη συλλογή των δεδομένων και χρειάζεται αφαίρεση τους. Αφαίρεση χαρακτηριστικών γνωρισμάτων τα οποία δεν έχουν καμία διαφορά σε όλα τα δείγματα μειώνει τον όγκο των δεδομένων για επεξεργασία χωρίς να χαθεί η πληροφορία. [1]

6.1.1 ΠΑΡΟΥΣΙΑΣΗ ΔΕΔΟΜΕΝΩΝ

Σύμφωνα με την καταγραφή των απαιτήσεων, την εισαγωγή και τη ψηφιοποίηση των στοιχείων όπως αναφέρεται σε προηγούμενο κεφάλαιο, δημιουργήθηκαν 111 συνολικά στήλες (χαρακτηριστικά – Παράρτημα Α) για κάθε ασθενή (record). Οι περισσότερες στήλες είχαν τιμές δυαδικού τύπου, αποτυπώνοντας την ύπαρξη και την μη ύπαρξη του στοιχείου, ενώ κάποιες άλλες είχαν τιμές τύπων αριθμών ή αλφαριθμητικών χαρακτήρων. Από τις 111 επιλέχθηκαν οι 106 (30 ασθενειών, 1 δημογραφικό, 6 γενικής περιγραφής, 2 των αποτελεσμάτων και 67 χαρακτηριστικά γνωρίσματα των επιχρισμάτων).

6.1.2 ΚΑΘΑΡΙΣΜΟΣ ΔΕΔΟΜΕΝΩΝ

Για τον καθαρισμό των δεδομένων σε συνεργασία με τον ειδικό παθολογοανατόμο ιατρό, συντάχθηκαν κανόνες για τη μοναδιαία κατάταξη των δειγμάτων σε μια κατηγορία σύμφωνα με τη σημαντικότητα κάθε ασθένειας. Έτσι αρχικά αφαιρέθηκαν από το δείγμα όλα τα ανεπαρκή και κατατάχθηκαν τα δείγματα με βάση τους κανόνες που εμφανίζονται στους πίνακες Πίνακας 6-1, Πίνακας 6-2, Πίνακας 6-3. Για το σκοπό αυτό στα επαρκή δεδομένα δημιουργήθηκαν νέες στήλες στοιχείων (ομαδοποίηση), κύριας κατηγορίας (MAGOR_CATEGORIES), με δυνατότητα να έχει τιμές LEMFATIC, OTHER, NO_DIAGNOSE, THYROID, αναλυτικής κατηγορίας (ANALYTIC_CATEGORIES), με δυνατότητα να έχει τιμές LEMFATIC, OTHER, NO_DIAGNOSE, ΥΠΟΝΟΙΑ, ΟΡΙΑΚΟ, ΝΕΟΠΛΑΣΜΑ, ΚΑΛΟΗΘΕΙΑ, ΚΑΚΟΗΘΕΙΑ και πλήρους κατηγορίας (FULL_CATEGORIES). με δυνατότητα να έχει τιμές LEMFATIC, ΥΠΟΝΟΙΑ, NO_DIAGNOSE, OTHER (όλες τις περιπτώσεις), ΚΑΛΟΗΘΕΙΑ (όλες τις περιπτώσεις), ΚΑΚΟΗΘΕΙΑ (όλες τις περιπτώσεις). Έτσι λάβαμε τελικά μόνο τα θυρεοειδικά με επαρκές υλικό.

Στον παρακάτω πίνακα δίνονται οι κανόνες για τον υπολογισμό κάθε στήλης

Πίνακας 6-1 Κανόνες Υπολογισμού Κύριας Κατηγορίας (MAGOR_CATEGORIES)

ΣΤΗΛΗ	ΚΑΝΟΝΕΣ ΥΠΟΛΟΓΙΣΜΟΥ	
	ΥΠΟΘΕΣΗ	ΑΠΟΤΕΛΕΣΜΑ
Κύριας Κατηγορίας (MAGOR_CATEGORIES)	Εάν είναι Λεμφαδένας και δεν ανήκει σε καμία άλλη περίπτωση.	Είναι Λεμφαδένας (LEMFATIC).
	Εάν δεν είναι Θυρεοειδικό και δεν είναι Λεμφαδένας αλλά είναι κάτι άλλο και δεν ανήκει σε καμία άλλη περίπτωση.	Είναι Άλλο (OTHER).

Εάν δεν ανήκει σε καμία Δεν είναι διαγνώσιμο
κατηγορία ασθένειας. (NO_DIAGNOSE).

Όλα όσα δεν εμπίπτουν στις Είναι αποκλειστικά
παραπάνω κατηγορίες. θυρεοειδικά
(THYROID).

Πίνακας 6-2 Κανόνες Υπολογισμού Αναλυτικής Κατηγορίας (ANALYTIC_CATEGORIES)

ΣΤΗΛΗ	ΚΑΝΟΝΕΣ ΥΠΟΛΟΓΙΣΜΟΥ	
	ΥΠΟΘΕΣΗ	ΑΠΟΤΕΛΕΣΜΑ
Αναλυτικής Κατηγορίας (ANALYTIC_CATEGORIES)	Εάν είναι Λεμφαδένας στην κύρια κατηγορία.	Μένει Λεμφαδένας (LEMFATIC).
	Εάν είναι Άλλο στην κύρια κατηγορία.	Μένει Άλλο (OTHER).
	Εάν είναι Μη διαγνώσιμο στην κύρια κατηγορία.	Μένει μη διαγνώσιμο (NO_DIAGNOSE).
	Εάν είναι στην κύρια κατηγορία Θυρεοειδικό και χαρακτηρίζεται ως Υπόνοια Κακοήθειας.	Είναι ΥΠΟΝΟΙΑ.
	Εάν είναι στην κύρια κατηγορία Θυρεοειδικό και χαρακτηρίζεται ως Θηλώδες ή Οξύφιλο Οριακό και δεν είναι Υπόνοια Κακοήθειας ή άλλη Κακοήθεια.	Είναι ΟΡΙΑΚΟ.
	Εάν είναι στην κύρια κατηγορία Θυρεοειδικό και χαρακτηρίζεται ως Θηλώδες ή Οξύφιλο Νεόπλασμα και δεν είναι Υπόνοια Κακοήθειας, Οριακό ή άλλη Κακοήθεια.	Είναι ΝΕΟΠΛΑΣΜΑ
	Εάν είναι στην κύρια κατηγορία Θυρεοειδικό και χαρακτηρίζεται ως κάποιου τύπου Κακοήθεια και δεν χαρακτηρίζεται ως Υπόνοια.	Είναι ΚΑΚΟΗΘΕΙΑ.
	Εάν είναι στην κύρια κατηγορία Θυρεοειδικό και χαρακτηρίζεται ως κάποιου τύπου Καλοήθεια και δεν χαρακτηρίζεται ως κάποιου τύπου Κακοήθεια, Υπόνοια,	Είναι ΚΑΛΟΗΘΕΙΑ.

Πίνακας 6-3 Κανόνες Υπολογισμού Πλήρους Κατηγορίας (FULL_CATEGORIES)

ΣΤΗΛΗ	ΚΑΝΟΝΕΣ ΥΠΟΛΟΓΙΣΜΟΥ	
	ΥΠΟΘΕΣΗ	ΑΠΟΤΕΛΕΣΜΑ
Πλήρους Κατηγορίας (FULL_CATEGORIES)	Εάν είναι Λεμφαδένας στην κύρια κατηγορία.	Μένει Λεμφαδένας (LEMFATIC).
	Εάν είναι στην κύρια κατηγορία Θυρεοειδικό και χαρακτηρίζεται ως Υπόνοια Κακοήθειας.	Είναι ΥΠΟΝΟΙΑ .
	Εάν είναι μη διαγνώσιμο στην κύρια κατηγορία.	Μένει μη διαγνώσιμο (NO_DIAGNOSE).
	Εάν είναι Άλλο στην κύρια και στην αναλυτική κατηγορία και είναι χαρακτηρισμένο με κάποια από τις δυνατές περιπτώσεις του άλλου.	Τότε είναι αυτό με το οποίο είναι χαρακτηρισμένο (ΠΑΡΑΘΥΡΕΟΕΙΔΗΣ, ΛΕΜΦΑΔΕΝΑΣ, ΒΡΑΧΕΙΑΚΗ ΚΥΣΤΗ, ΘΥΡΟΓΛΩΣΣΙΚΟΥ ΠΟΡΟΥ, ΦΑΓΟΚΟΚΚΥΩΜΑ, ΜΕΣΕΓΧΥΜΑΤΙΚΟ, ΟΓΚΟΣ ΚΑΡΟΤΥΔΙΚΟΥ, ΣΙΕΛΟΓΟΝΟΣ).
	Εάν είναι στην Κύρια κατηγορία Θυρεοειδικό, στην Αναλυτική Κατηγορία Οριακό και χαρακτηρίζεται ως Θηλώδες ή Οξύφιλο Οριακό.	Τότε είναι αντίστοιχα Θηλώδες ή Οξύφιλο Οριακό .
	Εάν είναι στην Κύρια κατηγορία Θυρεοειδικό, στην Αναλυτική Κατηγορία Νεόπλασμα και χαρακτηρίζεται ως Θηλώδες ή Οξύφιλο Νεόπλασμα.	Τότε είναι αντίστοιχα Θηλώδες ή Οξύφιλο ΝΕΟΠΛΑΣΜΑ .
	Εάν είναι στην κύρια κατηγορία Θυρεοειδικό και χαρακτηρίζεται ως Θηλώδες ή Οξύφιλο Νεόπλασμα και δεν είναι Υπόνοια Κακοήθειας, Οριακό ή άλλη Κακοήθεια.	Τότε είναι ΝΕΟΠΛΑΣΜΑ .

Εάν είναι στην Κύρια Κατηγορία Θυρεοειδικό, στην Αναλυτική Κακοήθεια και χαρακτηρίζεται ως κάποια από τους τύπους Κακοήθειας (ΘΗΛΩΔΕΣ ΚΑΡΚΙΝΩΜΑ Ca, ΑΝΑΠΛΑΣΤΙΚΟ ΚΑΡΚΙΝΩΜΑ Ca, ΜΥΕΛΟΕΙΔΕΣ ΚΑΡΚΙΝΩΜΑ Ca, ΑΤΑΞΙΝΟΜΗΤΑ, ΛΕΜΦΩΜΑ, ΛΟΠΙΑ, ΜΕΤΑΣΤΑΣΗ, ΚΑΡΚΙΝΩΜΑ ΧΑΜΗΛΗΣ ΔΙΑΦΟΡΟΠΟΙΗΣΗΣ, ΘΗΛΩΔΕΣ ΚΑΡΚΙΝΩΜΑ ΠΟΙΚΙΛΙΑ).	Τότε είναι αυτού του τύπου ΚΑΚΟΗΘΕΙΑ .
--	--

Εάν είναι στην Κύρια Κατηγορία Θυρεοειδικό, στην Αναλυτική Καλοήθεια και χαρακτηρίζεται ως κάποια από τους τύπους Καλοήθειας (ΟΖΩΔΗΣ ΥΠΕΡΠΛΑΣΙΑ, ΘΥΡΕΟΕΙΔΙΤΙΔΑ HASHIMOTO, ΟΞΥΦΙΛΟΣ ΟΖΟΣ, ΣΤΟΙΧΕΙΑ ΘΥΡΕΟΕΙΔΙΤΙΔΑΣ, ΥΠΟΞΕΙΑ ΘΥΡΕΟΕΙΔΙΤΙΣ, ΜΗ ΕΙΔΙΚΗ ΚΥΣΤΙΚΗ ΘΥΡΕΟΕΙΔΙΤΙΣ, ΜΗ ΤΑΞΙΝΟΜΟΥΜΕΝΗ ΘΥΡΕΟΕΙΔΙΤΗΣ, ΑΥΤΟΑΝΟΣΗ ΘΥΡΕΟΕΙΔΟΠΑΘΕΙΑ).	Τότε είναι αυτού του τύπου ΚΑΛΟΗΘΕΙΑ .
---	--

Στην τελική επιλογή χαρακτηριστικών γνωρισμάτων διατηρήθηκαν μόνο τα 60 χαρακτηριστικά (αφαίρεση) τα οποία περιγράφουν την δομή των επιχρισμάτων και δημιουργήθηκαν άλλα 7 (ομαδοποίηση), τα οποία αφορούν ομαδοποίηση άλλων χαρακτηριστικών με ποσοτικό προσδιορισμό πολύ και λίγο και για τα οποία χρειαζόταν και μια επιπλέον στήλη για την περίπτωση της ύπαρξης τους. Τέτοιο παράδειγμα είναι το 'ΚΟΛΛΟΕΙΔΕΣ ΠΟΛΥ', 'ΚΟΛΛΟΕΙΔΕΣ ΛΙΓΟ', 'ΥΠΑΡΞΗ ΚΟΛΛΟΕΙΔΟΥΣ'.

6.2 ΤΡΙΤΗ ΒΑΘΜΙΔΑ - ΕΥΡΕΣΗ ΚΑΤΑΛΛΗΛΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ (FEATURE SELECTION)

Η τρίτη βαθμίδα επεξεργασίας στις εφαρμογές εξόρυξης δεδομένων αφορά την επιλογή των κατάλληλων χαρακτηριστικών παραμέτρων. Αυτό το στάδιο είναι ιδιαίτερα σημαντικό αφού η λάθος επιλογή χαρακτηριστικών παραμέτρων μπορεί να δώσει διαφορετικά αποτελέσματα ταξινόμησης. Επίσης μπορεί να μας οδηγήσει σε εσφαλμένα συμπεράσματα [1]-[6].

Ο στόχος της επιλογής κατάλληλων χαρακτηριστικών παραμέτρων είναι να βρεθεί ένα υποσύνολο των χαρακτηριστικών με απόδοση πρόβλεψης συγκρίσιμη ή και καλύτερη από το πλήρες σύνολο χαρακτηριστικών. Λαμβάνοντας υπόψη ένα σύνολο χαρακτηριστικών, ο αριθμός των υποσυνόλων που δημιουργούνται είναι πεπερασμένος. Το υποσύνολο χαρακτηριστικών με την καλύτερη απόδοση επιλέγεται. Εντούτοις, υπάρχουν προφανείς δυσκολίες σε αυτήν την προσέγγιση στις εξής περιπτώσεις:

- Σε μεγάλους αριθμούς χαρακτηριστικών παραμέτρων, ο αριθμός των υποσυνόλων που μπορεί να δημιουργηθεί είναι αρκετά μεγάλος,
- Στην περίπτωση που η διαδικασία πρόβλεψης είναι λανθασμένη, οπότε αναζητούμε και το σφάλμα της κάθε λύσης [6].

6.2.1 ΜΕΘΟΔΟΣ ΕΠΙΛΟΓΗΣ ΚΑΤΑΛΛΗΛΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΠΑΡΑΜΕΤΡΩΝ

Για την επιλογή χαρακτηριστικών παραμέτρων εφαρμόζονται διάφορες τεχνικές.

6.2.1.1 ΜΕΣΟΣ ΟΡΟΣ ΚΑΙ ΔΙΑΣΠΟΡΑ

Στο κλασικό στατιστικό μοντέλο, οι περιπτώσεις ενός δείγματος περιγράφονται από κάποια κατανομή. Ο μέσος όρος και η διασπορά είναι τα δύο βασικά στοιχεία τα οποία μπορούν να χρησιμοποιηθούν για να περιγράψουν την κατανομή. Έτσι, εάν η πραγματική κατανομή είναι γνωστή, τα στοιχεία του δείγματος δεν είναι απαραίτητα για την περιγραφή της και αρκούν μόνο ο μέσος όρος και η διασπορά. Μπορούμε λοιπόν να εξετάσουμε την ανεξαρτησία των χαρακτηριστικών παραμέτρων με χρήση του μέσου όρου και της διασποράς συγκρίνοντας τους μέσους όρους των χαρακτηριστικών παραμέτρων των κατηγοριών για ένα πρόβλημα ταξινόμησης. Οι επόμενες δύο εξισώσεις συνοψίζουν τη δοκιμή, όπου το se είναι το τυποποιημένο λάθος και $significance$ (sig) είναι ίσο με 2, όπου το A και το B είναι το ίδιο χαρακτηριστικό που μετريέται για την κατηγορία 1 και κατηγορία 2. Επίσης το n_1 και το n_2 είναι το πλήθος των δειγμάτων ανά κατηγορία. Εάν η παρακάτω εξίσωση ικανοποιείται, η διαφορά των μέσων όρων των χαρακτηριστικών γνωρισμάτων θεωρείται στατιστικά σημαντική.

$$se(A - B) = \sqrt{\frac{var(A)}{n_1} + \frac{var(B)}{n_2}} \text{ και } \frac{|mean(A) - mean(B)|}{se(A - B)} > sig \quad (6-1)$$

$$\text{όπου } var(A \text{ και } B) = \frac{\sum_{i=1}^n (X_i - mean(X))^2}{n}$$

Ο μέσος όρος ενός χαρακτηριστικού γνωρίσματος συγκρίνεται και στις δύο κατηγορίες χωρίς να εξεταστεί η σχέση του με τα άλλα χαρακτηριστικά γνωρίσματα. Εάν η σύγκριση αποτύχει, το χαρακτηριστικό γνώρισμα μπορεί να αγνοηθεί. Απρόσμενα, πολλά χαρακτηριστικά γνωρίσματα μπορούν να αποτύχουν σε αυτή την απλή δοκιμή [1].

6.2.1.2 ΜΕΤΡΟ ΑΠΟΣΤΑΣΗΣ

Μια εναλλακτική προσέγγιση επιλογής χαρακτηριστικών είναι η χρήση του μέτρου της. Έστω M_1 (M_2) είναι το διάνυσμα των μέσων τιμών των χαρακτηριστικών παραμέτρων για την κατηγορία 1 (2), και C_1^{-1} (C_2^{-1}) είναι το αντίστροφο της μήτρας συνδιακύμανσης για την κατηγορία 1 (2). Η μέθοδος αυτή στηρίζεται πλήρως στα στοιχεία δειγμάτων χωρίς να απαιτείται η γνώση ή η προσέγγιση της κατανομής με βάση την εξίσωση:

$$D_M = (M_1 - M_2)(C_1 + C_2)^{-1}(M_1 - M_2)^T \quad (6-2)$$

Όπου T = Transpose

Αυτό είναι ένα γενικό μέτρο της απόστασης βασισμένο στο μέσο όρο και τη συνδιακύμανση. Το πρόβλημα του να βρεθεί ένα υποσύνολο των χαρακτηριστικών παραμέτρων, μπορεί να τεθεί ως αναζήτηση των k καλύτερων χαρακτηριστικών παραμέτρων μετρούμενων με βάση το D_M . Εάν οι χαρακτηριστικές παράμετροι είναι ανεξάρτητες, τότε όλα τα μη διαγώνια στοιχεία της αντίστροφης

μήτρας συνδιακύμανσης είναι μηδέν, και για το i χαρακτηριστικό γνώρισμα οι διαγώνιες τιμές του C^{-1} είναι $1/\text{var}(i)$. Το καλύτερο σύνολο των k ανεξάρτητων χαρακτηριστικών παραμέτρων είναι τα k χαρακτηριστικά γνώρισμα με τις μεγαλύτερες τιμές $(m_1(i) - m_2(i))^2 / (\text{var}_1(i) + \text{var}_2(i))$, όπου το $m_1(i)$ είναι ο μέσος όρος της χαρακτηριστικής παραμέτρου i στην κατηγορία 1, και $\text{var}_1(i)$ είναι η διασπορά ή διακύμανση της. Ως φίλτρο χαρακτηριστικών παραμέτρων, το μέτρο της απόστασης αποτελεί μια μικρή παραλλαγή της δοκιμής στατιστικής σημαντικότητας με τη μέθοδο ανεξάρτητων χαρακτηριστικών [1].

6.2.1.3 ΧΡΗΣΗ ΚΥΡΙΩΝ ΣΤΟΙΧΕΙΩΝ

Για να μειωθεί ο αριθμός των χαρακτηριστικών, ο απλούστερος τρόπος είναι να διαγραφεί κάποιο. Η διαγραφή συντηρεί τις αρχικές τιμές των υπολοίπων στοιχείων, το οποίο είναι ιδιαίτερα σημαντικό για την μετέπειτα εφαρμογή της οποιασδήποτε αλγοριθμικής μεθόδου. Οι μέθοδοι διαγραφής είναι φίλτρα τα οποία αφήνουν τους συνδυασμούς χαρακτηριστικών παραμέτρων για τις μεθόδους πρόβλεψης, οι οποίοι είναι περισσότερο σχετικοί με τη μέτρηση του πραγματικού λάθους και είναι περιεκτικότεροι στην αναζήτηση για λύση. Μια εναλλακτική προσέγγιση είναι να μειωθούν οι διαστάσεις των χαρακτηριστικών παραμέτρων με τη συγχώνευση τους, με αποτέλεσμα τη δημιουργία ενός νέου συνόλου λιγότερων παραμέτρων με νέες τιμές. Ο προφανέστερος μετασχηματισμός είναι γραμμικός, και αυτό είναι η βάση των μεθόδων κύριων στοιχείων. Λαμβάνοντας υπόψη τις m χαρακτηριστικές παραμέτρους, μπορούν να μετασχηματιστούν σε μια ενιαία νέα χαρακτηριστική παράμετρο f' , με την απλή εφαρμογή των βαρών (w) όπως στην παρακάτω εξίσωση [1].

$$f' = \sum_{j=1}^m (w(j) \times f(j)) \quad (6-3)$$

6.2.2 ΕΠΙΠΛΕΟΥΣΑ ΠΡΟΣΘΙΑ ΑΚΟΛΟΥΘΙΑΚΑ ΜΕΤΑΒΑΛΛΟΜΕΝΗ ΕΠΙΛΟΓΗ (SEQUENTIAL FLOAT FORWARD SEARCH TECHNIQUE - SFFS)

Οι διαδοχικοί αλγόριθμοι επιλογής κατάλληλων χαρακτηριστικών παραμέτρων (Sequential Features Selection Algorithms) είναι οι πλέον χρησιμοποιημένες μέθοδοι επιλογής χαρακτηριστικών παραμέτρων λόγω της απλότητας και της αποδοτικότητας τους. Οι πιο κοινοί διαδοχικοί αλγόριθμοι επιλογής χαρακτηριστικών γνωρισμάτων είναι η πρόσθια ακολουθιακά μεταβαλλόμενη επιλογή (Forward Sequential Selection - SFS) και η ανάστροφα ακολουθιακά μεταβαλλόμενη επιλογή (Backward Sequential Selection - SBS). Η πρόσθια ακολουθιακά μεταβαλλόμενη επιλογή (SFS) είναι ένας "άπληστος" αλγόριθμος αναζήτησης. Η SFS αρχίζει με ένα κενό υποσύνολο χαρακτηριστικών παραμέτρων, και διαδοχικά προσθέτει μια χαρακτηριστική παράμετρο σε αυτό το υποσύνολο, προερχόμενη από το υπερσύνολο χαρακτηριστικών παραμέτρων, εάν αυτή η προσθήκη αυξάνει την τιμή μιας αντικειμενικής συνάρτησης μέτρησης της απόδοσης $J(Ym)$. Αυτός ο κύκλος επαναλαμβάνεται έως ότου δεν παρουσιάζεται καμία βελτίωση από την προσθήκη χαρακτηριστικών παραμέτρων στο υποσύνολο. Η SBS λειτουργεί παρόμοια αλλά με ανάστροφη πορεία. Η αναζήτηση ξεκινάει από το πλήρες υπερσύνολο X , των D χαρακτηριστικών παραμέτρων και αφαιρεί μια χαρακτηριστική παράμετρο που βελτιώνει (ή επιδεινώνει ελάχιστα) την αντικειμενική συνάρτηση μέτρησης της απόδοσης, ώστε τελικά να λάβει το υποσύνολο των χαρακτηριστικών γνωρισμάτων με την καλύτερη απόδοση. Οι ψευδοκώδικες για τις SFS και SBS παρουσιάζονται παρακάτω.

Το κύριο πλεονέκτημα των διαδοχικών αλγορίθμων είναι το σχετικά χαμηλό υπολογιστικό κόστος τους, τάξης $O(d^2)$ [4]. Μια απλούστερη προσέγγιση στην επιλογή χαρακτηριστικών παραμέτρων θα ταξινομούσε τις χαρακτηριστικές παραμέτρους και θα επέλεγε τις κορυφαίες d για να δημιουργήσει το καλύτερο υποσύνολο χαρακτηριστικών γνωρισμάτων. Εν τούτης αυτή η προσέγγιση παραβλέπει την πιθανότητα ύπαρξης χαρακτηριστικών παραμέτρων με χαμηλές τιμές αντικειμενικής συνάρτησης μέτρησης της απόδοσης, οι οποίες αποδίδουν καλύτερα σε διαδοχική ανάλυση εξαιτίας συμπληρωματικών πληροφοριών συσχέτισης με άλλες χαρακτηριστικές παραμέτρους. Επίσης το

πρόβλημα με τις διαδοχικές προσεγγίσεις είναι πως τείνουν προς τοπικά ακρότατα εξαιτίας του γεγονότος ότι δεν επαναξιολογούν την χρησιμότητα χαρακτηριστικών παραμέτρων οι οποίες έχουν προστεθεί ή αφαιρεθεί. Για να αποφευχθούν αυτά τα προβλήματα ακολουθείται η τακτική επαναλαμβανόμενης πρόσθεσης l επιλογών συνδυασμένων με αφαίρεση r επιλογών μέχρι να βρεθεί ο επιθυμητός αριθμός των χαρακτηριστικών παραμέτρων. Το μειονέκτημα αυτής της προσέγγισης είναι ότι οι τιμές για τα l και r δεν είναι γνωστές εκ των προτέρων. Οι μέθοδοι επιπλέουσας διαδοχικά μεταβαλλόμενης επιλογής βελτιώνουν αυτή την τεχνική ορίζοντας δυναμικά τις καλύτερες τιμές για τις l και r ώστε να βελτιστοποιηθεί η τιμή της αντικειμενικής συνάρτησης μέτρησης της απόδοσης $J(Y_m)$. Ο ψευδοκώδικας για την επιπλέουσα πρόσθια ακολουθιακά μεταβαλλόμενη επιλογή (SFFS) παρατίθεται και αυτός παρακάτω [3]. Η επιλεγόμενη συνάρτηση ελέγχου της απόδοσης J πρέπει να ακολουθεί την συνθήκη μονοτονικότητας, $J(Y_s) \leq J(Y_t)$ για κάθε $Y_s \supseteq Y_t$.

Ψευδοκώδικας Πρόσθιας Ακολουθιακά Μεταβαλλόμενης Επιλογής (Sequential Forward Selection)

1. **Αρχικοποίηση** συνόλου χαρακτηριστικών παραμέτρων $Y_0 = \{ \emptyset \}; m = 0$
2. **Επιλέγεται** η επόμενη καλύτερη χαρακτηριστική παράμετρος $x^+ = \arg \max_{x \notin Y_m} [J(Y_m + x)]$
3. **Ανανεώνει** το σύνολο των χαρακτηριστικών παραμέτρων $Y_{m+1} = Y_m + x^+$
4. **Όσο** το $m < d$, $m = m + 1$
5. **Μετάβαση** στο βήμα 2

Ψευδοκώδικας Ανάστροφα Ακολουθιακά Μεταβαλλόμενης Επιλογής (Sequential Backward Selection)

1. **Αρχικοποίηση** συνόλου χαρακτηριστικών παραμέτρων $Y_0 = X; m = 0$
2. **Αφαιρεί** τη χειρότερη χαρακτηριστική παράμετρο $x^- = \arg \max_{x \in Y_m} [J(Y_m - x)]$
3. **Ανανεώνει** το σύνολο των χαρακτηριστικών παραμέτρων $Y_{m+1} = Y_m - x^-$
4. **Όσο** το $m < d$, $m = m + 1$
5. **Μετάβαση** στο βήμα 2

Ψευδοκώδικας Επιπλέουσας Πρόσθιας Ακολουθιακά Μεταβαλλόμενης Επιλογής (Sequential Floating Forward Selection).

1. **Αρχικοποίηση** συνόλου χαρακτηριστικών παραμέτρων $Y_0 = \{ \emptyset \}; m = 0$
2. **Βρίσκει** την επόμενη καλύτερη χαρακτηριστική παράμετρο και **ενημερώνει** το Y_m ,

$$x^+ = \arg \max_{x \notin Y_m} [J(Y_m + x)], \quad Y_{m+1} = Y_m + x^+, \quad m = m + 1 \quad (6-4)$$

3. **Βρίσκει** τη χειρότερη χαρακτηριστική παράμετρο $x^- = \arg \max_{x \in Y_m} [J(Y_m - x)]$
4. **Αν** $J(Y_m - x^-) > J(Y_m)$, **τότε** $Y_{m+1} = Y_m - x^-$, $m = m + 1$

Μετάβαση στο βήμα 3,
αλλιώς
Μετάβαση στο βήμα 2

6.2.3 ΕΥΡΕΣΗ ΚΑΤΑΛΛΗΛΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Λόγω του αρχικά υψηλού αριθμού χαρακτηριστικών παραμέτρων (67 ανά περίπτωση), ήταν απαραίτητο να εξαλειφθούν οι χαρακτηριστικές παράμετροι που συσχετίζονταν γραμμικά ή δεν είχαν καμία διαγνωστική πληροφορία. Η μέθοδος επιλογής χαρακτηριστικών παραμέτρων εφαρμόστηκε πριν από την ταξινόμηση, με στόχο την ανεύρεση ενός υποσυνόλου των χαρακτηριστικών παραμέτρων που βελτιστοποιούν σε ακρίβεια την ταξινόμηση. Λόγω των προαναφερόμενων πλεονεκτημάτων εφαρμόστηκε η τεχνική της επιπλέονσας πρόσθια ακολουθιακά μεταβαλλόμενης επιλογής (SFFS). Η μέση απόδοση 50 επαναλήψεων του k-Κοντινότερου γείτονα αποτέλεσε τη συνάρτηση εκτίμησης λόγω της αποτελεσματικότητας και της ευκολίας του εφαρμοζόμενου αλγορίθμου. Σε κάθε επανάληψη εκτέλεσης του αλγορίθμου τυχαία επιλεγμένο ποσοστό 50% των διαθέσιμων για feature selection δειγμάτων, χρησιμοποιούνταν για δεδομένα εκπαίδευσης του αλγορίθμου k-Κοντινότερου γείτονα. Το άλλο 50% χρησιμοποιούνταν για δεδομένα ελέγχου. Η απόδοση του αλγορίθμου ήταν το ποσοστό των ορθά ταξινομημένων δειγμάτων από το σύνολο ελέγχου. Το πλήθος των κοντινότερων γειτόνων k είναι περιττός αριθμός. Στην περίπτωση μας επιλέχθηκε $k=3$. Ο αριθμός των δειγμάτων που χρησιμοποιήθηκαν είναι 2.036 (1.886 καλοήθειες και 150 κακοήθειες). Εξ αυτών, όλες οι κακοήθειες είναι ιστολογικά επιβεβαιωμένες. Επίσης 140 καλοήθειες είναι ιστολογικά επιβεβαιωμένες με επάρκεια υλικού. Οι υπόλοιπες 1.726 καλοήθειες είναι επιβεβαιωμένες με συνεκτίμηση κλινικών, εργαστηριακών και απεικονιστικών ιατρικών εξετάσεων (υπέρηχοι κ.λπ.). Από τα 2.036 δείγματα το 25% χρησιμοποιήθηκε για την επιλογή χαρακτηριστικών παραμέτρων, δηλαδή 37 περιπτώσεις της κατηγορίας 2 (Malignant - κακοήθεια) και 472 περιπτώσεις της κατηγορίας 1 (Non Malignant – καλοήθεια).

6.2.4 ΠΑΡΟΥΣΙΑΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Από την εφαρμογή της τεχνικής SFFS για την επιλογή χαρακτηριστικών βρέθηκε σαν βέλτιστο υποσύνολο χαρακτηριστικών το ακόλουθο σύνολο των 12 χαρακτηριστικών.

ΨΑΜΜΩΔΗ (PSAMMOMA), ΘΥΛΑΚΙΩΔΕΙΣ ΣΧΗΜΑΤΙΣΜΟΙ ΥΠΑΡΞΗ (FOLLICULAR FORMATIONS EXIST), ΣΥΜΠΑΓΕΙΣ ΣΧΗΜΑΤΙΣΜΟΙ (SOLID FORMATIONS), ΔΙΠΥΡΗΝΑ ΚΥΤΤΑΡΑ (BINUCLEI CELLS), ΠΟΛΥΠΥΡΗΝΑ ΚΥΤΤΑΡΑ (MULTI NUCLEI CELLS), ΑΤΥΠΑ ΚΥΤΤΑΡΑ ΥΠΑΡΞΗ (ATYPICAL CELLS EXIST), ΟΞΥΦΙΛΑ ΚΥΤΤΑΡΑ (OXYPHILIC CELLS), ΣΑΦΗ ΟΡΙΑ ΚΥΤΤΑΡΩΝ (DISTINCT BORDERS), ΑΤΡΑΚΤΟΕΙΔΕΣ ΚΥΤΤΑΡΟΠΛΑΣΜΑ (SPINDLE), ΔΙΑΥΓΕΣ ΚΥΤΤΑΡΟΠΛΑΣΜΑ (CLEAR), ΕΚΚΕΝΤΡΟΣ ΠΥΡΗΝΑΣ (ECCENTRIC), ΕΓΚΛΕΙΣΤΑ (INCLUSIONS).

6.3 ΤΕΤΑΡΤΗ ΒΑΘΜΙΔΑ - ΣΥΣΤΗΜΑ ΑΠΟΦΑΣΕΩΝ ΚΑΙ ΕΦΑΡΜΟΖΟΜΕΝΕΣ ΤΕΧΝΙΚΕΣ

Όπως προαναφέραμε η **Τέταρτη βαθμίδα** (Εικόνα 17) επεξεργασίας αποτελεί την καρδιά του συστήματος και είναι η εφαρμογή της μεθόδου Ταξινόμησης με Χρήση Μεθόδων Εξόρυξης Δεδομένων.

Τα δεδομένα που χρησιμοποιήθηκαν για την ταξινόμηση ήταν 1.414 της κατηγορίας 1 (καλοήθεια) και 113 της κατηγορίας 2 (κακοήθεια), σύνολο 1.527. Διαχωρίστηκαν σε 754 της κατηγορίας 1 και 60 της κατηγορίας 2 για το σύνολο εκπαίδευσης και σε 660 κατηγορίας 1 και 53 κατηγορίας 2 για το σύνολο ελέγχου. Μέριμνα λήφθηκε ώστε στο σύνολο ελέγχου να περιέχονται 13 FNA-“suspicious” περιπτώσεις της κατηγορίας 1 και 13 FNA-“suspicious” περιπτώσεις της κατηγορίας 2.

6.4 ΔΕΝΤΡΑ ΑΠΟΦΑΣΕΩΝ - ΘΕΩΡΗΤΙΚΗ ΠΕΡΙΓΡΑΦΗ

Ένα δέντρο απόφασης είναι ένας ταξινομητής ο οποίος εμφανίζεται σαν διαδοχική εικόνα διαχωρισμού ενός συνόλου περιπτώσεων. Το δέντρο απόφασης αποτελείται από κόμβους που διαμορφώνουν ένα **δέντρο με ρίζα**, το οποίο σημαίνει ότι είναι ένα κατευθυνόμενο δέντρο με έναν κόμβο αποκαλούμενο "ρίζα" ο οποίος δεν έχει εισερχόμενη άκρη. Όλοι οι άλλοι κόμβοι έχουν ακριβώς μια εισερχόμενη άκρη. Ένας κόμβος με εξερχόμενες άκρες ονομάζεται **εσωτερικός** ή κόμβος δοκιμής. Όλοι οι άλλοι κόμβοι ονομάζονται φύλλα (επίσης γνωστοί ως κόμβοι τερματικοί ή απόφασης). Σε ένα δέντρο

απόφασης, κάθε εσωτερικός κόμβος χωρίζει το σύνολο των περιπτώσεων σε δύο ή περισσότερα υποσύνολα σύμφωνα με μια συγκεκριμένη συνάρτηση διαχωρισμού των τιμών εισόδου των χαρακτηριστικών παραμέτρων. Στην απλούστερη και συχνότερη περίπτωση, κάθε δοκιμή εξετάζει μια χαρακτηριστική παράμετρο, έτσι ώστε το σύνολο των δεδομένων να διαχωριστεί σύμφωνα με την τιμή αυτής της χαρακτηριστικής παραμέτρου. Στην περίπτωση χαρακτηριστικών παραμέτρων αριθμητικού τύπου, ο όρος τιμή της χαρακτηριστικής παραμέτρου αναφέρεται σε ένα διάστημα τιμών. Κάθε φύλλο του δέντρου κατατάσσεται σε μια κατηγορία που αντιπροσωπεύει την πιο κατάλληλη τιμή της κατηγορίας κατάταξης. Εναλλακτικά, το φύλλο μπορεί να κρατήσει ένα διάνυσμα πιθανοτήτων που δείχνει την πιθανότητα να έχει αυτό το φύλλο τη συγκεκριμένη τιμή κατηγορίας κατάταξης. Οι περιπτώσεις του δείγματος ταξινομούνται προχωρώντας από τη ρίζα του δέντρου μέχρι το κάθε φύλλο, σύμφωνα με το αποτέλεσμα των δοκιμών κατά μήκος της διαδρομής. Στην Εικόνα 18 παρατίθεται ένα παράδειγμα δέντρου απόφασης το οποίο απεικονίζει αν ένας δυνητικός πελάτης θα αποκριθεί ή όχι σε μια στοχευόμενη ταχυδρομική αποστολή. Οι εσωτερικοί κόμβοι απεικονίζονται ως κύκλοι, ενώ τα φύλλα απεικονίζονται ως τρίγωνα [8].

Αυτό το δέντρο απόφασης ενσωματώνει τόσο ποιοτικές όσο και αριθμητικές χαρακτηριστικές παραμέτρους. Λαμβάνοντας υπόψη αυτόν τον ταξινομητή, μπορεί να προβλεφθεί η πιθανότητα απάντησης του πελάτη και να κατανοηθούν τα συμπεριφοριστικά χαρακτηριστικά όλων των δυνητικών πελατών σε σχέση με την πιθανή απάντησή τους. Κάθε κόμβος ονομάζεται με την χαρακτηριστική παράμετρο που εξετάζει και οι κλάδοι του ονομάζονται με τις αντίστοιχες τιμές του.

Εικόνα 18 Παράδειγμα Δέντρο αποφάσεων

Στην περίπτωση χαρακτηριστικών παραμέτρων αριθμητικού τύπου, τα δέντρα απόφασης μπορούν να ερμηνευθούν γεωμετρικά σαν πολυδιάστατα επίπεδα με ένα ορθογώνιο σε κάθε ένα άξονα. Γενικά, οι ειδικοί προτιμούν τα λιγότερο σύνθετα δέντρα απόφασης, δεδομένου ότι μπορούν να γίνουν ευκολότερα κατανοητά. Επιπλέον, η πολυπλοκότητα των δέντρων απόφασης έχει μια κρίσιμη επίδραση στην ακρίβειά τους [9]. Η πολυπλοκότητα των δέντρων εξαρτάται από τα χρησιμοποιούμενα κριτήρια για την παύση ανάπτυξης του δέντρου καθώς και από τη μέθοδο περικοπής του δέντρου. Συνήθως η πολυπλοκότητα των δέντρων μετριέται με ένα από τα ακόλουθα μετρικά στοιχεία: συνολικός αριθμός κόμβων, συνολικός αριθμός φύλλων, βάθος δέντρων και αριθμός χρησιμοποιούμενων χαρακτηριστικών παραμέτρων. Η δημιουργία δέντρων απόφασης συσχετίζεται άμεσα με τη δημιουργία κανόνων απόφασης. Κάθε πορεία από τη ρίζα ενός δέντρου απόφασης προς ένα από τα φύλλα του μπορεί να μετασχηματιστεί σε έναν κανόνα απόφασης, αλλά με τη διασύνδεση των δοκιμών κατά μήκος της πορείας αυτής. Η κατηγορία του φύλλου που καταλήγει η σειρά των δοκιμών είναι η κατηγορία που δίνει αυτή η απόφαση [10].

6.4.1 ΑΛΓΟΡΙΘΜΙΚΟ ΠΛΑΙΣΙΟ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΕΩΝ

Ο στόχος της εφαρμογής αλγορίθμου δέντρου απόφασης είναι να βρεθεί το βέλτιστο δέντρο απόφασης ελαχιστοποιώντας το επίπεδο λάθος ταξινόμησης από το δέντρο απόφασης. Επίσης, πέραν της ελαχιστοποίησης του λάθους υπάρχουν και άλλες παράμετροι βελτιστοποίησης, όπως η ελαχιστοποίηση

του αριθμού των κόμβων ή η ελαχιστοποίηση του μέσου βάθους του δέντρου. Η χρησιμοποίηση των βέλτιστων αλγορίθμων δέντρων απόφασης είναι εφικτή μόνο σε μικρά προβλήματα [11], [12], [13]. Οι μέθοδοι ανάπτυξης δέντρων απόφασης μπορούν να διαιρεθούν σε δύο ομάδες: επίλυσης από επάνω προς τα κάτω (top-down) και από κάτω προς τα επάνω (bottom-up), με την πρώτη ομάδα να έχει ευρεία εφαρμογή σε έρευνες. Υπάρχουν διάφορες παραλλαγές των μεθόδων από επάνω προς τα κάτω (top-down) για την παραγωγή δέντρων απόφασης όπως επαναληπτικός διχοτομητής ID3 [9], C4.5, επέκταση του ID3 [10], ταξινόμησης και παλινδρόμησης (CART) [9]. Μερικές αποτελούνται από δύο εννοιολογικές φάσεις: ανάπτυξη και περικοπή (C4.5 και CART). Άλλες μέθοδοι εκτελούν μόνο τη φάση της ανάπτυξης. Παρακάτω παρουσιάζεται σε ψευδοκώδικα το αλγοριθμικό πλαίσιο για την από επάνω προς τα κάτω μέθοδο που χρησιμοποιεί την ανάπτυξη και την περικοπή.

Οι παραπάνω αναφερόμενοι αλγόριθμοι χαρακτηρίζονται ως άπληστοι και κατασκευάζουν το δέντρο απόφασης, με επαναληπτικό τρόπο της γνωστής μεθόδου "διαίρει και βασίλευε". Σε κάθε επανάληψη, ο αλγόριθμος εξετάζει τη διαίρεση του συνόλου εκπαίδευσης χρησιμοποιώντας το αποτέλεσμα της συνάρτησης διαχωρισμού των χαρακτηριστικών παραμέτρων εισόδου. Η επιλογή της καταλληλότερης συνάρτησης διαχωρισμού γίνεται σύμφωνα με ορισμένα μετρικά διαχωρισμού (splitting measures). Μετά από την επιλογή ενός κατάλληλου μετρικού διαχωρισμού, κάθε κόμβος υποδιαιρεί περαιτέρω το σύνολο εκπαίδευσης σε μικρότερα υποσύνολα, έως ότου να μην υπολογίζεται ικανή τιμή του κριτηρίου διαχωρισμού ή να ικανοποιείται το κριτήριο παύσης.

Ψευδοκώδικας

Δεδομένης μιας τυχαίας μεταβλητής x με k διακριτές τιμές, κατανεμημένες σύμφωνα με το $P = (p_1, p_2, \dots, p_k)$, ένα κριτήριο μη καθαρότητας είναι μια συνάρτηση $\phi: [0,1]^k \rightarrow \mathbf{R}$, η οποία ικανοποιεί τις ακόλουθες συνθήκες:

$$\phi(P) \geq 0.$$

$\phi(P)$ είναι ελάχιστη εάν υπάρχει i τέτοιο ώστε $p_i = 1$.

$\phi(P)$ είναι μέγιστη εάν για κάθε i , $1 \leq i \leq k$, $p_i = 1/k$.

$\phi(P)$ είναι συμμετρική σε σχέση με τους παράγοντες του P .

$\phi(P)$ είναι ομαλή σε όλο το εύρος.

Ο ψευδοκώδικας που αποδίδει την λειτουργία του αλγορίθμου Δέντρων αποφάσεων έχει ως εξής [7]:

Ανάπτυξη Δέντρου αποφάσεων (S,A,y)

Όπου :

S – Δεδομένα Εκπαίδευσης

A – Χαρακτηριστικές παράμετροι εισόδου

y – Διάνυσμα χαρακτηριστικών παραμέτρων στόχου

Δημιουργεί ένα νέο δέντρο T με ένα κόμβο ρίζα

Εάν ένα από τα κριτήρια παύσης (Stopping Criteria) πληρείται **τότε**

Σημαδεύει το κόμβο ρίζα στο δέντρο T σαν φύλλο με την ποιο συχνή τιμή του y στο S σαν ετικέτα

Αλλιώς

Βρίσκει μια συνάρτηση διάκρισης $f(A)$ των τιμών των χαρακτηριστικών παραμέτρων εισόδου ώστε η διάσπαση του S σύμφωνα με το αποτέλεσμα της $f(A)$ ($v_1, \dots, v_n$) να δίνει την καλύτερη τιμή για το μετρικό διάσπασης.

Αν το καλύτερο μετρικό διάσπασης $>$ ενός ορίου **τότε**

Θέτει t μαζί με $f(A)$

Για κάθε αποτέλεσμα v_i από $f(A)$:

Θέτει υποδέντρο $_i$ = Δημιουργία Δέντρου ($\sigma_{f(A)=v_i}, S, A, y$)

Συνδέει τη ρίζα του t_T στο

*Σημαδεύει το κόμβο ρίζα του T σαν φύλλο με συχνότερη τιμή του y στο S σαν
 ετικέτα
 Τέλος Αλλιώς
 Τέλος Αλλιώς
 Επανάληψη του T*

Περικοπή δέντρου αποφάσεων(S,T,y)

S – Δεδομένα Εκπαίδευσης

y – Διάνυσμα χαρακτηριστικών παραμέτρων στόχου

T-Δέντρο που θα περικοπεί

Κάνει

Διαλέγει ένα κόμβο t από το δέντρο T τέτοιου ώστε η περικοπή του να βελτιώνει στο μέγιστο βαθμό κάποια κριτήρια αξιολόγησης

Αν $t \neq \emptyset$ τότε T=αποκοπή(T,t)

Έως $t \neq \emptyset$

Δίνει το T μετά την αποκοπή

Σημειώνεται ότι εάν το διάνυσμα πιθανότητας έχει ένα στοιχείο (η μεταβλητή x παίρνει μόνο μια τιμή), τότε η μεταβλητή χαρακτηρίζεται ως καθαρή. Αντίθετα, εάν όλα τα στοιχεία είναι ίσα, το επίπεδο μη καθαρότητας φθάνει στο μέγιστο. Λαμβάνοντας υπόψη ένα σύνολο εκπαίδευσης S, το διάνυσμα πιθανότητας των χαρακτηριστικών παραμέτρων στόχων y ορίζεται ως [8]:

$$Py(S) = \left(\frac{|\sigma_{y=c1}S|}{|S|}, \dots, \frac{|\sigma_{y=c|\text{dom}(y)|}S|}{|S|} \right) \quad (6-5)$$

Η ποιότητα διαχωρισμού η οποία εξαρτάται από τη χαρακτηριστική παράμετρο a_i καθορίζεται σαν μείωση στην μη καθαρότητα της χαρακτηριστικής παραμέτρου στόχου αφού διαχωριστεί το S σύμφωνα με τις τιμές $v_{i,j} \in \text{dom}(a_i)$

$$\Delta\Phi(a_i, S) = \Phi(Py(S)) - \sum_{j=1}^{|\text{dom}(a_i)|} \frac{|\sigma_{a_i=u_{i,j}}S|}{|S|} * \Phi(Py(\sigma_{a_i=u_{i,j}}S)) \quad (6-6)$$

Για την εφαρμογή των δέντρων απόφασης χρησιμοποιούνται πολλά και διαφορετικά κριτήρια διαχωρισμού. Μια κατηγορία είναι τα μεταβλητά κριτήρια διαχωρισμού. Σύμφωνα με αυτά ένας εσωτερικός κόμβος χωρίζεται ανάλογα με την αξία μιας ενιαίας ιδιότητας. Συνεπώς, ο αλγόριθμος ψάχνει την καλύτερη ιδιότητα διαχωρισμού. Υπάρχουν διάφορα μεταβλητά κριτήρια διαχωρισμού. Σύμφωνα με την προέλευση του μέτρου: θεωρία πληροφοριών (information theory), εξάρτηση (dependence) και απόσταση (distance). Σύμφωνα με τη δομή μέτρου: βασισμένα στη μη καθαρότητα κριτηρίων (impurity based criteria), ομαλοποιημένα βασισμένα στη μη καθαρότητα κριτηρίων (normalized impurity based criteria), δείκτης gini (gini index) και δυαδικά κριτήρια (binary criteria).

Άλλος ένας σημαντικός παράγοντας για την εφαρμογή των δέντρων αποφάσεων, όπως αναφέρεται και στον ψευδοκώδικα είναι το κριτήριο παύσης. Η φάση ανάπτυξης του δέντρου συνεχίζεται μέχρι ένα κριτήριο παύσης να ικανοποιηθεί. Τέτοια κριτήρια είναι:

1. Όλες οι περιπτώσεις στο σύνολο εκπαίδευσης να ανήκουν σε μια ενιαία κατηγορία.
2. Το μέγιστο βάθος του δέντρου να έχει επιτευχθεί.
3. Ο αριθμός καταταγμένων δειγμάτων στον τελικό κόμβο να είναι μικρότερος από τον ελάχιστο αριθμό καταταγμένων δειγμάτων για τους γόνους κόμβους.
4. Εάν ο κόμβος χωρίζεται, ο αριθμός καταταγμένων δειγμάτων σε έναν ή περισσότερους κόμβους παιδιά να είναι μικρότερος από τον ελάχιστο αριθμό καταταγμένων δειγμάτων για τους κόμβους παιδιά.

5. Τα καλύτερα κριτήρια διαχωρισμού δεν είναι μεγαλύτερα από ένα ορισμένο κατώτατο όριο.

Η υιοθέτηση σκληρών κριτηρίων παύσης τείνει να δημιουργήσει μικρά και κακής ποιότητας δέντρα αποφάσεων. Σε αντίθεση, η χρησιμοποίηση χαλαρών κριτηρίων παύσης τείνει να παράγει μεγάλα δέντρα απόφασης που είναι υπερταξινομημένα στο σύνολο εκπαίδευσης. Οι μέθοδοι «κλαδέματος» αρχικά προτάθηκαν και αναπτύχθηκαν για την επίλυση του προβλήματος της υπερταξινόμησης (overfitting) [9]. Σύμφωνα με αυτήν την μεθοδολογία, ένα χαλαρό κριτήριο παύσης χρησιμοποιείται, αφήνοντας το δέντρο απόφασης να υπερταξινομήσει το σύνολο εκπαίδευσης. Κατόπιν το δέντρο κόβεται σε ένα μικρότερο δέντρο με την αφαίρεση των κλάδων οι οποίοι δεν συμβάλλουν στην ακρίβεια κατά τη γενικευμένη χρήση του δέντρου απόφασης. Σε διάφορες μελέτες υποστηρίζεται ότι η υιοθέτηση μεθόδων «κλαδέματος» μπορεί να βελτιώσει την ακρίβεια στη γενικευμένη εφαρμογή του δέντρου απόφασης, ειδικά στις περιπτώσεις με παρουσία θορύβου. Υπάρχουν διάφορες τεχνικές για το «κλάδεμα» των δέντρων απόφασης. Οι περισσότερες από αυτές εκτελούν από επάνω προς τα κάτω ή από κάτω προς τα επάνω περικοπή των κόμβων.

6.4.2 ΠΑΡΟΥΣΙΑΣΗ ΠΕΙΡΑΜΑΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ

Στα πλαίσια της παρούσας διδακτορικής διατριβής εξετάστηκε η απόδοση αλγοριθμικών μεθόδων δέντρων αποφάσεων εφαρμοζόμενων στο πρόβλημα ανεύρεσης μοντέλων ασθενών με κακοήθεια θυρεοειδούς. Έγιναν δοκιμές με τα στοιχεία όπως αυτά προέκυψαν από την εφαρμογή της μεθόδου επιλογής χαρακτηριστικών παραμέτρων (feature selection) SFFS. Το γεγονός ότι οι δύο κατηγορίες των δεδομένων είναι ανισοπληθείς, δηλαδή έχουμε περισσότερες καλοήθειες σε σχέση με κακοήθειες, θα μπορούσε να επηρεάζει την απόδοση των δέντρων απόφασης. Έτσι έγινε τυχαία επιλογή των καλοηθειών ώστε σε κάποιες από τις δοκιμές να επιλέγεται μέρος αυτών. Οι αναλογίες που δοκιμάστηκαν είναι: όλα τα στοιχεία εκπαίδευσης, στοιχεία εκπαίδευσης με αναλογία κακοήθειες προς καλοήθειες 1:1, 1:2 και 1:3.

6.4.3 ΑΠΟΤΕΛΕΣΜΑΤΑ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ

Η υλοποίηση πραγματοποιήθηκε στο MATLAB και για να επιτευχθεί αυτό αρχικά αναπτύχθηκε μια συνάρτηση, η οποία διαχωρίζει τα δεδομένα σε σύνολα εκπαίδευσης και ελέγχου και εφαρμόζει τις αναγκαίες αναλογίες για καλοήθειες και κακοήθειες. Έχει σαν όρισμα της τα στοιχεία εισόδου, την αναλογία συνόλων εκπαίδευσης και ελέγχου και την αναλογία καλοηθειών και κακοηθειών.

Η συνάρτηση classregtree που χρησιμοποιήθηκε έχει και κάποιες μη υποχρεωτικές παραμέτρους όπως: α) την 'categorical': Διάνυσμα από δείκτες στηλών οι οποίες πρέπει να χρησιμοποιηθούν σαν αταξινόμητες κατηγορικές μεταβλητές. β) Τη 'method': Έχει δύο τιμές 'classification' (είναι δεδομένη αν το y είναι κείμενο ή κατηγορική μεταβλητή όπως στην περίπτωση μας) ή 'regression' αν είναι αριθμητική μεταβλητή. γ) Τη 'names': Ένας πίνακας με τα ονόματα των χαρακτηριστικών παραμέτρων (features) τα οποία χρησιμοποιούνται για να εμφανίζονται στην απεικόνιση του δέντρου. δ) Την 'prune': 'on' (είναι το εξ ορισμού) για το κλάδεμα του δέντρου ή 'off' για τον υπολογισμό του πλήρους δέντρου. ε) Τη 'splitmin': Ο ελάχιστος αριθμός k δειγμάτων, που κάθε κόμβος πρέπει να έχει ώστε να μπορεί να διασπαστεί (τυπική τιμή είναι 10). ζ) Τη 'splitcriterion': Κριτήριο για την επιλογή του διαχωρισμού με επιλογές 'gdi': για το Gini's diversity index (εξ ορισμού), 'twoing': για τον κανόνα, 'deviance': για την μέγιστης κλίσης μείωση (maximum deviance reduction).

Μετά εφαρμόστηκαν οι συναρτήσεις αξιολόγησης (eval) και «κλαδέματος» δέντρου (prune) του MATLAB, καθώς και οι κατάλληλοι μαθηματικοί υπολογισμοί και υπολογίστηκε η ειδικότητα, η ευαισθησία και η συνολική απόδοση. Για να επιτευχθεί η καλύτερη γενίκευση των αποτελεσμάτων, ο αλγόριθμος έτρεξε 10 φορές με τυχαία διαφορετικά δείγματα και λάβαμε τη μέση τιμή των 10 αποτελεσμάτων. Το βέλτιστο αποτέλεσμα εμφανίστηκε, με χρήση κριτηρίου διαχωρισμού twoing και με χρήση όλων των στοιχείων κακοηθειών και καλοηθειών. Τα αποτελέσματα ήταν ευαισθησία 83,4%, ειδικότητα 99,2% και συνολική απόδοση 98%.

Το καλύτερο δέντρο το οποίο παράχθηκε από τη διαδικασία εκπαίδευσης παρουσιάζεται στην Εικόνα 19.

Τα δείγματα του συνόλου δοκιμών κατηγοριοποιήθηκαν με βάση τους κανόνες αυτού του δέντρου. Το σύμβολο <0.5 σημαίνει ότι η παράμετρος έχει τιμή ίση με το μηδέν και δηλώνει μη ύπαρξη της παραμέτρου σε αυτά τα δείγματα.

Εικόνα 19 Τυπικό Δέντρο απόφασης για ταξινόμηση

Ο Πίνακας 6-4 είναι μια τυπική περίπτωση πίνακα σύγχυσης ο οποίος σαν στήλες έχει τις δύο κατηγορίες «καλοήγη» και «κακοήγη» και το πλήθος των περιπτώσεων που έχει καταναίμει σε αυτές η FNA. Ενώ στις γραμμές του παρουσιάζεται το πλήθος των περιπτώσεων όπως τις έχει καταναίμει το δέντρο απόφασης.

Πίνακας 6-4

**Πίνακας Σύγχυσης (Confusion Matrix)
Για Μέσο Όρο 10 Ανεξάρτητων Επαναλήψεων**

Μέθοδος	Πραγματική Κλάση	Προβλεπόμενη κλάση	
		Class 1	Class 2
Twoing	Class 1	655	5
	Class 2	9	44

Ο Πίνακας 6-5 προκύπτει έμμεσα από τον προηγούμενο και παρουσιάζει την ευαισθησία 83,4% (sensitivity - fraction of true positives), την ειδικότητα 99,2% (specificity - fraction of true negatives) και την συνολική ακρίβεια 98%.

Πίνακας 6-5

**Μέση Ακρίβεια Ταξινόμησης
Για 10 Ανεξάρτητες Επαναλήψεις Του Αλγορίθμου**

Μέθοδος	Ευαισθησία (%)	Ειδικότητα (%)	Συνολική Ακρίβεια (%)
Twoing	83,4	99,2	98

Οι επόμενοι δύο πίνακες είναι ο πίνακας σύγχυσης (confusion matrix) και ο πίνακας των τιμών για την ευαισθησία, την ειδικότητα και τη συνολική ακρίβεια για το τμήμα των δεδομένων τα οποία χαρακτηρίζονται ως ύποπτα (suspicious).

Πίνακας 6-6

**Πίνακας Σύγχυσης (Confusion Matrix)
Για Τα Δείγματα Τα Οποία Χαρακτηρίζονται Ύποπτα,
Μέσος Όρος Για 10 Ανεξάρτητες Επαναλήψεις**

Μέθοδος	Πραγματική Κλάση	Προβλεπόμενη κλάση	
		Class 1	Class 2
Twoing	Class 1	8	5
	Class 2	5	8

Πίνακας 6-7
Ακρίβεια Ταξινόμησης
Για Τα Δείγματα Τα Οποία Χαρακτηρίζονται Ύποπτα,
Μέσος Όρος Για 10 Ανεξάρτητες Επαναλήψεις

Μέθοδος	Ευαισθησία (%)	Ειδικότητα (%)	Συνολική Ακρίβεια (%)
Twoing	60,8	62,3	61,5

Η συνάρτηση μεταφοράς (ROC) [59] για το δέντρο αποφάσεων παρουσιάζεται στην Εικόνα 20 όπου και αναδεικνύεται η μειωμένη ευαισθησία της μεθόδου.

Εικόνα 20 Συνάρτηση μεταφοράς (ROC) δέντρου απόφασης

6.5 ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ - ΘΕΩΡΗΤΙΚΗ ΠΕΡΙΓΡΑΦΗ

Τα νευρωνικά δίκτυα αποτελούνται από απλά στοιχεία (νευρώνες) που λειτουργούν παράλληλα και σε πολλά επίπεδα (layers). Αναπτύχθηκαν με βάση το βιολογικό νευρικό σύστημα. Όπως στη φύση έτσι και στην αλγοριθμική εφαρμογή η λειτουργία των νευρωνικών δικτύων καθορίζεται κατά ένα μεγάλο μέρος από τις συνδέσεις μεταξύ των στοιχείων. Ένα νευρωνικό δίκτυο μπορεί να εκπαιδευτεί ώστε να εκτελεί μια λειτουργία, ρυθμίζοντας τις τιμές των συνδέσεων (βαρών) μεταξύ των στοιχείων (νευρώνων). Συνήθως τα νευρωνικά δίκτυα εκπαιδεύονται ώστε μια συγκεκριμένη είσοδος να οδηγεί σε μια συγκεκριμένη απόκριση. Κάτι τέτοιο αποτυπώνεται στην Εικόνα 21. Το δίκτυο ρυθμίζεται, βασιζόμενο σε μια σύγκριση της εξόδου και της τιμής στόχου, μέχρι η έξοδος να γίνει ίση με την τιμή στόχου. Γενικά πολλά τέτοια ζεύγη εξόδου/στόχου χρησιμοποιούνται σε αυτή την επιβλεπόμενη εκπαίδευση του νευρωνικού δικτύου.

Δύο είναι οι μέθοδοι κατάρτισης των νευρωνικών δικτύων. Η μαζική (batch) κατάρτιση πραγματοποιεί αλλαγές στα βάρη και στις πολώσεις βασιζόμενη σε ένα ολόκληρο σύνολο (batch) διανυσμάτων εισόδου. Αντίθετα η επαυξητική (incremental) κατάρτιση αλλάζει τα βάρη και τις πολώσεις ενός δικτύου κάθε φορά που δέχεται μια νέα μεμονωμένη είσοδο. Η επαυξητική κατάρτιση αναφέρεται μερικές φορές ως "on line" ή "προσαρμοστική" (adaptive) κατάρτιση. Ο τομέας των νευρωνικών δικτύων έχει ιστορία

περίπου πέντε δεκαετιών αλλά έχει βρει εφαρμογές μόνο τα τελευταία δεκαπέντε χρόνια, και συνεχίζει ακόμα να αναπτύσσεται ταχύτατα [20]-[22].

Εικόνα 21 Σχηματικό διάγραμμα λειτουργίας νευρωνικού δικτύου

6.5.1 ΑΡΧΙΤΕΚΤΟΝΙΚΗ - ΜΟΝΤΕΛΟ ΝΕΥΡΩΝΑ

Η αρχιτεκτονική ενός βασικού νευρώνα ο οποίος χρησιμοποιείται σε αλγορίθμους οπισθοδιάδοσης (backpropagation) και σε πολυεπίπεδα νευρωνικά δίκτυα πρόσθιας τροφοδότησης (feedforward) παρουσιάζεται παρακάτω.

Ένα βασικός νευρώνας με R εισόδους φαίνεται στην Εικόνα 22. Κάθε είσοδος p σταθμίζεται με ένα κατάλληλο βάρος w . Το άθροισμα των σταθμισμένων εισόδων και η πόλωση b καθορίζουν την είσοδο της συνάρτησης μεταφοράς. Οι νευρώνες μπορούν να χρησιμοποιούν κάθε συνάρτηση μεταφοράς

Εικόνα 22 Βασικός νευρώνας με R εισόδους

f όπως την λογαριθμική-σιγμοειδή (logsig), την εφαπτομένη-σιγμοειδή (tansig), τη γραμμική συνάρτηση μεταφοράς purelin [22]. Εάν το τελευταίο επίπεδο ενός πολυεπίπεδου δικτύου έχει σιγμοειδή νευρώνα, τότε οι έξοδοι του δικτύου περιορίζονται σε μικρότερο εύρος τιμών. Εάν είναι γραμμικός ο νευρώνας εξόδου, τότε η έξοδος μπορεί να πάρει κάθε τιμή.

6.5.2 ΠΡΟΣΘΙΑΣ ΤΡΟΦΟΔΟΣΙΑΣ (FEEDFORWARD) ΔΙΚΤΥΑ

Ένα μονοεπίπεδο νευρωνικό δίκτυο από S logsig νευρώνες το οποίο έχει R εισόδους φαίνεται λεπτομερώς στην Εικόνα 23 μαζί με ένα διάγραμμα επιπέδου [21],[22].

Εικόνα 23 Σχηματική παράσταση δικτύου μονοεπίπεδου νευρωνικού δικτύου πρόσθιας τροφοδοσίας (feedforward)

Τα Πρόσθιας Τροφοδοσίας (feedforward) νευρωνικά δίκτυα έχουν συχνά ένα ή περισσότερα κρυμμένα επίπεδα σιγμοειδών νευρώνων, τα οποία ακολουθούνται από ένα επίπεδο εξόδου γραμμικών νευρώνων (Εικόνα 24). Τα πολλά επίπεδα νευρώνων με μη γραμμικές συναρτήσεις μεταφοράς επιτρέπουν στο δίκτυο να εκπαιδευτεί τόσο με μη γραμμικές όσο και με γραμμικές συσχετίσεις μεταξύ διανυσμάτων εισόδου και διανυσμάτων εξόδου. Το γραμμικό επίπεδο εξόδου επιτρέπει στο δίκτυο να παράγει τιμές πέραν του διαστήματος -1 έως +1. Από την άλλη, εάν απαιτείται να περιοριστούν οι εξόδοι ενός δικτύου (π.χ. μεταξύ 0 και 1), τότε το επίπεδο εξόδου θα πρέπει να χρησιμοποιήσει μια σιγμοειδή συνάρτηση μεταφοράς, όπως η logsig [22].

Εικόνα 24 Σχηματική παράσταση δικτύου πολλαπλών επιπέδων πρόσθιας τροφοδοσίας (feedforward)

Αυτό το δίκτυο μπορεί να χρησιμοποιηθεί σαν μια γενική συνάρτηση προσέγγισης. Μπορεί να προσεγγίσει αρκετά καλά κάθε συνάρτηση με πεπερασμένο αριθμό ασυνεχειών, δεδομένου ότι στα κρυμμένα επίπεδα υπάρχει επαρκής αριθμός νευρώνων.

6.5.2.1 ΕΚΠΑΙΔΕΥΣΗ ΔΙΚΤΥΟΥ

Κατά τη διάρκεια της εκπαίδευσης τα βάρη και οι πολώσεις του δικτύου τροποποιούνται διαρκώς ώστε να βελτιώσουν την απόδοση του δικτύου ή με άλλα λόγια να μειώσουν το σφάλμα που παρουσιάζεται. Όλοι αυτοί οι αλγόριθμοι χρησιμοποιούν την κλίση (gradient) της συνάρτησης απόδοσης για να αποφασίσουν πως θα τροποποιήσουν τα βάρη και τις πολώσεις, ώστε τελικά να βελτιώσουν την απόδοση του αλγορίθμου. Η κλίση (gradient) υπολογίζεται χρησιμοποιώντας την τεχνική *backpropagation*, η οποία περιλαμβάνει υπολογισμούς απόδοσης με βήματα προς τα πίσω και εντός του δικτύου.

6.5.2.2 ΑΛΓΟΡΙΘΜΟΣ ΟΠΙΣΘΟΔΙΑΔΟΣΗΣ (BACKPROPAGATION)

Υπάρχουν πολλές παραλλαγές του αλγορίθμου οπισθοδιάδοσης (*backpropagation*) [22]. Η απλούστερη υλοποίηση της εκπαίδευσης τύπου *backpropagation* ενημερώνει και μεταβάλλει τα βάρη και τις πολώσεις

του δικτύου έτσι ώστε η συνάρτηση απόδοσης (ή το σφάλμα) να μειώνεται γρήγορα προς τα αρνητικά της κλίσης της. Μια σχέση η οποία αποτυπώνει την επανάληψη αυτού του αλγορίθμου μπορεί να γραφεί ως εξής:

$$x_{k+1} = x_k - a_k g_k \quad (6-7)$$

Όπου x_k είναι ένα διάνυσμα των τρεχόντων βαρών και πολώσεων, g_k είναι η τρέχουσα κλίση και a_k είναι ο ρυθμός εκμάθησης.

Υπάρχουν δύο τρόποι με τους οποίους ο προαναφερόμενος αλγόριθμος μπορεί να υλοποιηθεί: σταδιακά (online) και μαζικά (batch). Στη σταδιακή εκδοχή η κλίση υπολογίζεται και τα βάρη ενημερώνονται μετά από κάθε είσοδο. Στη μαζική εκδοχή όλες οι εισοδοί δίνονται στο δίκτυο πριν από την ενημέρωση των βαρών. Υπάρχουν διάφορες τεχνικές και μέθοδοι για τον υπολογισμό των βαρών και των πολώσεων. Ενδεικτικά αναφέρουμε την μείωση κλίσης batch (traingd) και την μείωση κλίσης batch με momentum (traingdm). Στη βιβλιογραφία αναφέρονται και άλλες μέθοδοι backpropagation με ταχύτερα αποτελέσματα όπως, μεταβαλλόμενου ρυθμού εκπαίδευσης (variable learning rate - traingda, traingdx), ελαστικής backpropagation (trainrp), συζευγμένης βαθμίδωσης (conjugate gradient algorithms), fletcher-reeves update (traincgf), polak-ribière update (traincgp), powell-beale restarts (traincgb), scaled conjugate gradient (trainscg), bfgs (trainbfg), αλγόριθμος τέμνουσας ενός βήματος (one step secant algorithm - trainoss), levenberg-marquardt (trainlm), μειωμένης μνήμης levenberg-marquardt (trainlm) [22].

6.5.3 ΠΑΡΟΥΣΙΑΣΗ ΠΕΙΡΑΜΑΤΩΝ ΝΕΥΡΩΝΙΚΩΝ ΔΙΚΤΥΩΝ

Στα πλαίσια της παρούσας διδακτορικής διατριβής εξετάστηκε η απόδοση της εφαρμογής αλγοριθμικών μεθόδων νευρωνικών δικτύων στο πρόβλημα ανεύρεσης μοντέλων ασθενών με κακοήθεια θυρεοειδούς. Στις δοκιμές των νευρωνικών δικτύων έγιναν οι εξής επιλογές: Τα δίκτυα που δημιουργήθηκαν με τη βοήθεια της συνάρτησης newff του MATLAB, ήταν πρόσθιας τροφοδότησης με οπισθοδιάδοση (feed-forward backpropagation) τριών επιπέδων, με επίπεδο εισόδου 12 νευρώνων, κρυμμένο επίπεδο 10, 20, 40, 80 και 100 νευρώνων και επίπεδο εξόδου 2 νευρώνων. Έγιναν δοκιμές με τα στοιχεία όπως αυτά προέκυψαν από την μέθοδο επιλογής χαρακτηριστικών (feature selection) SFFS. Το γεγονός ότι οι δύο κατηγορίες των δεδομένων είναι ανισοπληθείς δηλαδή περισσότερες καλοήθειες σε σχέση με κακοήθειες, θα μπορούσε να επηρεάζει την απόδοση των νευρωνικών δικτύων. Έτσι πραγματοποιήθηκε τυχαία επιλογή των καλοηθειών ώστε σε κάποιες από τις δοκιμές να επιλέγεται μέρος αυτών. Οι αναλογίες που δοκιμάστηκαν είναι: όλα τα στοιχεία εκπαίδευσης (100%), στοιχεία εκπαίδευσης με αναλογία κακοήθειες προς καλοήθειες 1:1, 1:2 και 1:3. Αξιόλογα αποτελέσματα βρέθηκαν στην περίπτωση χρήσης όλων των στοιχείων κάτι που δείχνει ότι τα νευρωνικά δίκτυα δεν επηρεάζονται αναγκαίως από ανισοπληθείς κατηγορίες. Η συνάρτηση εκπαίδευσης βαρών και πολώσεων έμεινε σταθερή σε όλες τις περιπτώσεις όπως η αρχική learngdm. Ως συναρτήσεις εκπαίδευσης χρησιμοποιήθηκαν όλες οι δυνατές από τις υπάρχουσες συναρτήσεις traingd, traingdm, traingda, traingdx, trainrp, traincgf, traincgp, traincgb, trainscg, trainbfg, trainlm, trainoss. Επίσης ως συναρτήσεις μεταφοράς χρησιμοποιήθηκαν οι αρχικές tansig για τα κρυμμένα επίπεδα και purelin για το επίπεδο εξόδου.

Οι χρησιμοποιούμενες παράμετροι συνοψίζονται στον παρακάτω πίνακα

Πίνακας 6-8
Παράμετροι Που Χρησιμοποιήθηκαν Στα Νευρωνικά Δίκτυα

ΠΑΡΑΜΕΤΡΟΣ	ΤΙΜΗ
Αριθμός Επιπέδων (layers)	3
Αριθμός Νευρώνων Κρυφού Επιπέδου	10,20,40,80,100
Αριθμός Νευρώνων Εισόδου	12
Αριθμός Νευρώνων Εξόδου	2
Συνάρτηση Μεταφοράς κρυμμένων επιπέδων	Tansig
Συνάρτηση Μεταφοράς επιπέδου εξόδου	Purelin
Backpropagation Συνάρτηση Εκπαίδευσης Δικτύου	traingd, traingdm, traingda, traingdx, trainrp, traincgf, traincgp, traincgb, trainscg, trainbfg,

Backpropagation Συνάρτηση Εκπαίδευσης Βαρών και πολώσεων Συνάρτηση Απόδοσης Εκπαίδευσης	trainlm, trainoss Learnqdm Mse
---	--

Έτσι είχαμε 240 διαφορετικούς συνδυασμούς των παραπάνω παραμέτρων. Κάθε ένας από αυτούς τους συνδυασμούς έτρεξε 10 φορές. Από αυτές τις επαναλήψεις λήφθηκε ο μέσος όρος και υπολογίστηκε η ειδικότητα (specificity) και η ευαισθησία (sensitivity) για τον κάθε συνδυασμό.

6.5.4 ΑΠΟΤΕΛΕΣΜΑΤΑ ΝΕΥΡΩΝΙΚΩΝ ΔΙΚΤΥΩΝ

Η υλοποίηση έγινε στο MATLAB και για να επιτευχθεί αυτό αρχικά αναπτύχθηκε μια συνάρτηση, η οποία διαχωρίζει τα δεδομένα σε σύνολα εκπαίδευσης και ελέγχου και εφαρμόζει τις αναγκαίες αναλογίες για καλοήθειες και κακοήθειες. Έχει σαν όρισμα της τα στοιχεία εισόδου, την αναλογία συνόλων εκπαίδευσης και ελέγχου και την αναλογία καλοηθειών και κακοηθειών.

Το βέλτιστο αποτέλεσμα εμφανίστηκε, για ένα νευρωνικό δίκτυο 3-επιπέδων, με επίπεδο εισόδου 12 νευρώνων, κρυφό επίπεδο 10 νευρώνων και επίπεδο εξόδου 2 νευρώνων. Το νευρωνικό δίκτυο εκπαιδεύτηκε με χρήση backpropagation και συνάρτηση εκπαίδευσης δικτύου την traincgf (Fletcher-Reeves Update). Επίσης χρησιμοποιήθηκαν όλα τα στοιχεία κακοηθειών και καλοηθειών με τιμές ευαισθησίας 80,3%, ειδικότητας 99,2% και συνολικής απόδοσης 97,8%.

Πίνακας 6-9
Πινάκας Σύγχυσης (Confusion Matrix)
Για Μέσο Όρο 10 Ανεξάρτητων Επαναλήψεων

Μέθοδος	Πραγματική Κλάση	Προβλεπόμενη κλάση	
		Class 1	Class 2
Νευρωνικό Δίκτυο	Class 1	655	5
	Class 2	10	43

Ο επόμενος πίνακας προκύπτει έμμεσα από τον προηγούμενο και παρουσιάζει την ευαισθησία 80,3% (sensitivity - fraction of true positives), την ειδικότητα 99,2% (specificity - fraction of true negatives) και την συνολική ακρίβεια 97,8% .

Πίνακας 6-10
Μέση Ακρίβεια Ταξινόμησης
Για 10 Ανεξάρτητες Επαναλήψεις Του Αλγορίθμου

Μέθοδος	Ευαισθησία (%)	Ειδικότητα (%)	Συνολική Ακρίβεια (%)
Νευρωνικό Δίκτυο	80,3	99,2	97,8

Οι επόμενοι δύο πίνακες είναι ο πίνακας σύγχυσης (confusion matrix) και ο πίνακας των τιμών για την ευαισθησία, την ειδικότητα και τη συνολική ακρίβεια για το τμήμα των δεδομένων τα οποία χαρακτηρίζονται ως ύποπτα (suspicious).

Πίνακας 6-11
Πίνακας Σύγχυσης (Confusion Matrix)
Για Τα Δείγματα Τα Οποία Χαρακτηρίζονται Ύποπτα,
Μέσος Όρος Για 10 Ανεξάρτητες Επαναλήψεις

Μέθοδος	Πραγματική Κλάση	Προβλεπόμενη κλάση	
		Class 1	Class 2
Νευρωνικό	Class 1	9	4
	Class 2	3	10

Πίνακας 6-12
Ακρίβεια Ταξινόμησης
Για Τα Δείγματα Τα Οποία Χαρακτηρίζονται Ύποπτα,
Μέσος Όρος Για 10 Ανεξάρτητες Επαναλήψεις

Μέθοδος	Ευαισθησία (%)	Ειδικότητα (%)	Συνολική Ακρίβεια (%)
Νευρωνικό Δίκτυο	74,6	66,2	70,4

Η συνάρτηση μεταφοράς [59] για το παραπάνω νευρωνικό δίκτυο φαίνεται στην Εικόνα 25 και αναδεικνύει ότι για τη μέγιστη ευαισθησία της μεθόδου υπάρχει μεγάλος αριθμός ψευδώς θετικών δειγμάτων.

Εικόνα 25 Συνάρτηση μεταφοράς νευρωνικού δικτύου

Η συνάρτηση απόδοσης φαίνεται στην Εικόνα 26. Σε αυτήν φαίνεται ότι στις 50 επαναλήψεις το δείγμα ελέγχου εμφανίζεται να έχει αρχικά μικρότερο και στη συνέχεια αρκετά μεγαλύτερο μέσο τετραγωνικό σφάλμα σε σχέση με το δείγμα εκπαίδευσης. Αυτό σημαίνει ότι ο αλγόριθμος δεν έχει τέλεια προσαρμογή σε όλα τα δείγματα.

Εικόνα 26 Απόδοση αλγορίθμου νευρωνικού δικτύου

6.6 ΜΗΧΑΝΕΣ ΣΤΗΡΙΞΗΣ ΔΙΑΝΥΣΜΑΤΩΝ (SUPPORT VECTOR MACHINES – SVM'S)

Οι μηχανές στήριξης διανυσμάτων - ΜΣΔ (SUPPORT VECTOR MACHINES – SVM'S) είναι μια νέα και υποσχόμενη μέθοδος, προερχόμενη από τη θεωρία εκμάθησης της στατιστικής και η οποία αναπτύχθηκε από τους Vapnik και Chervonenkis. Είναι ένα σύνολο μεθόδων εποπτευόμενης εκμάθησης που χρησιμοποιούνται στην ταξινόμηση και στην αναγνώριση προτύπων. Μια μηχανή στήριξης διανυσμάτων κατασκευάζει ένα υπερεπίπεδο ή γενικότερα ένα σύνολο υπερεπιπέδων σε ένα πολύ-διάστατο χώρο. Διαισθητικά, ένας καλός διαχωρισμός επιτυγχάνεται από το υπερεπίπεδο το οποίο έχει τη μεγαλύτερη απόσταση από τα κοντινότερα σημεία εκπαίδευσης των κατηγοριών ταξινόμησης (αποκαλούμενο περιθώριο συνάρτησης). Γενικά όσο μεγαλύτερο το περιθώριο συνάρτησης τόσο χαμηλότερο είναι το παραγόμενο λάθος από τη χρήση του ταξινομητή [45].

Στην απλούστερη περίπτωση αναγνώρισης προτύπου, οι ΜΣΔ χρησιμοποιούν ένα γραμμικά διαχωρίσιμο υπερεπίπεδο για να δημιουργήσουν ένα ταξινομητή με το μέγιστο περιθώριο συνάρτησης. Προκειμένου να γίνει αυτό, το πρόβλημα εκμάθησης μετατρέπεται σε ένα πρόβλημα βελτιστοποίησης μη γραμμικού περιορισμού. Σε αυτή την περίπτωση, η συνάρτηση κόστους είναι τετραγωνική και οι περιορισμοί είναι γραμμικοί. Σε περιπτώσεις όπου οι κλάσεις δεν είναι γραμμικά χωρισμένες, η μέθοδος SVM πρώτα τις μετατρέπει σε γραμμικές με διάστημα τιμών χαρακτηριστικών παραμέτρων μεγαλύτερης διάστασης. Αυτή η μετατροπή μπορεί να επιτευχθεί χρησιμοποιώντας διάφορες μη γραμμικές μεθόδους, πολυωνυμικές, σιγμοειδείς, πολυτετραγωνικές, πολυεπίπεδους preceptors, RBF. Μετά από αυτό το μη γραμμικό βήμα μετατροπής η μέθοδος SVM είναι η κλασσική. [44], [45].

6.6.1 ΘΕΩΡΗΤΙΚΗ ΠΕΡΙΓΡΑΦΗ

Ο αλγόριθμος θεωρεί L σημεία εκπαίδευσης όπου κάθε είσοδος x_i έχει D χαρακτηριστικές παραμέτρους (με άλλα λόγια έχει διάσταση D) και ανήκει σε μια από τις δύο κατηγορίες $y_i = -1$ ή $+1$, με άλλα λόγια τα δεδομένα εκπαίδευσης είναι της μορφής:

$$\{x_i, y_i\}, \text{ όπου } i = 1 \dots L, y_i \in \{-1, 1\}, x \in \mathbb{R}^D$$

Γίνεται η υπόθεση ότι τα δεδομένα είναι γραμμικά διαχωρίσιμα, που σημαίνει ότι μπορεί να κατασκευαστεί μια γραμμή στην γραφική παράσταση των x_1 προς το x_2 η οποία να διαχωρίζει τις δύο κλάσεις όταν $D = 2$ και ένα υπερεπίπεδο διαχωρισμού μεταξύ των γραφικών παραστάσεων των $x_1, x_2, \dots, x_D$ όταν $D > 2$.

Αυτό το υπερεπίπεδο διαχωρισμού μπορεί να περιγραφεί ως $w \cdot x + b = 0$, όπου w είναι μια τιμή ομαλοποίησης του υπερεπιπέδου και $\frac{b}{\|w\|}$ είναι η κάθετη απόσταση από το υπερεπίπεδο στην αρχή των αξόνων.

Τα διανύσματα στήριξης είναι τα δείγματα τα οποία είναι πλησιέστερα στο υπερεπίπεδο διαχωρισμού και στόχος των μηχανών διανυσματικής στήριξης είναι να προσανατολίσουν αυτό το υπερεπίπεδο (γραμμή στην απλούστερη περίπτωση) με τέτοιο τρόπο ώστε να είναι όσο το δυνατόν μακρύτερα από τα κοντινότερα μέλη των δύο κλάσεων.

Εικόνα 27 Σχηματική παράσταση Μηχανών Διανυσματικής Υποστήριξης (SVM)

Στην Εικόνα 27, η υλοποίηση των ΜΣΔ είναι η επιλογή των μεταβλητών w και b έτσι ώστε τα δεδομένα εκπαίδευσης να περιγράφονται ως εξής:

$$x_i * w + b \geq +1 \text{ για } y_i = +1 \quad (6-8)$$

$$x_i * w + b \leq -1 \text{ για } y_i = -1 \quad (6-9)$$

Αυτές οι συναρτήσεις μπορούν να συνδυαστούν στην:

$$y_i * (x_i * w + b) - 1 \geq 0 \quad \forall i \quad (6-10)$$

Εάν θεωρηθούν τα σημεία τα όποια βρίσκονται πλησιέστερα στο υπερεπίπεδο διαχωρισμού, με άλλα λόγια τα διανύσματα στήριξης (φαίνονται σαν κύκλοι στην Εικόνα 27), τότε τα δύο επίπεδα (γραμμές) H_1 και H_2 πάνω στα οποία βρίσκονται τα κοντινότερα μέλη των δύο κατηγοριών μπορούν να περιγραφούν με τις ακόλουθες εξισώσεις:

$$x_i * w + b = +1 \text{ για } H_1 \quad (6-11)$$

$$x_i * w + b = -1 \text{ για } H_2 \quad (6-12)$$

Στην Εικόνα 27, ως d_1 ορίζεται η απόσταση από το H_1 ως το υπερεπίπεδο και ως d_2 η απόσταση από το H_2 ως το υπερεπίπεδο. Το υπερεπίπεδο ισαπέχει από τα H_1 και H_2 που σημαίνει ότι $d_1 = d_2$, μια ποσότητα γνωστή ως το όριο ΜΣΔ. Προκειμένου να προσανατολιστεί το υπερεπίπεδο ώστε να είναι όσο το δυνατόν μακρύτερα από τα διανύσματα στήριξης, πρέπει να μεγιστοποιηθεί το όριο ΜΣΔ. Απλή

διανυσματική γεωμετρία αποδεικνύει ότι το όριο είναι ίσο με $\frac{1}{\|w\|}$ και η μεγιστοποίησή του, με

δεδομένη και την παραπάνω ανισότητα (6.10), είναι ισοδύναμη με το να βρεθεί

$$\text{Ελάχιστο } \|w\| \text{ τέτοιο ώστε } y_i * (x_i * w + b) - 1 \geq 0 \quad \forall i \quad (6-13)$$

Το να ελαχιστοποιηθεί το $\|w\|$ είναι ισοδύναμο με το να ελαχιστοποιηθεί το $\frac{1}{2} \|W\|^2$ και η χρήση αυτού κάνει εφικτή την πραγματοποίηση τετραγωνικού προγραμματισμού (Quadratic Programmin – QP) βελτιστοποίησης. Το τι θα πρέπει να υπολογιστεί ακολουθεί στην επόμενη σχέση:

$$\min \frac{1}{2} \|W\|^2 \text{ ώστε } y_i * (x_i * w + b) - 1 \geq 0 \quad \forall i \quad (6-14)$$

Αντί να χρησιμοποιηθεί ένα διάνυσμα υποστήριξης x_s , αποδεικνύεται με τη χρήση πολλαπλασιαστών Lagrange [44] ότι είναι καλύτερα να υπολογιστεί ο μέσος όρος όλων των διανυσμάτων υποστήριξης στο S με την παρακάτω εξίσωση.

$$b = \frac{1}{N_S} \sum_{s \in S} \left(y_s - \sum_{m \in S} a_m y_m x_m x_s \right), \quad w = \sum_{i=1}^L a_i y_i x_i \quad (6-15)$$

όπου a_i είναι οι μη αρνητικοί πολλαπλασιαστές Lagrange, S το σύνολο των διανυσμάτων υποστήριξης και N_S είναι το πλήθος του συνόλου S . Έτσι έχουμε τις τιμές των μεταβλητών w και b οι οποίες ορίζουν το βέλτιστο προσανατολισμού υπερεπίπεδο και κατά συνέπεια τη μηχανή στήριξης διανυσμάτων. Στην απλούστερη περίπτωση, προκειμένου να χρησιμοποιηθούν οι ΜΣΔ και να επιλυθεί ένα γραμμικά διαχωρίσιμο πρόβλημα δυαδικής ταξινόμησης πρέπει [44]:

- Να δημιουργηθεί H , όπου $H_{ij} = y_i y_j x_i x_j$.
- Να βρεθεί a έτσι ώστε:

$$\sum_{i=1}^L a_i - \frac{1}{2} a^T H a \quad (6-16)$$

να είναι μεγιστοποιημένο, και να ισχύει η ανισότητα:

$$a_i \geq 0 \forall_i \text{ και } \sum_{i=1}^L a_i y_i = 0 \quad (6-17)$$

Αυτό γίνεται με επίλυση τετραγωνικού προγραμματισμού (QP) [44].

- Υπολογίζει το $w = \sum_{i=1}^L a_i y_i x_i$.
- Αποφασίζει το σύνολο των διανυσμάτων υποστήριξης S βρίσκοντας τους δείκτες ώστε $a_i > 0$.
- Υπολογίζει την παράμετρο

$$b = \frac{1}{N_S} \sum_{s \in S} \left(y_s - \sum_{m \in S} a_m y_m x_m x_s \right) \quad (6-18)$$

- Κάθε νέο σημείο x' κατηγοριοποιείται εκτιμώντας το $y' = \text{sgn}(w * x' + b)$

6.6.2 ΠΑΡΟΥΣΙΑΣΗ ΠΕΙΡΑΜΑΤΩΝ ΜΣΔ

Στα πλαίσια της παρούσας διδακτορικής διατριβής εξετάστηκε η απόδοση μηχανών στήριξης διανυσμάτων (ΜΣΔ) εφαρμοζόμενων στο πρόβλημα ανεύρεσης μοντέλων ασθενών με κακοήθεια θυρεοειδούς. Έγιναν δοκιμές με τα στοιχεία όπως προέκυψαν από την μέθοδο επιλογής χαρακτηριστικών παραμέτρων (feature selection) SFFS. Το γεγονός ότι δύο κατηγορίες των δεδομένων είναι ανισοπληθείς, δηλαδή περισσότερες καλοήθειες σε σχέση με κακοήθειες, θα μπορούσε να επηρεάζει την απόδοση μηχανών στήριξης διανυσμάτων. Έτσι έγινε τυχαία επιλογή των καλοηθειών ώστε σε κάποιες από τις δοκιμές να επιλέγεται μέρος αυτών. Οι αναλογίες που δοκιμάστηκαν είναι: όλα τα στοιχεία εκπαίδευσης, στοιχεία εκπαίδευσης με αναλογία καλοήθειες προς κακοήθειες 1:1, 1:2 και 1:3. Αξιόλογα αποτελέσματα βρέθηκαν στην περίπτωση χρήσης όλων των στοιχείων, κάτι που δείχνει ότι οι μηχανές στήριξης διανυσμάτων (ΜΣΔ) αποδίδουν πολύ καλά και σε προβλήματα με ανισοπληθείς κατηγορίες. Οι τιμές που χρησιμοποιήθηκαν για την συνάρτηση Kernel, η οποία χρειάζεται για την τροποποίηση και αποδοτικότερη επίλυση, ήταν γραμμική (linear), τετραγωνική (quadratic), γκαουσιανή βασισμένη στην ακτίνα (rbf - Gaussian Radial Basis), πολυωνυμική (polynomial). Για την συνάρτηση Kernel τύπου rbf το RBF Sigma Value δοκιμάστηκε με τιμές 1, 3, 5, 10. Για την πολυωνυμική συνάρτηση η μεταβλητή τιμή τάξης μεγέθους (Polyorder Value) δοκιμάστηκε με τιμές 0, 4, 5, 6. Ως μέθοδοι διαχωρισμού υπερεπιπέδων δοκιμάστηκαν ο τετραγωνικός προγραμματισμός QP (Quadratic Programming), η βελτιστοποίηση διαδοχικού ελαχίστου (SMO - Sequential Minimal Optimization), η μέθοδος ελαχίστων τετραγώνων (LS - Least-Squares).

Οι χρησιμοποιούμενες παράμετροι συνοψίζονται στο παρακάτω πίνακα

Πίνακας 6-13

Παράμετροι Που Χρησιμοποιήθηκαν Από Τις Μηχανές Στήριξης Διανυσμάτων (SVM's)

ΠΑΡΑΜΕΤΡΟΣ	ΤΙΜΗ
Συνάρτηση Kernel για μορφοποίηση των στοιχείων εισόδου	linear, quadratic, rbf (Gaussian Radial Basis), polynomial
Μέθοδος Διαχωρισμού Υπερεπιπέδων	QP (Quadratic Programming), SMO (Sequential Minimal Optimization), LS (Least-Squares)
Τιμή Sigma για τη συνάρτηση Kernel τύπου rbf	1, 3, 5, 10
Τιμή τάξης για τη συνάρτηση Kernel τύπου πολυωνυμικής (polynomial)	0, 4, 5, 6

Προέκυψαν 480 διαφορετικοί συνδυασμοί μεταξύ των προαναφερόμενων παραμέτρων. Κάθε ένας από αυτούς τους συνδυασμούς έτρεξε 10 φορές. Από αυτές τις επαναλήψεις λήφθηκε ο μέσος όρος και υπολογίστηκε η ειδικότητα (specificity), η ευαισθησία (sensitivity) και η συνολική ακρίβεια για κάθε συνδυασμό.

6.6.3 ΑΠΟΤΕΛΕΣΜΑΤΑ ΜΣΔ

Η υλοποίηση έγινε στο MATLAB και για να επιτευχθεί αυτό αρχικά χρησιμοποιήθηκε μια συνάρτηση που αναπτύχθηκε στα πλαίσια της διατριβής και η οποία διαχωρίζει τα δεδομένα σε σύνολα εκπαίδευσης και ελέγχου και εφαρμόζει τις αναγκαίες αναλογίες για καλοήθειες και κακοήθειες. Έχει σαν όρισμα τα στοιχεία εισόδου, την αναλογία συνόλων εκπαίδευσης και ελέγχου καθώς και την αναλογία καλοηθειών με τις κακοήθειες.

Το βέλτιστο αποτέλεσμα εμφανίστηκε, για συνάρτηση Kernel γραμμικής βάσης, με χρήση μεθόδου διαχωρισμού υπερειπέδων Quadratic Programming (QP), για μέγιστο αριθμό επαναλήψεων 10.000 και για χρήση όλων των στοιχείων κακοηθειών και καλοηθειών. Υπολογίστηκε ότι η ευαισθησία είναι 83,6%, η ειδικότητα 99,3% και η συνολική απόδοση 98,1%.

Πίνακας 6-14

Πίνακας Σύγχυσης (Confusion Matrix)**Για Μέσο Όρο 10 Ανεξάρτητων Επαναλήψεων**

Μέθοδος	Πραγματική Κλάση	Προβλεπόμενη κλάση	
		Class 1	Class 2
SVM	Class 1	655	5
	Class 2	9	44

Ο επόμενος Πίνακας 6-15 προκύπτει έμμεσα από τον προηγούμενο και παρουσιάζει την ευαισθησία 83,6% (sensitivity - fraction of true positives), την ειδικότητα 99,3% (specificity - fraction of true negatives) και την συνολική ακρίβεια 98,1%.

Πίνακας 6-15

Μέση Ακρίβεια Ταξινόμησης**Για 10 Ανεξάρτητες Επαναλήψεις Του Αλγορίθμου**

Μέθοδος	Ευαισθησία (%)	Ειδικότητα (%)	Συνολική Ακρίβεια (%)
SVM	83,6	99,3	98,1

Οι επόμενοι δύο πίνακες είναι ο πίνακας σύγχυσης (confusion matrix) και ο πίνακας των τιμών για την

ευαισθησία, την ειδικότητα και τη συνολική ακρίβεια για το τμήμα των δεδομένων τα οποία χαρακτηρίζονται ως ύποπτα (suspicious).

Πίνακας 6-16
Πίνακας Σύγχυσης (Confusion Matrix)
Για Τα Δείγματα Τα Οποία Χαρακτηρίζονται Ύποπτα,
Μέσος Όρος Για 10 Ανεξάρτητες Επαναλήψεις

Μέθοδος	Πραγματική Κλάση	Προβλεπόμενη κλάση	
		Class 1	Class 2
SVM	Class 1	8	5
	Class 2	4	9

Πίνακας 6-17
Ακρίβεια Ταξινόμησης
Για Τα Δείγματα Τα Οποία Χαρακτηρίζονται Ύποπτα,
Μέσος Όρος Για 10 Ανεξάρτητες Επαναλήψεις

Μέθοδος	Ευαισθησία (%)	Ειδικότητα (%)	Συνολική Ακρίβεια (%)
SVM	66,9	63,1	65

Η συνάρτηση μεταφοράς [59] για την παραπάνω ΜΣΔ παρουσιάζεται στην Εικόνα 28 η οποία αναδεικνύει την πολύ καλή απόκριση του αλγορίθμου αφού η καμπύλη απόκρισης προσεγγίζει την άνω αριστερή γωνία της βέλτιστης ROC.

Εικόνα 28 Συνάρτηση μεταφοράς αλγορίθμου SVM

6.7 Αλγόριθμος *k*- KONTINOTEΡΟΥ ΓΕΙΤΟΝΑ (*k* NEAREST NEIGHBOR ALGORITHM)

Ο πιο συχνά χρησιμοποιούμενος αλγόριθμος για ταξινόμηση είναι ο *k*-κοντινότερου γείτονα (*k*-nearest neighbour). Αυτός ο αλγόριθμος είναι ένα παράδειγμα εκμάθησης βασισμένης στην απόσταση (distance-based learning), όπου το σύνολο δεδομένων εκπαίδευσης είναι αποθηκευμένο έτσι ώστε η

ταξινόμηση μιας νέας, αταξινομήτης περίπτωσης, να βρεθεί μέσα από τη σύγκριση της με τις περισσότερες όμοιες περιπτώσεις του συνόλου εκπαίδευσης. Για παράδειγμα, ας πάρουμε την περίπτωση όπου θέλουμε να βρούμε τον τύπο του φαρμάκου που πρέπει να συνταγογραφηθεί για κάποιον ασθενή, με βάση κάποια χαρακτηριστικά στοιχεία για τον κάθε ασθενή, όπως η ηλικία, οι τιμές των ηλεκτρολυτών K/Na κ.α.. Για ένα δείγμα 200 ασθενών, στην Εικόνα 29 [51] φαίνεται το γράφημα διασποράς των ασθενών για τη μέτρηση των ηλεκτρολυτών σε σχέση με την ηλικία. Το συγκεκριμένο φάρμακο το οποίο έχει συνταγογραφηθεί φαίνεται με τις διαφορετικές αποχρώσεις του γκρι. Έστω ότι τα ανοικτά γκρι σημεία δείχνουν το φάρμακο Y, τα ελαφρά γκρι σημεία το φάρμακο A ή X, τα πολύ έντονα γκρι σημεία δείχνουν το φάρμακο B ή C. Ας υποθεθεί ότι μια νέα περίπτωση, χωρίς ταξινόμηση φαρμάκου, πρέπει να κατηγοριοποιηθεί για να λάβει την καλύτερη συνταγή, με βάση το τι έχουν πάρει άλλοι ασθενείς με παρόμοια χαρακτηριστικά. Η περίπτωση του ασθενούς 1, 40 ετών και με μέτρηση των ηλεκτρολυτών 29, τοποθετείται στο κέντρο του κύκλου που αντιστοιχεί στον ασθενή 1, όπως φαίνεται στην Εικόνα 29. Όλες οι γειτονικές περιπτώσεις είναι ασθενείς που έλαβαν το φάρμακο Y, κατηγοριοποιώντας έτσι εύκολα τον ασθενή 1 στο φάρμακο Y. Ένα ακόμα παράδειγμα είναι αυτό της περίπτωσης του ασθενούς 2 (Εικόνα 29), ο οποίος έχει μικρότερες τιμές ηλεκτρολυτών και μικρότερη ηλικία. Στην Εικόνα 29 παρέχεται μια άποψη για τα δείγματα εκπαίδευσης στην «γειτονιά» του ασθενή 2. Ας υποθεθεί ότι $k = 1$, έτσι ώστε ο ασθενής 2 να κατηγοριοποιηθεί σύμφωνα με τον ένα γείτονα που είναι κοντύτερα σε αυτόν. Σε αυτή την περίπτωση ο ασθενής 2 θα κατηγοριοποιηθεί για το φάρμακο B ή C (σκουρό γκρι).

Εικόνα 29 Γράφημα διασποράς μετρήσεων ηλεκτρολυτών ασθενών σε σχέση με την ηλικία και την επίδραση του φαρμάκου

Ας υποθεθεί ότι $k = 2$, έτσι ώστε ο ασθενής 2 να κατηγοριοποιηθεί σύμφωνα με τους δύο κοντινότερους γείτονες του. Το ένα από αυτά τα σημεία είναι έντονο γκρι και το άλλο μέτριο γκρι. Έτσι ο αλγόριθμος θα έπρεπε να αποφασίσει ανάμεσα σε ποια από τις δύο περιπτώσεις (φάρμακο B και C - έντονο γκρι ή φάρμακο A και X - μέτριο γκρι) να κατατάξει τον ασθενή. Δεν υπάρχει κάποιος τρόπος να επικρατήσει μια από τις δύο επιλογές. Έτσι σε αλγορίθμους k-κοντινότερου γείτονα επιλέγεται πάντα για τον k περιττός αριθμός. Στην περίπτωση αυτή αν επιλεγεί $k = 3$, ο ασθενής 2 έχει πλέον δύο κοντινά σημεία του μέτρια γκρι έναντι ενός έντονου γκρι και μπορεί να κατηγοριοποιηθεί. Η κατηγορία που θα εμπίπτει

ο ασθενής 2, και κάθε ασθενής αναλόγου περίπτωσης, θα διαφέρει με βάση το τι τιμή θα επιλεγεί για το k . Τέλος ως θεωρήσουμε τον ασθενή 3, ο οποίος είναι 47 ετών και έχει μέτρηση ηλεκτρολυτών 13,5. Αν επιλεγεί $k = 1$, ο αλγόριθμος θα μας δώσει σκούρο γκρι (φάρμακο B και C). Η επιλογή του $k = 2$ δεν βοηθάει στην ταξινόμηση της περίπτωσης του ασθενούς 3. Αλλά ακόμα και η επιλογή του $k = 3$ δεν βοηθάει στην προκειμένη περίπτωση, αφού οι τρεις κοντινότεροι γείτονες του ασθενούς 3 έχουν καταταχθεί για θεραπεία με τρία διαφορετικά φάρμακα. Για την εφαρμογή του αλγορίθμου k -κοντινότερου γείτονα θα πρέπει λοιπόν να αποφασιστεί το πόσοι γείτονες θα συνυπολογιστούν για την ταξινόμηση (τιμή του k). Όμως πέραν αυτού υπάρχει και μια άλλη σημαντική παράμετρος για την εφαρμογή του αλγορίθμου που αφορά στο πώς ορίζεται ο κοντινότερος γείτονας και είναι το μετρικό απόστασης (συνάρτηση απόστασης) που θα χρησιμοποιηθεί [50],[51].

Μετρικό απόστασης (συνάρτηση απόστασης) ορίζεται συνήθως μια πραγματική συνάρτηση d , έτσι ώστε για κάθε σημείο συντεταγμένων x, y, z :

1. $d(x,y) \geq 0$, και $d(x,y) = 0$ μόνο όταν $x = y$
2. $d(x,y) = d(y,x)$
3. $d(x,z) \leq d(x,y) + d(y,z)$

Η πρώτη ιδιότητα εξασφαλίζει ότι έχουμε μόνο θετικές αποστάσεις, και η μόνη περίπτωση να είναι μηδέν το μέτρο είναι να έχουμε ίδιες συντεταγμένες. Η ιδιότητα 2 εξασφαλίζει την αντιμεταθετικότητα και η ιδιότητα τρία την τριγωνική ανισότητα, η οποία ορίζει ότι αν θέσουμε ένα τρίτο σημείο δεν θα μικρύνουμε την απόσταση μεταξύ δύο άλλων σημείων. Η πιο κοινή συνάρτηση απόστασης είναι η Ευκλείδεια απόσταση, η οποία αναπαριστάται με τον κλασικό τρόπο [51].

$$d = \sqrt{(x_i - y_i)^2} \quad (6-19)$$

όπου $x = x_1, x_2, \dots, x_m$, και $y = y_1, y_2, \dots, y_m$ είναι οι m τιμές των χαρακτηριστικών παραμέτρων των δύο περιπτώσεων. Έστω για παράδειγμα ότι ο ασθενής A είναι $x_1 = 20$ ετών και έχει μέτρηση ηλεκτρολυτών $x_2 = 12$, ενώ ο ασθενής B έχει $y_1 = 30$ και $y_2 = 8$. Η Ευκλείδεια απόσταση ανάμεσα στα δύο αυτά σημεία είναι:

$d = \sqrt{(20-30)^2 + (12-8)^2} = 10.77$. Αυτό φυσικά μπορεί να γενικευθεί και για την περίπτωση πολυδιάστατων παραμέτρων.

Όταν γίνεται μέτρηση της απόστασης, κάποιες παράμετροι που από τη φύση τους έχουν μεγαλύτερη κλίμακα μέτρησης μπορεί να επικαλύψουν την επιρροή άλλων παραμέτρων με μικρότερη κλίμακα. Για να αποφευχθεί αυτό το φαινόμενο, εφαρμόζεται η μέθοδος της κανονικοποίησης των τιμών, είτε σαν κανονικοποίηση ελαχίστων – μεγίστων, είτε σαν κανονικοποίηση με Z -συνάρτηση. Οι τύποι για τις δύο μεθόδους κανονικοποίησης δίνονται παρακάτω:

Κανονικοποίηση μεγίστων - ελαχίστων

$$X^* = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (6-20)$$

Κανονικοποίηση Z -συνάρτηση:

$$X^* = \frac{X - \text{mean}(X)}{SD(X)} \quad (6-21)$$

Για κατηγορικά δεδομένα, η Ευκλείδεια απόσταση δεν είναι κατάλληλη και αντί αυτής ορίζουμε μια συνάρτηση «διαφορετικότητας» για να συγκρίνουμε τις τιμές της i th παραμέτρου ενός ζεύγους περιπτώσεων του συνόλου ως εξής:

«Διαφορετικότητα» $(x_i, y_i) = 0$, αν $x_i = y_i$, αλλιώς είναι 1, όπου x_i, y_i είναι κατηγορικά δεδομένα.

Για την εφαρμογή του αλγορίθμου k-κοντινότερου γείτονα είναι σημαντικό να υπάρχουν πολλά στοιχεία στο σύνολο εκπαίδευσης με διαφορετικές τιμές στις διάφορες χαρακτηριστικές παραμέτρους, έτσι ώστε να περιγράφεται καλύτερα και αναλυτικότερα ο χώρος του συνόλου των δειγμάτων. Επίσης, θα πρέπει να αντιπροσωπεύονται ενδεικτικά και οι σπάνιες περιπτώσεις, ώστε ο αλγόριθμος να μην βρίσκει μόνο τις συχνά συναντώμενες. Έτσι το σύνολο των δειγμάτων θα πρέπει να περιέχει ικανό αριθμό περιπτώσεων εξαίρεσης. Μια μέθοδος για να επιτευχθεί αυτό είναι η μείωση του αριθμού των συχνά συναντώμενων δειγμάτων.

Σημαντικός παράγοντας για την εφαρμογή του αλγορίθμου k-κοντινότερου γείτονα είναι η επιλογή του k. Αν επιλεγεί μια μικρή τιμή του k, τότε το αποτέλεσμα επηρεάζεται εύκολα από τιμές που είναι πολύ μακριά (outliers) ή ασυνήθιστες τιμές θορύβου ("noise"). Αυτό θα οδηγήσει και στο φαινόμενο του overfitting, τείνοντας το τελικό μοντέλο να απομνημονεύσει το σύνολο εκπαίδευσης και να μην μπορεί να εφαρμοστεί γενικά. Αντίθετα, η επιλογή μεγάλης τιμής k τείνει να εξομαλύνει ιδιοσυγκρασιακές συμπεριφορές του συνόλου εκπαίδευσης. Όμως αν επιλεγεί πολύ μεγάλη τιμή του k θα χαθεί η τοπικά ενδιαφέρουσα πληροφορία. Έτσι απαιτείται εξισορρόπηση μεταξύ των δύο. Αρκετές φορές ακολουθείται διαδικασία δοκιμών με διάφορες τιμές για το k ώστε να βρεθεί το βέλτιστο αποτέλεσμα.

6.7.1 ΠΑΡΟΥΣΙΑΣΗ ΠΕΙΡΑΜΑΤΩΝ k-KONTINOTEROY ΓΕΙΤΟΝΑ

Στα πλαίσια της παρούσας διδακτορικής διατριβής εξετάστηκε η απόδοση αλγοριθμικών μεθόδων k-κοντινότερου γείτονα (k-nearest neighbor algorithm) εφαρμοζόμενων στο πρόβλημα ανεύρεσης μοντέλων ασθενών με κακοήθεια θυρεοειδούς. Έγιναν δοκιμές με τα στοιχεία όπως προέκυψαν από τη μέθοδο επιλογής χαρακτηριστικών (feature selection) SFFS. Το γεγονός ότι οι δύο κατηγορίες των δεδομένων μας είναι ανισοπλήθεις (περισσότερες καλοήθειες σε σχέση με κακοήθειες), θα μπορούσε να επηρεάζει την απόδοση των αλγοριθμικών μεθόδων k-κοντινότερου γείτονα. Έτσι έγινε τυχαία επιλογή των καλοηθειών ώστε σε κάποια από τις δοκιμές να επιλέγεται μέρος. Οι αναλογίες που δοκιμάστηκαν είναι: όλα τα στοιχεία εκπαίδευσης, στοιχεία εκπαίδευσης με αναλογία καλοήθειες προς κακοήθειες, 1:1, 1:2 και 1:3. Αξιόλογα αποτελέσματα βρέθηκαν στην περίπτωση που χρησιμοποιήθηκαν όλα τα στοιχεία, κάτι που δείχνει ότι η μέθοδος k-κοντινότερου γείτονα δεν επηρεάζεται αναγκαστικά από ανισοπληθείς κατηγορίες.

6.7.2 ΑΠΟΤΕΛΕΣΜΑΤΑ k-KONTINOTEROY ΓΕΙΤΟΝΑ

Η υλοποίηση έγινε στο MATLAB και για να επιτευχθεί αυτό αρχικά χρησιμοποιήθηκε μια συνάρτηση που αναπτύχθηκε στα πλαίσια της διατριβής, η οποία διαχωρίζει τα δεδομένα σε σύνολα εκπαίδευσης και ελέγχου και εφαρμόζει τις αναγκαίες αναλογίες για καλοήθειες και κακοήθειες. Έχει σαν ορίσματα τα στοιχεία εισόδου, την αναλογία των συνόλων εκπαίδευσης και ελέγχου και την αναλογία καλοηθειών και κακοηθειών.

Η συνάρτηση knnclassify του MATLAB έχει την ακόλουθη σύνταξη:

Class = knnclassify(Sample, Training, Group, k, distance, rule)

Κατηγοριοποιεί τα δεδομένα σε μια γραμμή (περίπτωση) βασιζόμενη στα αντίστοιχα στοιχεία από το δείγμα εκπαίδευσης.

Η συνάρτηση knnclassify του MATLAB έχει τις ακόλουθες παραμέτρους που δοκιμάστηκαν:

- Το πλήθος των k κοντινότερων γειτόνων που θα χρησιμοποιηθούν για την ταξινόμηση
- Το μετρικό της απόστασης
Αυτό μπορεί να είναι:
 - euclidean — Ευκλείδεια απόσταση (εξ ορισμού)
 - cityblock — άθροισμα απολύτων διαφορών
 - cosine — 1 μείον το συνημίτονο της γωνίας που περικλείεται από τα σημεία

- correlation — 1 μείον το δείγμα της συσχέτισης μεταξύ των σημείων
- hamming — ποσοστό των bits τα οποία διαφέρουν (κατάλληλο για δυαδικά δεδομένα)
- Τον κανόνα διαχωρισμού που μπορεί να είναι:
 - nearest — Κανόνας πλειοψηφίας για διαχωρισμό με βάση τους κοντινότερους γείτονες
 - random — Κανόνας πλειοψηφίας για διαχωρισμό με βάση τυχαίους γείτονες
 - consensus — Κανόνας κοινής συναίνεσης

Για να επιτευχθεί καλύτερη γενίκευση των αποτελεσμάτων, ο αλγόριθμος έτρεξε 10 φορές με τυχαία διαφορετικά δείγματα και λάβαμε τη μέση τιμή των 10 αποτελεσμάτων. Το βέλτιστο αποτέλεσμα εμφανίστηκε με το συνδυασμό των εξής παραμέτρων: λήψη 9 κοντινότερων γειτόνων, μετρικό τύπου correlation και κανόνα διαχωρισμού nearest. Τα αποτελέσματα ήταν για την ευαισθησία 88,1%, ειδικότητα 99% και συνολική απόδοση 98,2%.

Πίνακας 6-18
Πίνακας Σύγχυσης (Confusion Matrix)
Για Μέσο Όρο 10 Ανεξάρτητων Επαναλήψεων

Μέθοδος	Πραγματική Κλάση	Προβλεπόμενη κλάση	
		Class 1	Class 2
k-NN	Class 1	653	7
	Class 2	6	47

Ο επόμενος πίνακας προκύπτει έμμεσα από τον προηγούμενο και παρουσιάζει την ευαισθησία (sensitivity - fraction of true positives), την ειδικότητα (specificity - fraction of true negatives) και την συνολική ακρίβεια.

Πίνακας 6-19
Μέση Ακρίβεια Ταξινόμησης
Για 10 Ανεξάρτητες Επαναλήψεις Του Αλγορίθμου

Μέθοδος	Ευαισθησία (%)	Ειδικότητα (%)	Συνολική Ακρίβεια (%)
k-NN	88,1	99	98,2

Οι επόμενοι δύο πίνακες είναι ο πίνακας σύγχυσης (confusion matrix) και ο πίνακας των τιμών για την ευαισθησία, την ειδικότητα και τη συνολική ακρίβεια για το τμήμα των δεδομένων τα οποία χαρακτηρίζονται ως ύποπτα (suspicious).

Πίνακας 6-20
Πίνακας Σύγχυσης (Confusion Matrix)
Για Τα Δείγματα Τα Οποία Χαρακτηρίζονται Ύποπτα,
Μέσος Όρος Για 10 Ανεξάρτητες Επαναλήψεις

Μέθοδος	Πραγματική Κλάση	Προβλεπόμενη κλάση	
		Class 1	Class 2
k-NN	Class 1	8	5
	Class 2	5	8

Πίνακας 6-21
Ακρίβεια Ταξινόμησης
Για Τα Δείγματα Τα Οποία Χαρακτηρίζονται Ύποπτα,
Μέσος Όρος Για 10 Ανεξάρτητες Επαναλήψεις

Μέθοδος	Ευαισθησία (%)	Ειδικότητα (%)	Συνολική Ακρίβεια (%)
k-NN	63,9	60	61,9

Η συνάρτηση μεταφοράς [59] για το παραπάνω k-NN είναι όπως φαίνεται στο παρακάτω σχήμα.

Εικόνα 30 Συνάρτηση μεταφοράς για το παραπάνω k-NN

6.8 ΑΛΓΟΡΙΘΜΟΣ ΠΛΕΙΟΨΗΦΙΚΟΥ ΚΑΝΟΝΑ

Ο αλγόριθμος πλειοψηφικού κανόνα θεωρείται από τις απλούστερες και αποτελεσματικότερες μεθόδους που χρησιμοποιούνται για να συνδυαστούν τα αποτελέσματα πολλών και διαφορετικών ταξινομητών [53], [54]. Ο αλγόριθμος πλειοψηφικού κανόνα μετράει τις «ψήφους» για κάθε κατηγορία από κάθε αλγόριθμο που συνδυάζει και επιλέγει την πλειοψηφούσα κατηγορία, όπως φαίνεται και στον παρακάτω πίνακα.

Πίνακας 6-22
Ταξινόμηση Σύμφωνα Με Τις Αρχές Του Αλγόριθμου Πλειοψηφικού Κανόνα
(MAJORITY VOTE PROCEDURE)

Αλγόριθμος 1	Αλγόριθμος 2	Αλγόριθμος 3	 Αλγόριθμος n	Τελική Απόφαση Βασισμένη στον αλγόριθμο Πλειοψηφίας
Κατηγορία 1	Κατηγορία 1	Κατηγορία 1		Κατηγορία 1
Κατηγορία 1	Κατηγορία 1	Κατηγορία 2		Κατηγορία 1
Κατηγορία 1	Κατηγορία 2	Κατηγορία 1		Κατηγορία 1
Κατηγορία 1	Κατηγορία 2	Κατηγορία 2		Κατηγορία 2
Κατηγορία 2	Κατηγορία 1	Κατηγορία 1		Κατηγορία 1
Κατηγορία 2	Κατηγορία 1	Κατηγορία 2		Κατηγορία 2
Κατηγορία 2	Κατηγορία 2	Κατηγορία 1		Κατηγορία 2
Κατηγορία 2	Κατηγορία 2	Κατηγορία 2		Κατηγορία 2

Θεωρητικά, αν οι αλγόριθμοι έχουν ανεξάρτητα λάθη ταξινόμησης, ο αλγόριθμος πλειοψηφικού κανόνα επιλέγει τον καλύτερο αλγόριθμο. Η ιδέα του συνδυασμού πολλών και διαφορετικών αλγόριθμων ταξινόμησης θεωρείται μια σημαντική σύλληψη και έχει παρουσιαστεί με την μορφή πολλών και διαφορετικών παραλλαγών σε διάφορες έρευνες [54]. Στη [56], συστήνεται μια μέθοδος με την επιλογή

ενός αριθμού νευρωνικών δικτύων με γενετικούς αλγορίθμους. Άλλες προσεγγίσεις είναι βασισμένες σε μετρήσεις της διαφορετικότητας. Η εφαρμογή της μεθόδου πλειοψηφικού κανόνα έχει και κάποια όρια, ως προς το ανώτερο και το κατώτερο όριο ακρίβειας σε σχέση με την εξατομικευμένη ακρίβεια κάθε αλγορίθμου «ψηφοφόρου» και το πλήθος αυτών [56]. Μια ευρέως χρησιμοποιούμενη εκδοχή του είναι με τη χρήση βαρών για τον κάθε αλγόριθμο «ψηφοφόρο» [57]-[58].

6.8.1 ΠΑΡΟΥΣΙΑΣΗ ΠΕΙΡΑΜΑΤΩΝ – ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Στα πλαίσια της παρούσας διδακτορικής διατριβής εξετάστηκε η απόδοση αλγοριθμικής μεθόδου πλειοψηφικού κανόνα εφαρμοζόμενης στο πρόβλημα ανεύρεσης μοντέλων ασθενών με κακοήθεια θυρεοειδούς. Στις δοκιμές της μεθόδου πλειοψηφικού κανόνα χρησιμοποιήθηκαν για την «ψηφοφορία» οι αλγόριθμοι Νευρωνικών Δικτύων, Μηχανών Διανυσμάτων Στήριξης, k-κοντινότερου γείτονα. Στη μέθοδο πλειοψηφικού κανόνα χρησιμοποιείται περιττός αριθμός «ψηφοφόρων» αλγορίθμων ώστε να μην υπάρχει η περίπτωση της ισοψηφίας.

6.8.2 ΑΠΟΤΕΛΕΣΜΑΤΑ

Η υλοποίηση της μεθόδου έγινε στο MATLAB με απλή μαθηματική συνάρτηση, για την ταξινόμηση κάθε περίπτωσης του συνόλου ελέγχου με βάση την ψήφο που λαμβάνει από κάθε ένα αλγόριθμο «ψηφοφόρο» που χρησιμοποιήθηκε. Το τελικό αποτέλεσμα είναι ένα διάνυσμα με τις κατηγορίες που κάθε μια περίπτωση κατατάσσεται σύμφωνα με τον κανόνα πλειοψηφίας. Σε καμία περίπτωση δεν έχουμε ισοψηφία γιατί έχουμε περιττό αριθμό ψηφοφόρων (3), έτσι κάθε περίπτωση του συνόλου ελέγχου ταξινομείται σε μια κατηγορία κατά μέγιστο με 3 ψήφους. Για την καλύτερη γενίκευση των αποτελεσμάτων ο αλγόριθμος «έτρεξε» 10 φορές με τυχαία διαφορετικά δείγματα και λήφθηκε ο μέσος όρος. Παρακάτω φαίνονται ο πίνακας σύγχυσης καθώς και ο πίνακας των τιμών της ευαισθησίας, της ειδικότητας και της συνολικής ακρίβειας.

Πίνακας 6-23
Πίνακας Σύγχυσης (Confusion Matrix)
Για Μέσο Όρο 10 Ανεξάρτητων Επαναλήψεων

Μέθοδος	Πραγματική Κλάση	Προβλεπόμενη Κλάση	
		Class 1	Class 2
Πλειοψηφίας (CMV)	Class 1	656	4
	Class 2	6	47

Πίνακας 6-24
Μέση Ακρίβεια Ταξινόμησης
Για 10 Ανεξάρτητες Επαναλήψεις Του Αλγορίθμου

Μέθοδος	Ευαισθησία (%)	Ειδικότητα (%)	Συνολική Ακρίβεια (%)
Πλειοψηφίας (CMV)	89,1	99,4	98,6

Ιδιαίτερα για την ταξινόμηση των επιχρισμάτων, τα οποία χαρακτηρίζονται ύποπτα «suspicious» από τους ειδικούς η χρήση του αλγορίθμου πλειοψηφίας φαίνεται να είναι εξαιρετικά αποδοτική. Τα αποτελέσματα για αυτά τα δείγματα φαίνονται στους παρακάτω πίνακες.

Πίνακας 6-25
Πίνακας Σύγχυσης (Confusion Matrix)
Για Τα Δείγματα Τα Οποία Χαρακτηρίζονται Ύποπτα,
Μέσος Όρος Για 10 Ανεξάρτητες Επαναλήψεις

Μέθοδος	Πραγματική Κλάση	Προβλεπόμενη κλάση	
		Class 1	Class 2
Πλειοψηφίας (CMV)	Class 1	10	3
	Class 2	1	12

Πίνακας 6-26
Ακρίβεια Ταξινόμησης
Για Τα Δείγματα Τα Οποία Χαρακτηρίζονται Ύποπτα,
Μέσος Όρος Για 10 Ανεξάρτητες Επαναλήψεις

Μέθοδος	Ευσισθησία (%)	Ειδικότητα (%)	Συνολική Ακρίβεια(%)
Πλειοψηφίας (CMV)	93,9	75,4	84,6

6.9 ΒΙΒΛΙΟΓΡΑΦΙΑ

1. **Bao, HO Tu.,** *Introduction to knowledge discovery and data mining*, course, Institute of Information Technology National Center for Natural Science and Technology, <http://www.ebook.edu.vn/?page=1.9&view=1694>, [Web Accessed 22th January 2008].
2. **MacLennan, ZhaoHui Tang and Jamie.** *Data Mining with SQL Server 2005*. s.l. : Wiley Publishing, Inc., 2005.
3. **Lauren S. Burrell, Otis L. Smart, George Georgoulas, Eric Marsh, and Vachtsevanos J. George,** *Evaluation of Feature Selection Techniques for Analysis of Functional MRI and EEG*, International Conference on Data Mining, Monte Carlo Resort, Las Vegas, Nevada, June 25-28 2007.
4. **Doak, J.,** *An Evaluation of Feature Selection Methods and Their Application to Computer Security* (Tech. Rep. CSE-92-18). University of California at Davis, 1992.
5. **Guyon and Elisseeff,** "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157- 1182, 2003.
6. **Mitra, P., Murthy, C.A., Pal., S.K.,** *Unsupervised feature selection using feature similarity*, *IEEE Trans. Pat. Anal. Mach. Intell.*, 2002, Τόμ. 24, σσ. 301-312.
7. **Pudil P., Nonovocova J., Kittler J.,** *Floating search methods in feature selection*. *Patt. Recog. Let.*, 1994, Τόμ. 15, σσ. 1119-1125.
8. **Oded Z. Maimon, Lior Rokach,** *Data mining and knowledge discovery handbook*, Springer 2005.
9. **Breiman L., Friedman J., Olshen R., and Stone C.,** *Classification and Regression Trees*. Wadsworth Int. Group, 1984.
10. **Quinlan, J.R.,** *Simplifying decision trees*, *International Journal of Machine Studies*, 27, 221-234, 1987.
11. **Hyafil L. and Rivest R.L.,** *Constructing optimal binary decision trees is NP-complete*, *Information Processing Letters*, 5(1):15-17, 1976.
12. **Zantema H, and Bodlaender H.L.,** *Finding Small Equivalent Decision Trees is Hard*, *International Journal of Foundations of Computer Science*, 11(2):343-354, 2000.
13. **Naumov G.E.,** *NP-completeness of problems of construction of optimal decision trees*. *Soviet Physics: Doklady*, 36(4):270-271, 1991.
14. **Quinlan, J.R.,** *Induction of decision Trees*, *Machine Learning* 1, 81-106, 1986.
15. **Quinlan, J.R.,** *C4.5: Programs for Machine Learning*, Morgan Kaufmann, Los Altos, 1993.

16. **Mugambi E.M., Hunter Andrew, Oatley Giles, Kennedy Lee.** *Polynomial-fuzzy decision tree structures for classifying medical data.* Knowledge-Based Systems. 2004, Τόμ. 17, σσ. 81–87.
17. **Wen-Jia Kuo, Ruey-Feng Chang, Dar-Ren Chen, and Cheng Chun Lee.** *Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images.* 2001, Τόμ. 66, σσ. 51–57.
18. **Breiman, L., et al.** *Classification and Regression Trees.* s.l. : Chapman & Hall, Boca Raton, 1993.
19. **Gelfand S. B., Ravishankar C. S., and Delp E. J.** *An iterative growing and pruning algorithm for classification tree design* 1991, IEEE Transaction on Pattern Analysis and Machine Intelligence, Τόμ. 13(2), σσ. 163-174.
20. **Caudill, M., and C. Butler.** *Understanding Neural Networks:Computer Explorations.* Τόμ. 1 and 2 Cambridge, MA: the MIT Press, 1992.
21. **M. T., and M.Menhaj.,** *Training feedforward networks with.* Hagan, 6, 1994, IEEE Transactions on Neural Networks, Τόμ. 5, σσ. 989–993.
22. **Demuth Howard, Beale Mark,** *Neural Network Toolbox User's Guide, Neural Network Toolbox, For Use with MATLAB®, COPYRIGHT 1992 - 2000 by The MathWorks, Inc.*
23. **Nguyen, D., and Widrow B.. s.l. :***The truck backer-upper: An example of self-learning in neural networks.* International Joint Conference on Neural Networks, 1989. Τόμ. 2, σσ. 357-363.
24. **Chen, S., Cowan C. F. N., and Grant P. M..** *Orthogonal least squares learning algorithm for radial basis function networks.* IEEE Transactions on Neural Networks. 1991, Τόμ. 2, 2, σσ. 302-309.
25. **Haykin, S.** *Neural Networks: A comprehensive foundation.* Upper Saddle River, NJ, USA : Prentice Hall, 1999.
26. **Caudill, M.** *Neural Networks Primer.* [επιμ.] CA: Miller. San Francisco : Freeman Publications, 1989.
27. **Demuth Howard, Beale Mark.** *Neural Network Toolbox, For Use with MATLAB, User's Guide.* s.l. : MATLAB.
28. **Hagan, M. T., Demuth H. B., and Beale M. H..** *Neural Network.* Boston, MA : PWS Publishing, 1996.
29. **Nguyen, D., and Widrow B.. s.l. :** *Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights.* International Joint Conference on Neural Networks, 1990. Τόμ. 3, σσ. 21-26.
30. **Battiti, R.** *First and second order methods for learning: Between.* Neural Computation. 1992, Τόμ. 4, 2, σσ. 141–166.
31. **Lexington, MA:M.I.T. Lincoln.** DARPA Neural Network Study. 1988.
32. **Charalambous, C. s.l. :** *Conjugate gradient algorithm for efficient training of artificial neural networks.* IEEE Proceedings, 1992. Τόμ. 139, σσ. 301–310.
33. **Quinlan, J. R.** *C4.5: Programs for Machine Learning.* Los Altos : Morgan Kaufmann, 1993.
34. **MacKay, D. J. C.** *Bayesian interpolation.* 3, s.l. : Neural Computation, 1992, Τόμ. 4, σσ. 415-447.
35. **Himmelblau, D. M.** *Applied Nonlinear Programming.* New York : McGraw-Hill, 1972.
36. **Li, J., A. N. Michel, and W. Porod.** *Analysis and synthesis of a class of neural networks: linear systems operating on a closed hypercube*11, s.l. : IEEE Transactions on Circuits and Systems, 1989, Τόμ. 36, σσ. 1405-1422.
37. **Lippman, R. P. s.l. :** *An introduction to computing with neural nets.* IEEE ASSP Magazine, 1987, σσ. 4-22.
38. **Wasserman, P. D.** *Advanced Methods in Neural Computing.* New York : Van Nostrand Reinhold, 1993.
39. **Moller, M. F., s.l. :** *A scaled conjugate gradient algorithm for fast supervised learning*Neural Networks, Τόμ. 6, σσ. 525-533, 1993.
40. **Vogl, T. P., J. Mangis K., Rigler A. K., Zink W. T., and Alkon D. L.. s.l. :***Accelerating the convergence of the backpropagation method.* Biological Cybernetics, Τόμ. 59, σσ. 256-264, 1988.

41. **McCulloch, W. S., and Pitts W. H. s.l. :** *A logical calculus of ideas immanent in nervous activity*. Bulletin of Mathematical Biophysics, Τόμ. 5, σσ. 115-133.
42. **Riedmiller, M., and Braun H. s.l. :** *A direct adaptive method for faster backpropagation learning: The RPROP algorithm*. IEEE International Conference on Neural Networks, 1993.
43. **Wasserman, P. D.** *Advanced Methods in Neural Computing*. New York : Van Nostrand Reinhold, 1993.
44. **Fletcher Tristan,** *Support Vector Machines Explained*, UCLA, March 1, 2009.
45. **Cristianini N., Shawe-Taylor J..** *An introduction to support Vector Machines and other kernel-based learning methods*. New York, NY, USA : Cambridge University Press, 2000.
46. **Goldberg, D.** *Genetic Algorithms in optimization, search and machine learning*. s.l. : Addison-Wesley, 1989.
47. **Foresee, F. D., and Hagan M. T..** *Gauss-Newton approximation*. Proceedings of the 1997 International Joint. σσ. 1930-1935.
48. **Scholkopf B., Sung K., Burges C., Girosi F., Niyogi P., Poggio T. and Vapnik V.** *Comparing support vector machines with gaussian kernels, to radial basis function classifiers.*, IEEE Trans. Sign Processing, Τόμ. 45, σσ. 2758-2765 1997.
49. **Cristianini N., Shawe-Taylor J..** *An introduction to support Vector Machines: and other kernel-based learning methods*. New York, NY, USA : Cambridge University Press, SVM, 2000.
50. **David Hand, Heikki Mannila, and Padhraic Smyth,** *Principles of Data Mining*, MIT, Press, Cambridge, MA, 2001.
51. **Daniel T. Larose,** *Discovering Knowledge in data, An Introduction to Data Mining* by John Wiley & Sons, Inc. 2005.
52. **Narasimhamurthy, A.:** *Evaluation of diversity measures for binary classifier ensembles*. In: Oza, N.C., Polikar, R., Kittler, J., Roli, F. (eds.) MCS 2005. LNCS, 2005, vol. 3541, pp. 267–277. Springer, Heidelberg.
53. **Kuncheva, L.I.:** *Combining Pattern Classifiers: Methods and Algorithms*. Wiley InterScience, Chichester, 2004.
54. **Oh, S.:** *On the relationship between majority vote accuracy and dependency in multiple classifier systems*. Pattern Recog. Letters, 2003, 24, 359–363.
55. **Zhou, Z., Wu, J., Tang, W.:** *On the Relation between Dependence and Diversity in Multiple Classifier Systems*. In: Proceedings of the International Conference on Information Technology, ITCC 2005, Las Vegas, USA, pp. 134–139.
56. **Kuncheva L.I., Whitaker C.J., C.A. Shipp.** *Limits on the Majority Vote Accuracy in Classier Fusion*. Pattern Analysis and Applications, 2003, 6, 22-31.
57. **Littlestone Nick, Manfred K. Warmuth.** *The weighted Majority Algorithm*. UCSC CRL 91-28. October 26, 1992.
58. **Aliasgar Gangardiwala, Robi Polikar.** *Dynamically Weighted Majority Voting for Incremental Learning and Comparison of Three Boosting Based Approaches*. Proceedings of International Joint Conference on Neural Networks, Montreal, Canada, July 31 - August 4, 2005.
59. **Fawcett, T.** *An introduction to ROC analysis*. 2006, Pattern Recogn Lett, Τόμ. 27, σσ. 861-874.
60. **DANIEL T. LAROSE,** *Discovering knowledge in data. An Introduction to Data Mining*, John Wiley & Sons, Inc. 2005.
61. **Mahfouf, Linkens D.A.** *The Intelligent systems in Biomedicine laboratory in the Department of Automatic Control and Systems Engineering at the University of Sheffield, UK*. 2001, Artificial Intelligence in Medicine, Τόμ. 21, σσ. 171-176.
62. **Grossberg, S.** *Studies of the Mind and Brain*. Dordrecht, Holland : Reidel Press, 1982.
63. **Powell, M. J. D. s.l. :** *Restart procedures for the conjugate gradient method* Mathematical Programming, 1977. Τόμ. 12, σσ. 241-254.
64. **Rosenblatt, F.** *Principles of Neurodynamics*. Washington D.C. : Spartan Press, 1961.
65. **Jolliffe, I. T.** *Principal Component Analysis*. New York : Springer-Verlag, 1986.
66. **Gill, P. E., Murray W., and Wright M. H..** *Practical Optimization*. s.l. : New York: Academic Press, 1981.

67. **Bishop, C. M.** *Pattern Recognition and Machine Learning (Information Science and Statistics)*. s.l. : Springer, 2006.
68. **Dennis, J. E., and Schnabel R. B.** *Numerical Methods for*. Englewood Cliffs, NJ : Prentice-Hall, 1983.
69. **Chen, G. Leroy and H.** *Introduction to the special issue on decision support in medicine*. 2007, Decision Support Systems, Τόμ. 43, σσ. 1203–1206.
70. **Scales, L. E.** *Introduction to Non-Linear Optimization*. New York : Springer-Verlag, 1985.
71. **Fletcher, R., and Reeves C. M.** s.l. : *Function minimization by conjugate*. Computer Journal, 1964, Τόμ. 7, σσ. 149-154.
72. **Elman, J. L.** s.l. : *Finding structure in time*. Cognitive Science, 1990, Τόμ. 14, σσ. 179-211.
73. **Burges, C. J. C.** *Data Mining and Knowledge Discovery*. 1998.
74. **Breault J.L., Goodall C.R., P.J. Fos.** *Data mining a diabetic data warehouse*. Artificial Intelligence in Medicine 26 (2002) 37-54, Breast Cancer Research and Treatment 66: 51–57. 2001.
75. **Steurer J, Fischer JE, Bachmann LM, Koller M, ter Riet G.** *Communicating accuracy of tests to general practitioners: a controlled study*. BMJ 2002;324:824 .
76. **Reid MC, Lane DA, Feinstein AR.** s.l. : *Academic calculations versus clinical judgments: practicing physicians' use of quantitative measures of test accuracy*. Am. J. Med. 104 (4), 1998.
77. **Brent, R. P.** *Algorithms for Minimization Without Derivatives*. [επιμ.] Englewood Cliffs. s.l. : Prentice-Hall, 1973.
78. **Widrow, B., and M. E. Hoff.** *Adaptive switching circuits*. New York IRE : IRE WESCON Convention Record, 1960. σσ. 96-104.
79. **Widrow, B., and S. D. Sterns.** *Adaptive Signal Processing New York*. s.l. : Prentice-Hall, 1985.
80. **Puhan MA, Steurer J, Bachmann LM, ter Riet G.** *A randomized trial of ways to describe test accuracy: the effect on physicians' post-test probability estimates*. s.l. : Ann. Intern. Med. 143 (3), 2005.
81. **Papageorgiou, P.P. Groumpos.** *A new hybrid method using evolutionary algorithms to train Fuzzy Cognitive Maps*. E.I. 2005, Applied Soft Computing, Τόμ. 5, σσ. 409–431.

7 ΤΕΧΝΗΤΑ ΑΝΟΣΟΠΟΙΗΤΙΚΑ ΣΥΣΤΗΜΑΤΑ (ARTIFICIAL IMMUNE SYSTEMS)

Τα τελευταία χρόνια το ενδιαφέρον της επιστημονικής κοινότητας για τα τεχνητά ανοσοποιητικά συστήματα (Auto Immune Systems - AIS) αυξάνεται συνεχώς [1]. Ένα AIS είναι ένα υπολογιστικό μοντέλο που η λειτουργία του βασίζεται στη δομή και λειτουργία του βιολογικού ανοσοποιητικού συστήματος. Το βιολογικό ανοσοποιητικό σύστημα έχει τη φυσική ικανότητα να αναγνωρίζει και να διακρίνει ανάμεσα στα διάφορα κύτταρα, δικά του και ξένα, ποιο είναι υγιές ή παθογόνο και κατά συνέπεια αποδεκτό ή μη. Με αυτό τον τρόπο, αναγνωρίζοντας το ως ξένο, αντιμετωπίζει με επιτυχία κάθε παθογόνο μικροοργανισμό που εισβάλλει στο σώμα. Επίσης διαθέτει «μνήμη», μια σημαντικότητα ικανότητα που του επιτρέπει να αναγνωρίσει και να ανταποκριθεί πολύ πιο άμεσα στην εισβολή ενός παθογόνου μικροοργανισμού που έχει εισβάλλει και παλαιότερα στον οργανισμό. Αυτές οι σημαντικές ιδιότητες του ανοσοποιητικού συστήματος καθώς και άλλες όπως η ευρωστία, η προσαρμοστικότητα, η εξελιξιμότητα, η δυνατότητα διαφορετικής αλληλεπίδρασης σε διαφορετικές χρονικές στιγμές είναι ιδιότητες επιθυμητές στο χώρο των συστημάτων πληροφορικής. Είναι μια πραγματική πρόκληση για τον κόσμο της πληροφορικής να αναπτυχθούν συστήματα που να μπορούν να αντιμετωπίσουν όλο και πιο σύνθετα προβλήματα [2].

Η προέλευση του AIS έχει τις ρίζες της στις θεωρητικές εργασίες ανοσολογίας των Farmer, Perelson και Varela [3]-[5]. Σε αυτές τις εργασίες παρουσιάστηκαν διάφορα θεωρητικά πρότυπα ανοσοποιητικών δικτύων που χρησιμοποίησαν τις ιδιότητες μάθησης και διατήρησης μνήμης του βιολογικού ανοσοποιητικού συστήματος. Μετέπειτα πολλές μελέτες ασχολήθηκαν με το αντικείμενο και αποτέλεσαν την βάση για την εφαρμογή των AIS [6],[7]. Σήμερα τα AIS έχουν πολλές εφαρμογές κυρίως στην αναγνώριση προτύπων [8]-[9], στην ανίχνευση παρείσφρησης δικτύων υπολογιστών [10], στην ανίχνευση σφαλμάτων [11]-[12], στην βιοπληροφορική [13], στην εκμάθηση μηχανών (machine learning) [14]-[15], στην ρομποτική [16], στην πολύμορφη βελτιστοποίηση [17]-[18], ενώ έχουν περιγραφεί και ανοσοβασισμένες τεχνικές εκμάθησης χωρίς επιτήρηση [18]. Ειδικότερα για τις εφαρμογές στην αναγνώριση προτύπων και την ταξινόμηση γίνεται χρήση της ιδιότητας της κατανεμημένης επεξεργασίας και της μακροπρόθεσμης μνήμης του ανοσοποιητικού συστήματος που δρα ως ταξινομητής αναγνωρίζοντας συνεχώς «οικεία» από τα «μη οικεία» κύτταρα [8]-[9].

7.1 ΒΑΣΙΚΕΣ ΜΕΘΟΔΟΙ ΤΕΧΝΗΤΩΝ ΑΝΟΣΟΠΟΙΗΤΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ (ARTIFICIAL IMMUNE SYSTEMS - AIS)

7.1.1 ΑΡΝΗΤΙΚΗΣ ΕΠΙΛΟΓΗΣ

Το βιολογικό ανοσοποιητικό σύστημα έχει όπως προαναφέραμε την ιδιότητα του διαχωρισμού «οικείου» – «μη οικείου». Τα μόρια του ίδιου οργανισμού αντιμετωπίζονται και αυτά ως αντιγόνα (ίδια ή αυτο-αντιγόνα). Η ιδιότητα του ανοσοποιητικού συστήματος να μην αντιδρά στα ίδια αντιγόνα ονομάζεται ανοσολογική ανοχή (self tolerance). Για να επιτευχθεί ο διαχωρισμός «οικείου» – «μη οικείου», το ανοσοποιητικό σύστημα χρησιμοποιεί την τεχνική της αρνητικής επιλογής. Δεν πολλαπλασιάζονται όλα τα λεμφοκύτταρα που αναγνωρίζουν ένα αντιγόνο. Αντίθετα, τα λεμφοκύτταρα που αναγνωρίζουν ίδια αντιγόνα, υποβάλλονται σε έναν ελεγχόμενο θάνατο (απόπτωση) ή αδρανοποιούνται. Η αρνητική επιλογή ισχύει τόσο για τα T όσο και για τα B λεμφοκύτταρα και συμβαίνει κυρίως στο θύμο αδένα, ο οποίος παρέχει ένα περιβάλλον πλούσιο σε παρουσία αυτοαντιγόνων.

Οι αλγόριθμοι αρνητικής επιλογής εμπνέονται ακριβώς από αυτόν τον κύριο μηχανισμό του ανοσοποιητικού συστήματος να μην αντιδρά με τα ίδια αντιγόνα (αυτοαντιγόνα). Ο πρώτος αλγόριθμος αρνητικής επιλογής προτάθηκε για να ανιχνεύσει την επίδραση και τα αποτελέσματα ενός ιού σε ένα ηλεκτρονικό υπολογιστή [22]. Η έναρξη αυτού του αλγορίθμου γίνεται με το να παραχθεί ένα σύνολο αυτοσειρών S , το οποίο ορίζει την κανονική κατάσταση του συστήματος. Κατόπιν ο στόχος είναι να παραχθεί ένα σύνολο ανιχνευτών D , οι οποίοι δεσμεύουν/αναγνωρίζουν μόνο το συμπλήρωμα του S . Αυτοί οι ανιχνευτές μπορούν να εφαρμοστούν στα νέα δεδομένα προκειμένου να τα ταξινομήσουν ως αυτοαντιγόνα ή μη. Ο παρακάτω ψευδοκώδικας αποτυπώνει τον αλγόριθμο.

Είσοδος: S_{seen} = σύνολο γνωστών αυτό αντιγόνων

Εξοδος: D = σύνολο παραγόμενων ανιχνευτών

Έναρξη

Επανάληψη

*Τυχαία και δυναμικά παραγόμενοι ανιχνευτές τα τοποθετούν σε σύνολα P
Υπολογισμός της συγγένειας κάθε μέλους του P με κάθε μέλος του S_{seen}*

*Εάν τουλάχιστον ένα στοιχείο του S αναγνωρίζει έναν ανιχνευτή στο P σύμφωνα με κάποιο
κατώφλι*

*τότε ο ανιχνευτής απορρίπτεται, διαφορετικά προστίθεται στο σύνολο των διαθέσιμων
ανιχνευτών D*

Μέχρι να ικανοποιηθεί ένα κριτήριο για να σταματήσει

Τέλος

7.1.2 ΕΠΙΛΟΓΗΣ ΚΛΩΝΟΥ

Η αρχή επιλογής κλώνων [23] διέπει τη διαδικασία παραγωγής των Β-λεμφοκυττάρων και έχει ως εξής: η ενεργοποίηση των λεμφοκυττάρων συμβαίνει με τη δέσμευση ενός κατάλληλου αντιγόνου. Με την ενεργοποίηση παράγονται κλώνοι του λεμφοκυττάρου, οι οποίοι έχουν ίδιους υποδοχείς με το αρχικό λεμφοκύτταρο που συλλαμβάνει το αντιγόνο. Αυτό εξασφαλίζει ότι μόνο τα λεμφοκύτταρα που συνδέονται με το συγκεκριμένο αντιγόνο πολλαπλασιάζονται σε μεγάλους αριθμούς. Σύμφωνα με την θεωρία επιλογής κλώνου, κάθε λεμφοκύτταρο το οποίο έχει υποδοχείς αντιγόνων ειδικούς για τα μόρια του οργανισμού στον οποίο βρίσκεται, πρέπει να διαγραφεί (πεθάνει) κατά την διάρκεια ανάπτυξης των λεμφοκυττάρων. Αυτό εξασφαλίζει ότι μόνο τα ξένα προς τον οργανισμό αντιγόνα μπορούν να προκαλέσουν την κλωνοποίηση λεμφοκυττάρων και κατά συνέπεια να αποκλειστεί η πιθανότητα μιας καταστρεπτικής αυτοάνοσης αντίδρασης. Από αυτή την άποψη, το ανοσοποιητικό σύστημα μπορεί να ειπωθεί σαν ένας ταξινομητής (classifier) των αντιγόνων σε αυτοαντιγόνα και μη αυτοαντιγόνα, τα δεύτερα εκ των οποίων θεωρούνται παθογόνα και πρέπει να εκδιωχθούν από το σώμα.

Κατά τη διάρκεια της κλωνοποίησης των Β-λεμφοκυττάρων (αλλά όχι των Τ-λεμφοκυττάρων), η μέση ειδικότητα των αντισωμάτων αυξάνεται για το αντιγόνο που προκάλεσε την κλωνοποίηση. Αυτό το φαινόμενο ονομάζεται ωρίμανση συγγένειας (affinity maturation) και είναι υπεύθυνο για το γεγονός ότι σε μια επόμενη έκθεση στο ίδιο αντιγόνο, η ανοσολογική απάντηση του οργανισμού είναι αμεσότερη και αποτελεσματικότερη εξαιτίας της υψηλής ειδικότητας των αντισωμάτων που έχουν μεγαλύτερη συγγένεια με το αντιγόνο. Η ωρίμανση συγγένειας συμβαίνει μέσω της υπερμετάλλαξης (hypermutation) και του μηχανισμού επιλογής στον οποίο υποβάλλονται τα Β-λεμφοκύτταρα κατά την διάρκεια της κλωνικής τους επέκτασης. Η υπερμετάλλαξη αλλάζει τα Β-λεμφοκύτταρα με την εισαγωγή τυχαίων αλλαγών στο γενετικό τους υλικό.

Η θεωρία επιλογής κλώνου έχει χρησιμοποιηθεί ως έμπνευση για την ανάπτυξη των AIS ώστε να βοηθήσει σε θέματα βελτιστοποίησης και αναγνώρισης προτύπων. Συγκεκριμένα η ιδέα είναι εμπνευσμένη από την διαδικασία ωρίμανσης των B-λεμφοκυττάρων η οποία ενεργοποιείται από τα αντιγόνα, μαζί με το σχετικό μηχανισμό υπερμετάλλαξης (hypermutation). Καθώς και από την ιδιότητα της κυτταρικής μνήμης. Ειδικότερα υπάρχουν δύο κύριες ιδιότητες των B-λεμφοκυττάρων που χρησιμοποιούνται. Η πρώτη είναι ότι ο πολλαπλασιασμός των B-λεμφοκυττάρων, είναι ευθέως ανάλογος προς τη συγγένεια του αντιγόνου που δεσμεύουν, κατά συνέπεια όσο υψηλότερη είναι η συγγένεια τόσο περισσότεροι κλώνοι παράγονται. Η δεύτερη ιδιότητα είναι ότι οι μεταλλάξεις του B-λεμφοκυττάρου, είναι αντιστρόφως ανάλογες προς τη συγγένεια με το αντιγόνο το οποίο δεσμεύει. Χρησιμοποιώντας αυτές τις δύο ιδιότητες [17] αναπτύχθηκε μια από τις δημοφιλέστερες και ευρέως χρησιμοποιούμενες μεθόδους AIS, αυτή της επιλογής κλώνου η οποία λέγεται CLONALG. Έχει χρησιμοποιηθεί σε αναγνώριση προτύπων και βελτιστοποίηση πολυπαραγοντικών συναρτήσεων. Ο παρακάτω ψευδοκώδικας αποτυπώνει τον αλγόριθμο.

Είσοδος: S = σύνολο προτύπων προς αναγνώριση, n αριθμός των χειρότερων στοιχείων που θα επιλεγούν για διαγραφή

Έξοδος: M = σύνολο των ανιχνευτών μνήμης, ικανών να κατηγοριοποιήσουν κρυφά πρότυπα

Έναρξη

Δημιουργία αρχικού συνόλου αντισωμάτων, A

Για όλα τα πρότυπα στο S :

Ορίζει τη συγγένεια κάθε αντιγόνου του συνόλου A

Δημιουργεί κλώνους ενός υποσυνόλου των αντισωμάτων του συνόλου A με τη μεγαλύτερη συγγένεια

Ο αριθμός των κλώνων για κάθε αντίσωμα είναι ανάλογος με τη συγγένεια
 Διαμορφώνει τις ιδιότητες των κλώνων στο σύνολο A , και **βάζει** ένα αντίγραφο με τα αντισώματα με τη μεγαλύτερη συγγένεια στο A μέσα στο σύνολο μνήμης, M
Αντικαθιστά τα αντισώματα n με την μικρότερη συγγένεια στο A με νέα τυχαία παραγόμενα αντισώματα

Τέλος

Τέλος

7.1.3 ΑΝΟΣΟΠΟΙΗΤΙΚΑ ΔΙΚΤΥΑ

Το 1974 παρουσιάστηκε η θεωρία των ανοσοποιητικών δικτύων βασισμένη στις ιδιότητες της μάθησης και μνήμης του ανοσοποιητικού συστήματος [17]. Η βασική αρχή της θεωρίας είναι ότι κάθε υποδοχής λεμφοκυττάρων μέσα σε έναν οργανισμό μπορεί να αναγνωριστεί από έναν υποσύνολο του συνολικού πληθυσμού των υποδοχέων. Οι υποδοχείς αυτού του συνόλου αναγνώρισης έχουν το δικό τους σύνολο αναγνώρισης κοκ., έτσι ώστε να δημιουργείται ένα ανοσοποιητικό δίκτυο από αλληλεπιδράσεις. Στην περίπτωση απουσίας ξένων αντιγόνων το ανοσοποιητικό σύστημα θα επιδείξει μια συμπεριφορά ή δραστηριότητα η οποία θα προκύψει από την αλληλεπίδραση με τον εαυτό του και από αυτή την αλληλεπίδραση θα αναδυθεί μια ανοσολογική συμπεριφορά όπως η ανοχή και η μνήμη.

Είσοδος: S = Σύνολο προτύπων προς αναγνώριση, nt κατώφλι συγγένειας δικτύου, ct κατώφλι κλωνοποίησης, h αριθμός των κλώνων μεγαλύτερης συγγένειας, a αριθμός νέων αντιγόνων
Έξοδος: N = σύνολο ανιχνευτών μνήμης ικανών να κατηγοριοποιήσουν κρυφά πρότυπα

Έναρξη

Δημιουργεί ένα αρχικό τυχαίο σύνολο αντισωμάτων δικτύου, N

Για όλα τα πρότυπα στο S :

Ορίζει την συγγένεια με κάθε αντίσωμα στο N

Παράγει κλώνους ενός υποσυνόλου των αντισωμάτων στο N με την μεγαλύτερη συγγένεια. Ο αριθμός των κλώνων για κάθε αντίσωμα είναι ανάλογος με τη συγγένεια - ειδικότητα

Διαμορφώνει τις ιδιότητες αυτών των κλώνων στο σύνολο A , και βάζει h πλήθος κλώνων με τη μεγαλύτερη συγγένεια μέσα σε ένα σύνολο κλώνων μνήμης, C

Αφαιρεί όλα τα στοιχεία του συνόλου C των οποίων η συγγένεια με τα αντιγόνα είναι μικρότερη από ένα προκαθορισμένο κατώφλι ct

Ορίζει τη συγγένεια μεταξύ όλων των αντισωμάτων στο C και **αφαιρεί** τα αντισώματα των οποίων η συγγένεια μεταξύ τους είναι μικρότερη από το κατώφλι, c_i

Ενσωματώνει τους εναπομείναντες κλώνους του C στο N

Τέλος

Ορίζει την συγγένεια μεταξύ κάθε ζεύγους αντισωμάτων στο N και **αφαιρεί** όλα τα αντισώματα των οποίων η συγγένεια είναι μικρότερη από το κατώφλι n_i

Διαλέγει ένα τυχαίο αριθμό από τυχαία δημιουργούμενα αντισώματα και βάζει αυτά στο N

Τέλος ώσπου να ικανοποιηθεί το κριτήριο παύσης

Τέλος

7.1.4 ΠΡΟΤΕΙΝΟΜΕΝΗ ΤΕΧΝΙΚΗ

Στην παρούσα εργασία παρουσιάζεται μια διαδικασία δύο βημάτων για την ταξινόμηση των επιχρισμάτων FNA ως κακοήγη και καλοήγη. Αρχικά, πραγματοποιείται επιλογή χαρακτηριστικών γνωρισμάτων, προκειμένου να μειωθεί ο αριθμός των κυτταρολογικών χαρακτηριστικών παραμέτρων. Στη συνέχεια πραγματοποιείται εποπτευόμενη ταξινόμηση χρησιμοποιώντας έναν αλγόριθμο βασισμένο στη μέθοδο AIS, ως τροποποίηση του προτεινόμενου αλγορίθμου στο [24] AIRS και ο οποίος είναι βασισμένος στην πληθυσμιακή τεχνική (population-based technique). Ο ταξινομητής εκπαιδεύεται και εξετάζεται σε σχέση με στοιχεία επιβεβαιωμένα από μετεγχειρητικές ιστολογικές διαγνώσεις. Ο πυρήνας του προτεινόμενου αλγορίθμου είναι το αποκαλούμενο κουτί - κελί (BoxCell), το οποίο ορίζεται ως το σύνολο των παραλληλεπιπέδων στο χώρο των χαρακτηριστικών γνωρισμάτων. Οι κατάλληλα ορισμένοι τελεστές ενεργούν στα BoxCells προκειμένου αυτά να μετατραπούν σε μεμονωμένους, στοιχειώδεις ταξινομητές, καθένας εκ των οποίων είναι εξειδικευμένος στην εκτέλεση της ταξινόμησης σε μια συγκεκριμένη περιοχή του χώρου των χαρακτηριστικών γνωρισμάτων.

Η βασική διαφορά του προτεινόμενου αλγορίθμου AIS από τον AIRS [24] είναι ότι ο δεύτερος είναι ένας αλγόριθμος λογικής μιας χρήσης των δεδομένων, με την έννοια ότι κάθε αντιγόνο (σημείο εκπαίδευσης) εμφανίζεται στο σύστημα μόνο μια φορά κατά τη φάση εκπαίδευσης. Στην συγκεκριμένη εφαρμογή ο προτεινόμενος αλγόριθμος εμφανίζεται υπολογιστικά αποδοτικότερος, αφού οριακά υπερτερεί του AIRS, ενώ διατηρεί σημαντικά μικρότερο αριθμό BoxCells από τον αριθμό των κυττάρων μνήμης που παράγει ο AIRS. Για τα αποτελέσματα της συγκεκριμένης εφαρμογής ο μέσος αριθμός κυττάρων μνήμης του AIRS είναι 18.4, ενώ του AIS 13.4 BoxCells.

Μια ακόμα σημαντική διαφορά είναι ο ορισμός των BoxCells σαν παραλληλεπίπεδα στο χώρο των χαρακτηριστικών παραμέτρων σε σύγκριση με τα κύτταρα μνήμης (ARB) του AIRS, καθώς επιτρέπει διαφορετικό ορισμό της συνάρτησης απόδοσης. Στο [24] η απόδοση του BoxCell ορίζεται χρησιμοποιώντας τον όρο τιμή διέγερσης, η οποία ορίζεται ως το μέτρο συγγένειας ενός BoxCell με ένα συγκεκριμένο αντιγόνο. Ο ορισμός των BoxCells στην προτεινόμενη μεθοδολογία επιτρέπει μόνο δύο πιθανές περιπτώσεις για το BoxCell. Η μία να είναι το αντιγόνο (δεδομένο) μέσα στο BoxCell

(αντίσωμα) και η άλλη να είναι εκτός του BoxCell. Στην πρώτη το αντιγόνο χαρακτηρίζεται ως ορθά ή λανθασμένα ταξινομημένο ανάλογα με την κατηγορία του BoxCell, ενώ στην δεύτερη δεν επηρεάζει την απόδοση του BoxCell.

Ο προτεινόμενος αλγόριθμος διατηρεί το σύνολο των BoxCells τα οποία μπορούν να θεωρηθούν σαν ισοδύναμο των κυττάρων μνήμης του AIRS και χρησιμοποιεί απλούστερη διαχείριση της κλωνοποίησης σε σχέση με τον AIRS αφού διατηρούνται μόνο οι κλώνοι του τελευταίου κλωνοποιημένου BoxCell. Οι κλώνοι δεν επιτρέπεται να παράγουν κλώνους όπως στον AIRS, ενώ το κριτήριο για την επιλογή καλύτερου κλώνου είναι απλούστερο. Ο αριθμός των κλώνων οι οποίοι παράγονται είναι σταθερός σε όλη την διαδικασία εκπαίδευσης και ανεξάρτητος από την απόδοση των BoxCells, αλλά η μέγιστη επιτρεπόμενη διαφορά μεταξύ του γόνου και των κλώνων αυξάνεται όσο η απόδοση μειώνεται. Αξίζει να σημειωθεί ότι ο προτεινόμενος αλγόριθμος ακολουθεί και μια διαφορετική τακτική στο θέμα της αφαίρεσης BoxCell σε σχέση με τον AIRS. Η τακτική της αφαίρεσης των BoxCells είναι μια τροποποιημένη προσέγγιση των [1],[25]. Με αυτόν τον τρόπο η ελαχιστοποίηση του αριθμού των BoxCells και των ακμών τους καλύπτει όλο το χώρο των χαρακτηριστικών παραμέτρων.

7.2 ΤΑΞΙΝΟΜΗΣΗ ΜΕ ΤΗΝ ΧΡΗΣΗ ΤΕΧΝΗΤΩΝ ΑΝΟΣΟΠΟΙΗΤΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ (ARTIFICIAL IMMUNE SYSTEMS - AIS)

Η προτεινόμενη αλγοριθμική προσέγγιση χρησιμοποιεί ένα σύνολο στοιχείων εκπαίδευσης και ένα σύνολο στοιχείων δοκιμών για να εξεταστεί η ακρίβεια ταξινόμησης του αλγορίθμου. Έστω $A = \{a_1, a_2, \dots, a_n\}$ είναι ένα σετ από χαρακτηριστικές παραμέτρους ($n=12$ στην περίπτωση μας) και $d_j = (a_{j1}, a_{j2}, \dots, a_{jn})$ ($j=1, 2, \dots, N$) είναι το διάνυσμα των τιμών των χαρακτηριστικών παραμέτρων το οποίο αντιστοιχεί στο j th δείγμα (επίχρισμα στην περίπτωση μας). Στην περίπτωση μας, d_j είναι ένα διάνυσμα από δυαδικές τιμές με διάσταση 12. Επιπλέον, έστω $c_j \in C = \{1, 2, \dots, N_c\}$ χαρακτηρίζει την ταμπέλα της κατηγορίας του d_j . Στην περίπτωση μας έχουμε δύο κατηγορίες, $N_c = 2$.

Αρχικά ένας πληθυσμός από K BoxCells, B_k , ($k=1, 2, \dots, K$) δημιουργείται τυχαία. Κάθε BoxCell ενεργεί ως μεμονωμένος ταξινομητής στοιχείων. Ένα BoxCell ορίζεται ως $2n$ - διάστατο διάνυσμα που καθορίζει ένα n - διάστατο παραλληλεπίπεδο στο διάστημα χαρακτηριστικών παραμέτρων: οι πρώτες n παράμετροι καθορίζουν τη θέση της χαμηλότερης αριστερής γωνίας (κορυφής) στο διάστημα χαρακτηριστικών παραμέτρων, ενώ οι επόμενες παράμετροι n καθορίζουν τη θέση της ανώτερης δεξιά γωνίας (κορυφής) του παραλληλεπιπέδου, δηλαδή $B_k = (b_1^k, b_2^k, \dots, b_n^k, b_{n+1}^k, \dots, b_{2n}^k)$. Παραδείγματος χάριν, για ένα δυσδιάστατο διάστημα χαρακτηριστικών παραμέτρων, ένα BoxCell ορίζεται ως ένα παραλληλόγραμμο στο διάστημα χαρακτηριστικών παραμέτρων, με (b_1^k, b_2^k) τις συντεταγμένες για τη χαμηλότερη αριστερή κορυφή και (b_3^k, b_4^k) τις συντεταγμένες για την ανώτερη δεξιά κορυφή. Κάθε BoxCell φέρνει επίσης μια ετικέτα κατηγορίας $L_k \in C$. Ο αριθμός των BoxCells K , ποικίλλει καθώς εξελίσσεται η εκμάθηση του αλγορίθμου. Σε κάθε επανάληψη, κάθε BoxCell B_k εξετάζεται σε σχέση με όλα τα δεδομένα των στοιχείων d_j του συνόλου εκπαίδευσης. Ένα διάνυσμα χαρακτηριστικών παραμέτρων d_j καθορίζεται ότι είναι μέσα στο B_k εάν βρίσκεται στο παραλληλεπίπεδο που καθορίζεται από το B_k :

$$a_{ji} \in [b_i^k, b_{i+n}^k] \quad \text{Για κάθε } i=1, 2, \dots, n. \quad (7-1)$$

Ένα διάνυσμα χαρακτηριστικών d_j (δείγμα) καθορίζεται ότι είναι ορθά αναγνωρισμένο από το BoxCell B_k , εάν είναι μέσα στο B_k και $L_k = c_j$, ανήκει στην ίδια κατηγορία με το BoxCell. Αλλιώς, εάν το d_j

είναι μέσα στο B_k και $L_k \neq c_j$, τότε το διάνυσμα των χαρακτηριστικών παραμέτρων (δείγμα) έχει λανθασμένη ταξινόμηση.

Για κάθε B_k καθορίζεται ένας παράγοντας ορθής ταξινόμησης, f_k :

$$f_k = \begin{cases} \frac{cr_k - wr_k}{cr_k + wr_k}, & cr_k + wr_k \neq 0 \\ -\infty & , cr_k + wr_k = 0 \end{cases} \quad (7-2)$$

όπου cr_k και wr_k είναι αντίστοιχα ο αριθμός των διανυσμάτων των χαρακτηριστικών παραμέτρων που αναγνωρίστηκαν σωστά και λανθασμένα αντίστοιχα και βρίσκονται κοντά στην περιοχή του παραλληλεπίπεδου B_k .

Η τιμή του f_k είναι ένας πραγματικός αριθμός στο διάστημα $[-1,1]$, εκτός από την περίπτωση που το B_k δεν περιέχει διάνυσμα χαρακτηριστικών παραμέτρων (κενό BoxCell) οπότε σε αυτή την περίπτωση το αρνητικό άπειρο δίνεται σαν τιμή ικανότητάς ορθής ταξινόμησης.

Η φάση εκπαίδευσης του προτεινόμενου αλγορίθμου ταξινόμησης AIS συνοψίζεται ως εξής:

Αρχικοποίηση πληθυσμού των BoxCells

Για $i=1$ έως το μέγιστο αριθμό επαναλήψεων

Όσο (% μεταβολής μέσης απόδοσης πληθυσμού $> T$)

Υπολογίζει την απόδοση όλων των BoxCells

Κλωνοποίηση των BoxCells

Διογκώνει τα BoxCells τα οποία χαρακτηρίζονται ως κατηγορία 2

Ξεφουσκώνει τα μη κενά BoxCells με τουλάχιστον ένα ψευδώς θετικό διάνυσμα χαρακτηριστικών γνωρισμάτων, τα οποία χαρακτηρίζονται ως κατηγορία 1.

Υπολογίζει τη μέση απόδοση του πληθυσμού

Τέλος όσο

Αφαιρεί BoxCells από τον πληθυσμό του AIS.

Συγχωνεύει BoxCells

Προσθέτει νέα BoxCells

Αφαιρεί τα κενά BoxCells

Τέλος επανάληψης για

Η αναλυτική περιγραφή των διαδικασιών και των παραμέτρων δίνεται στην επόμενη παράγραφο.

7.3 ΠΕΡΙΓΡΑΦΗ ΠΡΟΤΕΙΝΟΜΕΝΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΑΞΙΝΟΜΗΣΗΣ AIS

7.3.1 ΑΡΧΙΚΟΠΟΙΗΣΗ ΠΛΗΘΥΣΜΟΥ BOXCELLS

Αρχικά, ένα σύνολο K_0 BoxCells δημιουργούνται τυχαία, χρησιμοποιώντας μια ομοιόμορφη κατανομή από τις επιτρεπόμενες τιμές παραμέτρων. Το εύρος των επιτρεπόμενων τιμών παραμέτρων ορίζεται από τις ελάχιστες και μέγιστες τιμές για κάθε χαρακτηριστική παράμετρο (στην περίπτωση μας, 0 και 1, αντίστοιχα), υπολογιζόμενες από τα δεδομένα εκπαίδευσης. Η κατηγορία L_k για κάθε BoxCell ορίζεται τυχαία, λαμβάνοντας μια τιμή από το σύνολο C των πιθανών κατηγοριών κατάταξης, στην περίπτωση μας κακώθες ή καλοήθες.

Προσοχή πρέπει να ληφθεί στην περίπτωση του αραιού πολυδιάστατου διαστήματος χαρακτηριστικών γνωρισμάτων, έτσι ώστε η πλειοψηφία των τυχαία παραγόμενων BoxCells να μην είναι κενά. Για αυτόν τον λόγο, σε περίπτωση διαστήματος χαρακτηριστικών γνωρισμάτων με μεγάλο εύρος τιμών, έχει αποδειχθεί χρήσιμο να δημιουργούνται οι συντεταγμένες της χαμηλότερης αριστερής (άνωτερης δεξιάς)

κορυφής του παραλληλεπιπέδου χρησιμοποιώντας το ελάχιστο (μέγιστο) μιας σειράς τυχαίων αριθμών για κάθε μια συντεταγμένη.

7.3.2 ΚΛΩΝΟΠΟΙΗΣΗ ΤΩΝ BOXCELLS

Μετά από τον υπολογισμό της αποδοτικότητας του πληθυσμού των BoxCells, κάθε BoxCell εισέρχεται σε μια φάση υπερμετάλλαξης, δηλαδή μια φάση αναπαραγωγής με την κλωνοποίηση, με μια αυξανόμενη πιθανότητα της μετάλλαξης για κάθε παράμετρο BoxCell. Ειδικότερα, για κάθε BoxCell (γόνος BoxCell), παράγεται τυχαία ένας σταθερός αριθμός κλώνων BoxCells N_{cl} , και ο κλώνος με την καλύτερη απόδοση αντικαθιστά το γονέα. Αν δεν υπάρχει BoxCell με καλύτερη απόδοση διατηρείται ο γονέας. Η διαδικασία κλωνοποίησης πραγματοποιείται σύμφωνα με τα εξής:

1) για ένα BoxCell με αποδοτικότητα ίση με 1 (μόνο ορθά αναγνωρισμένα διανύσματα χαρακτηριστικών γνωρισμάτων είναι μέσα στο BoxCell), δεν παράγονται κλώνοι.

2) για ένα BoxCell με την αποδοτικότητα μεταξύ $[0,1)$, οι κλώνοι BoxCells παράγονται χρησιμοποιώντας μια γκαουσιανή κατανομή πιθανότητας, με τη μέση τιμή ίση με τις τιμές παραμέτρων του BoxCell και τυπική απόκλιση σ_c , γραμμικά εξαρτώμενη από την τιμή απόδοσης του BoxCell, δηλαδή $\sigma_c(x) = 0.5 \sigma_0(1-x) + 0.5 \sigma_1(1+x)$, όπου $\sigma_0 = \sigma_c(-1)$, $\sigma_1 = \sigma_c(1)$.

3) για ένα κενό BoxCell (δηλ. η ικανότητα είναι ίση με $-\infty$), οι κλώνοι BoxCells παράγονται χρησιμοποιώντας μια ομοιόμορφη κατανομή από το επιτρεπτό πεδίο τιμών.

4) για ένα BoxCell με την αποδοτικότητα μεταξύ $[-1,0)$ (περισσότερα λάθος αναγνωρισμένα διανύσματα χαρακτηριστικών παραμέτρων σε σχέση με τα ορθά αναγνωρισμένα υπάρχουν μέσα στο BoxCell), η κατηγορία του BoxCell επανεκχωρείται στην κατηγορία της πλειοψηφίας των διανυσμάτων χαρακτηριστικών γνωρισμάτων που είναι μέσα στο BoxCell. Στην περίπτωση, όπου ο αριθμός κατηγοριών είναι δύο ($N_c=2$), η αλλαγή κατηγορίας θα οδηγήσει και σε τροποποίηση της ικανότητας του BoxCell, σύμφωνα με:

$$f_k^{new} = 1 + f_k^{old} \quad (7-3)$$

υπό τον όρο ότι $f_k^{old} < 0$. Κατόπιν, οι κλώνοι παράγονται όπως στην περίπτωση 2.

Η διαδικασία κλωνοποίησης, δημιουργεί έναν νέο πληθυσμό με μια αυξανόμενη μέση αποδοτικότητα. Εάν η μέση αποδοτικότητα του νέου πληθυσμού των BoxCells είναι σημαντικά καλύτερη από τη μέση αποδοτικότητα του προηγούμενου πληθυσμού, τότε η προαναφερθείσα διαδικασία επαναλαμβάνεται.

7.3.3 ΔΙΑΣΤΟΛΗ - ΣΥΡΡΙΚΝΩΣΗ ΤΩΝ BOXCELLS

Προκειμένου να εξασφαλιστεί ότι η κατανομή των BoxCells θα μεγιστοποιήσει την κάλυψη του διαστήματος χαρακτηριστικών παραμέτρων χρησιμοποιείται ο τελεστής της διαστολής - συρρίκνωσης. Η μέθοδος λειτουργεί σε κάθε BoxCell ως εξής:

Για $i=1$ n_I (n_I : προκαθορισμένος αριθμός επαναλήψεων)

Μεταβάλλει τη διάσταση του BoxCell σύμφωνα με έναν προκαθορισμένο παράγοντα, s_{id}

δέχεται το νέο BoxCell, μόνο εάν η νέα αποδοτικότητα είναι μεγαλύτερη ή ίση με την προηγούμενη

End for

Εάν $s_{id} > 1$ ($s_{id} < 1$) το BoxCell διαστέλλεται (συρρικνώνεται)

7.3.4 ΑΦΑΙΡΕΣΗ BOXCELL

Αυτός ο τελεστής αφαιρεί από τον πληθυσμό των BoxCells αυτά που παραμένουν κενά μετά από την εφαρμογή της κλωνοποίησης, δεδομένου ότι δεν συμβάλλουν στη γενική διαδικασία εκπαίδευσης. Επιπλέον, ένα γεμάτο BoxCell $\mathbf{B}_j$ αφαιρείται από τον πληθυσμό, εάν υπάρχει ένα BoxCell $\mathbf{B}_k$ τέτοιο ώστε:

- 1) $L_k = L_j$ και
- 2) 99% των σωστά ταξινομημένων διανυσμάτων χαρακτηριστικών παραμέτρων τα οποία περιλαμβάνονται στο $\mathbf{B}_j$ είναι μέσα και στο $\mathbf{B}_k$ και
- 3) $f_k > f_j$

Το αποτέλεσμα αυτής της διεργασίας είναι να προσδιοριστούν BoxCells τα οποία έχουν σημαντική επικάλυψη παραλληλεπίπεδων και να διατηρήσει το καλύτερο από αυτά. Ο σκοπός της αφαίρεσης BoxCell είναι να βελτιώσουν την τοποθέτηση των BoxCells και να διατηρήσουν την ποικιλομορφία του πληθυσμού τους.

7.3.5 ΣΥΓΧΩΝΕΥΣΗ BOXCELLS

Δύο BoxCells συγχωνεύονται εάν:

- 1) η Ευκλείδεια απόσταση δύο κεντροειδών είναι κάτω από ένα καθορισμένο όριο και
- 2) η αποδοτικότητα του BoxCell που συγχωνεύει δεν είναι χειρότερη από την αποδοτικότητα κάθε ενός από τα συγχωνευόμενα BoxCells.

Το Κεντροειδές (Centroid) ενός BoxCell καθορίζεται κατά τρόπο απλό στο διάστημα χαρακτηριστικών παραμέτρων, με τον υπολογισμό του μέσου όρου των n -διάστατων διανυσμάτων της "χαμηλότερης αριστερής" κορυφής και της "ανώτερης δεξιάς" κορυφής.

7.3.6 ΠΡΟΣΘΗΚΗ ΝΕΟΥ BOXCELL

Προκειμένου να βελτιωθεί η αναζήτηση του διαστήματος χαρακτηριστικών παραμέτρων, καθώς επίσης και για να αποφευχθεί η έλλειψη ποικιλομορφίας του πληθυσμού των BoxCells, που μπορεί να είναι η αιτία για την περιορισμένη βελτίωση της μέσης αποδοτικότητας των Boxcells, ένας αριθμός νέων BoxCells προστίθενται στον πληθυσμό, σύμφωνα με τον ακόλουθο αλγόριθμο:

Τα μη καταταγμένα διανύσματα χαρακτηριστικών παραμέτρων προσδιορίζονται από το σύνολο διανυσμάτων χαρακτηριστικών παραμέτρων. Ένα διάνυσμα χαρακτηριστικών παραμέτρων χαρακτηρίζεται ως μη αναγνωρισμένο, εάν:

- 1) δεν είναι μέσα σε οποιοδήποτε BoxCell, ή
- 2) κατατάσσεται λάθος από περισσότερα BoxCells, σε σχέση με τον αριθμό των BoxCells που το ταξινομούν ορθά.

Κάθε μη αναγνωρισμένο διάνυσμα χαρακτηριστικών παραμέτρων χρησιμεύει ως ένας "σπόρος" για να δημιουργηθεί τυχαία ένα νέο BoxCell, χρησιμοποιώντας μια κανονική τυχαία κατανομή με τη μέση τιμή ίση με το διάνυσμα χαρακτηριστικών παραμέτρων και μια αρκετά μικρή τυπική απόκλιση σ_a , που ορίζεται ως ένα μέρος του διανύσματος των μέγιστων τιμών χαρακτηριστικών παραμέτρων, μείον το διάνυσμα των ελάχιστων τιμών χαρακτηριστικών παραμέτρων.

Ο αριθμός των πρόσφατα εισαχθέντων BoxCells ορίζεται ως ένα σταθερό ποσοστό Pa του πληθυσμού των BoxCells. Εάν ο αριθμός των μη αναγνωρισμένων διανυσμάτων χαρακτηριστικών παραμέτρων είναι

μεγαλύτερος από τον αναμενόμενο αριθμό των νέων BoxCells, τότε τα εναπομείναντα μη αναγνωρισμένα διανύσματα χαρακτηριστικών παραμέτρων αγνοούνται. Διαφορετικά, τα υπόλοιπα των BoxCells που πρέπει να προστεθούν στον πληθυσμό δημιουργούνται χρησιμοποιώντας τη διαδικασία έναρξης BoxCell, με χρήση μιας ομοιόμορφης τυχαίας κατανομής από το σύνολο του διαστήματος χαρακτηριστικών παραμέτρων.

7.3.7 ΔΙΑΔΙΚΑΣΙΑ ΤΑΞΙΝΟΜΗΣΗΣ ΒΑΣΙΣΜΕΝΗ ΣΤΟ AIS

Αφότου έχει εκπαιδευθεί ο ταξινομητής AIS, εφαρμόζεται στα μη ταξινομημένα διανύσματα χαρακτηριστικών παραμέτρων ως ακολούθως. Ένα μη ταξινομημένο διάνυσμα χαρακτηριστικών παραμέτρων $\mathbf{d}_j$ παρουσιάζεται σε όλα τα BoxCells $\mathbf{B}_k$ και τα BoxCells, $\mathbf{B}_{k_j}$ ($k_j \in \{1, \dots, K\}$) των οποίων τα παραλληλεπίπεδα περιέχουν το $\mathbf{d}_j$ προσδιορίζονται. Υπολογίζεται η απόσταση του $\mathbf{d}_j$ από το κεντροειδές $\mathbf{cn}_{k_j}$, για κάθε BoxCell που περιέχει το $\mathbf{d}_j$, με τη βοήθεια του μετρικού D^∞ . Το μετρικό supremum (νόρμα άπειρο) για δύο n - διάστατα διανύσματα $x=(x_1, \dots, x_n)$, $y=(y_1, \dots, y_n)$ ορίζεται από την εξίσωση

$$D^\infty(x, y) = \max_{i=1, \dots, n} \{|x_i - y_i|\} \quad (7-4)$$

Εάν η τιμή D^∞ λαμβάνεται για $i=i^*$, τότε η υπολογιζόμενη απόσταση ομαλοποιείται από το μισό του πλάτους του παραλληλεπίπεδου κατά μήκος αυτής της διάστασης. Επομένως, για κάθε $\mathbf{d}_j$ ορίζεται ένας

$$\text{πραγματικός αριθμός } G_{jk} = \text{sgn}(1.5 - L_{k_j}) \left[1 - 2 \frac{D^\infty(\mathbf{d}_j, \mathbf{cn}_{k_j})}{b_{i^*+n}^{k_j} - b_{i^*}^{k_j}} \right] \text{ που παρουσιάζει πόσο εσωτερικά αυτό}$$

βρίσκεται στο $\mathbf{B}_{k_j}$, ή με άλλα λόγια, πόσο βέβαιος είναι ο ταξινομητής για να αναθέσει σε αυτό το BoxCell το διάνυσμα χαρακτηριστικών παραμέτρων $\mathbf{d}_j$ και να το ταξινομήσει στην κλάση L_{k_j} . Ο ορισμός του G_{jk} υπονοεί ότι $G_{jk} \in [0, 1)$ εάν $L_k=1$ (έτσι το BoxCell ανήκει στην κατηγορία 1) και $G_{jk} \in [-1, 0]$ εάν $L_k=2$ (έτσι το BoxCell ανήκει στην κατηγορία 2). Το διάνυσμα χαρακτηριστικών παραμέτρων $\mathbf{d}_j$ κατατάσσεται στην κατηγορία 1 εάν $\frac{1}{n_1} \sum G_{jk} > \text{thr}$ (n_1 είναι ο αριθμός των BoxCells που περιέχουν το

$\mathbf{d}_j$), διαφορετικά αυτό κατατάσσεται στην κατηγορία 2. Η παράμετρος thr καθορίζει τη συμπεριφορά του ταξινομητή. Εάν $\text{thr} = 0$, ο ταξινομητής δεν είναι προκατειλημμένος προς οποιαδήποτε από τις δύο κατηγορίες. Εάν $\text{thr} > 0$, τότε ο ταξινομητής ευνοεί τα ψευδώς θετικά έναντι των ψευδώς αρνητικών, ενώ εάν $\text{thr} < 0$, τότε ο ταξινομητής ευνοεί τα ψευδώς αρνητικά έναντι των ψευδώς θετικών. Εάν το $\mathbf{d}_j$ δεν είναι μέσα σε κανένα παραλληλεπίπεδο BoxCell, τότε θεωρείται αταξινόμητο. Μεταβάλλοντας την thr παράμετρο μεταξύ $[-0.5, 0.5]$, το αληθώς θετικό ποσοστό του ταξινομητή (ευαισθησία) σχεδιάζεται σε σχέση με το ψευδώς θετικό ποσοστό (1-ειδικότητα), παράγοντας την καμπύλη απόκρισης (ROC) [26], η οποία παρουσιάζεται στην Εικόνα 31. Στην παρούσα μελέτη επιλέχθηκε $\text{thr}=0$.

Εικόνα 31 Συνάρτηση μεταφοράς αλγορίθμου AIS

7.3.8 ΠΑΡΟΥΣΙΑΣΗ ΠΕΙΡΑΜΑΤΩΝ ΤΕΧΝΗΤΩΝ ΑΝΟΣΟΠΟΙΗΤΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ (ARTIFICIAL IMMUNE SYSTEMS)

Στα πλαίσια της παρούσας διδακτορικής διατριβής εξετάστηκε η απόδοση αλγοριθμικών μεθόδων τεχνητών ανοσοποιητικών συστημάτων (Artificial Immune Systems) στην εφαρμογή τους στο πεδίο ανεύρεσης μοντέλων που να περιγράφουν κακοήθεια θυρεοειδούς. Η υλοποίηση έγινε στο MATLAB και για να επιτευχθεί αυτό αρχικά χρησιμοποιήθηκε μια συνάρτηση που αναπτύχθηκε στα πλαίσια της διατριβής, η οποία διαχωρίζει τα δεδομένα σε σύνολα εκπαίδευσης, εκμάθησης και εφαρμόζει τις αναγκαίες αναλογίες για καλοήθη και κακοήθη. Έχει σαν ορίσματα τα στοιχεία εισόδου, την αναλογία συνόλου εκπαίδευσης και ελέγχου και την αναλογία καλοηθειών και κακοηθειών. Στις δοκιμές των τεχνητών αυτοάνοσων συστημάτων έγιναν οι εξής επιλογές των παραμέτρων που φαίνονται στον Πίνακα 7-1.

Πίνακας 7-1
Αρχικές Τιμές Για Τις Παραμέτρους Αλγορίθμου Ταξινόμησης

	ΠΑΡΑΜΕΤΡΟΣ	ΤΙΜΗ
Αρχικός αριθμός Boxcells	θ	50
Πλήθος κλώνων σε κλωνοποίηση	cl	100
	Μέγιστες επαναλήψεις	50
Όριο για % μεταβολή του μέσου όρου της απόδοσης του πληθυσμού	T	0.05
Σταθερά απόκλισης για κλωνοποίηση	σ_0	0.75
Σταθερά απόκλισης για κλωνοποίηση	σ_1	0.05
Ποσοστό πρόσφατα εισαχθέντων BoxCells	p_a	0.5
Σταθερά απόκλισης για προσθήκη	σ_a	0.02
Αριθμός επαναλήψεων διαστολής – συρρίκνωσης	n_I	5
Σταθερά διαστολής - συρρίκνωσης	s_{id}	1.5 (διόγκωση) 0.5 (ξεφούσκωμα)

7.3.9 ΑΠΟΤΕΛΕΣΜΑΤΑ

Τα βέλτιστα αποτελέσματα παρουσιάζονται στην περίπτωση όπου χρησιμοποιούνται όλα τα στοιχεία χωρίς καμία μείωση στην αναλογία κακοηθειών και καλοηθειών. Ο παρακάτω πίνακας σύγκρισης παρουσιάζει τα αποτελέσματα τα οποία απέφερε η δοκιμή του αλγορίθμου ταξινόμησης.

Πίνακας 7-2 Πίνακας Σύγχυσης (Confusion Matrix) Για Μέσο Όρο 10 Ανεξάρτητων Επαναλήψεων			
Μέθοδος	Πραγματική Κλάση	Προβλεπόμενη κλάση	
		Class 1	Class 2
AIS	Class 1	655	5
	Class 2	5	48

Ο επόμενος πίνακας προκύπτει έμμεσα από τον προηγούμενο και παρουσιάζει την ευαισθησία (sensitivity - fraction of true positives), την ειδικότητα (specificity - fraction of true negatives) και την συνολική ακρίβεια.

Πίνακας 7-3 Μέση Ακρίβεια Ταξινόμησης Για 10 Ανεξάρτητες Επαναλήψεις Του Αλγορίθμου			
Μέθοδος	Ευαισθησία (%)	Ειδικότητα (%)	Συνολική Ακρίβεια (%)
AIS	90	99,2	98,5

Οι επόμενοι δύο πίνακες είναι ο πίνακας σύγχυσης (confusion matrix) και ο πίνακας των τιμών για την ευαισθησία, την ειδικότητα και τη συνολική ακρίβεια για το τμήμα των δεδομένων τα οποία χαρακτηρίζονται ως ύποπτα (suspicious).

Πίνακας 7-4 Πίνακας Σύγχυσης (Confusion Matrix) Για Τα Δείγματα Τα Οποία Χαρακτηρίζονται Ύποπτα, Μέσος Όρος Για 10 Ανεξάρτητες Επαναλήψεις			
Μέθοδος	Πραγματική Κλάση	Προβλεπόμενη κλάση	
		Class 1	Class 2
AIS	Class 1	8	5
	Class 2	2	11

Πίνακας 7-5 Ακρίβεια Ταξινόμησης Για Τα Δείγματα Τα Οποία Χαρακτηρίζονται Ύποπτα, Μέσος Όρος Για 10 Ανεξάρτητες Επαναλήψεις			
Μέθοδος	Ευαισθησία (%)	Ειδικότητα (%)	Συνολική Ακρίβεια (%)
AIS	86,9	64,6	75,8

Η συνάρτηση μεταφοράς για τον παραπάνω AIS αλγόριθμο είναι όπως φαίνεται στην Εικόνα 32.

Εικόνα 32 Συνάρτηση μεταφοράς AIS

7.3.10 ΣΥΝΟΛΙΚΗ ΠΑΡΟΥΣΙΑΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΣΥΓΚΡΙΣΗ ΜΕ ΤΗΝ FNA ΚΑΙ ΜΕΤΑΞΥ ΑΛΓΟΡΙΘΜΙΚΩΝ ΜΕΘΟΔΩΝ

Στην παρούσα διδακτορική διατριβή παρουσιάστηκε ένα πλήρες σύστημα συλλογής δεδομένων, επιλογής των κατάλληλων δεδομένων και αξιολόγησης των στοιχείων τους για την ανάπτυξη ενός ολοκληρωμένου συστήματος υποβοήθησης απόφασης (Decision Support System - DSS) για την ταξινόμηση επιχρισμάτων βιοψίας FNA σε καλοήγη και κακοήγη. Επίσης βασικός στόχος που επετεύχθη ήταν η βελτίωση της ορθής ταξινόμησης των ύποπτων επιχρισμάτων (suspicious), για τα οποία είναι γνωστή η αδυναμία της μεθόδου FNA να τα κατηγοριοποιήσει ορθά. Σύμφωνα με την προτεινόμενη μεθοδολογία, αρχικά εφαρμόστηκε επιλογή των παραμέτρων ώστε να μειωθεί το πλήθος αυτών. Στη συνέχεια αναπτύχθηκε τόσο ο αλγόριθμος AIS όσο και η μέθοδος Πλειοψηφίας (majority vote - CMV) με χρήση τεσσάρων γνωστών αλγορίθμων Decision Trees, Artificial Neural Network, Support Vector Machine, και k-NN. Το σύστημα εκπαιδεύτηκε και ελέγχθηκε και για τις δύο μεθόδους σε σχέση με το δείγμα για το οποίο είχαμε ιστολογικές επιβεβαιώσεις - κατηγοριοποιήσεις (ground truth). Για περιπτώσεις οι οποίες χαρακτηρίστηκαν ως μη κακοήγεις από την FNA, και για τις οποίες δεν είχαμε ιστολογικές επιβεβαιώσεις, το δείγμα προέκυψε από την συνεκτίμηση και άλλων κλινικών, εργαστηριακών και απεικονιστικών εξετάσεων.

Στον Πίνακα 7-6 παρουσιάζεται συγκεντρωτικά πλέον ο πίνακας συσχέτισης όλων των μεθόδων ταξινόμησης, των δύο προτεινόμενων και της FNA. Τα αποτελέσματα αυτού του πίνακα χρησιμοποιούνται με στόχο να υπολογιστεί η ευαισθησία (ποσοστό των αληθώς θετικών), η ειδικότητα (ποσοστό των αληθώς αρνητικών) και η συνολική ακρίβεια (ποσοστό των αληθώς θετικών και αληθώς αρνητικών) όλων των μεθόδων και της βιοψίας FNA. Οι τιμές για τα παραπάνω παρουσιάζονται στον Πίνακα 7-7.

Όπως μπορούμε να δούμε από τους πίνακες Πίνακας 7-6 και Πίνακας 7-7 η συνολική ακρίβεια για τον προτεινόμενο αλγόριθμο βασισμένο στην μέθοδο της πλειοψηφίας (majority vote procedure - CMV) στο σύνολο ελέγχου είναι η μεγαλύτερη (98,6%) σε σχέση με όλους τους άλλους και με την FNA. Επίσης ο προτεινόμενος αλγόριθμος AIS είναι πολύ κοντά στο προηγούμενο ποσοστό (98,5%). Αναλυτικότερα, η ευαισθησία της μεθόδου CVM είναι αισθητά καλύτερη από όλους τους άλλους αλγορίθμους (89,1%), αλλά οριακά μικρότερη σε σχέση με αυτή της FNA (89,3%). Είναι όμως σημαντικά μικρότερη από την ευαισθησία του AIS, ο οποίος παρουσιάζει την υψηλότερη τιμή ευαισθησίας (90%). Σχετικά με την ειδικότητα η μέθοδος CMV παρουσιάζεται οριακά καλύτερη (99,4%) από τις μεθόδους AIS, ANN, SVM, k-NN, Δέντρων απόφασης (99,2% - 99,2% - 99,3% - 99% - 99,2% αντίστοιχα) και αισθητά καλύτερος από την FNA (98%).

Πίνακας 7-6
Πίνακας Σύγχυσης (Confusion Matrix)
Για Μέσο Όρο 10 Ανεξάρτητων Επαναλήψεων

Μέθοδος	Πραγματική Κλάση	Προβλεπόμενη Κλάση	
		Class 1	Class 2
Βιοψία FNA	Class 1	647	13
	Class 2	6	47
Πλειοψηφίας (CMV)	Class 1	656	4
	Class 2	6	47
AIS	Class 1	655	5
	Class 2	5	48
Δέντρα Αποφάσεων	Class 1	655	5
	Class 2	9	44
ANN	Class 1	655	5
	Class 2	10	43
SVM	Class 1	655	5
	Class 2	9	44
k-NN	Class 1	653	7
	Class 2	6	47

Πίνακας 7-7
Μέση Ακρίβεια Ταξινόμησης
Για 10 Ανεξάρτητες Επαναλήψεις Του Αλγορίθμου

Μέθοδος	Ευαισθησία (%)	Ειδικότητα (%)	Συνολική Ακρίβεια (%)
Βιοψία FNA	89,3	98,0	97,3
Πλειοψηφίας (CMV)	89,1	99,4	98,6
AIS	90	99,2	98,5
Δέντρα Αποφάσεων	83,4	99,2	98
ANN	80,3	99,2	97,8
SVM	83,6	99,3	98,1
k-NN	88,1	99,0	98,2

Στα πλαίσια της παρούσας διδακτορικής διατριβής οι μέθοδοι Πλειοψηφίας και AIS προτάθηκαν ώστε να βοηθήσουν στην διάγνωση της κακοήθειας θυρεοειδούς. Αρχικός στόχος ήταν να παρουσιαστεί μια ολοκληρωμένη μέθοδος ταξινόμησης για ιατρικές εφαρμογές. Ειδικότερα για το πεδίο της βιοψίας FNA ο στόχος ήταν να βελτιωθούν τα αποτελέσματα της. Ιδιαίτερα στον τομέα της ταξινόμησης των

επιχρισμάτων, τα οποία χαρακτηρίζονται ύποπτα «suspicious» από τους ειδικούς και το οποίο αποτελεί το αδύναμο σημείο της μεθόδου. Αυτό επετεύχθη με την ανάλυση των δεδομένων τόσο με την μέθοδο πλειοψηφίας όσο και με τον αλγόριθμο AIS.

Το τελευταίο και πλέον ενδιαφέρον σημείο προκύπτει από την ικανότητα των προτεινόμενων αλγορίθμων, στην ταξινόμηση των δειγμάτων τα οποία χαρακτηρίζονται ως ύποπτα από την βιοψία FNA. Η ορθή κατάταξη αυτών των δειγμάτων αποτελεί ένα πολύ δύσκολο κομμάτι, ακόμα και για τους πλέον ειδικούς. Στο σύνολο με επιβεβαιωμένα δείγματα (“golden” σετ) 13 από τα 26 ύποπτα επιχρίσματα με ιστολογική επιβεβαίωση αποδείχτηκαν ότι ήταν καλοήγη. Όλα τα περιστατικά που χαρακτηρίζονται ως ύποπτα από την FNA αντιμετωπίζονται κλινικά σαν κακοήθειες. Όπως μπορούμε να καταλάβουμε η βιοψία FNA δείχνει ευαισθησία 100%, αλλά ειδικότητα 0% (στην μέθοδο FNA υποθέτουμε ότι ένα περιστατικό που χαρακτηρίζεται ύποπτο είναι αληθώς θετικό). Η εφαρμογή όποιου άλλου μεμονωμένου αλγορίθμου βελτιώνει αισθητά την διαχείριση αυτών των περιπτώσεων, όπως φαίνεται στους παρακάτω πίνακες. Ιδιαίτερα όμως ο προτεινόμενος αλγόριθμος Πλειοψηφίας βελτιώνει εξαιρετικά την ταξινόμηση παρουσιάζοντας ευαισθησία 93,9% συγκρίσιμη με το 100% της FNA και ειδικότητα και συνολική ακρίβεια μεγαλύτερες, 75,4% και 84,6% αντίστοιχα, σε σχέση με τα 0% και 50% της FNA. Έτσι με τη χρήση του προτεινόμενου αλγορίθμου πλειοψηφίας το 75,4% από τις υπόνοιες μπορούν ασφαλώς να χαρακτηριστούν σαν μη κακοήθειες. Επιπλέον οι αλγόριθμοι Δέντρων Αποφάσεων, SVM και k-NN έχουν μεγαλύτερη ειδικότητα (60,8 – 66,9 – 63,9 αντίστοιχα) σε σχέση με την FNA αλλά έχουν και αισθητά μικρότερη ευαισθησία (62,3 - 63,1 – 60 αντίστοιχα) και συνολική ακρίβεια (61,5 – 65 – 61,9 αντίστοιχα). Οι αλγόριθμοι AIS και ANN παρουσιάζουν καλύτερες τιμές σε σχέση με τους άλλους ανεξάρτητους αλλά υποδεέστερες σε σχέση με τον αλγόριθμο πλειοψηφίας. Από όλες τις ανεξάρτητες μεθόδους η AIS έχει την καλύτερη απόδοση ακολουθούμενος από τα Νευρωνικά δίκτυα, ενώ την καλύτερη απόδοση συνολικά για τα ύποπτα έδωσε η μέθοδος της πλειοψηφίας.

Πίνακας 7-8
Πίνακας Σύγχυσης (Confusion Matrix)
Για Τα Δείγματα Τα Οποία Χαρακτηρίζονται Ύποπτα,
Μέσος Όρος Για 10 Ανεξάρτητες Επαναλήψεις

Μέθοδος	Πραγματική Κλάση	Προβλεπόμενη κλάση	
		Class 1	Class 2
Βιοψία FNA	Class 1	0,0	13,0
	Class 2	0,0	13,0
Πλειοψηφίας (CMV)	Class 1	10	3
	Class 2	1	12
AIS	Class 1	8	5
	Class 2	2	11
Δέντρα Αποφάσεων	Class 1	8	5
	Class 2	5	8
ANN	Class 1	9	4
	Class 2	3	10
SVM	Class 1	8	5
	Class 2	4	9
k-NN	Class 1	8	5
	Class 2	5	8

Πίνακας 7-9
Ακρίβεια Ταξινόμησης
Για Τα Δείγματα Τα Οποία Χαρακτηρίζονται Ύποπτα,
Μέσος Όρος Για 10 Ανεξάρτητες Επαναλήψεις

Μέθοδος	Ευαισθησία (%)	Ειδικότητα (%)	Συνολική Ακρίβεια(%)
Βιοψία FNA	100,0	0,0	50,0
Πλειοψηφίας (CMV)	93,9	75,4	84,6
AIS	86,9	64,6	75,8
Δέντρα Αποφάσεων	60,8	62,3	61,5
ANN	74,6	66,2	70,4
SVM	66,9	63,1	65,0
k-NN	63,9	60,0	61,9

7.4 ΒΙΒΛΙΟΓΡΑΦΙΑ

1. **de Castro L.N. and Timmis J.**, “An artificial immune network for multimodal function optimization”, Proc. of the IEEE Congress on Evolutionary Computation, (CEC’02), vol. 1, pp. 699-674, 2002.
2. **Newborough J. and Stepney S.** *A Generic Framework for Population-Based Algorithms, Implemented on Multiple FPGAs*. Published in the proceedings of ICARIS, 4th International Conference on Artificial Immune Systems, Banff, Canada, 2005.
3. **Farmer J., Packard N. and Perelson A.** *The immune system, adaptation, and machine learning*. Physica D, 22:187-204, 1986.
4. **Perelson A.** *Immune network theory*. Immunological Reviews, (10): 5-36, 1989.
5. **Varela F., A. Coutinho, B. Dupire and N. Vaz.** *Cognitive Networks: Immune and Neural and Otherwise*. Theoretical Immunology: Part Two, SFI Studies in the science of Complexity, 2, pp. 359 – 371. 1988.
6. **Bersini H. and Varela F.** *The Immune Recruitment Mechanism: A Selective Evolutionary Strategy*. Published in the proceedings of the 4th International Conference on Genetic Algorithms - R. Belew and L. Booker (Eds.) - Morgan Kaufman - pp. 520-526, 1991.
7. **Hightower R., Forrest S. and Perelson A.** *The evolution of emergent organization in immune system gene libraries*. In L. J. Eshelman (Ed.) Published in the proceedings of the Sixth International Conference on Genetic Algorithms, Morgan Kaufmann, San Francisco, CA, pp. 344-350, 1995.
8. **Carter J.**, “The immune systems as a model for pattern recognition and classification”, J. Am. Med. Inform. Assoc., vol. 7, no. 1, pp. 28-41, 2000.
9. **Watkins A. and Timmis J.**, “Artificial Immune Recognition System (AIRS): Revisions and Refinements”, Proceedings of the 1st International Conference on Artificial Immune Systems (ICARIS), J. Timmis and P. J. Bentley (eds.), University of Kent at Canterbury, pp. 173–181, 2002.
10. **D’haeseleer P., Forrest S. and Helman P.** *A distributed approach to anomaly detection*. Submitted to ACM Transactions on Information System Security, 1997.
11. **Canham and Tyrrell.** *A Multilayered Immune System for Hardware Fault Tolerance within an Embryonic Array*. Published in the proceedings of 1st International Conference on Artificial Immune Systems (ICARIS), University of Kent at Canterbury, UK, September 9th-11th, 2002.

12. **Modupe Ayara.** *Dissertation Title: An Immune Inspired Approach For Adaptable Error Detection in Embedded Systems*, University of Kent, Canterbury, UK, 2005.
13. **Cutello V., Nicosia G. and Pavone M.** *Exploring the Capability of Immune Algorithms: A Characterization of Hypermutation Operators*. Published in the proceedings of the Third International Conference on Artificial Immune Systems (ICARIS), Catania, Italy. September 13-16, 2004.
14. **Kim J. and Bentley P.** *Toward an Artificial Immune System for Network Intrusion Detection: An Investigation of Dynamic Clonal Selection*. Published in the proceedings of the special sessions on artificial immune systems in Congress on Evolutionary Computation, IEEE World Congress on Computational Intelligence, Honolulu, Hawaii, May 2002.
15. **Watkins Andrew, Timmis Jon, and Boggess Lois.** *Artificial Immune Recognition System (AIRS): An Immune Inspired Supervised Machine Learning Algorithm*. <http://ais.cs.memphis.edu/files/papers/AIS-bibliography-Dec07.pdf> 61 Genetic Programming and Evolvable Machines, 5(1), March 2004.
16. **Krohling, Zhou & Tyrrell.** *Evolving FPGA-based robot controllers using an evolutionary algorithm*. Published in the proceedings of the 1st International Conference on Artificial Immune Systems (ICARIS), University of Kent at Canterbury, UK, September 9-11, 2002.
17. **Jerne N. K.,** *Clonal selection in a lymphocyte network*. In *Cellular Selection and Regulation in the Immune Response*, Edelman, G. M. ed, Raven Press, New York, op. cit. p. 39-48, (1974).
18. **Timmis J., Neal M., and Hunt J.,** "An Artificial Immune System for data analysis", *Biosystems*, vol. 55, no. 1-3, pp. 143–150, 2000.
19. **Latifoglu F., Sahan S., S Kara., Gunes S.,** "Diagnosis of atherosclerosis from carotid artery Doppler signals as a real-world medical application of artificial immune systems", *Expert Systems with Applications*, vol. 33, pp. 786–793, 2007.
20. **Polat K. and Gunes S.,** "Medical decision support system based on artificial immune recognition immune system (AIRS), fuzzy weighted pre-processing and feature selection", *Expert Systems with Applications*, vol. 33, pp. 484–490, 2007.
21. **Freitas Alex and Timmis Jon.** *Revisiting the Foundations of Artificial Immune Systems: A Problem-oriented Perspective*. Published in the proceedings of Second International Conference on Artificial Immune Systems (ICARIS), September 1-3, 2003, Napier University, Edinburgh, UK.
22. **Forrest S., Perleson S., Allen L., and Cherukuri R.,** "Self-Nonself discrimination in a computer", *Proceedings of IEEE Symposium on Research in Security and Privacy*, Oakland, USA, pp. 202–212, 1994.
23. **De Castro Leandro, Timmis Jonathan** *Artificial immune systems: a new computational intelligence approach*, Springer, 2002.
24. **Watkins A., Timmis J., Boggess L.** *Artificial Immune Recognition System (AIRS): An Immune-Inspired Supervised Learning Algorithm*. *Genetic Programming and Evolvable*. 2004, vol. 5, 3, pp. 291-317.
25. **de Castro L.N. and von Zuben F.J.,** *Learning and optimization using the clonal selection principle*, *IEEE Trans. Evolutionary Computation*, Special Issue on Artificial Immune Systems, vol. 6, no. 3, pp 239-251, 2002.
26. **Fawcett, T.** *An introduction to ROC analysis*. 2006, *Pattern Recogn Lett*, Τόμ. 27, σσ. 861-874.

8 ΣΤΟΧΟΙ ΠΟΥ ΕΠΕΤΕΥΧΘΗΣΑΝ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΕΠΕΚΤΑΣΕΙΣ

Στην παρούσα διδακτορική διατριβή έγινε μια εκτενής ανάλυση δεδομένων από βιοψία θυρεοειδούς με λεπτή βελόνα FNA. Πέραν της επίτευξης των στόχων της η διατριβή αυτή μπορεί να αποτελέσει έναυσμα για μελλοντική εργασία πάνω στο συγκεκριμένο αντικείμενο, έχοντας προοπτικές επέκτασης τόσο σε επίπεδο τεχνολογικό όσο και ιατρικό.

8.1 ΣΤΟΧΟΙ ΠΟΥ ΕΠΕΤΕΥΧΘΗΣΑΝ

Οι στόχοι οι οποίοι επετεύχθησαν στην παρούσα διδακτορική διατριβή εκτείνονται σε όλο το φάσμα του διεπιστημονικού αντικειμένου της ιατρικής πληροφορικής. Πρώτος στόχος που επετεύχθη ήταν η αξιοποίηση του έντυπου αρχειακού υλικού, το οποίο δεν είχε υποβληθεί σε επεξεργασία ούτε με σύγχρονες μεθόδους ούτε σε αυτή την έκταση. Επίσης, με την μεθοδολογική ανάλυση που παρουσιάστηκε στην παρούσα διδακτορική διατριβή έγινε πλήρης ανάλυση και εύρεση των καλύτερων χαρακτηριστικών παραμέτρων τα οποία συμβάλλουν στην αξιολόγηση της FNA. Ο ειδικός παθολογοανατόμος ο οποίος συνεργάστηκε είχε μεγάλη εμπειρία και γνώση του αντικειμένου, και ως εκ τούτου είχε εντοπίσει σε υψηλό βαθμό τις σχετικές χαρακτηριστικές παραμέτρους κυρίως από εμπειρική παρατήρηση. Με την εφαρμογή των μεθόδων επιλογής χαρακτηριστικών παραμέτρων επιτεύχθηκε η επιστημονική επιβεβαίωση και ανεύρεση τόσο ενός εκάστου χαρακτηριστικού αλλά κυρίως συντέθηκε ένα σύνολο χαρακτηριστικών παραμέτρων τα οποία είναι κατάλληλα για την αξιολόγηση της FNA. Άλλο ένα σημαντικό επίτευγμα της παρούσας διδακτορικής διατριβής είναι η εφαρμογή μεθόδων εξόρυξης δεδομένων σε δεδομένα FNA. Εφαρμόστηκαν ανεξάρτητα τέσσερις γνωστές μέθοδοι εξόρυξης δεδομένων μεγάλης αξιοπιστίας, Δέντρα αποφάσεων (Decision Trees), Τεχνικά νευρωνικά δίκτυα (Artificial Neural Network), Μηχανές διανυσμάτων στήριξης (Support Vector Machine), και k - κοντινότερου γείτονα (k-NN). Τα αρχικά αποτελέσματα ήταν συγκρίσιμα με αυτά της FNA. Περαιτέρω βελτίωση όμως των αποτελεσμάτων επιτεύχθηκε με την εφαρμογή της μεθόδου πλειοψηφικού κανόνα (Majority Vote - CMV) συνδυάζοντας τα αποτελέσματα από την εφαρμογή τριών από τους προαναφερόμενους αλγορίθμους (Artificial Neural Network, Support Vector Machine και k-NN). Σαφέστατα όμως η προφανέστερη και μεγαλύτερη βελτίωση επετεύχθη με την εφαρμογή της τροποποιημένης μεθόδου τεχνητών αυτοάνοσων συστημάτων (Artificial Immune Systems – AIS) παρουσιάζοντας ιδιαίτερα βελτιωμένα αποτελέσματα στην ταξινόμηση των επιχρισμάτων, τα οποία χαρακτηρίζονται ως ύποπτα (suspicious) από τους ειδικούς και αποτελούν το αδύναμο σημείο της μεθόδου FNA. Μεγάλη σημασία έχει η αυξημένη ικανότητα ταξινόμησης των υπόπτων (suspicious) επιχρισμάτων με τη μέθοδο AIS σε σύγκριση με την ικανότητα του ειδικού, ειδικά αν συνυπολογιστεί ότι ο ειδικός παθολογοανατόμος είχε μεγάλη εμπειρία στο αντικείμενο και άρα μεγάλη επιτυχία στην αξιολόγηση.

8.2 ΜΕΛΛΟΝΤΙΚΕΣ ΕΠΕΚΤΑΣΕΙΣ

Μια πρώτη προοπτική επέκτασης θα μπορούσε να είναι η ανάπτυξη μιας βάσης δεδομένων στην οποία θα συλλέγονται όχι μόνο στοιχεία FNA αλλά και άλλων εξετάσεων του θυρεοειδούς. Έτσι η συστηματική συλλογή αποτελεσμάτων ανάλυσης αίματος για μέτρηση των ορμονών που ελέγχει ή παράγονται από το θυρεοειδή (TSH, T4, T3), υπερηχογραφήματα του θυρεοειδούς και του τραχήλου γενικότερα, σπινθηρογραφήματα με ιώδιο το οποίο χρησιμοποιείται για το διαχωρισμό «θερμού - ψυχρού» όζου, αξονικές ή μαγνητικές τομογραφίες και φυσικά βιοψίας κατόπιν εγχείρισης αφαιρέσεως του θυρεοειδούς μπορεί να εμπλουτίσει τα δεδομένα προς επεξεργασία και να οδηγήσει σε ακόμα

καλύτερα αποτελέσματα. Στην παρούσα διατριβή έγινε συνυπολογισμός παράλληλων εργαστηριακών μετρήσεων για την επιλογή του μεγαλύτερου και πλέον αξιόπιστου δείγματος όχι όμως για όλους τους ασθενείς, αφού τα δεδομένα δεν είχαν συλλεχθεί με ένα συστηματικό τρόπο. Η ύπαρξη ιατρικών εξετάσεων και κυρίως απεικονιστικών μπορεί να δώσει πολλές πληροφορίες οι οποίες θα αξιοποιηθούν με προχωρημένες μεθόδους αποφάσεων για δεδομένα εικόνων. Επίσης, η ύπαρξη των δεδομένων μετεγχειρητικής ιστολογικής επιβεβαίωσης μπορεί να αυξήσει σημαντικά τον αριθμό των περιπτώσεων δεδομένων τα οποία αποτελούν το «χρυσό δείγμα» (golden standard) και είναι οδηγός για κάθε μελέτη συμπεριφοράς των κυτταρολογικών παραμέτρων.

Μια άλλη προοπτική μελλοντικής επέκτασης αποτελεί και η περαιτέρω ανάλυση των δεδομένων με άλλες μεθόδους, πέραν αυτών που παρουσιάζονται στην παρούσα διδακτορική διατριβή για την εύρεση των πιθανών πηγών λανθασμένης διάγνωσης. Ένα ψευδώς αρνητικό αποτέλεσμα μπορεί να έχει θανατηφόρες συνέπειες ενώ τα ψευδώς θετικά οδηγούν τόσο σε ανώφελη επιβάρυνση του ασθενούς θέτοντας σε κίνδυνο την υγεία του, όσο και σε ανώφελο ψυχολογικό βάρος αλλά και σε οικονομική επιβάρυνση. Επίσης θα μπορούσαν να εξεταστούν και οι περιπτώσεις οι οποίες χαρακτηρίζονται ως ύποπτες και διαχειρίζονται ως κακοήθειες καταλήγοντας σε χειρουργικές επεμβάσεις.

Πέρα από τις ασθένειες που διαγιγνώσκονται σχετικά καλά με την μέθοδο FNA, ένα σύστημα ιατρικής υποβοήθησης μπορεί να αυξήσει την αποτελεσματικότητα στη διάγνωση. Υπάρχουν ασθένειες του θυρεοειδούς στις οποίες τόσο η FNA όσο και οι άλλες εξετάσεις θυρεοειδούς έχουν μέτρια έως χαμηλή διαγνωστική αξία και ακρίβεια. Οι παρακάτω νόσοι μπορούν να αποτελέσουν κάθε μια ένα ξεχωριστό αντικείμενο μελλοντικής έρευνας:

- Όζοι με εκτεταμένη κυστική εκφύλιση, ίνωση, αποτιτάνωση (δυσχέρεια λήψης)
- Νόσος Graves με λήψη αντιθυρεοειδικών
- Νόσος Hashimoto με επιθηλιακή ατυπία, εκτεταμένη μετάπλαση (οξύφιλη ή πλακώδη) ή έντονη ανοσολογική διέγερση (συχνά ψευδώς θετικά επιχρίσματα)
- Νόσος hashimoto με ανάπτυξη λεμφώματος (απαραίτητη και συχνά επίπονη η διαπίστωση μονοκλωνικού πληθυσμού στις ιστολογικές τομές)
- Εστία θηλώδους Ca σε έδαφος νόσου Hashimoto (συχνά ψευδώς αρνητικά επιχρίσματα)

Η νόσος Hashimoto αποτελεί συχνά συγχυτικό διαγνωστικό παράγοντα προς αμφότερες τις κατευθύνσεις.

Έναυσμα για μελλοντική επέκταση μπορεί να αποτελέσει το μειονέκτημα της μεθόδου FNA για ταξινόμηση στοιχείων που ανήκουν στην **"Γκρίζα ζώνη" της βιοψίας του θυρεοειδούς**. Οι οντότητες που αποτελούν τη "γκρίζα ζώνη" από διαγνωστική άποψη στο θυρεοειδή είναι:

- a) "Αδενωματώδεις" υπερπλαστικοί όζοι
- b) Θυλακιώδη αδενώματα (ή οξύφιλα από κύτταρα Huerthle), (οριακά και νεοπλάσματα)
- c) Θυλακιώδη καρκινώματα (ή οξύφιλα από κύτταρα Huerthle), (οριακά και νεοπλάσματα)

Στη γκρίζα αυτή ζώνη κινούμαστε ανάλογα με κάποια πενιχρά μορφολογικά κριτήρια και τη γενικότερη εμπειρία του ειδικού σ' ένα διαγνωστικό μεταίχμιο. Πολύ χρήσιμη για το χειρισμό των ασθενών που υπάγονται σ' αυτές τις ομάδες είναι και η προφορική επικοινωνία με τον Ενδοκρινολόγο, ώστε να συνεκτιμηθούν και λοιπές διαγνωστικές παράμετροι (λ.χ. μέγεθος όζου, ηλικία ασθενούς, υπερηχογραφική παρουσία κάψας, συνύπαρξη άλλων όζων, αντιθυρεοειδικά αντισώματα, κύηση κ.λπ.). Τόσο το σύστημα υποβοήθησης ιατρικών αποφάσεων που παρουσιάζεται στην διδακτορική διατριβή όσο και κάποιο πιο εξελιγμένο μπορεί να βοηθήσει ουσιαστικά στην περαιτέρω βελτίωση των αποτελεσμάτων της μεθόδου FNA για την **"Γκρίζα ζώνη"**.

Μελλοντική προέκταση θα μπορούσε να είναι η ανάπτυξη μιας νεότερης μεθοδολογίας για τη μείωση του μεγάλου αριθμού στοιχείων και παραμέτρων, τη συρρίκνωση του μεγέθους με την εξάλειψη εγγραφών αλλά και παραμέτρων σε περιπτώσεις που:

- τα στοιχεία μπορεί να είναι πάρα πολλά για κάποια μέθοδο εξόρυξης δεδομένων.
- ή ο αναμενόμενος χρόνος για την απόδοση μιας λύσης μπορεί να είναι πάρα πολύ μεγάλος.

Τέλος, στο επίπεδο της τέταρτης βαθμίδας επεξεργασίας η οποία αποτελεί και την καρδιά του συστήματος, υπάρχει μεγάλο πεδίο για περαιτέρω δοκιμές και άλλων μεθόδων εξόρυξης δεδομένων, ή και των ίδιων με παραλλαγές, από τις πολυπληθείς που υπάρχουν επιστημονικά.

ΥΠΟΤΙΤΛΟΙ ΕΙΚΟΝΩΝ

Εικόνα 1. Θέση Θυρεοειδούς.....	17
Εικόνα 2. Δομή και Σχήμα Θυρεοειδούς	17
Εικόνα 3. Παραθυρεοειδείς Αδένες	18
Εικόνα 4. Έκτοπες θέσεις Θυρεοειδούς.....	18
Εικόνα 5. Ιστολογία Θυρεοειδούς.....	19
Εικόνα 6. Θυρεοειδικό κυστίδιο	19
Εικόνα 7. Κύτταρα C	19
Εικόνα 8 Πίνακας Διαγνώσεων FNA	22
Εικόνα 9 Παράδειγμα καμπύλης ROC	30
Εικόνα 10 Μπλοκ διάγραμμα βάσης εφαρμογής συλλογής δεδομένων FNA.....	35
Εικόνα 11 Σχεδιασμός Βάσης Δεδομένων με εργαλεία SQL Server 2005.....	35
Εικόνα 12 Οθόνη Δημογραφικών Στοιχείων Ασθενή.....	36
Εικόνα 13 Οθόνη Καρτέλας Εξετάσεων Ασθενούς - Γενικά στοιχεία εξετάσεων.....	37
Εικόνα 14 Οθόνη Καρτέλας Εξετάσεων Ασθενούς – Αποτελέσματα κυτταρολογικής εξέτασης	38
Εικόνα 15 Οθόνη Καρτέλας Εξετάσεων Ασθενούς – Επιθηλιακά στοιχεία.....	38
Εικόνα 16 Οθόνη Καρτέλας Εξετάσεων Ασθενούς - Κυτταρολογικά στοιχεία.....	39
Εικόνα 17 Βαθμίδες Συστήματος Ιατρικής Υποβοήθησης.....	55
Εικόνα 18 Παράδειγμα Δέντρο αποφάσεων	64
Εικόνα 19 Τυπικό Δέντρο απόφασης για ταξινόμηση.....	68
Εικόνα 20 Συνάρτηση μεταφοράς (ROC) δέντρου απόφασης	69
Εικόνα 21 Σχηματικό διάγραμμα λειτουργίας νευρωνικού δικτύου	70
Εικόνα 22 Βασικός νευρώνας με R εισόδους	70
Εικόνα 23 Σχηματική παράσταση δικτύου μονοεπίπεδου νευρωνικού δικτύου πρόσθιας τροφοδοσίας (feedforward).....	71
Εικόνα 24 Σχηματική παράσταση δικτύου πολλαπλών επιπέδων πρόσθιας τροφοδοσίας (feedforward) .	71
Εικόνα 25 Συνάρτηση μεταφοράς νευρωνικού δικτύου	74
Εικόνα 26 Απόδοση αλγορίθμου νευρωνικού δικτύου.....	74
Εικόνα 27 Σχηματική παράσταση Μηχανών Διανυσματικής Υποστήριξης (SVM).....	76
Εικόνα 28 Συνάρτηση μεταφοράς αλγορίθμου SVM.....	79
Εικόνα 29 Γράφημα διασποράς μετρήσεων ηλεκτρολυτών ασθενών σε σχέση με την ηλικία και την επίδραση του φαρμάκου	80
Εικόνα 30 Συνάρτηση μεταφοράς για το παραπάνω k-NN	84
Εικόνα 31 Συνάρτηση μεταφοράς αλγορίθμου AIS	99
Εικόνα 32 Συνάρτηση μεταφοράς AIS	101

ΥΠΕΡΤΙΤΛΟΙ ΠΙΝΑΚΩΝ

Πίνακας 3-1 Confusion Matrix	29
Πίνακας 5-1 Ορισμοί στατιστικών όρων ευαισθησίας, ειδικότητας, αρνητικής αξίας πρόβλεψης, θετικής προφητικής αξίας	42
Πίνακας 5-2 Κατανομή κύριων κυτταρολογικών διαγνώσεων σε ασθενείς με μονήρεις ή πολυοζώδεις όζους	43
Πίνακας 5-3 Κατανομή Καλόηθων Κυτταρολογικών Διαγνώσεων	43
Πίνακας 5-4 Κατανομή κακόηθων κυτταρολογικών διαγνώσεων	44
Πίνακας 5-5 Κυτταρολογική Ανάλυση διαφόρων κατηγοριών Ασθενειών	45
Πίνακας 5-6 Κυτταρολογικές παράμετροι συσχετιζόμενες με Αληθώς Αρνητικά, Ψευδώς Αρνητικά, Αληθώς Θετικά και Ψευδώς Υποπτα	49
Πίνακας 5-7 Σύγκριση διαφόρων κυτταρολογικών χαρακτηριστικών μεταξύ Αληθώς Θετικών και Ψευδώς Θετικών/Υπόπτων	50
Πίνακας 5-8 Αποτελέσματα Κυτταρολογικών προς Ιστολογικές και κατανομή τους.....	51
Πίνακας 6-1 Κανόνες Υπολογισμού Κύριας Κατηγορίας (MAGOR_CATEGORIES).....	56
Πίνακας 6-2 Κανόνες Υπολογισμού Αναλυτικής Κατηγορίας (ANALYTIC_CATEGORIES).....	57
Πίνακας 6-3 Κανόνες Υπολογισμού Πλήρους Κατηγορίας (FULL_CATEGORIES).....	58
Πίνακας 6-4 Πίνακας Σύγχυσης (Confusion Matrix) Για Μέσο Όρο 10 Ανεξάρτητων Επαναλήψεων	68
Πίνακας 6-5 Μέση Ακρίβεια Ταξινόμησης Για 10 Ανεξάρτητες Επαναλήψεις Του Αλγορίθμου	68
Πίνακας 6-6 Πίνακας Σύγχυσης (Confusion Matrix) Για Τα Δείγματα Τα Οποία Χαρακτηρίζονται Υποπτα,.....	68
Πίνακας 6-7 Ακρίβεια Ταξινόμησης Για Τα Δείγματα Τα Οποία Χαρακτηρίζονται Υποπτα,.....	69
Πίνακας 6-8 Παράμετροι Που Χρησιμοποιήθηκαν Στα Νευρωνικά Δίκτυα	72
Πίνακας 6-9 Πινάκας Σύγχυσης (Confusion Matrix) Για Μέσο Όρο 10 Ανεξάρτητων Επαναλήψεων	73
Πίνακας 6-10 Μέση Ακρίβεια Ταξινόμησης Για 10 Ανεξάρτητες Επαναλήψεις Του Αλγορίθμου	73
Πίνακας 6-11 Πίνακας Σύγχυσης (Confusion Matrix) Για Τα Δείγματα Τα Οποία Χαρακτηρίζονται Υποπτα,.....	74
Πίνακας 6-12 Ακρίβεια Ταξινόμησης Για Τα Δείγματα Τα Οποία Χαρακτηρίζονται Υποπτα,.....	74
Πίνακας 6-13 Παράμετροι Που Χρησιμοποιήθηκαν Από Τις Μηχανές Στήριξης Διανυσμάτων (SVM's)	78
Πίνακας 6-14 Πίνακας Σύγχυσης (Confusion Matrix) Για Μέσο Όρο 10 Ανεξάρτητων Επαναλήψεων ..	78
Πίνακας 6-15 Μέση Ακρίβεια Ταξινόμησης Για 10 Ανεξάρτητες Επαναλήψεις Του Αλγορίθμου	78
Πίνακας 6-16 Πίνακας Σύγχυσης (Confusion Matrix) Για Τα Δείγματα Τα Οποία Χαρακτηρίζονται Υποπτα,.....	79
Πίνακας 6-17 Ακρίβεια Ταξινόμησης Για Τα Δείγματα Τα Οποία Χαρακτηρίζονται Υποπτα,.....	79
Πίνακας 6-18 Πίνακας Σύγχυσης (Confusion Matrix)	83
Πίνακας 6-19 Μέση Ακρίβεια Ταξινόμησης Για 10 Ανεξάρτητες Επαναλήψεις Του Αλγορίθμου	83
Πίνακας 6-20 Πίνακας Σύγχυσης (Confusion Matrix) Για Τα Δείγματα Τα Οποία Χαρακτηρίζονται Υποπτα,.....	83
Πίνακας 6-21 Ακρίβεια Ταξινόμησης Για Τα Δείγματα Τα Οποία Χαρακτηρίζονται Υποπτα,.....	84
Πίνακας 6-22 Ταξινόμηση Σύμφωνα Με Τις Αρχές Του Αλγόριθμου Πλειοψηφικού Κανόνα (MAJORITY VOTE PROCEDURE)	84
Πίνακας 6-23 Πίνακας Σύγχυσης (Confusion Matrix)	85
Πίνακας 6-24 Μέση Ακρίβεια Ταξινόμησης Για 10 Ανεξάρτητες Επαναλήψεις Του Αλγορίθμου	85
Πίνακας 6-25 Πίνακας Σύγχυσης (Confusion Matrix) Για Τα Δείγματα Τα Οποία Χαρακτηρίζονται Υποπτα,.....	86
Πίνακας 6-26 Ακρίβεια Ταξινόμησης Για Τα Δείγματα Τα Οποία Χαρακτηρίζονται Υποπτα,.....	86
Πίνακας 7-1 Αρχικές Τιμές Για Τις Παραμέτρους Αλγορίθμου Ταξινόμησης	99
Πίνακας 7-2 Πίνακας Σύγχυσης (Confusion Matrix)	100
Πίνακας 7-3 Μέση Ακρίβεια Ταξινόμησης Για 10 Ανεξάρτητες Επαναλήψεις Του Αλγορίθμου	100
Πίνακας 7-4 Πίνακας Σύγχυσης (Confusion Matrix) Για Τα Δείγματα Τα Οποία Χαρακτηρίζονται Υποπτα,.....	100
Πίνακας 7-5 Ακρίβεια Ταξινόμησης Για Τα Δείγματα Τα Οποία Χαρακτηρίζονται Υποπτα,.....	100

Πίνακας 7-6 Πίνακας Σύγχυσης (Confusion Matrix)	102
Πίνακας 7-7 Μέση Ακρίβεια Ταξινόμησης Για 10 Ανεξάρτητες Επαναλήψεις Του Αλγορίθμου	102
Πίνακας 7-8 Πίνακας Σύγχυσης (Confusion Matrix) Για Τα Δείγματα Τα Οποία Χαρακτηρίζονται Ύποπτα,	103
Πίνακας 7-9 Ακρίβεια Ταξινόμησης Για Τα Δείγματα Τα Οποία Χαρακτηρίζονται Ύποπτα,	104

ΠΑΡΑΡΤΗΜΑ Α

Πεδία Βάσης Δεδομένων και περιγραφή τους

A/A	ΕΛΛΗΝΙΚΗ ΠΕΡΙΓΡΑΦΗ	ΑΓΓΛΙΚΗ ΠΕΡΙΓΡΑΦΗ ΠΕΔΙΟΥ	ΟΝΟΜΑ ΠΕΔΙΟΥ ΒΑΣΗΣ	ΕΠΙΤΡΕΠΟ ΜΕΝΕΣ ΤΙΜΕΣ	ΚΑΤΗΓΟΡΙΑ ΣΤΟΙΧΕΙΩΝ
1	ΚΩΔΙΚΟΣ ΑΣΘΕΝΗ	PATIENT CODE	CODE	5 DIGIT AUTOMATIC CONTINUOUS	ΔΗΜΟΓΡΑΦΙΚΑ (DEMOGRAPHICS)
2	ΟΝΟΜΑ	FIRSTNAME	FIRSTNAME	STRING OF 30 CHARACTERS	
3	ΕΠΩΝΥΜΟ	SURNAME	SURNAME	STRING OF 30 CHARACTERS	
4	ΦΥΛΛΟ	GENT	GENT	MEN OR WOMEN	
5	ΗΛΙΚΙΑ	AGE	AGE	NUMBER 0 - 100	
6	ΗΜΕΡΟΜΗΝΙΑ ΓΕΝΝΗΣΗΣ	BIRTHDAY	BIRTHDAY	DATE	
7	ΙΣΤΟΡΙΚΟ	HISTORY	HISTORY	STRNG OF 100 CHARACTERS	
8	ΚΩΔΙΚΟΣ ΕΠΙΣΚΕΨΗΣ	VISIT CODE	CODE_VISIT	5 DIGIT AUTOMATIC CONTINUOUS	(ΠΕΡΙΓΡΑΦΗ ΕΠΙΣΚΕΨΗΣ) VISIT DESCRIPTION
9	ΗΜΕΡΟΜΗΝΙΑ ΕΠΙΣΚΕΨΗΣ	DATE OF VISIT	DATE_VISIT	DATE	
10	ΕΠΑΡΚΕΣ ΥΛΙΚΟ	ENOUGH MATERIAL	STAFF_YES	BINARY VALUE 0 OR 1	ΣΤΟΙΧΕΙΑ ΓΕΝΙΚΗΣ ΠΕΡΙΓΡΑΦΗΣ (GENERAL DESCRIPTION ELEMENTS)
11	ΜΗ ΕΠΑΡΚΕΣ ΥΛΙΚΟ	NOT ENOUGH MATERIAL	STAFF_NO	BINARY VALUE 0 OR 1	
12	ΑΜΦΙΒΟΛΟ ΥΛΙΚΟ	INDETERMINATE MATERIAL	STAFF_NOT_SURE	BINARY VALUE 0 OR 1	
13	ΜΟΝΗΡΗΣ ΟΖΟΣ	SOLITARY NODULE	OZOS_NUM_ONE	BINARY VALUE 0 OR 1	
14	ΠΟΛΥΟΖΩΔΗΣ	MULTI NODULE	OZOS_NUM_MANY	BINARY VALUE 0 OR 1	
15	ΘΕΣΗ ΟΖΟΥ	OZOS POSITION	OZOS_POS	RIGHT, LEFT, ISTHMOS, BOTH SIDES, RIGHT & ISTHMOS, LEFT & ISTHMOS,	

A/A	ΕΛΛΗΝΙΚΗ ΠΕΡΙΓΡΑΦΗ	ΑΓΓΛΙΚΗ ΠΕΡΙΓΡΑΦΗ ΠΕΔΙΟΥ	ΟΝΟΜΑ ΠΕΔΙΟΥ ΒΑΣΗΣ	ΕΠΙΤΡΕΠΟ ΜΕΝΕΣ ΤΙΜΕΣ	ΚΑΤΗΓΟΡΙΑ ΣΤΟΙΧΕΙΩΝ
				BOTH SIDES & ISTHMOS	
16	ΚΥΣΤΙΚΟΣ ΟΖΟΣ	CYSTIC	OZOS_TYPE_CYST	BINARY VALUE 0 OR 1	
17	ΣΥΜΠΑΓΗΣ ΟΖΟΣ	SOLID NODULE	OZOS_TYPE_COMP	BINARY VALUE 0 OR 1	
18	ΠΟΣΟΤΗΤΑ ΟΖΟΥ	VOLUME OF CYSTIC OZOS	OZOS_VOL	NUMBER MAX 10 DIGITS	
19	ΧΡΩΜΑ ΟΖΟΥ	COLOR OF CYSTIC OZOS	OZOS_COLOR	DARK COLOR, LIGHT COLOR	
20	ΚΥΤΤΑΡΟΛΟΓΙΚΟ ΣΥΓΚΡΙΣΗ ΜΕ ΙΣΤΟΛΟΓΙΚΟ	CYTOLOGIC COMPARED TO HISTOLOGIC DIAGNOSIS	OZOS_RESULT	TRUE, FALSE NEGATIVE, FALSE POSITIVE, FALSE SUSPICIOUS, WE DO NOT KNOW	ΑΠΟΤΕΛΕΣΜΑΤΑ (RESULTS)
21	ΙΣΤΟΛΟΓΙΚΟ	HISTOLOGIC DIAGNOSIS	OZOS_RESULT_NAME	Βλέπε πίνακα Error! Reference source not found.	
22	ΟΖΩΔΗΣ ΥΠΕΡΠΛΑΣΙΑ	NODULAR HYPERPLASIA	D_G_OZODIS_IPER	BINARY VALUE 0 OR 1	ΘΥΡΕΟΕΙΔΙΚΕΣ ΑΣΘΕΝΕΙΕΣ (THYROID DISEASES)
23	HASHIMOTO	HASHIMOTO	D_G_HASHI	BINARY VALUE 0 OR 1	
24	ΟΞΥΦΙΛΟΣ ΟΖΟΣ	OXIPHILIC NODULE	D_G_OXI	BINARY VALUE 0 OR 1	
25	ΣΤΟΙΧΕΙΑ ΘΥΡΕΟΕΙΔΙΤΙΔΑΣ	THYROIDITIS ELEMENTS	D_G_ST_THYRO	BINARY VALUE 0 OR 1	
26	ΥΠΟΞΕΙΑ ΘΥΡΕΟΕΙΔΙΤΙΣ	SUBACUTE THYROIDITIS	D_G_IPOXI	BINARY VALUE 0 OR 1	
27	ΜΗ ΕΙΔΙΚΗ ΚΥΣΤΙΚΗ	NON SPECIFIC CYSTIC	D_G_EIDIKI	BINARY VALUE 0 OR 1	
28	ΜΗ ΤΑΞΙΝΟΜΟΥΜΕΝΗ ΘΥΡΕΟΕΙΔΙΤΙΣ	NOT CLASSIFIED THYROIDITIS	D_G_NO_TAKSI	BINARY VALUE 0 OR 1	
29	ΑΥΤΟΑΝΟΣΗ ΘΥΡΕΟΕΙΔΟΠΑΘΕΙΑ	AUTOIMMUNE THYROIDITIS	D_G_AYTO	BINARY VALUE 0 OR 1	
30	ΚΑΡΚΙΝΩΜΑ ΧΑΜΗΛΗΣ ΔΙΑΦΟΡΟΠΟΙΗ	POURLY DIFFERENTIATED CARCINOMA	D_B_LOW	BINARY VALUE 0 OR 1	

A/A	ΕΛΛΗΝΙΚΗ ΠΕΡΙΓΡΑΦΗ	ΑΓΓΛΙΚΗ ΠΕΡΙΓΡΑΦΗ ΠΕΔΙΟΥ	ΟΝΟΜΑ ΠΕΔΙΟΥ ΒΑΣΗΣ	ΕΠΙΤΡΕΠΟ ΜΕΝΕΣ ΤΙΜΕΣ	ΚΑΤΗΓΟΡΙΑ ΣΤΟΙΧΕΙΩΝ
	ΣΗΣ				
31	ΜΕΤΑΣΤΑΣΗ	METASTASIS	D_B_META	BINARY VALUE 0 OR 1	
32	ΘΗΛΩΔΕΣ ΚΑΡΚΙΝΩΜΑ Ca	PAPILLARY CARCINOMA	D_B_THILO	BINARY VALUE 0 OR 1	
33	ΘΗΛΩΔΕΣ ΚΑΡΚΙΝΩΜΑ ΠΟΙΚΙΛΙΑ	PAPILLARY CARCINOMA VARIANT	D_B_PIKI	BINARY VALUE 0 OR 1	
34	ΑΝΑΠΛΑΣΤΙΚΟ ΚΑΡΚΙΝΩΜΑ	ANAPLASTIC CARCINOMA	D_B_ANAPL	BINARY VALUE 0 OR 1	
35	ΜΥΕΛΟΕΙΔΕΣ ΚΑΡΚΙΝΩΜΑ	MEDULLARY CARCINOMA	D_B_MIELO	BINARY VALUE 0 OR 1	
36	ΑΤΑΞΙΝΟΜΗΤΑ	UNCLASSIFIED	D_B_ATAXI	BINARY VALUE 0 OR 1	
37	ΛΕΜΦΩΜΑ	LYMPHOMA	D_B_LEMFO	BINARY VALUE 0 OR 1	
38	ΛΟΙΠΑ	OTHER	D_B_OTHER	BINARY VALUE 0 OR 1	
39	ΘΥΛΑΚΙΩΔΕΣ ΝΕΟΠΛΑΣΜΑ	FOLLICULAR NEOPLASM	D_B_NEO_THILO	BINARY VALUE 0 OR 1	
40	ΟΞΥΦΙΛΟ ΝΕΟΠΛΑΣΜΑ	OXIPHILIC NEOPLASM	D_B_NEO_OXI	BINARY VALUE 0 OR 1	
41	ΥΠΟΝΟΙΑ ΚΑΚΟΗΘΕΙΑΣ	SUSPICIOUS	D_SUSPICIOUS	BINARY VALUE 0 OR 1	
42	ΟΡΙΑΚΟ ΘΥΛΑΚΙΩΔΕΣ	BORDER FOLLICULAR	D_LIMIT_THILO	BINARY VALUE 0 OR 1	
43	ΟΡΙΑΚΟ ΟΞΥΦΙΛΟ	BORDER OXYPHILIC	D_LIMIT_OXI	BINARY VALUE 0 OR 1	
44	ΠΑΡΑΘΥΡΕΟΕΙΔΗΣ	PARATHYROID	O_PARA	BINARY VALUE 0 OR 1	ΑΛΛΕΣ ΑΣΘΕΝΕΙΕΣ (OTHER DISEASES)
45	ΛΕΜΦΑΔΕΝΑΣ	LYMPH NODE	O_LEMFO	BINARY VALUE 0 OR 1	
46	ΒΡΑΧΕΙΑΚΗ ΚΥΣΤΗ	BRANCHIAL CYST	O_BRAXO	BINARY VALUE 0 OR 1	
47	ΘΥΡΟΓΛΩΣΣΙΚΟ Υ ΠΟΡΟΥ	THYROGLOSSIC DUCT	O_THYRO	BINARY VALUE 0 OR 1	
48	ΦΑΓΟΚΟΚΚΥΩΜΑ	FAGOGRANULOMA	O_FAGO	BINARY VALUE 0 OR 1	
49	ΜΕΣΕΓΧΥΜΑΤΙΚΟ	MESENCHYMAL	O_MESE	BINARY VALUE 0 OR 1	
50	ΟΓΚΟΣ ΚΑΡΟΤΥΔΙΚΟΥ	CAROTID BODY TUMOR	O_KAROTI	BINARY VALUE 0	

A/A	ΕΛΛΗΝΙΚΗ ΠΕΡΙΓΡΑΦΗ	ΑΓΓΛΙΚΗ ΠΕΡΙΓΡΑΦΗ ΠΕΔΙΟΥ	ΟΝΟΜΑ ΠΕΔΙΟΥ ΒΑΣΗΣ	ΕΠΙΤΡΕΠΟ ΜΕΝΕΣ ΤΙΜΕΣ	ΚΑΤΗΓΟΡΙΑ ΣΤΟΙΧΕΙΩΝ
				OR 1	
51	ΣΙΕΛΟΓΟΝΟΣ	SALLIVARY	O_SIELO	BINARY VALUE 0 OR 1	
52	ΚΟΛΛΟΕΙΔΕΣ ΠΟΛΥ	COLLOID NUMERUS	EPI_SUB_CO L_BIG	BINARY VALUE 0 OR 1	
53	ΚΟΛΟΕΙΔΕΣ ΛΙΓΟ	COLLOID FEW	EPI_SUB_CO L_SMALL	BINARY VALUE 0 OR 1	
54	ΑΙΜΑ	BLOOD	EPI_SUB_BLOOD	BINARY VALUE 0 OR 1	ΣΤΟΙΧΕΙΑ ΥΠΟΣΤΡΩΜΑ ΤΟΣ (SUBSTRATE ELEMENTS)
55	ΑΜΥΛΟΕΙΔΕΣ	AMYLOID	EPI_SUB_AMI	BINARY VALUE 0 OR 1	
56	ΜΙΚΡΗ ΚΥΤΤΑΡΟΒΡΙΘ ΕΙΑ	LOW CELLULARITY	EPI_CYTO_S MALL	BINARY VALUE 0 OR 1	
57	ΜΕΓΑΛΗ ΚΥΤΤΑΡΟΒΡΙΘ ΕΙΑ	HIGH CELLULARITY	EPI_CYTO_BI G	BINARY VALUE 0 OR 1	
58	ΛΕΜΦΟΚΥΤΤΑΡ Α ΠΟΛΛΑ	LYMFOCYTES NUMERUS	EPI_SUB_AD D_LEMFO_BI G	BINARY VALUE 0 OR 1	ΠΡΟΣΘΕΤΑ ΣΤΟΙΧΕΙΑ ΥΠΟΣΤΡΩΜΑ ΤΟΣ (ADDITIONAL SUBSTRATE ELEMENTS)
59	ΛΕΜΦΟΚΥΤΤΑΡ Α ΛΙΓΑ	LYMFOCYTES FEW	EPI_SUB_AD D_LEMFO_S MALL	BINARY VALUE 0 OR 1	
60	ΝΕΚΡΩΤΙΚΟ	NEKROTIC	EPI_SUB_AD D_NEKRO	BINARY VALUE 0 OR 1	
61	ΑΣΒΕΣΤΙΟ	CALCIUM	EPI_SUB_AD D_ASB	BINARY VALUE 0 OR 1	
62	ΛΙΠΟΕΙΔΗ	LIPOID	EPI_SUB_AD D_LIPO	BINARY VALUE 0 OR 1	
63	ΦΑΓΟΚΥΤΤΑΡΑ ΠΟΛΛΑ	PHAGOCYTES NUMERUS	EPI_SUB_AD D_FAGO_BIG	BINARY VALUE 0 OR 1	
64	ΦΑΓΟΚΥΤΤΑΡΑ ΛΙΓΑ	PHAGOCYTES FEW	EPI_SUB_AD D_FAGO_S MALL	BINARY VALUE 0 OR 1	
65	ΠΟΛΥΜΟΡΦΟΠ ΥΡΗΝΑ	POLIMORPHS	EPI_SUB_AD D_POLI	BINARY VALUE 0 OR 1	
66	ΓΙΓΑΝΤΟΚΥΤΤΑ ΡΑ	MULTINUCLEATED GIANT CELLS	EPI_SUB_AD D_GIGA	BINARY VALUE 0 OR 1	
67	ΆΛΛΟ	OTHER	EPI_SUB_AD D_OTHER	BINARY VALUE 0 OR 1	
68	DETRITUS	DETRITUS	EPI_SUB_AD D_DET	BINARY VALUE 0 OR 1	
69	ΣΥΝΔΕΤΙΚΟ ΥΠΟΣΤΡΩΜΑ	CONNECTED TISSUE BACKGROUND	EPI_SUB_AD D_SIND	BINARY VALUE 0 OR 1	
70	ΨΑΜΜΩΔΗ	PSAMMOMA	EPI_SUB_AD D_PSAM	BINARY VALUE 0	

A/A	ΕΛΛΗΝΙΚΗ ΠΕΡΙΓΡΑΦΗ	ΑΓΓΛΙΚΗ ΠΕΡΙΓΡΑΦΗ ΠΕΔΙΟΥ	ΟΝΟΜΑ ΠΕΔΙΟΥ ΒΑΣΗΣ	ΕΠΙΤΡΕΠΟ ΜΕΝΕΣ ΤΙΜΕΣ	ΚΑΤΗΓΟΡΙΑ ΣΤΟΙΧΕΙΩΝ
				OR 1	
71	ΙΝΟΒΛΑΣΤΕΣ	FIBROBLAST	EPI_SUB_AD D_INO	BINARY VALUE 0 OR 1	
72	ΙΣΤΙΟΚΥΤΤΑΡΑ	ISTIOCYTES	EPI_SUB_AD D_ISTIO	BINARY VALUE 0 OR 1	
73	ΚΑΛΗ ΣΥΝΟΧΗ	TIGHT COHESION	EPI_CELL_SI NO_GOOD	BINARY VALUE 0 OR 1	ΣΥΝΟΧΗ (COHESION)
74	ΧΑΛΑΡΗ ΣΥΝΟΧΗ	LOOSE COHESION	EPI_CELL_SI NO_LOOSE	BINARY VALUE 0 OR 1	
75	ΜΟΝΟΣΤΟΙΒΑΔΩΣΗ	MONOLAYER	EPI_ARCHI_M ONO	BINARY VALUE 0 OR 1	ΑΡΧΙΤΕΚΤΟΝΙΚΑ ΣΤΟΙΧΕΙΑ (ARCHITECTURAL ELEMENTS)
76	ΜΙΚΡΟΘΥΛΑΚΙΑ	MICROFOLLICLES	EPI_ARCHI_M ICRO	BINARY VALUE 0 OR 1	
77	ΘΥΛΑΚΙΩΔΕΙΣ ΣΧΗΜΑΤΙΣΜΟΙ ΠΟΛΛΟΙ	FOLLICULAR FORMATIONS NUMERUS	EPI_ARCHI_T HILA_BIG	BINARY VALUE 0 OR 1	
78	ΘΥΛΑΚΙΩΔΕΙΣ ΣΧΗΜΑΤΙΣΜΟΙ ΛΙΓΟΙ	FOLLICULAR FORMATIONS FEW	EPI_ARCHI_T HILA_SMALL	BINARY VALUE 0 OR 1	
79	ΣΥΜΠΑΓΕΙΣ ΣΧΗΜΑΤΙΣΜΟΙ	SOLID FORMATIONS	EPI_ARCHI_S IMPA	BINARY VALUE 0 OR 1	
80	ΘΗΛΩΔΕΙΣ ΣΧΗΜΑΤΙΣΜΟΙ ΠΟΛΛΟΙ	PAPILLARY FORMATIONS NUMERUS	EPI_ARCHI_T HILO_BIG	BINARY VALUE 0 OR 1	
81	ΘΗΛΩΔΕΙΣ ΣΧΗΜΑΤΙΣΜΟΙ ΛΙΓΟΙ	PAPILLARY FORMATIONS FEW	EPI_ARCHI_T HILO_SMALL	BINARY VALUE 0 OR 1	
82	ΕΛΕΥΘΕΡΑ	FREE	EPI_ARCHI_F REE	BINARY VALUE 0 OR 1	
83	ΜΟΝΟΠΥΡΗΝΑ ΚΥΤΤΑΡΑ	MONONUCLEAR CELL	CYTO_CELL_ MONO	BINARY VALUE 0 OR 1	ΚΥΤΤΑΡΟΛΟΓΙΚΑ ΣΤΟΙΧΕΙΑ (CYTOLOGICAL ELEMENTS)
84	ΔΙΠΥΡΗΝΑ ΚΥΤΤΑΡΑ	BINUCLEI CELLS	CYTO_CELL_ DIP	BINARY VALUE 0 OR 1	
85	ΠΟΛΥΠΥΡΗΝΑ ΚΥΤΤΑΡΑ	MULTI NUCLEI CELLS	CYTO_CELL_ POLI	BINARY VALUE 0 OR 1	
86	ΘΥΛΑΚΙΚΑ ΚΥΤΤΑΡΑ	FOLLICULAR CELLS	CYTO_CELL_ THILA	BINARY VALUE 0 OR 1	
87	ΑΤΥΠΑ ΚΥΤΤΑΡΑ ΠΟΛΛΑ	ATYPICAL CELLS NUMERUS	CYTO_CELL_ ATI_BIG	BINARY VALUE 0 OR 1	
88	ΑΤΥΠΑ ΚΥΤΤΑΡΑ ΛΙΓΑ	ATYPICAL CELLS FEW	CYTO_CELL_ ATI_SMALL	BINARY VALUE 0 OR 1	
89	ΕΠΙΔΕΡΜΟΕΙΔΗ ΚΥΤΤΑΡΑ	EPIDERMOID CELLS	CYTO_CELL_ EPI	BINARY VALUE 0 OR 1	
90	ΟΞΥΦΙΛΑ ΚΥΤΤΑΡΑ	OXYPHILIC CELLS	CYTO_CELL_ OXI	BINARY VALUE 0	

A/A	ΕΛΛΗΝΙΚΗ ΠΕΡΙΓΡΑΦΗ	ΑΓΓΛΙΚΗ ΠΕΡΙΓΡΑΦΗ ΠΕΔΙΟΥ	ΟΝΟΜΑ ΠΕΔΙΟΥ ΒΑΣΗΣ	ΕΠΙΤΡΕΠΟ ΜΕΝΕΣ ΤΙΜΕΣ	ΚΑΤΗΓΟΡΙΑ ΣΤΟΙΧΕΙΩΝ
				OR 1	
91	ΑΝΙΣΟΚΥΤΤΑΡΩΣΗ	ANISOKITAROSIS	CYTO_CELL_MORFI_ANI	BINARY VALUE 0 OR 1	
92	ΟΜΑΛΗ	SMOOTH	CYTO_CELL_MORFI_OMA	BINARY VALUE 0 OR 1	
93	ΠΟΛΥΜΟΡΦΙΑ ΚΥΤΤΑΡΩΝ	POLIMORPHI OF CELLS	CYTO_CELL_MORFI_POLI	BINARY VALUE 0 OR 1	
94	ΣΑΦΗ ΟΡΙΑ	DISTINCT BORDERS	CYTO_CELL_LIMIT_DEF	BINARY VALUE 0 OR 1	
95	ΜΗ ΣΑΦΗ ΟΡΙΑ	NON DISTICT BORDERS	CYTO_CELL_LIMIT_NOT_DEF	BINARY VALUE 0 OR 1	
96	ΑΤΡΑΚΤΟΕΙΔΗ	SPINDLE	CYTO_CELL_PLASMA_ATRA	BINARY VALUE 0 OR 1	
97	ΚΛΑΣΣΙΚΟ	CLASSIC	CYTO_CELL_PLASMA_KLAS	BINARY VALUE 0 OR 1	
98	ΠΛΑΣΜΟΚΥΤΟΕΙΔΕΣ	PLASMACYTOID	CYTO_CELL_PLASMA_PLASMA	BINARY VALUE 0 OR 1	
99	ΔΙΑΥΓΕΣ	CLEAR	CYTO_CELL_PLASMA_DIA	BINARY VALUE 0 OR 1	
100	ΠΟΛΥΓΩΝΙΚΑ	POLIGONAL	CYTO_CELL_PLASMA_POLIG	BINARY VALUE 0 OR 1	
101	ΚΟΚΚΙΩΣΗ	CRANULATION	CYTO_CELL_PLASMA_KOKK	BINARY VALUE 0 OR 1	
102	ΚΑΙΝΟΤΟΠΙΩΣΗ	VACULISATION	CYTO_CELL_PLASMA_KAINO	BINARY VALUE 0 OR 1	
103	ΕΚΚΕΝΤΡΟΣ	ECCENTRIC	CYTO_NUCL_POS_NOT_CENTER	BINARY VALUE 0 OR 1	
104	ΚΕΝΤΡΙΚΟΣ	CENTRAL	CYTO_NUCL_POS_CENTER	BINARY VALUE 0 OR 1	
105	ΑΝΙΣΟΚΑΡΥΩΣΗ	ANISOKARYOSIS	CYTO_NUC_MORFI_ANI	BINARY VALUE 0 OR 1	
106	ΕΓΚΛΕΙΣΤΑ	INCLUSIONS	CYTO_NUC_MORFI_ESO	BINARY VALUE 0 OR 1	
107	ΠΥΚΝΩΤΙΚΟΣ ΠΥΡΗΝΑΣ	PYCNOTIC NUCLEUS	CYTO_NUC_MORFI_PICNO	BINARY VALUE 0 OR 1	
108	ΙΣΟΠΥΡΗΝΩΣΗ	ISOPIRINOSIS	CYTO_NUC_MORFI_ISO	BINARY VALUE 0 OR 1	
109	ΠΥΡΗΝΙΚΟΣ ΣΥΝΩΣΤΙΣΜΟΣ	NUCLEAR CROUDING	CYTO_NUC_MORFI_SINO	BINARY VALUE 0 OR 1	
110	ΠΥΡΗΝΙΑ	NUCLEOLI	CYTO_NUC_MORFI_PIRI	BINARY VALUE 0	

A/A	ΕΛΛΗΝΙΚΗ ΠΕΡΙΓΡΑΦΗ	ΑΓΓΛΙΚΗ ΠΕΡΙΓΡΑΦΗ ΠΕΔΙΟΥ	ΟΝΟΜΑ ΠΕΔΙΟΥ ΒΑΣΗΣ	ΕΠΙΤΡΕΠΟ ΜΕΝΕΣ ΤΙΜΕΣ	ΚΑΤΗΓΟΡΙΑ ΣΤΟΙΧΕΙΩΝ
				OR 1	
111	ΠΟΛΥΜΟΡΦΙΑ ΠΥΡΗΝΑ	NUCLEAR POLIMORPHY	CYTO_NUC_MORFI_POLI	BINARY VALUE 0 OR 1	

ΠΕΔΙΟ ΙΣΤΟΛΟΓΙΚΗΣ ΔΙΑΓΝΩΣΗΣ	NOT DONE, NODULAR HYPERPLASIA, HASHIMOTO, OXIPHILIC NODULE, THYROIDITIS ELEMENTS, SUBACUTE THYROIDITIS, NON SPECIFIC CYSTIC, NOT CLASSIFIED, AUTOIMMUNE THYROIDITIS THYROIDITIS, POURLY DIFFERENTIATED CARCINOMA, METASTASIS, PAPILLARY CARCINOMA, PAPILLARY CARCINOMA VARIANT, ANAPLASTIC CARCINOMA, MEDULLARY CARCINOMA, UNCLASSIFIED, LYMPHOMA, OTHER, FOLLICULAR NEOPLASM, OXIPHILIC NEOPLASM, SUSPICIOUS, BORDER LINE FOLLICULAR, BORDER LINE OXYPHILIC, PARATHYROID, LYMPH NODE, BRANCHIAL CYST, THYROGLOSIC DUCT, FAGOGRANULOMA, MESENCHYMAL, CAROTID BODY TUMOR, SALLIVARY, ΑΠΟΣΤΗΜΑ, FOLLICULAR ADENOMA, FOLLICULAR CARCINOMA, OXIPHYLIC ADENOMA, OXIPHYLIC CARCINOMA
-----------------------------	--

Δυνατές τιμές του πεδίου Ιστολογικής Διάγνωσης

ΠΑΡΑΡΤΗΜΑ Β

ΚΥΤΤΑΡΟΛΟΓΙΚΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ

A/ A	ΕΛΛΗΝΙΚΗ ΠΕΡΙΓΡΑΦΗ	ΑΓΓΛΙΚΗ ΠΕΡΙΓΡΑΦΗ ΠΕΔΙΟΥ	
1	ΚΟΛΛΟΕΙΔΕΣ ΠΟΛΥ	COLLOID NUMERUS	ΣΤΟΙΧΕΙΑ ΥΠΟΣΤΡΩΜΑΤΟΣ (SUBSTRATE ELEMENTS)
2	ΚΟΛΟΕΙΔΕΣ ΛΙΓΟ	COLLOID FEW	
3	ΚΟΛΟΕΙΔΕΣ ΥΠΑΡΞΗ	COLLOID EXIST	
4	ΑΙΜΑ	BLOOD	
5	ΑΜΥΛΟΕΙΔΕΣ	AMYLOID	
6	ΜΙΚΡΗ ΚΥΤΤΑΡΟΒΡΙΘΕΙΑ	LOW CELLULARITY	
7	ΜΕΓΑΛΗ ΚΥΤΤΑΡΟΒΡΙΘΕΙΑ	HIGH CELLULARITY	
8	ΚΥΤΤΑΡΟΒΡΙΘΕΙΑ ΥΠΑΡΞΗ	CELLULARITY EXIST	
9	ΛΕΜΦΟΚΥΤΤΑΡΑ ΠΟΛΛΑ	LYMFOCYTES NUMERUS	ΠΡΟΣΘΕΤΑ ΣΤΟΙΧΕΙΑ ΥΠΟΣΤΡΩΜΑΤΟΣ (ADDITIONAL SUBSTRATE ELEMENTS)
10	ΛΕΜΦΟΚΥΤΤΑΡΑ ΛΙΓΑ	LYMFOCYTES FEW	
11	ΛΕΜΦΟΚΥΤΤΑΡΑ ΥΠΑΡΞΗ	LYMFOCYTES EXIST	
12	ΝΕΚΡΩΤΙΚΟ	NEKROTIC	
13	ΑΣΒΕΣΤΙΟ	CALCIUM	
14	ΛΙΠΟΕΙΔΗ	LIPOID	
15	ΦΑΓΟΚΥΤΤΑΡΑ ΠΟΛΛΑ	PHAGOCYTES NUMERUS	
16	ΦΑΓΟΚΥΤΤΑΡΑ ΛΙΓΑ	PHAGOCYTES FEW	
17	ΦΑΓΟΚΥΤΤΑΡΑ ΥΠΑΡΞΗ	PHAGOCYTES EXIST	
18	ΠΟΛΥΜΟΡΦΟΠΥΡΗ ΝΑ	POLIMORPHS	
19	ΓΙΓΑΝΤΟΚΥΤΤΑΡΑ	MULTINUCLEATED GIANT CELLS	
20	ΆΛΛΟ	OTHER	
21	DETRITUS	DETRITUS	
22	ΣΥΝΔΕΤΙΚΟ ΥΠΟΣΤΡΩΜΑ	CONNECTED TISSUE BACKGROUND	
23	ΨΑΜΜΩΔΗ	PSAMMOMA	
24	ΙΝΟΒΛΑΣΤΕΣ	FIBROBLAST	
25	ΙΣΤΙΟΚΥΤΤΑΡΑ	ISTIOCYTES	
26	ΚΑΛΗ ΣΥΝΟΧΗ	TIGHT COHESION	ΣΥΝΟΧΗ (COHESION)
27	ΧΑΛΑΡΗ ΣΥΝΟΧΗ	LOOSE COHESION	
28	ΜΟΝΟΣΤΟΙΒΑΔΩΣΗ	MONOLAYER	ΑΡΧΙΤΕΚΤΟΝΙΚΑ ΣΤΟΙΧΕΙΑ (ARCHITECTUAL ELEMENTS)
29	ΜΙΚΡΟΘΥΛΑΚΙΑ	MICROFOLLICLES	
30	ΘΥΛΑΚΙΩΔΕΙΣ ΣΧΗΜΑΤΙΣΜΟΙ	FOLLICULAR FORMATIONS NUMERUS	

A/ A	ΕΛΛΗΝΙΚΗ ΠΕΡΙΓΡΑΦΗ	ΑΓΓΛΙΚΗ ΠΕΡΙΓΡΑΦΗ ΠΕΔΙΟΥ	
	ΠΟΛΛΟΙ		
31	ΘΥΛΑΚΙΩΔΕΙΣ ΣΧΗΜΑΤΙΣΜΟΙ ΛΙΓΟΙ	FOLLICULAR FORMATIONS FEW	
32	ΘΥΛΑΚΙΩΔΕΙΣ ΣΧΗΜΑΤΙΣΜΟΙ ΥΠΑΡΞΗ	FOLLICULAR FORMATIONS EXIST	
33	ΣΥΜΠΑΓΕΙΣ ΣΧΗΜΑΤΙΣΜΟΙ	SOLID FORMATIONS	
34	ΘΗΛΩΔΕΙΣ ΣΧΗΜΑΤΙΣΜΟΙ ΠΟΛΛΟΙ	PAPILLARY FORMATIONS NUMERUS	
35	ΘΗΛΩΔΕΙΣ ΣΧΗΜΑΤΙΣΜΟΙ ΛΙΓΟΙ	PAPILLARY FORMATIONS FEW	
36	ΘΗΛΩΔΕΙΣ ΣΧΗΜΑΤΙΣΜΟΙ ΥΠΑΡΞΗ	PAPILLARY FORMATIONS EXIST	
37	ΕΛΕΥΘΕΡΑ	FREE	ΚΥΤΤΑΡΟΛΟΓΙΚΑ ΣΤΟΙΧΕΙΑ (CYTOLOGICAL ELEMENTS)
38	ΜΟΝΟΠΥΡΗΝΑ ΚΥΤΤΑΡΑ	MONONUCLEAR CELL	
39	ΔΙΠΥΡΗΝΑ ΚΥΤΤΑΡΑ	BINUCLEI CELLS	
40	ΠΟΛΥΠΥΡΗΝΑ ΚΥΤΤΑΡΑ	MULTI NUCLEI CELLS	
41	ΘΥΛΑΚΙΚΑ ΚΥΤΤΑΡΑ	FOLLICULAR CELLS	
42	ΑΤΥΠΑ ΚΥΤΤΑΡΑ ΠΟΛΛΑ	ATYPICAL CELLS NUMERUS	
43	ΑΤΥΠΑ ΚΥΤΤΑΡΑ ΛΙΓΑ	ATYPICAL CELLS FEW	
44	ΑΤΥΠΑ ΚΥΤΤΑΡΑ ΥΠΑΡΞΗ	ATYPICAL CELLS EXIST	
45	ΕΠΙΔΕΡΜΟΕΙΔΗ ΚΥΤΤΑΡΑ	EPIDERMOID CELLS	
46	ΟΞΥΦΙΛΑ	OXYPHILIC CELLS	
47	ΑΝΙΣΟΚΥΤΤΑΡΩΣΗ	ANISOKYTAROSIS	
48	ΟΜΑΛΗ	SMOOTH	
49	ΠΟΛΥΜΟΡΦΙΑ ΚΥΤΤΑΡΩΝ	POLIMORPHI OF CELLS	
50	ΣΑΦΗ ΟΡΙΑ	DISTINCT BORDERS	
51	ΜΗ ΣΑΦΗ ΟΡΙΑ	NON DISTICT BORDERS	
52	ΑΤΡΑΚΤΟΕΙΔΗ	SPINDLE	
53	ΚΛΑΣΣΙΚΟ	CLASSIC	
54	ΠΛΑΣΜΟΚΥΤΟΕΙΔΕΣ	PLASMACYTOID	
55	ΔΙΑΥΓΕΣ	CLEAR	
56	ΠΟΛΥΓΩΝΙΚΑ	POLIGONAL	
57	ΚΟΚΚΙΩΣΗ	CRANULATION	
58	ΚΑΙΝΟΤΟΠΙΩΣΗ	VACULISATION	
59	ΕΚΚΕΝΤΡΟΣ	ECCENTRIC	
60	ΚΕΝΤΡΙΚΟΣ	CENTRAL	

A/ Α	ΕΛΛΗΝΙΚΗ ΠΕΡΙΓΡΑΦΗ	ΑΓΓΛΙΚΗ ΠΕΡΙΓΡΑΦΗ ΠΕΔΙΟΥ	
61	ΑΝΙΣΟΚΑΡΥΩΣΗ	ANISOKARYOSIS	
62	ΕΓΚΛΕΙΣΤΑ	INCLUSIONS	
63	ΠΥΚΝΩΤΙΚΟΣ ΠΥΡΗΝΑΣ	RYCNOTIC NUCLEUS	
64	ΙΣΟΠΥΡΗΝΩΣΗ	ISOPIRINOSIS	
65	ΠΥΡΗΝΙΚΟΣ ΣΥΝΩΣΤΙΣΜΟΣ	NUCLEAR CROUDING	
66	ΠΥΡΗΝΙΑ	NUCLEOLI	
67	ΠΟΛΥΜΟΡΦΙΑ ΠΥΡΗΝΑ	NUCLEAR POLIMORPHY	

ΔΗΜΟΣΙΕΥΣΕΙΣ

Δημοσιεύσεις σε Διεθνή Επιστημονικά Περιοδικά με κριτές

[Π1] **Emmanouil A. Zoulas**, Pantelis A. Asvestas, George K. Matsopoulos, Sofia Tseleni-Balafouta, *A Decision Support System for Assisting FNA Diagnosis of Thyroid Malignancy*, (έχει γίνει αποδεκτό για δημοσίευση από το περιοδικό "Analytical and Quantitative Cytology and Histology"

[Π2] K.K. Delibasis, P.A. Asvestas, G.K. Matsopoulos, **E. Zoulas**, and S. Tseleni-Balafouta, "Computer-Aided Diagnosis of Thyroid Malignancy Using an Artificial Immune System Classification Algorithm", presented at IEEE Transactions on Information Technology in Biomedicine, 2009, pp.680-686.

Δημοσιεύσεις σε Ελληνικά και Διεθνή Συνέδρια με κριτές

[Σ1]**E. Zoulas**, P.A. Asvestas, G.K. Matsopoulos, N. Uzunoglu, S. Tseleni-Balafouta, H. Gakiopoulou, *A data mining approach for classifying FNA thyroid data*, 6TH EUROPEAN SYMPOSIUM ON BIOMEDICAL ENGINEERING, "Center for Mediterranean Architecture". 19 - 21 Ιουνίου 2008, Chania, Crete, Greece.

[Σ2]**Emmanouil. Zoulas**, Harikleia. Gakiopoulou, Efstratios Patsouris, George Matsopoulos, Nikolaos Uzunoglu, Sofia Tseleni-Balafouta, H. Gakiopoulou, *Analysis of the cytological features of 9054 FNA of thyroid and correlation to the cytological and histological diagnoses*, Volume 451, Number 2, August 2007, 21st European Congress of Pathology, 8-13 September 2007, Constantinopolis, Turkey.