ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΝΙΚΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Διπλωματική Εργασία

ΑΝΑΠΤΥΞΗ ΑΛΓΟΡΙΘΜΩΝ ΓΙΑ ΕΞΑΓΩΓΗ ΓΟΝΙΔΙΑΚΩΝ ΥΠΟΓΡΑΦΩΝ ΣΤΗΝ ΟΣΤΕΟΑΡΘΡΙΤΙΔΑ

ΜΠΑΚΑΛΗ ΚΩΝΣΤΑΝΤΙΝΑ

Επιβλέπων Καθηγητής: Καθηγητής Ζερβάκης Μιχάλης

Εξεταστική Επιτροπή: Καθηγητής Ζερβάκης Μιχάλης

Καθηγητής Λιάβας Αθανάσιος

Καθηγητής Πετράκης Ευριπίδης

Χανιά, Απρίλιος 2013

Θα ήθελα να ευχαριστήσω

Τον επιβλέποντα καθηγητή κ. Μιχάλη Ζερβάκη για τη καθοδήγηση και τη πολύτιμη βοήθειά του καθ’ όλη τη διάρκεια της παρούσας διπλωματικής εργασίας.

Την κα. Αικατερίνη Μπέη για τη μεγάλη βοήθεια που μου προσέφερε στο βιολογικό κομμάτι της εργασίας καθώς και για τη διαρκή υποστήριξη και τις πολύτιμες συμβουλές τις σε όλη τη περίοδο της εργασίας μέχρι και την εκπλήρωσή της.

Τους καθηγητές κ. Λιάβα Αθανάσιο και κ. Πετράκη Ευριπίδη για τη συμμετοχή τους στη παρουσίαση και αξιολόγηση αυτής της εργασίας.

Την οικογένεια μου για την αμέριστη συμπαράστασή τους όλα αυτά τα χρόνια των προπτυχιακών μου σπουδών.

Περίληψη

Οι μικροσυστοιχίες DNA καθιστούν δυνατή τη ποσοτική ανάλυση της έκφρασης χιλιάδων γονιδίων με ένα παράλληλο και διεξοδικό τρόπο. Το πρότυπο της γονιδιακής έκφρασης που παράγεται, γνωστό ως προφίλ έκφρασης, απεικονίζει το υποσύνολο των μεταγράφων των γονιδίων που εκφράζονται σε ένα κύτταρο ή σε έναν ιστό. Η ανάλυση και επεξεργασία των δεδομένων της γονιδιακής έκφρασης πραγματοποιείται με τη βοήθεια εργαλείων της Βιοπληροφορικής και εξάγονται συμπεράσματα σχετικά με διαγνωστικούς ή προγνωστικούς γενετικούς βιοδείκτες (γονίδια) που συνδέονται με κάποια συγκεκριμένη ασθένεια. Στη παρούσα διπλωματική εργασία γίνεται μελέτη και υλοποίηση αλγορίθμων γονιδιακής ανάλυσης με στόχο τη μείωση των διαστάσεων και τη ταξινόμηση γονιδιακών δεδομένων. Συγκεκριμένα, χρησιμοποιούνται τρείς διαφορετικές τεχνικές ((1)RFE-LNW, (2)LASSO, (3)συνδυασμός RFE-LNW και FSMLP) για τη δημιουργία υποσυνόλων με πιθανά επικρατέστερα γονίδια καθώς και ο γραμμικός SVM ταξινομητής για την εκτίμηση της προγνωστικής δύναμης των διαφορετικών υποσυνόλων γονιδίων που παράγονται. Η διαδικασία της επιλογής των γονιδίων είναι ενσωματωμένη σε ένα External Cross Validation σύστημα, με σκοπό να ενισχυθεί η εμπιστοσύνη στα αποτελέσματα. Η πειραματική διαδικασία διεξάγεται πάνω σε μια βάση δεδομένων που σχετίζεται με την οστεοαρθρίτιδα, ασθένεια η οποία οδηγεί σε σημαντική νοσηρότητα και υστερεί στη διαθεσιμότητα αποτελεσματικών θεραπειών. Τα δεδομένα γονιδιακής έκφρασης που χρησιμοποιήθηκαν προέρχονται από δείγματα ιστού αρθρικού υμένα ασθενών με οστεοαρθρίτιδα και συγκρίθηκαν με αντίστοιχα δείγματα υγιών ατόμων.

Οι γονιδιακές υπογραφές στις οποίες καταλήγουμε μετά την επεξεργασία των γονιδιακών δεδομένων αξιολογούνται με όρους στατιστικούς και βιολογικούς. Τα αποτελέσματα που προκύπτουν δείχνουν ότι η μέθοδος LASSO παράγει ένα σύνολο από γενετικούς βιοδείκτες (γονιδιακή υπογραφή) με υψηλή προγνωστική ακρίβεια. Ο συνδυασμός των μεθόδων RFE-LNW & FSMLP καθώς και η τεχνική RFE-LNW ακολουθούν με αρκετά ικανοποιητικά αποτελέσματα. Τέλος, εξετάζοντας το βιολογικό περιεχόμενο των γονιδιακών υπογραφών με τη βοήθεια του συστήματος ταξινόμησης WebGestalt παρατηρούμε ότι τα γονίδια των υπογραφών, παρόλο που δεν είναι τα ίδια, συμμετέχουν στις ίδιες βιολογικές διεργασίες καθώς και σε αρκετά κοινά μοριακά μονοπάτια. Επίσης με το εργαλείο Genotator υπολογίστηκε η συσχέτιση των γονιδίων που περιλαμβάνονται στις υπογραφές με την ασθένεια της οστεοαρθρίτιδας.

Περιεχόμενα

Λίστα Εικόνων

Λίστα Πινάκων

ΚΕΦΑΛΑΙΟ 1: ΕΙΣΑΓΩΓΗ

1.1 Γενική Εισαγωγή

Τα τελευταία χρόνια, η τεχνολογία μικροσυστοιχίων DNA [1] χαρακτηρίζεται ως μια από τις πιο σημαντικές μεθόδους για την ανάλυση της γονιδιακής έκφρασης, αφού επιτρέπει τη ταυτόχρονη παρακολούθηση των επιπέδων έκφρασης χιλιάδων γονιδίων. Η πληροφορία που λαμβάνεται από τις μικροσυστοιχίες DNA βρίσκεται στη μορφή συνόλων δεδομένων των οποίων ο αριθμός των χαρακτηριστικών (γονιδίων) ξεπερνά σε πολύ μεγάλο βαθμό το μέγεθος των δειγμάτων. Το γεγονός αυτό μπορεί να αντιμετωπιστεί με διάφορες τεχνικές της επιστήμης της Βιοπληροφορικής οι οποίες επιτρέπουν τη μείωση των διαστάσεων του μεγάλου όγκου των δεδομένων και την επιλογή των σημαντικών χαρακτηριστικών (γονιδίων) που είναι απαραίτητα για την εκτέλεση της σωστής ταξινόμησης.

Στη παρούσα εργασία εστιάζουμε στην εύρεση ενός συνόλου από σημαντικά γονίδια, τα οποία σχετίζονται με την ασθένεια της οστεοαρθρίτιδας και εκφράζονται διαφορετικά ανάμεσα σε υγιείς και οστεοαρθριτιδικούς ιστούς. Προς αυτή τη κατεύθυνση οδηγούμαστε με τη βοήθεια διάφορων τεχνικών επιλογής και ταξινόμησης γονιδιακών δεδομένων. Στις ενότητες που ακολουθούν, παρατίθενται μια λεπτομερής περιγραφή για την ασθένεια της οστεοαρθρίτιδας, τις μικροσυστοιχίες DNA και τη βάση δεδομένων που χρησιμοποιήσαμε.

1.2 Οστεοαρθρίτιδα

1.2.1 Ορισμός

Η οστεοαρθρίτιδα (ΟΑ) είναι μια εκφυλιστική νόσος των αρθρώσεων [2], και είναι αποτέλεσμα βιολογικών, χημικών και ιξωδοελαστικών αλλοιώσεων

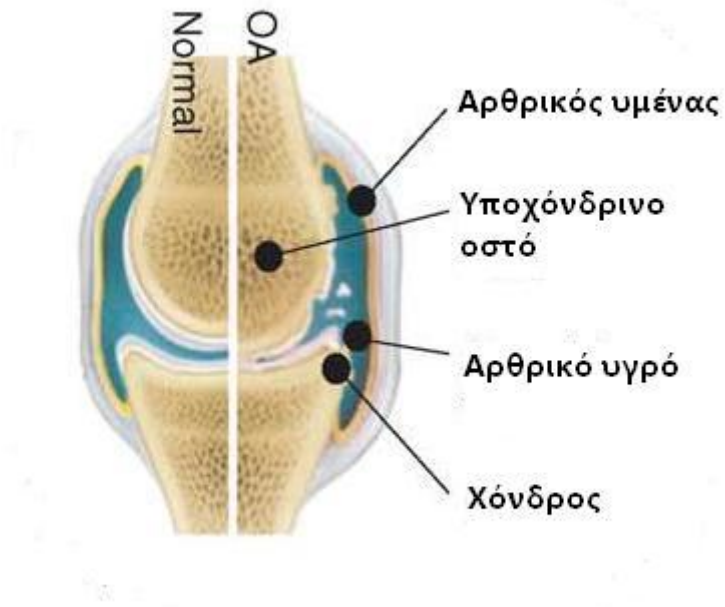
- του χόνδρου,
- του αρθρικού υμένα,
- του υποχόνδρινου οστού,
- και του αρθρικού υγρού.

Χαρακτηρίζεται παθολογικά [3] από εστιασμένες περιοχές βλάβης στον αρθρικό χόνδρο, επικεντρώνεται στις περιοχές συσσώρευσης του φορτίου. Προσβάλει πρωτίστως τις αρθρώσεις στα γόνατα (Εικόνα 1), το ισχίο, την σπονδυλική στήλη, τα χέρια και τα πόδια. Σχετίζεται κυρίως με την αύξηση της ηλικίας, και προσβάλλει περίπου διπλάσιο αριθμό γυναικών από εκείνον των ανδρών. Αναλυτικότερα, η ΟΑ σχετίζεται με την ανάπτυξη νέων οστών περιαρθρικά (οστεοφύτωση), με αλλοιώσεις στο υποχόνδρινο οστό, με την πάχυνση του αρθρικού ινώδους θυλάκου και με ποικίλου βαθμού ήπια αρθροθυλακίτιδα (υμενίτιδα). Η υμενίτιδα (synovitis) [4], ή αλλιώς αρθροθυλακίτιδα, ορίζεται ως η φλεγμονή του αρθρικού υμένα (αρθρικής μεμβράνης). Στις μέρες μας, όλο και περισσότερο αναγνωρίζεται ότι η υμενίτιδα παίζει σημαντικό ρόλο στην παθογένεση της ΟΑ. Η υμενίτιδα σε προχωρημένο στάδιο της ΟΑ μπορεί να είναι εξίσου σοβαρή όπως και στην ρευματοειδή αρθρίτιδα (RA). Στη πραγματικότητα, η σοβαρότητα της υμενίτιδας μπορεί να θεωρηθεί ως ένας δείκτης της δομικής αλλαγής και της κλινικής έκβασης στην ΟΑ. Από την άλλη πλευρά, η αρθρική παθολογία στην ΟΑ φαίνεται να είναι ένα πολύπλοκο ζήτημα. Για παράδειγμα, ο αρθρικός υμένας μιας επώδυνης άρθρωσης σε πρώιμο στάδιο ΟΑ μπορεί να είναι ιστολογικά υγιής , ενώ η υμενίτιδα μπορεί να εμφανίζεται σε αρθρώσεις χωρίς πόνο. Δηλαδή, παρόλο που η υμενίτιδα εντοπίζεται συχνά και μπορεί να είναι ασυμπτωματική, αρθροσκοπικές έρευνες εισηγούνται ότι οι εντοπισμένες πολλαπλασιαστικές και φλεγμονώδεις αλλοιώσεις του αρθρικού υμένα εμφανίζονται μέχρι και στο 50% των ασθενών με ΟΑ, και ότι ο ενεργοποιημένος αρθρικός υμένας μπορεί να παράγει πρωτεάσες και κυτοκίνες οι οποίες επιταχύνουν την εξέλιξη της νόσου.

Στα πλαίσια της παρούσας διπλωματικής, χρησιμοποιήθηκαν δεδομένα γονιδιακής έκφρασης [5] που προέρχονται από δείγματα ιστού αρθρικού υμένα ασθενών με ΟΑ και συγκρίθηκαν με αντίστοιχα δείγματα υγιών ατόμων. Όπως αναφέρθηκε, ο αρθρικός

υμένας (αρθρική μεμβράνη) είναι ένας από τους ιστούς των αρθρώσεων που συμμετέχει στην αρθριτική εξέλιξη της νόσου.

Όπως φαίνεται και στην Εικόνα 1, ο αρθρικός υμένας [6] είναι ο πλησιέστερος ιστός στον αρθρικό χόνδρο και μπορεί εύκολα να συλλεχθεί μέσω της αρθροσκόπησης (διαδικασία ρουτίνας) όταν κλινικοί γιατροί επιβεβαιώσουν μια διάγνωση για βλάβη του αρθρικού χόνδρου.



Εικόνα 1: Άρθρωση από ανθρώπινο γόνατο [7].

Έρευνες έχουν αναφερθεί στην κυτταρική ανάκτηση από τον αρθρικό υμένα για την αποκατάσταση της βλάβης στον χόνδρο. Ειδικότερα, αναφέρεται ότι ο αρθρικός υμένας [6] είναι ο μόνος ιστός ο οποίος μπορεί να παράγει υαλώδη χόνδρο σε καλοήθεις συνθήκες, όπως στην αρθρική χονδρομάτωση και οστεοχόνδρινα ίχνη στην οστεοαρθρίτιδα, από τα οποία προκύπτει ότι ο αρθρικός υμένας λειτουργεί ως μια πηγή κυττάρων για την αποκατάσταση του αρθρικού χόνδρου.

1.2.2 Διαγνωστικές μέθοδοι και αποτελεσματικότητα

Η ΟΑ είναι η πιο κοινή μορφή αρθρίτιδας [3], [8], [9], η οποία οδηγεί σε σημαντική νοσηρότητα και ανικανότητα στην τρίτη ηλικία. Δυστυχώς, η συγκεκριμένη νόσος υστερεί στην διαθεσιμότητα αποτελεσματικών θεραπειών σε σχέση με άλλες σκελετικές παθήσεις όπως η οστεοπόρωση. Παρόλο που σήμερα υπάρχουν διαθέσιμα φάρμακα που καταπραΰνουν τον πόνο και βελτιώνουν την λειτουργία, δεν υπάρχουν φάρμακα που μπορούν να θεραπεύσουν την ΟΑ, κυρίως επειδή δεν υπάρχει αξιόπιστη μέθοδος η οποία μπορεί να χρησιμοποιηθεί για να εντοπίσει πρώιμες αλλαγές της ΟΑ. Οι υπάρχουσες διαγνωστικές μέθοδοι είναι:

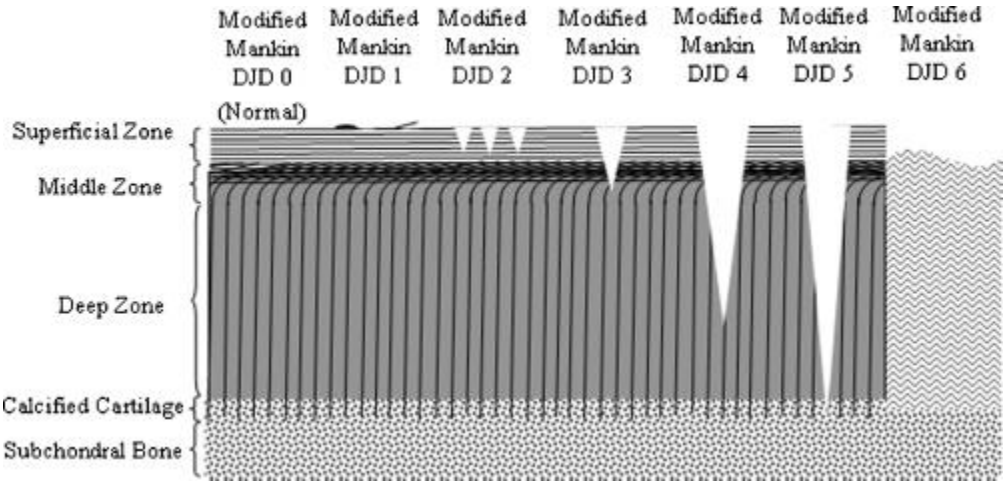
- η απλή ακτινογραφία, η οποία είναι η κατεξοχήν διαγνωστική μέθοδος για την ΟΑ δίνοντας ξεκάθαρη κλινική εικόνα. Ωστόσο, εξαιτίας της ημιποσοτικής κλίμακας διαβάθμισης και της χαμηλής ευαισθησίας της χαρακτηρίζεται ως ανεπαρκής στο να προσδιορίζει την εξέλιξη και το αποτέλεσμα των νέων θεραπειών σε σύντομα χρονοδιαγράμματα. Επιπλέον, οι οστεοαρθριτιδικές αλλοιώσεις που μπορούν να εντοπιστούν στις ακτινογραφίες βρίσκονται συνήθως σε πιο προχωρημένο στάδιο της ασθένειας, με αποτέλεσμα η βλάβη στους ιστούς των αρθρώσεων να θεωρείται μη αναστρέψιμη.

- η μαγνητική τομογραφία, η οποία έχει την δυνατότητα να απεικονίζει ταυτόχρονα όλους τους ιστούς των αρθρώσεων και χρησιμοποιείται όλο και περισσότερο στις ερευνητικές μελέτες για την ΟΑ. Όμως, οι παράμετροι που μπορούν να χρησιμοποιηθούν για την πρόωρη διάγνωση και τις κλινικές δοκιμές είναι ακόμα ασαφείς.

- ο υπέρηχος, ο οποίος μπορεί να καταδείξει δομικές αλλοιώσεις στο χόνδρο, στο μηνίσκο, στην επιφάνεια του οστού, στον αρθρικό υμένα, στους τένοντες, στους συνδέσμους, στον αρθρικό ινώδη θύλακα και στους ορογόνους θύλακες σε πρώιμο αλλά και σε προχωρημένο στάδιο της ΟΑ.

- η βιοψία – κλίμακα Mankin's. Ο Mankin το 1971 πρότεινε [10] μια ιστολογική – ιστοχημική κλίμακα αξιολόγησης του βαθμού σοβαρότητας της οστεοαρθρίτιδας στα διάφορα ιστολογικά δείγματα, η οποία βασίζεται σε δομικές αλλαγές της θεμέλιας ουσίας και των χονδροκυττάρων στη χρώση με safranin-O και στην

ακεραιότητα του μεταιχμίου. Βέβαια γίνεται αντιληπτό ότι, με την πληθώρα των υπαρχουσών απεικονιστικών τεχνικών, η κλίμακα του Mankin δεν χρησιμοποιείται πλέον για την διάγνωση της νόσου, αλλά χρησιμοποιείται στο καθορισμό της σοβαρότητας της νόσου στις ιστολογικές τομές. Στην Εικόνα 2 περιγράφεται η κλίμακα του Mankin. Σύμφωνα με αυτή υπάρχουν 7 βαθμοί (από 0 έως 6). Ο φυσιολογικός χόνδρος βαθμολογείται με 0 (DJD 0). Αλλαγές της επιφανειακής στοιβάδας του χόνδρου βαθμολογούνται με 1 (DJD 1). Με 2 βαθμολογείται η παρουσία σχισμών στην επιφανειακή ζώνη (DJD 2). Σχισμές μέσα στην ενδιάμεση ζώνη αντιστοιχούν στο βαθμό 3 (DJD 3). Σχισμές στην εν τω βάθει ζώνη : βαθμός 4 (DJD 4). Σχισμές στην ασβεστοποιημένη ζώνη : βαθμός 5 (DJD 5). Τέλος, πλήρης αποδιοργάνωση του χόνδρου αντιστοιχεί στον βαθμό 6 (DJD 6).



Εικόνα 2 : Σχηματική απεικόνιση της κλίμακας Mankin's. (DJD : Degenerative joint disease) [10].

Παρατηρούμε ότι η συγκεκριμένη νόσος επιφέρει σημαντικές κοινωνικές και οικονομικές επιπτώσεις τόσο από την άποψη των επώδυνων συμπτωμάτων για τους πάσχοντες όσο και από την άποψη της χρήσης των υπηρεσιών υγείας. Αυτές οι κοινωνικοοικονομικές επιδράσεις αναμένεται να ενταθούν παγκοσμίως με την αυξανόμενη επικράτηση της παχυσαρκίας και την αύξηση του μέσου όρου ζωής. Καθίσταται λοιπόν επιτακτική η ανάγκη να αναπτυχθούν αξιόπιστοι βιοδείκτες, οι οποίοι θα μπορούσαν να δώσουν πολύτιμες πληροφορίες για την διαδικασία αλλοίωσης των

αρθρώσεων στην ΟΑ. Τέτοιοι βιοδείκτες θα μπορούσαν να βοηθήσουν στην κατεύθυνση της ανάπτυξης νέων φαρμάκων με το να αναγνωρίζουν τους παράγοντες της γρήγορης εξέλιξης της νόσου και να ανακαλύπτουν μια πρώιμη θεραπευτική απάντηση, με αποτέλεσμα να μειώνεται ο αριθμός των ασθενών και ο χρόνος που απαιτείται για κλινικές δοκιμές.

1.2.3 Βιοδείκτες

1.2.3.1 Βιοχημικοί Δείκτες

Όταν μιλάμε για βιοχημικούς δείκτες [3], [11], αναφερόμαστε σε μόρια, όπως ορμόνες, ένζυμα, αντισώματα, πρωτεΐνες, μέταλλα, ιχνοστοιχεία ή οποιαδήποτε άλλη ουσία η οποία ανιχνεύεται στο αίμα, τα ούρα ή άλλα υγρά και ιστούς του οργανισμού και αποτελούν ενδείξεις καλής ή διαταραγμένης λειτουργίας του οργανισμού.

Τα τελευταία χρόνια, οι ραγδαίες εξελίξεις στην κατανόηση της βιοχημείας του αρθρικού χόνδρου έχουν οδηγήσει στην διερεύνηση πολλών πρωτεϊνών ως πιθανών βιοδεικτών της οστεοαρθρίτιδας. Αν και η γενική θεώρηση για τους βιοδείκτες περιλαμβάνει τις βιολογικές ουσίες, ωστόσο αρκετοί ερευνητές αναγνωρίζουν ως βιοδείκτες (α) τους παραδοσιακούς παράγοντες κινδύνου της ασθένειας και (β) την απεικόνιση. Ένας «ηχηρός» βιοδείκτης σημαίνει εξειδίκευση για την νόσο της ΟΑ, ανακλά την πραγματική ανάπτυξη και εξέλιξη της νόσου, διευκολύνει την πρόωρη διάγνωση, είναι ευαίσθητος στις αλλαγές που προκαλούνται εξαιτίας της θεραπευτικής παρέμβασης, και τέλος μπορεί να προβλέψει την έκβαση της νόσου.

Η προοδευτική απώλεια του αρθρικού χόνδρου είναι ένα βασικό χαρακτηριστικό της ΟΑ. Ο αρθρικός χόνδρος, ένας μη αγγειακός ιστός (δεν περιέχει αιμοφόρα αγγεία), αποτελείται από χονδροκύτταρα ενσωματωμένα σε μια εξωκυττάρια ουσία, η οποία παρέχει εμβιομηχανικά και φυσιολογικά χαρακτηριστικά που είναι απαραίτητα για την κίνηση των αρθρώσεων. Η ανισορροπία στην σύνθεση και την αποδόμηση του χόνδρου είναι κεντρικής σημασίας για την απώλεια του χόνδρου στην ΟΑ. Για αυτό τον λόγο, οι βιοδείκτες που αποτυπώνουν αυτές τις μεταβολικές διαδικασίες στην εξωκυττάρια ουσία βρίσκονται στο επίκεντρο της έρευνας. Επιπλέον, η ΟΑ είναι ευρέως αποδεκτή ως μια νόσος που αφορά ολόκληρη την άρθρωση, προσβάλλοντας όχι μόνο τον χόνδρο αλλά

ακόμα και το υποχόνδριο οστό καθώς και τον αρθρικό υμένα. Αλλοιώσεις στις παραπάνω δομές έχουν βρεθεί ότι σχετίζονται με την ασθένεια και αποτελούν μια ενδιαφέρουσα κατεύθυνση στην έρευνα για τους βιοδείκτες της οστεοαρθρίτιδας.

Στη συνέχεια παραθέτουμε μερικούς βιοχημικούς δείκτες που σχετίζονται με την νόσο της ΟΑ:

1. Υπάρχουν οχτώ βιοχημικοί δείκτες που αφορούν στον μεταβολισμό του κολλαγόνου τύπου ΙΙ (CII). Έξι από αυτούς αφορούν στην αποδόμηση του CII (CTX-II, Helix-II, C2C, Coll2-1, Coll2-1 NO_2 ,TIINE) και δύο την σύνθεση του CII (PIIANP, PIICP). Το κολλαγόνο τύπου ΙΙ (CII) παρέχει το μεγαλύτερο μέρος των οργανικών συστατικών στον εξωκυττάρια ουσία (15%-22%), ακολουθούμενο από την αγκρεκάνη (4%-7%), και άλλες μη κολλαγόνες πρωτεΐνες (0.5%-1%), συμπεριλαμβανομένης της ολιγομερούς πρωτεΐνης θεμέλιας ουσίας του χόνδρου (COMP).

2. Άλλοι βιοχημικοί_δείκτες για την οστεοαρθρίτιδα που έχουν μελετηθεί λιγότερο περιλαμβάνουν αυτούς που αφορούν στην σύνθεση του κολλαγόνου τύπου Ι (PICP,PINP), την σύνθεση του κολλαγόνου τύπου ΙΙΙ (PIIINP), την αποδόμηση του κολλαγόνου τύπου Ι και ΙΙ (C1, 2C),την οστεοκαλσίνη, την οστική σιαλοπρωτεΐνη, τη θειική κεράτινη, τη θειική χονδροϊτίνη 846 (CS846), την ανθρώπινη γλυκοπρωτεΐνη του αρθρικού χόνδρου 39 (YKL-40), τις διασταυρωμένες συνδέσεις του κολλαγόνου (πυριδινολίνη [Pyr], δεοξυπυριδινολίνη [D-Pyr], γλυκόσυλο- γαλακτόσυλο-πυριδινολίνη [Glc-Gal-Pyr], και πεντοζιδίνη).

Η μέτρηση των παραπάνω βιοχημικών δεικτών [8] πραγματοποιείται μέσω εξειδικευμένων μεθόδων προσδιορισμού (π.χ. ενζυμικός ανοσοπροσροφητικός προσδιορισμός - ELISAs, υγρή χρωματογραφία-φασματομετρία μάζας - LC-MS) οι οποίες χρησιμοποιούνται για την ανίχνευσή τους στο αίμα, τα ούρα και το αρθρικό υγρό. Δυστυχώς υπάρχει μια μεταβλητότητα στη μέτρηση των βιοχημικών δεικτών που προέρχεται από ένα μεγάλο αριθμό παραγόντων που δεν σχετίζονται με την ασθένεια της ΟΑ (ημερήσιος ρυθμός, φυσική δραστηριότητα) και μπορούν να επηρεάσουν τα εργαστηριακά ευρήματα οδηγώντας σε επισφαλή αποτελέσματα. Παρά τους περιορισμούς,

οι βιοχημικοί δείκτες μπορούν να βοηθήσουν στη κλινική αξιολόγηση της ΟΑ καταδεικνύοντας ταυτόχρονα την ανάγκη για εύρεση νέων πιο “αξιόπιστων” δεικτών.

1.2.3.2 Γενετικοί βιοδείκτες

Η εφαρμογή των τεχνολογιών όπως η μεταβολωμική, η πρωτεομική και η γονιδιωματική [3], [8] στην ΟΑ παράγουν ενδεχομένως επιπρόσθετους βιοδείκτες που θα μπορούσαν να βοηθήσουν στον εντοπισμό πρόωρων οστεοαρθριτιδικών αλλοιώσεων.

Στοιχεία δείχνουν ότι οι γενετικοί δείκτες παίζουν σημαντικό ρόλο στην ΟΑ, παρόλο που μπορεί να διαφέρουν ανάλογα με τον τόπο και το φύλο. Από μελέτες σε διδύμους, αυτή η γενετική επιρροή έχει εκτιμηθεί ότι κυμαίνεται μεταξύ 40% και 65%, και ότι ο κίνδυνος για τους πρώτου βαθμού συγγενείς είναι διπλάσιος έως και τριπλάσιος. Η φύση της γενετικής επιρροής στην ΟΑ είναι ακόμα αδιευκρίνιστη, αλλά είναι πιθανόν να περιλαμβάνει ένα συνδυασμό από επιδράσεις στην δομή (π.χ. κολλαγόνο), αλλοιώσεις στον χόνδρο, μεταβολισμό των οστών ή φλεγμονή. Θεωρείται ότι η ταυτοποίηση συγκεκριμένων γενετικών παραγόντων (γενετικοί βιοδείκτες) για την ΟΑ καθώς και ο συνδυασμός τους (γονιδιακές υπογραφές) μπορεί να βοηθήσει στο να κατανοήσουμε τη παθογένεση της νόσου, να αναγνωρίσουμε εγκαίρως ανθρώπους και οικογένειες με υψηλό κίνδυνο για εμφάνιση ΟΑ και να προσαρμόσουμε κατάλληλα τις διάφορες θεραπευτικές στρατηγικές. Η παρούσα διπλωματική εργασία στοχεύει στην εύρεση ενός συνόλου από (πιθανούς) σημαντικούς γενετικούς δείκτες (γονίδια) οι οποίοι μπορούν να αναγνωρίσουν με μεγάλη ακρίβεια άτομα που πάσχουν από την ασθένεια της ΟΑ.

Από αρκετές έρευνες σύνδεσης που έχουν πραγματοποιηθεί, έχουν βρεθεί μεγάλες περιοχές χρωμοσωμάτων που σχετίζονται με την ΟΑ, αλλά αυτές είναι περιορισμένης αξίας για την ανίχνευση εξειδικευμένων γονιδίων. Ένας αριθμός υποψήφιων γονιδίων έχει αναφερθεί για την ΟΑ.

1.2.3.3 Αναγκαιότητα εύρεσης κατάλληλων βιοδεικτών για την οστεοαρθρίτιδα

Η παθογένεια, η θεραπεία, και οι βιολογικοί δείκτες (βιοχημικοί, γενετικοί, απεικονιστικοί), είναι τρείς τομείς που αναπτύσσονται και εξελίσσονται παράλληλα στην ΟΑ, και αποτελούν αντικείμενα έντονου ερευνητικού ενδιαφέροντος. Στα πλαίσια αυτά, σημαντική βαρύτητα δίνεται στη μεγάλη ανάγκη εύρεσης βιοδεικτών, όπως επισημαίνεται από την Παγκόσμια Πρωτοβουλία για Βιοδείκτες της ΟΑ (OA Biomakers Global Initiative) [3]. Σκοπός αυτού του δικτύου επιστημόνων [12] είναι να αναπτύξει και να χαρακτηρίσει νέους και να επανεξετάσει παλιότερους βιολογικούς δείκτες για την ΟΑ. Η ομάδα αυτή πρότεινε το 2006 μια ταξινόμηση των βιοχημικών δεικτών με τα αρχικά BIPED η οποία είχε ευρεία αποδοχή. Τα αρχικά BIPED υποδεικνύουν τα 5 χαρακτηριστικά που μπορεί να έχει ένας βιολογικός δείκτης. Έτσι, κάθε βιολογικός δείκτης για την ΟΑ ταξινομείται πλέον σε μια ή περισσότερες από τις παρακάτω κατηγορίες:

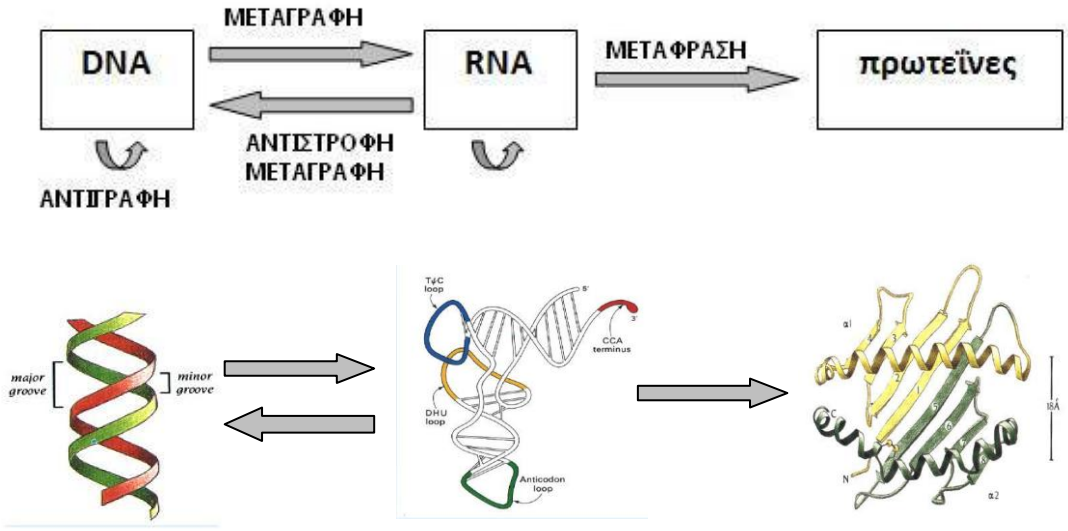
- δείκτης σοβαρότητας της νόσου (burden of disease)
- ερευνητικός δείκτης (investigative)
- προγνωστικός δείκτης (prognostic)
- δείκτης αποτελεσματικότητας της παρέμβασης (efficacy of intervention)
- διαγνωστικός δείκτης (diagnostic)

Υπάρχει ένα σημαντικό κομμάτι δουλειάς πάνω στους βιοδείκτες για την ασθένεια της ΟΑ, και δεν υπάρχει αμφιβολία ότι επιτρέπει την καλύτερη κατανόηση της εξέλιξης της νόσου. Ωστόσο, μέχρι σήμερα, κανένας από τους προτεινόμενους βιοδείκτες δεν μπόρεσε να χρησιμοποιηθεί στην καθημερινή κλινική πρακτική για διάγνωση, παρακολούθηση, πρόγνωση και κλινικές δοκιμές. Αυτό οφείλεται κυρίως στην έλλειψη της πληροφορίας σχετικά με την ευαισθησία, την ιδιαιτερότητα, τα φυσιολογικά όρια και τις κλινικά σημαντικές διαφορές. Φυσικά, απαιτείται περισσότερη έρευνα για να χαρακτηριστούν περαιτέρω οι ήδη ταυτοποιημένοι βιοδείκτες και να ανακαλυφθούν νέοι βιολογικοί δείκτες. Η εργασία μας κινείται προς αυτή τη κατεύθυνση, αφού με τη βοήθεια διάφορων εργαλείων της Βιοπληροφορικής στοχεύει στην ανίχνευση μοριακών υπογραφών και δυνητικών γενετικών δεικτών.

1.3 Γονιδιακή έκφραση στο κύτταρο

Εκτός από λίγες εξαιρέσεις, κάθε κύτταρο στο σώμα περιέχει ένα ολόκληρο σετ από χρωμοσώματα και παρόμοια γονίδια. Όμως, μόνο ένα ποσοστό αυτών των γονιδίων ενεργοποιείται και αυτό το ποσοστό που «εκφράζεται» δίνει μοναδικές ιδιότητες σε κάθε κύτταρο. Η «γονιδιακή έκφραση» είναι ο όρος που χρησιμοποιείται για να εκφράσει τη μεταγραφή της πληροφορίας που περιέχεται στο DNA, την αποθήκη της γενετικής πληροφορίας, σε μόρια αγγελιοφόρου RNA (mRNA). Αυτά τα μόρια χαρακτηρίζονται ως τα ενδιάμεσα μόρια που μεταφέρουν την πληροφορία για τη πρωτεϊνοσύνθεση. Στον μηχανισμό που συνθέτει πρωτεΐνες συμμετέχουν και άλλα μόρια RNA όπως μεταφορικό RNA (tRNA) και ριβωσομικό RNA (rRNA). Η διαδικασία μεταγραφής (transcription) ακολουθείται από τη διαδικασία μετάφρασης (translation), δηλαδή τη σύνθεση πρωτεϊνών, οι οποίες φέρουν εις πέρας τις σημαντικές λειτουργίες του κυττάρου. [13],[14],[15]

Η ροή των γενετικών πληροφοριών στα φυσιολογικά κύτταρα απεικονίζεται στην Εικόνα 3:



Εικόνα 3: Ροή γενετικής πληροφορίας [16].

Οι επιστήμονες μελετούν [15] το είδος και την ποσότητα των mRNAs που παράγονται από ένα κύτταρο για να μάθουν ποια γονίδια εκφράζονται, γεγονός που δίνει πληροφορίες για το πώς το κύτταρο απαντά στις μεταβαλλόμενες ανάγκες του. Η γονιδιακή έκφραση είναι μια πολύπλοκη και αυστηρά ελεγχόμενη διεργασία που επιτρέπει σε ένα κύτταρο να απαντά δυναμικά στα περιβαλλοντικά ερεθίσματα και στις δικές του ανάγκες.

Αυτός ο μηχανισμός δρα σαν ένας διακόπτης ανοικτού/κλειστού για να ελέγξει ποια γονίδια θα εκφραστούν στο κύτταρο, και ως ένας διακόπτης «ελέγχου ροής», που αυξάνει ή ελαττώνει το επίπεδο έκφρασης συγκεκριμένων γονιδίων, όταν αυτό κριθεί απαραίτητο.

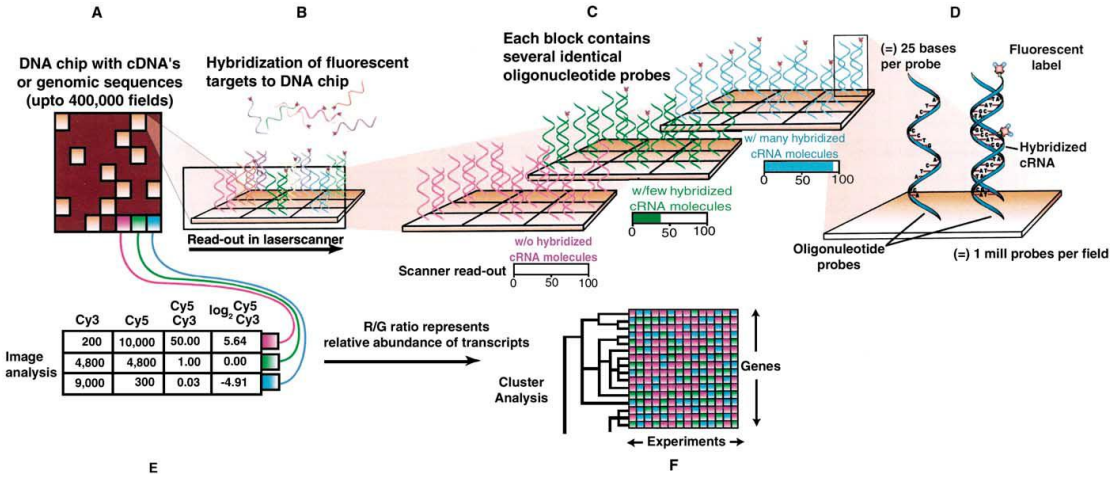
Η μελέτη της γονιδιακής έκφρασης [15] ενέχει τη παρατήρηση των ποσοτήτων mRNA ή πρωτεϊνών που παράγονται από ένα κύτταρο μια δεδομένη στιγμή. Η αρχή πίσω από την ανάλυση της γονιδιακής έκφρασης βασίζεται στη σύγκριση δειγμάτων, για παράδειγμα νέων και γερασμένων ιστών για τη μελέτη της ανάπτυξης και της γήρανσης, απλών και πολύπλοκων οργανισμών για τη μελέτη της εξέλιξης, ασθενών και υγιών ιστών για τη μελέτη συγκεκριμένων ασθενειών.

1.4 Τεχνικές Ανάλυσης της Γονιδιακής Έκφρασης με Μικροσυστοιχίες Γονιδίων (DNA Microarrays)

Πρόσφατα, αναπτύχθηκαν τεχνικές που βασίζονται σε συστοιχίες αλληλουχιών DNA [13],[14],[17] και καθιστούν δυνατή την ποσοτική ανάλυση της έκφρασης χιλιάδων γονιδίων με ένα παράλληλο και διεξοδικό τρόπο. Έτσι, αντί να μελετάμε κάθε γονίδιο ξεχωριστά, τώρα έχουμε τη δυνατότητα σφαιρικής θεώρησης της γονιδιακής έκφρασης, δηλαδή μπορούμε να διερευνήσουμε το κύτταρο συνολικά ως πολύπλοκο δίκτυο ρυθμιστικών μηχανισμών γονιδίων και να παρακολουθήσουμε μεταβολές της έκφρασης πλειάδας γονιδίων εξαιτίας μιας νόσου ή ως συνέπεια της επίδρασης διαφόρων παραγόντων. Μια από τις πρόσφατες τεχνικές που χρησιμοποιούνται για την ανάλυση της γονιδιακής έκφρασης βασίζεται σε συστοιχία μεγάλου αριθμού διαφορετικών ολιγοδεοξυριβονουκλεοτιδίων που είναι ακινητοποιημένα σε γυάλινη επιφάνεια. Ένα τέτοιο σύστημα αποτελεί ένα βιολογικό chip ή αλλιώς μια μικροσυστοιχία γονιδίων (DNA chip/ DNA microarray/ gene chip).

Το πρότυπο της γονιδιακής έκφρασης που παράγεται, γνωστό ως προφίλ έκφρασης (expression profile), απεικονίζει το υποσύνολο των μεταγράφων (transcripts) των γονιδίων που εκφράζονται σε ένα κύτταρο ή σε έναν ιστό. Οι DNA μικροσυστοιχίες, όπως αναφέρθηκε παραπάνω, μπορούν ταυτόχρονα να μετρούν το επίπεδο έκφρασης χιλιάδων γονιδίων ενός συγκεκριμένου δείγματος mRNA (Εικόνα 4). Στο πιο θεμελιώδες επίπεδό του, το προφίλ έκφρασης μπορεί να αποδώσει ποιοτικά ποια γονίδια εκφράζονται σε κατάσταση ασθένειας.

Όπως φαίνεται στην Εικόνα 3 για να πραγματοποιηθεί ένα πείραμα μικροσυστοιχίας [18],[19] ακολουθείται η παρακάτω διαδικασία:



Εικόνα 4: Σχήμα απεικόνισης πειράματος μικροσυστοιχίας DNA [18].

- Πρότυπα των γονιδίων ενδιαφέροντος βρίσκονται σε μικροσκοπικές διαφάνειες με επικάλυψη από γυαλί. (Στάδιο Α)

- mRNA από τα κύτταρα προς μελέτη (π.χ από αρθρικό ιστό που έχει προσβληθεί από οστεοαρθρίτιδα) καθώς και από κύτταρα αναφοράς (π.χ από υγιή αρθρικό ιστό) απομονώνεται και πολλαπλασιάζεται με χρήση της αλυσιδωτής αντίδρασης πολυμεράσης (PCR). Το mRNA μετατρέπεται σε cDNA μέσω της αντίστροφης μεταγραφής. Έπειτα, το cDNA του δείγματος προς μελέτη σημαίνεται με κόκκινη φθορίζουσα χρωστική (Cy5) και το cDNA του δείγματος αναφοράς με πράσινη χρωστική (Cy3). Τα δύο δειγμάτα cDNA αναμιγνύονται και υβριδοποιούνται[1] με τους DNA στόχους του πλακιδίου. (Στάδιο Β)

- Ανάλογα με το μέγεθος της γονιδιακής έκφρασης , οι ακινητοποιημένοι ανιχνευτές (probes) στο τσιπ γονιδίων (gene chip) υβριδοποιούνται με μηδενικά (w/0), με λίγα (w/few) ή με πολλά (w/many) cDNA μόρια. (Στάδιο C)

[1] **Υβριδοποίηση:** σύνδεση δύο μονόκλωνων συμπληρωματικών αλυσίδων DNA ή συμπληρωματικών DNA-RNA με υδρογονικούς δεσμούς, σύμφωνα με τον κανόνα της συμπληρωματικότητας των βάσεων**.** [20]

- Ακολουθεί η σάρωση της μικροσυστοιχίας με σαρωτή laser ή συνεστιακό μικροσκόπιο. Η διέγερση με λέιζερ των ενσωματωμένων στόχων παράγει εκπομπή με χαρακτηριστικά φάσματα. Οι μονόχρωμες εικόνες που προκύπτουν από τη διαδικασία της σάρωσης εισάγονται σε λογισμικό στο οποίο οι εικόνες ψευδοχρωματίζονται και συγχωνεύονται.

- Μετά τον καθορισμό της έντασης φθορισμού κάθε χρωστικής σε κάθε spot, σχηματίζονται οι λόγοι log(Cy5/Cy3). Θετική τιμή του log(Cy5/Cy3) υποδεικνύει ένα σχετικό πλεόνασμα των μεταγράφων (transcripts) του συγκεκριμένου γονιδίου στο υπό μελέτη δείγμα, ενώ αρνητική τιμή υποδεικνύει πλεόνασμα στο δείγμα αναφοράς. (Στάδιο E)

- Τα δεδομένα ομαδοποιούνται με τη χρήση τεχνικών ομαδοποίησης (cluster analysis) και παρουσιάζονται με τη μορφή πίνακα. Ο κατακόρυφος άξονας του πίνακα αντιστοιχεί σε διαφορετικά γονίδια και ο οριζόντιος σε διαφορετικά πειράματα, ενώ χρησιμοποιούνται διαφορετικά χρώματα για θετικό log(Cy5/Cy3) (κόκκινο), αρνητικό log(Cy5/Cy3) (πράσινο) και log(Cy5/Cy3) ίσο με μηδέν (μαύρο). (Στάδιο F)

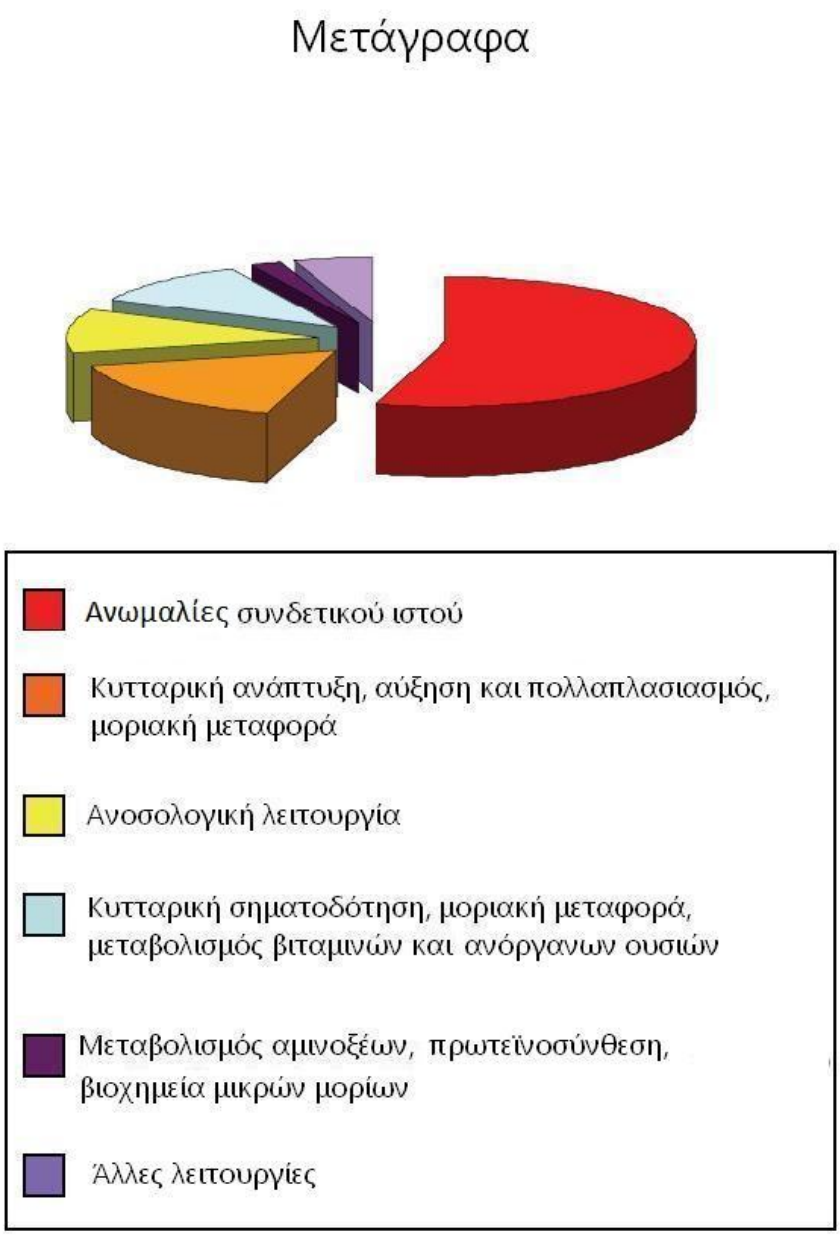
Το υψηλής απόδοσης προφίλ έκφρασης που προκύπτει από τις μικροσυστοιχίες DNA μπορεί να χρησιμοποιηθεί για να συγκρίνει το επίπεδο της γονιδιακής μεταγραφής σε κλινικές συνθήκες με σκοπό να:

- αναγνωριστούν/ ταυτοποιηθούν διαγνωστικοί ή προγνωστικοί βιοδείκτες
- ταξινομηθούν ασθένειες (π.χ. όγκοι με διαφορετική πρόγνωση που είναι δυσδιάκριτοι με μικροσκοπική εξέταση)
- παρακολουθηθεί η απόκριση στην θεραπεία
- κατανοηθούν οι μηχανισμοί που εμπλέκονται στις διεργασίες αιτιοπαθογένεσης.

Συνεπώς , η μελέτη του προφίλ της γονιδιακής έκφρασης των κυττάρων και των ιστών και κατ' επέκταση οι μικροσυστοιχίες DNA θεωρούνται [21] σημαντικά εργαλεία για την ανακάλυψη στη κλινική ιατρική. Η ανάλυση και επεξεργασία [17], [22] των δεδομένων της γονιδιακής έκφρασης μπορεί να πραγματοποιηθεί με τη βοήθεια εργαλείων της Βιοπληροφορικής (bioinformatics tools) όπως η ανάλυση ταξινόμησης (classification analysis), η ανάλυση ομαδοποίησης (cluster analysis), οι χάρτες αυτό-οργάνωσης (self-organizing maps, SOM) και η ανάλυση κυρίων συνιστωσών (principle component analysis,

PCA), που θα επιτρέψουν την εξαγωγή συμπερασμάτων μετά από κριτική επεξεργασία της πρωτογενούς πληροφορίας.

Μέχρι στιγμής έρευνες [23] για την ανάλυση της γονιδιακής έκφρασης στην ΟΑ με μικροσυστοιχίες γονιδίων έχουν καταδείξει διαταραχές σε διάφορες βιολογικές διεργασίες, όπως φαίνεται στο διάγραμμα της Εικόνας 5.



Εικόνα 5 : Γράφημα-πίτα (pie-chart) που δείχνει τη κατανομή των βιολογικών λειτουργιών στο επίπεδο των μεταγράφων (mRNAs) για την ασθένεια της ΟΑ [23].

1.5 Περιγραφή του Συνόλου Δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήσαμε στην παρούσα διπλωματική εργασία προέρχεται από την εργασία των Huber και συνεργατών [5] και έχει ως αντικείμενο μελέτης της την οστεοαρθρίτιδα. Τα δείγματα (SM) της βάσης δεδομένων προήλθαν από αρθρικό ιστό ατόμων που πάσχουν από οστεοαρθρίτιδα και υγιών ατόμων αντίστοιχα. Πιο αναλυτικά, τα δείγματα (SM) λήφθησαν από 10 ασθενείς με οστεοαρθρίτιδα (n=10 OA) κατά την αντικατάσταση της πάσχουσας άρθρωσης καθώς και από 9 υγιή άτομα (n=9 NC) κατά τη χειρουργική επέμβαση τραύματος σε άρθρωση. Οι παραπάνω επεμβάσεις πραγματοποιήθηκαν στο τμήμα Ορθοπεδικής του Πανεπιστημιακού Νοσοκομείου Jena, Waldkrankenhaus 'Rudolf Elle'(Άιζενμπεργκ, Γερμανία) .

Η ανάλυση της γονιδιακής έκφρασης των παραπάνω δειγμάτων (10 OA και 9 NC) πραγματοποιήθηκε με τη βοήθεια των ολιγονουκλεοτιδικών συστοιχιών Affymetrix U133A/B. Έτσι έχουμε στην διάθεσή μας δύο επιμέρους συστοιχίες [24] από τις οποίες και προκύπτουν δύο διαφορετικά σύνολα δεδομένων:

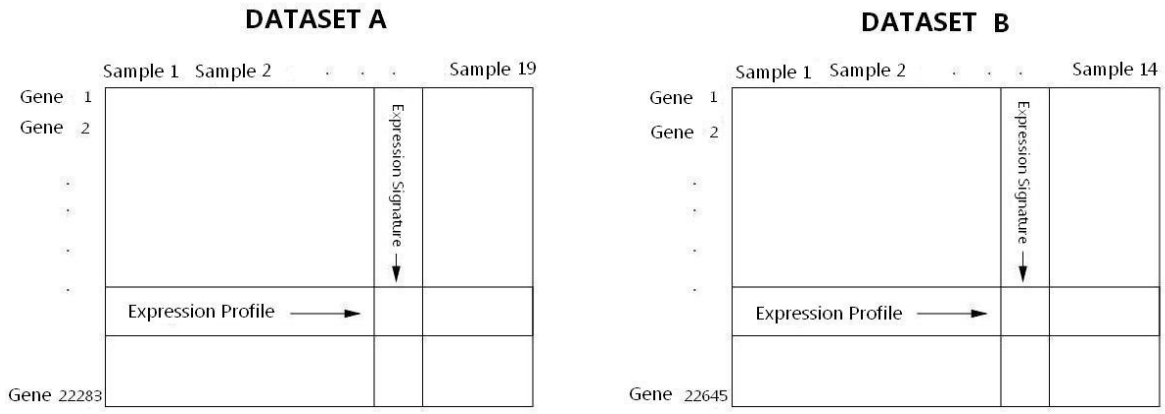
1)Platform GPL96 [HG-U133A] Affymetrix Human Genome U133A Array, από την οποία προέκυψε το πρώτο σύνολο δεδομένων (Dataset A) που χρησιμοποιήσαμε στην εργασία μας. Το Dataset A αποτελείται από 22283 γονίδια και 19 δείγματα (10 δείγματα OA, 9 δείγματα NC).

2)Platform GPL97[HG-U133B] Affymetrix Human Genome U133B Array, από την οποία προέκυψε το δεύτερο σύνολο δεδομένων (Dataset B) που χρησιμοποιήσαμε στην εργασία μας. Το Dataset B αποτελείται από 22645 γονίδια και 14 δείγματα (10 δείγματα OA, 4 δείγματα NC).

Τα σύνολα δεδομένων Dataset A και Dataset B βρίσκονται στη μορφή πινάκων, όπως φαίνεται στην Εικόνα 6. Οι γραμμές τους αναπαριστούν τα γονίδια από τα οποία έχουν συγκεντρωθεί τα δεδομένα και οι στήλες αντιπροσωπεύουν τα διαφορετικά πειραματικά δείγματα. Κάθε κελί αποτελεί την έκφραση ενός συγκεκριμένου γονιδίου για ένα συγκεκριμένο δείγμα και ισούται με τη μετρούμενη ένταση φθορισμού του λόγου log(Cy5/Cy3). Όπως έχουμε αναφέρει, οι Cy5 και Cy3 είναι φθορίζουσες χρωστικές ουσίες με τις οποίες σημαίνονται αντίστοιχα τα μόρια cDNA του δείγματος προς μελέτη και τα μόρια cDNA του δείγματος αναφοράς. Οι τιμές των εντάσεων έχουν κανονικοποιηθεί έτσι

ώστε να διευκολύνουν τις συγκρίσεις μεταξύ των διαφορετικών δειγμάτων. Επίσης για το σκοπό αυτό, τα δεδομένα από διαφορετικά πειραματικά δείγματα ομαδοποιούνται σύμφωνα με τις κατηγορίες ασθενείς (OA) / υγιείς δότες (NC).

Στη παρούσα διπλωματική εργασία γίνεται μελέτη και υλοποίηση αλγορίθμων γονιδιακής ανάλυσης με στόχο τη μείωση των διαστάσεων (RFE-LNW, LASSO, συνδυασμός RFE-LNW & FSMLP) και τη ταξινόμηση (linear SVM) των γονιδιακών δεδομένων των δύο Datasets. Οι ετικέτες NC και OA χρησιμοποιούνται ως αρνητική και θετική κλάση αντίστοιχα για την υλοποίηση της δυαδικής ταξινόμησης. Το target των δειγμάτων που ανήκουν στην πρώτη κλάση (NC) χαρακτηρίζεται από το -1, ενώ το target των δειγμάτων της δεύτερης κλάσης (OA) από το +1. Πρέπει να επισημάνουμε ότι τα σύνολα δεδομένων (Dataset A και Dataset B), που χρησιμοποιούμε στη παρούσα διπλωματική εργασία (Εικόνα 6), δεν διαθέτουν κοινά γονίδια και κατά συνέπεια μπορούν να θεωρηθούν και να επεξεργαστούν ως δύο ανεξάρτητα σύνολα.



Εικόνα 6: Πίνακες Γονιδιακής Έκφρασης, που αντιστοιχούν στα σύνολα δεδομένων της εργασίας μας. Κάθε θέση στους πίνακες περιλαμβάνει πληροφορία για την έκφραση ενός συγκεκριμένου γονιδίου σε ένα συγκεκριμένο δείγμα [5],[25].

Βιολογική ερμηνεία σύμφωνα με το σύστημα ταξινόμησης WebGestalt και το συστήμα Genotator

Οι γονιδιακές υπογραφές στις οποίες καταλήξαμε μετά την επεξεργασία των δεδομένων των δύο Datasets εξετάστηκαν ως προς το βιολογικό τους περιεχόμενο με τη βοήθεια των εργαλείων WebGestalt και Genotator.

Το WebGestalt (WEB-based GEne SeT AnaLysis Toolkit) [26] είναι ένα σύστημα εξόρυξης δεδομένων από σύνολα γονιδίων. Συγκεκριμένα μας βοηθά να διαχειριστούμε τα

γονιδιακά δεδομένα, να ανακτήσουμε τη πληροφορία που περιέχουν μέσω διάφορων ενσωματωμένων πηγών, να οργανώσουμε, να οπτικοποιήσουμε και να αναλύσουμε στατιστικά το βιολογικό πλαίσιο στο οποίο συμμετέχουν.

Με το σύστημα WebGestalt η οργάνωση των γονιδίων που περιέχονται στις υπογραφές πραγματοποιείται ως προς τα κοινά λειτουργικά χαρακτηριστικά, όπως είναι οι κατηγορίες (π.χ. βιολογικές διεργασίες) που περιλαμβάνονται στην οντολογία γονιδίων (Gene Ontology, GO) και τα μοριακά μονοπάτια που περιλαμβάνονται στην εγκυκλοπαίδεια KEGG των γονιδίων και γονιδιωμάτων (Kyoto Encyclopedia of Genes and Genomes, KEGG). Οι κατηγορίες που περιλαμβάνονται στην οντολογία γονιδίων (GO) [27] καλύπτουν τρεις περιοχές: (α) **τα κυτταρικά συστατικά**, δηλαδή τα μέρη ενός κυττάρου ή το εξωκυττάριο περιβάλλον του, (β) **τις μοριακές λειτουργίες**, που συνιστούν τις στοιχειώδεις δραστηριότητες ενός γονιδιακού προϊόντος στο μοριακό επίπεδο και (γ) **τις βιολογικές διεργασίες**, που περιλαμβάνουν τις λειτουργίες ή ένα σύνολο από μοριακά συμβάντα με μια καθορισμένη αρχή και τέλος, και τα οποία σχετίζονται με τη λειτουργία ολοκληρωμένων έμβιων μονάδων όπως κύτταρα, ιστοί, όργανα και οργανισμοί. Όσο αναφορά τα μοριακά μονοπάτια, η εγκυκλοπαίδεια KEGG [28] αποτελεί μια συλλογή από χάρτες με μονοπάτια τα οποία απεικονίζουν τις γνώσεις μας σχετικά με τις κυτταρικές διεργασίες, την επεξεργασία γενετικής πληροφορίας, την επεξεργασία περιβαλλοντικής πληροφορίας, τα συστήματα των οργανισμών και τις ανθρώπινες ασθένειες πάνω σε μοριακά δίκτυα αλληλεπίδρασης και αντίδρασης.

Το σύνολο των γονιδίων που περιλαμβάνονται στις γονιδιακές υπογραφές αναλύθηκε επίσης και με το σύστημα Genotator, όπως αναφέραμε προηγουμένως. Το Genotator [29] είναι μια μετά-βάση δεδομένων η οποία συγκεντρώνει και βαθμολογεί τις συσχετίσεις μεταξύ γονιδίων και διάφορων ασθενειών. Το συγκεκριμένο εργαλείο ενσωματώνει αυτόματα δεδομένα από 11 εξωτερικά προσβάσιμες κλινικές πηγές και χρησιμοποιεί αυτά τα δεδομένα με τη βοήθεια μιας φόρμουλας με σκοπό να ταξινομήσει τα γονίδια ανάλογα με τη σχετικότητα τους ως προς κάποια συγκεκριμένη ασθένεια. Ο βαθμός συσχέτισης γονιδίου – ασθένειας υπολογίζεται ως ένα σκορ (Genotator Score, GS) μέσω του τύπου:

$$GS = GAD_Y - GAD_N + \varphi(GPS) + \frac{1}{\gamma}(DB) + \frac{1}{\kappa}(REF) \qquad (5.1)$$

όπου

- GAD_Y = ο συνολικός αριθμός των θετικών (“Yes” labeled) συσχετίσεων μεταξύ γονιδίου και ασθένειας στην Genetic Association Database
- GAD_N = ο συνολικός αριθμός των αρνητικών (“No” labeled) συσχετίσεων μεταξύ γονιδίου και ασθένειας στην Genetic Association Database
- GPS = το Gene Prospector’s score για τη σχέση γονιδίου – ασθένειας
- DB = ο συνολικός αριθμός των βάσεων δεδομένων (από τις 11 συνολικά) που εμφανίστηκε το συγκεκριμένο γονίδιο
- REF = ο συνολικός αριθμός αναφορών για το συγκεκριμένο γονίδιο

Οι σταθερές φ, γ και κ χρησιμοποιούνται για να ρυθμίσουν την συνεισφορά των παραμέτρων GPS, DB, REF στο τελικό σκορ (GS).

1.6 Σκοπός της εργασίας

Αντικειμενικός σκοπός της παρούσας εργασίας είναι η εφαρμογή αλγορίθμων γονιδιακής ανάλυσης στο σύνολο δεδομένων του Huber [5], το οποίο σχετίζεται με την ασθένεια της οστεοαρθρίτιδας. Με τη χρήση τριών διαφορετικών τεχνικών ταξινόμησης και επιλογής γονιδίων καταλήγουμε σε τρεις διαφορετικές γονιδιακές υπογραφές για το κάθε Dataset. Ως γονιδιακή υπογραφή (gene signature) ορίζουμε την ομάδα από τα πιο σημαντικά και άρα πληροφοριακά γονίδια από το σύνολο δεδομένων των οποίων η από κοινού έκφραση χαρακτηρίζει μοναδικά κάποια πάθηση. Οι μέθοδοι που χρησιμοποιήσαμε για την εύρεση και επιλογή των κατάλληλων βιοδεικτών από το σύνολο δεδομένων της οστεοαρθρίτιδας είναι:

- Για τη πρώτη γονιδιακή υπογραφή εφαρμόσαμε : RFE-LNW και linear SVM
- Για τη δεύτερη γονιδιακή υπογραφή εφαρμόσαμε : LASSO και linear SVM
- Για τη τρίτη γονιδιακή υπογραφή εφαρμόσαμε : RFE-LNW, FSMLP και linear SVM

Οι αλγόριθμοι που αφορούν τις μεθόδους RFE-LNW, LASSO, FSMLP χρησιμοποιούνται για τη δημιουργία υποσυνόλων με πιθανά επικρατέστερα γονίδια και ο αλγόριθμος του γραμμικού SVM (linear SVM) ταξινομητή χρησιμοποιείται για την εκτίμηση της προγνωστικής δύναμης των διαφορετικών υποσυνόλων γονιδίων που παράγονται. Η διαδικασία της επιλογής των γονιδίων είναι ενσωματωμένη σε ένα External Cross Validation σύστημα, με σκοπό να ενισχυθεί η εμπιστοσύνη στα αποτελέσματα. Σκοπός μας λοιπόν

είναι η μείωση των διαστάσεων του αρχικού συνόλου δεδομένων με την επιλογή ενός μικρού υποσυνόλου σχετικών γονιδίων για το κάθε μοντέλο, μέσω του οποίου επιτυγχάνεται υψηλή ακρίβεια ταξινόμησης.

Οι γονιδιακές υπογραφές στις οποίες καταλήξαμε εξετάζονται ως προς τον αριθμό των γονιδίων που περιέχουν καθώς και ως προς την ακρίβεια ταξινόμησης που επιτυγχάνουν. Επίσης εξετάζουμε τη βιολογική περιγραφή των επιλεγμένων γονιδίων με τη βοήθεια του συστήματος ταξινόμησης WebGestalt και του συστήματος Genotator.

1.7 Δομή της εργασίας

Η οργάνωση των κεφαλαίων που ακολουθούν , η οποία βασίστηκε στο τρόπο ανάπτυξης της παρούσας εργασίας, έχει ως εξής:

Στο Κεφάλαιο 2 περιγράφεται το μεθοδολογικό υπόβαθρο της εργασίας. Αναλύονται οι διάφοροι τύποι μηχανικής μάθησης, οι μέθοδοι επιλογής χαρακτηριστικών καθώς και η τεχνική cross-validation.

Στο Κεφάλαιο 3 γίνεται λεπτομερής αναφορά στους αλγορίθμους με τους οποίους υλοποιούνται τα μοντέλα μάθησης και πρόβλεψης που χρησιμοποιήθηκαν. Παρουσιάζονται τα νευρωνικά δίκτυα και το παραλλαγμένο νευρωνικό δίκτυο (FSMLP) που χρησιμοποιήσαμε. Γίνεται περιγραφή των Μηχανών Διανυσμάτων Υποστήριξης (SVMs), της μεθόδου RFE-LNW για την αναδρομική εξάλειψη των χαρακτηριστικών, καθώς και της μεθόδου LASSO. Επιπλέον παρουσιάζονται διάφορα μέτρα αξιολόγησης της απόδοσης ενός ταξινομητή.

Στο Κεφάλαιο 4 περιγράφεται η προτεινόμενη μεθοδολογία για τη επιλογή των γονιδιακών υπογραφών. Παρουσιάζονται αναλυτικά όλα τα βήματα που ακολουθήθηκαν για την επεξεργασία του συνόλου δεδομένων με τη βοήθεια τριών διαφορετικών τεχνικών και τα αποτελέσματα που προέκυψαν από την εφαρμογή τους.

Στο Κεφάλαιο 5 ακολουθεί η στατιστική και βιολογική σύγκριση των γονιδιακών υπογραφών.

Στο Κεφάλαιο 6 παρουσιάζονται τα συμπεράσματα που προέκυψαν και γίνεται αναφορά στις μελλοντικές επεκτάσεις της εργασίας.

Τέλος, το Παράρτημα Α περιλαμβάνει τα ονόματα και τα σύμβολα των γονιδίων που αποτελούν τις γονιδιακές υπογραφές, το Παράρτημα Β τη βιολογική επεξεργασία των ευρημάτων της μελέτης των Huber και συνεργατών [5] καθώς και εκείνης των Davis και συνεργατών[65] με σκοπό τη σύγκριση τους με τα ευρήματα της παρούσας εργασίας, ενώ στο παράρτημα Γ παραθέτουμε ένα λεξικό των όρων που χρησιμοποιήθηκαν στη διπλωματική εργασία.

Συνεισφορά της εργασίας αποτελεί η σύγκριση αλγορίθμων ταξινόμησης και επιλογής γονιδίων με βάση τη στατιστική τους συμπεριφορά καθώς και η σύγκριση των αποτελεσμάτων των μεθόδων αλλά και των βιολογικών ευρημάτων με την υπάρχουσα βιβλιογραφία.

ΚΕΦΑΛΑΙΟ 2: ΜΕΘΟΔΟΛΟΓΙΚΟ ΥΠΟΒΑΘΡΟ

2.1 Τεχνικές Μάθησης

Η Μηχανική Μάθηση είναι το πεδίο της έρευνας το οποίο σχετίζεται με την επίσημη μελέτη των συστημάτων μάθησης [30],[31]. Χαρακτηρίζεται ως κλάδος της Τεχνητής Νοημοσύνης, ο οποίος σχεδιάστηκε στις αρχές της δεκαετίας του '60 με στόχο να σχεδιάσει και να αναπτύξει αλγορίθμους και τεχνικές οι οποίες εφαρμόζουν διάφορους τύπους μάθησης, μηχανισμούς ικανούς να συμπεριλάβουν γνώση από παραδείγματα/στιγμιότυπα δεδομένων. Είναι ένα διεπιστημονικό πεδίο το οποίο δανείζεται και στηρίζεται σε ιδέες από τον τομέα της στατιστικής, της επιστήμης υπολογιστών, της μηχανικής, της θεωρίας βελτιστοποίησης και πολλών άλλων κλάδων της επιστήμης και των μαθηματικών. Η Μηχανική Μάθηση διαθέτει ένα ευρύ φάσμα εφαρμογών στους τομείς της βιοπληροφορικής, της ιατρικής διάγνωσης και της ταξινόμησης DNA ακολουθιών.

Οι κύριοι τύποι μηχανικής μάθησης είναι : (α)μάθηση με επίβλεψη (supervised learning), (β)μάθηση χωρίς επίβλεψη (unsupervised learning), (γ)μάθηση με ημί-επίβλεψη (semi-supervised learning).

- Η μάθηση με επίβλεψη είναι μια τεχνική της Μηχανικής Μάθησης για την δημιουργία μιας συνάρτησης από δεδομένα εκπαίδευσης, τα οποία αποτελούνται από ζεύγη αντικειμένων εισόδου και επιθυμητών εξόδων. Στην μάθηση με επίβλεψη, η μηχανή λαμβάνει μια ακολουθία εισόδων $x_1,x_2,x_3,..$ καθώς και μια ακολουθία επιθυμητών εξόδων $y_1,y_2,y_3,...$. Η έξοδος της συνάρτησης μπορεί να είναι είτε ένας πραγματικός αριθμός (παλινδρόμηση - regression) ή μπορεί να προβλέπει την ετικέτα κλάσης (class label) των δεδομένων εισόδου (ταξινόμηση – classification). Ο σκοπός της μηχανής είναι να μάθει να παράγει την σωστή έξοδο μιας καινούριας εισόδου έχοντας προηγουμένως εκπαιδευτεί πάνω σε κάποια δεδομένα (ζεύγη εισόδου και στόχου εξόδου, $[x_i, y_i]$). Για να επιτευχθεί αυτό, η διαδικασία θα πρέπει να μπορεί να γενικεύεται από τα δεδομένα εκπαίδευσης σε νέες αθέατες καταστάσεις (ανεξάρτητα σύνολα δεδομένων ελέγχου) .

- Η μάθηση χωρίς επίβλεψη είναι η μέθοδος της Μηχανικής Μάθησης στην οποία ένα μοντέλο προσαρμόζεται στις παρατηρήσεις. Διαφέρει από την μάθηση με επίβλεψη στο γεγονός ότι δεν διαθέτει a priori έξοδο. Στην μάθηση χωρίς επίβλεψη, η μηχανή λαμβάνει μόνο την ακολουθία εισόδων $x_1,x_2,x_3,..$ χωρίς την ακολουθία εξόδων. Τα δεδομένα αυτά αντιμετωπίζονται σαν ένα σύνολο από τυχαίες μεταβλητές. Ο σκοπός της μηχανής αυτής είναι να κατασκευάσει αναπαραστάσεις της εισόδου οι οποίες μπορούν να χρησιμοποιηθούν για τη λήψη αποφάσεων και τη πρόβλεψη μελλοντικών εισόδων. Κατά μία έννοια, η μάθηση χωρίς επίβλεψη μπορεί να θεωρηθεί ως μια διαδικασία εύρεσης προτύπων σε δεδομένα. Ένα παράδειγμα μάθησης χωρίς επίβλεψη αποτελεί η τεχνική της ομαδοποίησης (clustering).

Συχνά, η εύρεση και απόκτηση δεδομένων με ετικέτα είναι μια δύσκολη και χρονοβόρα διαδικασία σε αντίθεση με εκείνη για τα δεδομένα χωρίς ετικέτα.[32],[33]

- Η μάθηση με ημί-επίβλεψη αντιμετωπίζει το συγκεκριμένο πρόβλημα συνδυάζοντας κάποια χαρακτηριστικά των παραπάνω δυο μεθόδων. Πιο αναλυτικά, ένας μικρός αριθμός δεδομένων με ετικέτα χρησιμοποιείται σε συνδυασμό με ένα μεγαλύτερο σύνολο δεδομένων χωρίς ετικέτα για να κατασκευαστεί ένα καλύτερος ταξινομητής. Έτσι για παράδειγμα, ένα σύνολο δεδομένων **X** χωρίζεται σε δύο μέρη: τα δεδομένα $\mathbf{X}_m:=(x_1, \dots , x_m)$ για τα οποία δίνονται οι αντίστοιχες ετικέτες $\mathbf{Y}_m:=(y_1 , \dots , y_m)$, και τα δεδομένα $\mathbf{X}_n:=(x_{m+1}, \dots , x_n)$ για τα οποία οι ετικέτες δεν είναι γνωστές. Σημαντικές τεχνικές για την εύρεση των ετικετών των μη επιγεγραμμένων δεδομένων και τη κατασκευή ταξινομητών υψηλής ακρίβειας, χρησιμοποιώντας τη πληροφορία και από τα δύο σύνολα $\mathbf{X}_m$ και $\mathbf{X}_n$, είναι το Self-Training, το Co-Training και το Transductive SVMs (T-SVMs) [33]. Η ημι-επιβλεπόμενη μάθηση παρουσιάζει σημαντικό ενδιαφέρον , τόσο θεωρητικό όσο και πρακτικό, εξαιτίας της βελτιωμένης απόδοσής της σε σχέση με τους δυο προηγούμενους τύπους μάθησης που περιγράφηκαν (με επίβλεψη, χωρίς επίβλεψη), όταν ο αριθμός των δεδομένων με ετικέτα που προσφέρεται είναι περιορισμένος.

2.2 Επιλογή Χαρακτηριστικών (Feature Selection)

Τα τελευταία χρόνια η επιλογή χαρακτηριστικών [34] έχει γίνει το επίκεντρο πολλών ερευνητικών τομέων. Με την ταχεία πρόοδο και εξέλιξη της τεχνολογίας των ηλεκτρονικών υπολογιστών και των βάσεων δεδομένων, σύνολα δεδομένων που αποτελούνται από εκατοντάδες ή και χιλιάδες μεταβλητές ή χαρακτηριστικά συναντώνται συχνά στους τομείς της αναγνώρισης προτύπων, της εξόρυξης δεδομένων και της μηχανικής μάθησης. Η επεξεργασία τέτοιων μεγάλων συνόλων δεδομένων αποτελεί πρόκληση εξαιτίας του γεγονότος ότι οι παραδοσιακές τεχνικές μηχανικής μάθησης λειτουργούν καλά μόνο για μικρά σύνολα δεδομένων.

Ιδιαίτερα στο χώρο της βιοπληροφορικής [35], οι μαζικές μετρήσεις της γονιδιακής έκφρασης στην τεχνολογία DNA μικροσυστοιχίων έχει οδηγήσει σε δεδομένα με μεγάλο αριθμό γονιδίων (της τάξης των μερικών χιλιάδων) ο οποίος υπερέχει κατά πολύ του αριθμού των βιολογικών δειγμάτων (συνήθως μικρότερος από 100). Πολλές έρευνες δείχνουν ότι τα περισσότερα γονίδια που προκύπτουν από ένα πείραμα DNA μικροσυστοιχίας δεν είναι ικανά να διαχωρίσουν με ακρίβεια τις διαφορετικές κλάσεις του προβλήματος, και αρκετοί επιστήμονες πιστεύουν ότι απλοί ταξινομητές με λίγα γονίδια (λιγότερα από 15-20) επιτυγχάνουν καλύτερες ακρίβειες. Η επιλογή χαρακτηριστικών αντιμετωπίζει το παραπάνω πρόβλημα (γνωστό και ως «κατάρα των διαστάσεων») αφαιρώντας τα άσχετα ή περιττά δεδομένα. Με αυτό τον τρόπο, βελτιώνεται η απόδοση του αλγορίθμου εκπαίδευσης, μειώνεται το υπολογιστικό κόστος και παρέχεται καλύτερη κατανόηση των δεδομένων.

Οι μέθοδοι επιλογής δεικτών (χαρακτηριστικών) μπορούν να χωριστούν [34],[36] σε δύο επιμέρους κατηγορίες: (α) τις filter και (β)wrapper μεθόδους.

- Οι filter μέθοδοι εστιάζουν στα εγγενή χαρακτηριστικά των δεδομένων χρησιμοποιώντας διάφορες στοχαστικές μετρικές όπως Fisher's ratio, T-statistics, x^2 statistic, information gain και πολλές άλλες [35]. Τα χαρακτηριστικά (γονίδια) κατατάσσονται ανάλογα με την τιμή τους σε μια από τις παραπάνω μετρικές και τα χαρακτηριστικά (γονίδια) εκείνα με τις υψηλότερες τιμές τα οποία παρουσιάζουν την καλύτερη ικανότητα διάκρισης σε συνδυασμό με την υψηλότερη ακρίβεια ταξινόμησης, επιλέγονται ως τα πιο σημαντικά. Το υπολογιστικό κόστος των filter μεθόδων είναι σχετικά μικρό αφού είναι ανεξάρτητες από τον αλγόριθμο μάθησης που εφαρμόζεται. Παρόλα αυτά, ενέχουν τον κίνδυνο επιλογής υποσυνόλων

χαρακτηριστικών τα οποία πιθανώς να μην ταιριάζουν με τον αλγόριθμο εκπαίδευσης που έχει επιλεχθεί.

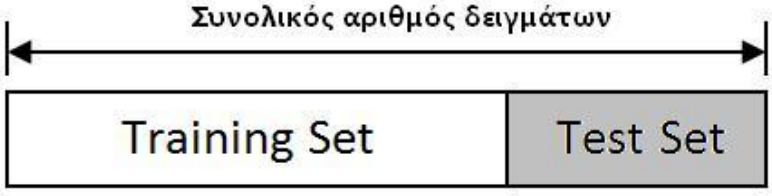
- Οι wrapper μέθοδοι, από την άλλη, χρησιμοποιούν τον αλγόριθμο μάθησης για να αξιολογήσουν τα υποσύνολα χαρακτηριστικών [37],[38]. Αναλυτικότερα, δουλεύουν με έναν αναδρομικό τρόπο, όπου ένας ταξινομητής χρησιμοποιείται για να εκχωρήσει ένα σχετικό βάρος σε κάθε χαρακτηριστικό και στην συνέχεια το ή τα χαρακτηριστικά με το μικρότερο βάρος εξαλείφονται. Στον επόμενο κύκλο επανάληψης, τα βάρη επανεκτιμούνται και προσαρμόζονται δυναμικά, ενώ η διαδικασία συνεχίζεται αναδρομικά. Τέλος, το μικρότερο σύνολο χαρακτηριστικών που επιτυγχάνει την υψηλότερη ακρίβεια ταξινόμησης επιλέγεται ως το σύνολο των πιο σημαντικών χαρακτηριστικών. Οι wrapper μέθοδοι υπερτερούν των filter όσον αναφορά την ακρίβεια πρόβλεψης, αλλά σε γενικές γραμμές το υπολογιστικό τους κόστος είναι μεγαλύτερο.

2.3 Αξιολόγηση αποτελεσμάτων μέσω Cross Validation

Το cross validation [39] είναι μια στατιστική μέθοδος αξιολόγησης και σύγκρισης αλγορίθμων εκμάθησης. Χωρίζει τα διαθέσιμα δεδομένα (dataset) σε δύο υποσύνολα: ένα που χρησιμοποιείται για την εκμάθηση ή εκπαίδευση ενός μοντέλου (training set) , και ένα που χρησιμοποιείται για την αξιολόγηση του μοντέλου (test / validation set). Μετά την ολοκλήρωση της εκπαίδευσης του μοντέλου με το training set , ακολουθεί η πρόβλεψη για τα δεδομένα που περιέχονται στο test set. Η απόδοση του αλγορίθμου εκμάθησης (μοντέλου) που εξετάζεται μπορεί να υπολογιστεί χρησιμοποιώντας κάποιες προκαθορισμένες μετρικές απόδοσης, όπως για παράδειγμα η ακρίβεια ταξινόμησης. Πολλές φορές, για τη μείωση της μεταβλητότητας, πραγματοποιούνται αρκετές επαναλήψεις του cross validation. Σε αυτή τη περίπτωση, ως συνολικό μέτρο αξιολόγησης του αλγόριθμου εκπαίδευσης υπολογίζεται ο μέσος όρος των τιμών της μετρικής (π.χ. ακρίβεια) που έχει χρησιμοποιηθεί σε κάθε κύκλο.

Οι πιο συνηθισμένες τεχνικές για cross validation είναι οι εξής:

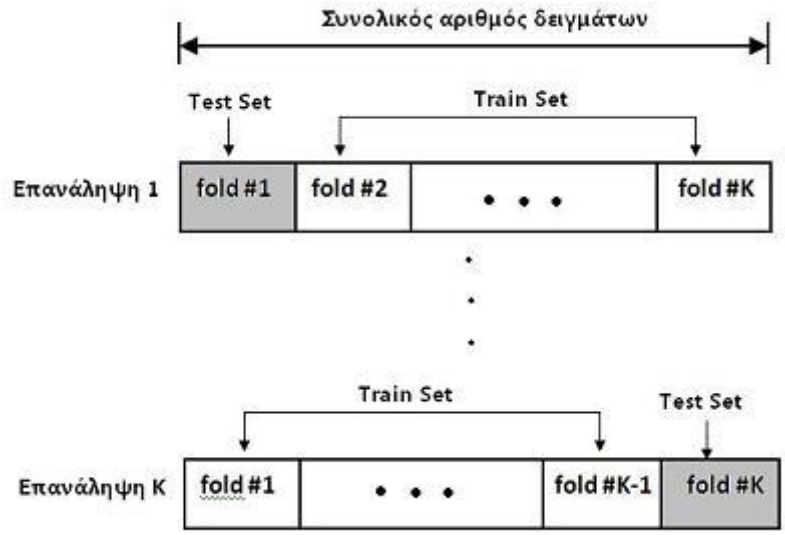
- **Hold-Out Validation:** Στην hold-out μέθοδο τα διαθέσιμα δεδομένα χωρίζονται σε δύο μη επικαλυπτόμενα υποσύνολα: το ένα για την εκπαίδευση (Training set) και το άλλο για την αξιολόγηση του επιθυμητού μοντέλου (Test set), όπως φαίνεται στην Εικόνα 7. Το Test set δεν χρησιμοποιείται καθόλου στην διαδικασία της εκπαίδευσης. Αυτό έχει ως αποτέλεσμα η hold-out μέθοδος να αποφεύγει την επικάλυψη μεταξύ των δεδομένων εκπαίδευσης και ελέγχου, αποδίδοντας κατά αυτό τον τρόπο μια ακριβέστερη εκτίμηση της απόδοσης του αλγορίθμου. Το μειονέκτημα αυτής της μεθόδου είναι ότι δεν χρησιμοποιεί όλα τα διαθέσιμα δεδομένα και αυτό έχει ως συνέπεια τα αποτελέσματα να εξαρτώνται σε μεγάλο βαθμό από την επιλογή για τον διαχωρισμό των δεδομένων σε Training set και Test set.



Εικόνα 7: Holdout Μέθοδος [40].

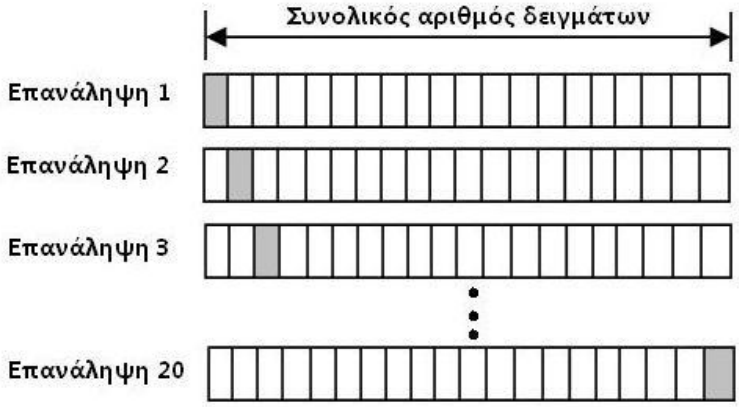
Οι μέθοδοι k-fold και Leave-One-Out Cross validation χαρακτηρίζονται ως κλασσικές μέθοδοι cross validation [39],[41], οι οποίες χρησιμοποιούν τα δεδομένα εκπαίδευσης (training data) με σκοπό να εκτιμήσουν την ικανότητα πρόβλεψης ενός μοντέλου ταξινόμησης. Στις δύο αυτές τυπικές μορφές του cross validation, τα training και test sets πρέπει να αναμιχθούν σε διαδοχικές επαναλήψεις έτσι ώστε κάθε στοιχείο του dataset να έχει την ευκαιρία να αξιολογηθεί ομοίως με τα υπόλοιπα.

- **K-fold Cross Validation**: Η k-fold cross validation μέθοδος χωρίζει τυχαία το σύνολο δεδομένων σε k μη επικαλυπτόμενα υποσύνολα (folds) περίπου ίδιου μεγέθους (N/k, όπου N: ο αριθμός δειγμάτων του συνόλου δεδομένων). Ένα μοντέλο ταξινόμησης εκπαιδεύεται στα k-1 υποσύνολα και στην συνέχεια ταξινομεί τα δείγματα του υποσυνόλου εκείνου που δεν χρησιμοποιήθηκε στην διαδικασία εκπαίδευσης. Αυτή η διαδικασία επαναλαμβάνεται k φορές, έτσι ώστε κάθε δείγμα που ταξινομείται από το μοντέλο να μην έχει συμπεριληφθεί στην διαδικασία εκπαίδευσής του. Συνήθως το k παίρνει τις τιμές 5 ή 10. Στην Εικόνα 8 απεικονίζεται η μέθοδος k-fold cross validation. Όταν το k=N, όπου N ο συνολικός αριθμός δειγμάτων του dataset, έχουμε την περίπτωση του Leave-One-Out cross validation.



Εικόνα 8 : Μέθοδος K-fold cross validation. Με γκρι χρώμα σημειώνονται τα test set σε κάθε επανάληψη [42].

- **<u>Leave-One-Out Cross Validation:</u>** Η μέθοδος leave one out cross validation (LOOCV)- (γνωστή και ως jackknifed classification [41]) - είναι μια ειδική περίπτωση του k-fold cross validation όπου το k είναι ίσο με τον αριθμό των σημείων δεδομένων του dataset. Με άλλα λόγια, σε κάθε επανάληψη το μοντέλο εκπαιδεύεται πάνω σε όλα τα σημεία εκτός από ένα, ενώ η αξιολόγησή του γίνεται πάνω σε αυτό το σημείο. Για αυτό το λόγο, η εκτιμώμενη ακρίβεια (accuracy) που προκύπτει χρησιμοποιώντας την μέθοδο LOOCV είναι σχεδόν αμερόληπτη. Η συγκεκριμένη μέθοδος χρησιμοποιείται ευρέως στον κλάδο της Βιοπληροφορικής, όπου συνήθως τα δεδομένα αποτελούνται από μερικές δεκάδες δειγμάτων. Στην Εικόνα 9 παρουσιάζεται η μέθοδος LOOCV όπως εφαρμόζεται σε ένα σύνολο δεδομένων που αποτελείται από N=20 δείγματα.



Εικόνα 9 : Μέθοδος Leave-One-out cross validation. Με γκρι χρώμα σημειώνονται τα test set σε κάθε επανάληψη [40].

Στον Πίνακα 1 που ακολουθεί εμφανίζονται συνοπτικά τα πλεονεκτήματα και τα μειονεκτήματα των μεθόδων Hold-out, k-fold και Leave-One-Out cross validation που αναλύθηκαν παραπάνω.

Πίνακας 1: Πλεονεκτήματα και μειονεκτήματα των μεθόδων αξιολόγησης [39]

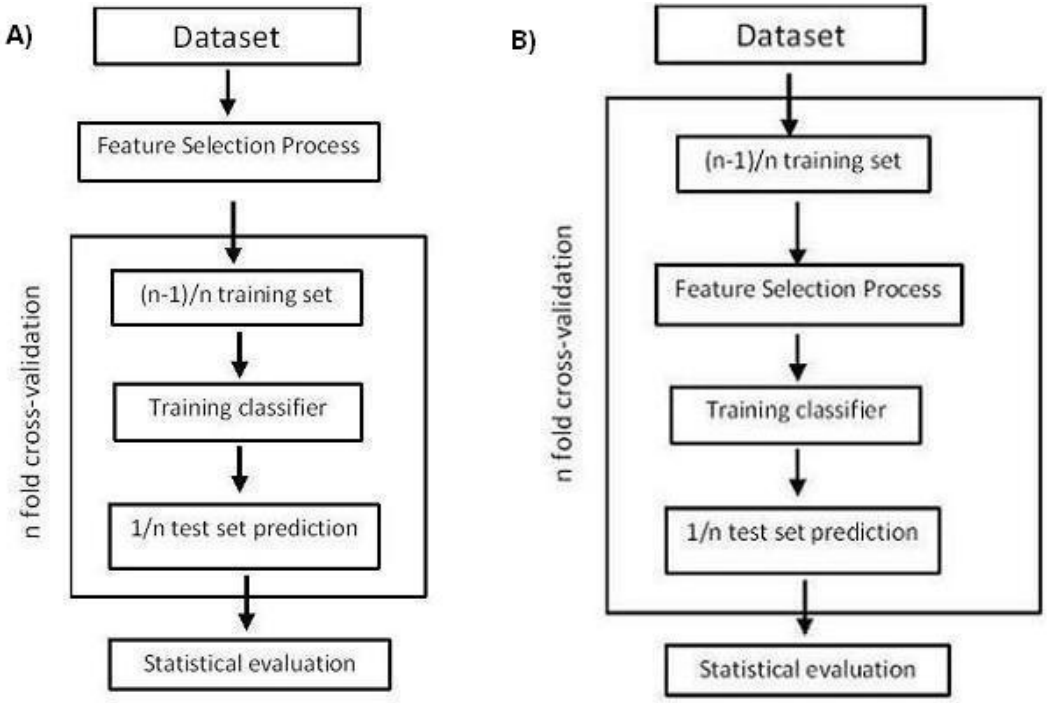
Μέθοδος Validation	Πλεονεκτήματα	Μειονεκτήματα
Hold-out validation	-Ανεξάρτητο σύνολο δεδομένων εκπαίδευσης και ελέγχου -Αρκετά γρήγορη μέθοδος	- Μειωμένα δεδομένα για εκπαίδευση και αξιολόγηση. -Μεγάλη διακύμανση
k-fold cross validation	-Ακριβής εκτίμηση της απόδοσης του μοντέλου	-Επικαλυπτόμενα δεδομένα εκπαίδευσης -Αργή υπολογιστικά σε σχέση με την Hold-out μέθοδο
Leave-One- Out cross validation	-Αμερόληπτη εκτίμηση της απόδοσης του μοντέλου -Ευρεία χρήση στον κλάδο της βιοπληροφορικής	-Υψηλό υπολογιστικό κόστος (το πιο υψηλό σε σχέση με τις δύο προηγούμενες μεθόδους)

Οι μέθοδοι Leave-One- Out και k-fold Cross Validation μπορούν να υλοποιηθούν ως εσωτερικό (Εικόνα 10(A)) ή εξωτερικό (Εικόνα 10(B)) cross validation [41]. Αναλυτικότερα:

- **Εσωτερικό (internal) cross validation:** Χωρίζουμε τα δεδομένα μας σε training και test set χρησιμοποιώντας την Leave-one-out ή την k-fold μέθοδο μετά την διαδικασία επιλογής χαρακτηριστικών (γονιδίων). Έπειτα εκπαιδεύουμε έναν αριθμό ταξινομητών (N για την Leave-one out και k για την k-fold). Κάθε ένας από τους παραπάνω ταξινομητές εκπαιδεύεται σε ένα υποσύνολο δειγμάτων εκπαίδευσης(N-1 για την LOOCV και k-1 για την K-fold μέθοδο αντίστοιχα) και έπειτα χρησιμοποιείται για να ταξινομήσει τα εναπομείναντα δείγματα.

- **Εξωτερικό (external) cross validation:** Χωρίζουμε τα δεδομένα μας σε training και test set με την βοήθεια των μεθόδων LOOCV και k-fold CV πριν από την διαδικασία επιλογής χαρακτηριστικών. Χρησιμοποιούμε κάθε ένα από τα N ή k υποσύνολα που προέκυψαν για να εκτελέσουμε την διαδικασία επιλογής χαρακτηριστικών (independent feature selection) και εκπαίδευσης του ταξινομητή. Στην συνέχεια ταξινομούμε τα δείγματα (test set) που δεν περιλαμβάνονται στο υποσύνολο εκπαίδευσης που χρησιμοποιήθηκε.

Στο εσωτερικό cross validation, κάθε ένας από τους N ή k ταξινομητές κατασκευάζεται από ένα υποσύνολο χαρακτηριστικών που έχουν επιλεχθεί χρησιμοποιώντας ολόκληρο το σύνολο δεδομένων (με όλα τα δείγματα). Κατά συνέπεια, τα δείγματα που επιλέγονται μέσω cross validation για την αξιολόγηση του ταξινομητή δεν μπορούν να θεωρηθούν ως ένα ανεξάρτητο σύνολο δεδομένων ελέγχου (independent test set). Αντίθετα, στο εξωτερικό cross validation, τα δείγματα για την αξιολόγηση του ταξινομητή, δεν συμμετέχουν στην διαδικασία επιλογής χαρακτηριστικών. Έτσι μπορούν καλύτερα να θεωρηθούν ως ένα ανεξάρτητο test set. Σε σύνολα δεδομένων γονιδιακής έκφρασης με μεγάλο αριθμό χαρακτηριστικών (γονιδίων) και σχετικά μικρό αριθμό βιολογικών δειγμάτων δεν προτείνεται η χρήση του εσωτερικού cross validation. Ο λόγος είναι ότι για σύνολα δεδομένων αυτής της μορφής με μικρό αριθμό δειγμάτων, η χρήση εσωτερικού CV επιτυγχάνει μηδενικά ή σχεδόν μηδενικά ποσοστά σφάλματος ταξινόμησης, αλλά παράλληλα ο ταξινομητής που κατασκευάζεται προσαρμόζεται τόσο πολύ στα δεδομένα εκπαίδευσης που αδυνατεί να ταξινομήσει σωστά νέα δείγματα (φαινόμενο overfitting).



Εικόνα 10: Εσωτερικό (A) και Εξωτερικό (B) Cross validation [43].

ΚΕΦΑΛΑΙΟ 3: ΥΠΟΛΟΓΙΣΤΙΚΕΣ ΜΕΘΟΔΟΙ ΓΙΑ ΤΑΞΙΝΟΜΗΣΗ

3.1 Νευρωνικά Δίκτυα

Η έρευνα σχετικά με τα τεχνητά νευρωνικά δίκτυα [44], [45] (χάριν συντομίας αποκαλούνται συνήθως «νευρωνικά δίκτυα») είναι εμπνευσμένη από την δομή και την λειτουργία του εγκεφάλου. Ο εγκέφαλος είναι ένας εξαιρετικά πολύπλοκος, μη γραμμικός, παράλληλος υπολογιστής (σύστημα επεξεργασίας πληροφοριών). Έχει την δυνατότητα να οργανώνει τα δομικά του στοιχεία, γνωστά ως νευρώνες, με τρόπο ώστε να εκτελούν συγκεκριμένους υπολογισμούς (π.χ. αναγνώριση προτύπων, αντίληψη και έλεγχο της κίνησης) με ταχύτητα πολλαπλάσια από αυτή του γρηγορότερου υπολογιστή που υπάρχει σήμερα. Κίνητρο για την μελέτη του νευρώνα και των νευρωνικών δικτύων είναι η ελπίδα ανακάλυψης ενός νέου υπολογιστικού μοντέλου βασισμένου σε μια δικτυακή δομή παρόμοια με αυτή του εγκεφάλου

Τα συνήθη Τεχνητά Νευρωνικά Δίκτυα (ΤΝΔ) χρησιμοποιούν πολύ απλοποιημένα μοντέλα νευρώνων τέτοια ώστε να διατηρούν μόνο τα πολύ αδρά χαρακτηριστικά των λεπτομερών μοντέλων που χρησιμοποιούνται στην νευρολογία. Μπορούμε λοιπόν να διατυπώσουμε τον ορισμό ενός ΤΝΔ ως εξής [Alexander & Morton 1990]:

Ένα νευρωνικό δίκτυο είναι ένας τεράστιος παράλληλος επεξεργαστής με κατανεμημένη αρχιτεκτονική, ο οποίος αποτελείται από απλές μονάδες επεξεργασίας και έχει από τη φύση του τη δυνατότητα να αποθηκεύει εμπειρική γνώση και να την καθιστά διαθέσιμη για χρήση. Μοιάζει με τον ανθρώπινο εγκέφαλο σε δύο σημεία:

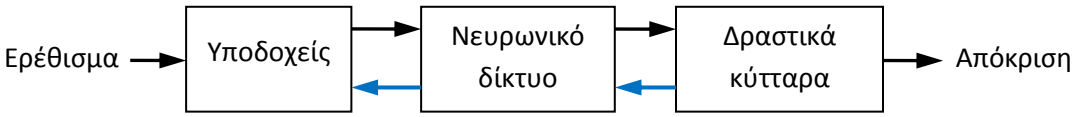
1. *Το δίκτυο προσλαμβάνει την γνώση από το περιβάλλον του, μέσω μιας διαδικασίας μάθησης (ιδιότητα γνωστή ως πλαστικότητα των νευρώνων).*
2. *Η ισχύς των συνδέσεων μεταξύ των νευρώνων, που αποκαλείται συναπτικό βάρος χρησιμοποιείται για την αποθήκευση της γνώσης που αποκτιέται.*

Η διαδικασία μέσω της οποίας επιτυγχάνεται η μάθηση αποκαλείται αλγόριθμος μάθησης και η λειτουργία του είναι να τροποποιεί τα συναπτικά βάρη του δικτύου με τον κατάλληλο τρόπο για την επίτευξη του επιθυμητού στόχου. Η πρόκληση που αντιμετωπίζει η θεωρία των ΤΝΔ είναι η εύρεση κατάλληλων αλγορίθμων εκπαίδευσης των δικτύων και ανάκλησης της πληροφορίας που αυτά περιέχουν. Για την επίτευξη αυτού του στόχου απαιτείται ο ορισμός του κατάλληλου περιβάλλοντος εκπαίδευσης, πχ. αν ο δίκτυο θα εκπαιδεύεται με επίβλεψη ή αν το δίκτυο θα αφήνεται μόνο του να αυτό-οργανωθεί (χωρίς επίβλεψη) και με ποιο συγκεκριμένο κριτήριο και στόχο.

3.2 Η Βιολογική Έμπνευση των Τεχνητών Νευρωνικών Δικτύων

3.2.1 Ο Ανθρώπινος Εγκέφαλος

Το ανθρώπινο νευρικό σύστημα [45] μπορεί να αντιμετωπιστεί ως ένα σύστημα τριών σταδίων, όπως απεικονίζεται στην Εικόνα 11.



Εικόνα 11: Σχηματική Αναπαράσταση Νευρικού Συστήματος [45].

Το κέντρο του συστήματος είναι ο εγκέφαλος, ο οποίος αναπαριστάνεται από το νευρωνικό (νευρικό) δίκτυο. Στην εικόνα παρουσιάζονται δύο ομάδες βελών. Αυτά με κατεύθυνση από αριστερά προς τα δεξιά υποδεικνύουν την μετάδοση των σημάτων πληροφορίας προς τα εμπρός (πρόσθια τροφοδότηση του συστήματος). Αυτά με κατεύθυνση από δεξιά προς τα αριστερά σηματοδοτούν την παρουσία ανάδρασης

(feedback) στο σύστημα. Οι υποδοχείς μετατρέπουν τα ερεθίσματα που προέρχονται από το ανθρώπινο σώμα ή το εξωτερικό περιβάλλον σε ηλεκτρικά σήματα (ώσεις) που μεταφέρουν πληροφορία στο νευρικό δίκτυο (τον εγκέφαλο). Τα δραστικά κύτταρα μετατρέπουν τα ηλεκτρικά σήματα που παράγονται από το νευρωνικό δίκτυο σε αισθητές αποκρίσεις (οι έξοδοι του συστήματος).

3.2.2 Βιολογικός Νευρώνας

Το νευρικό κύτταρο ή νευρώνας (Εικόνα 12) είναι το βασικό δομικό στοιχείο του εγκεφάλου [44] τόσο στον άνθρωπο όσο και στα ζώα. Ο νευρώνας είναι ένα μεγάλο σε μέγεθος κύτταρο το οποίο, ανατομικά, αποτελείται από τα εξής τμήματα:

a. το **σώμα**,
b. τους **δενδρίτες**,
c. τον **άξονα**, και
d. τις **συνάψεις** που συνδέουν τις διακλαδώσεις του άξονα με τους δενδρίτες άλλων νευρώνων δημιουργώντας έτσι ένα νευρωνικό δίκτυο.

Λειτουργικά, τα τμήματα του νευρώνα παίζουν διαφορετικούς ρόλους:

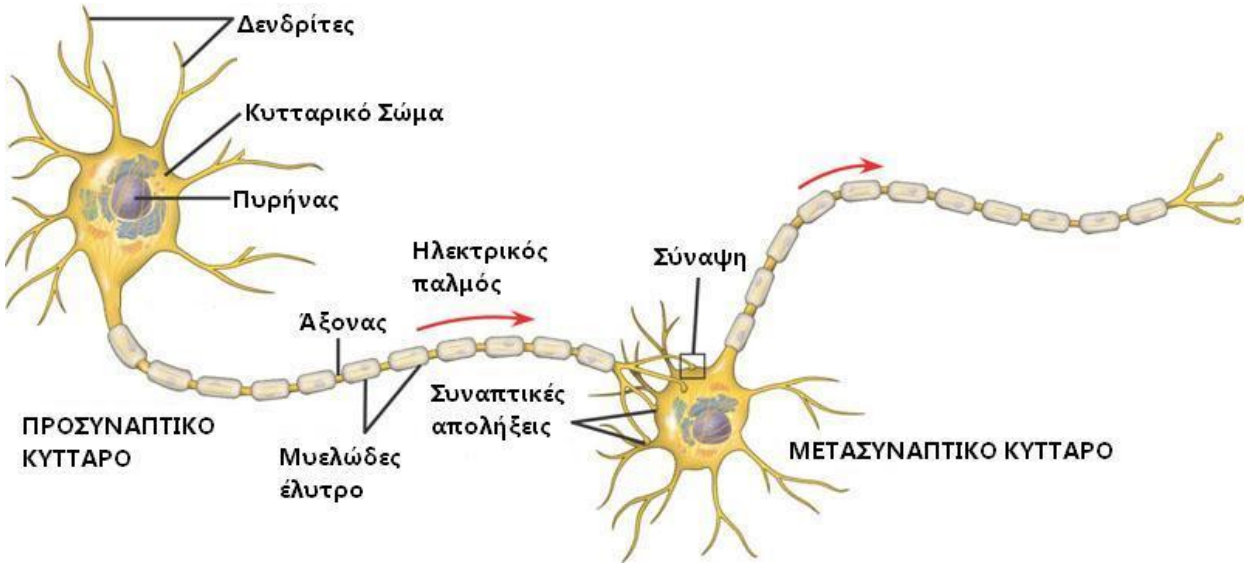
- Οι **δενδρίτες** είναι οι πύλες εισόδου του νευρώνα. Δέχονται ηλεκτρικά σήματα από άλλους νευρώνες.
- Ο **άξονας** είναι η πύλη εξόδου του νευρώνα. Ο άξονας στέλνει σήματα προς άλλους νευρώνες υπό μορφή ηλεκτρικών παλμών σταθερού πλάτος αλλά μεταβλητής συχνότητας.
- Οι **συνάψεις** είναι τα σημεία ένωσης μεταξύ διακλαδώσεων του άξονα ενός νευρώνα και των δενδριτών από άλλους νευρώνες. Είναι κύστες με ηλεκτροχημικό υλικό – ιόντα, κυρίως Νατρίου και Καλίου (Na^+, K^+). Το υλικό αυτό μεταδίδει την ηλεκτρική δραστηριότητα του άξονα – αποστολέα στους δενδρίτες – παραλήπτες. Το ποσοστό της ηλεκτρικής δραστηριότητας που μεταδίδεται τελικά στον δενδρίτη λέγεται συναπτικό βάρος. Οι συνάψεις χωρίζονται σε ενισχυτικές (exitatory) και σε ανασταλτικές (inhibitory) ανάλογα με το αν το φορτίο που εκλύεται από την σύναψη ερεθίζει το νευρώνα προς το να παράγει παλμούς με μεγαλύτερη συχνότητα ή αντίθετα αν τον καταστέλλει εμποδίζοντάς τον να παράγει παλμούς.

3.2.3 Πώς λειτουργεί ο βιολογικός νευρώνας;

Στους βιολογικούς νευρώνες [44], φορείς πληροφορίας είναι οι ηλεκτρικοί παλμοί που ταξιδεύουν στον άξονα κάθε νευρώνα και μέσω των συνάψεων διαδίδονται στους δενδρίτες των παραληπτών νευρώνων (Εικόνα 12). Κάθε νευρώνας Α συλλέγει όλο το ηλεκτρικό φορτίο που δέχεται από κάθε σύναψη στους δενδρίτες του ζυγίζοντας το εισερχόμενο φορτίο με το αντίστοιχο συναπτικό βάρος. Έτσι, όσο πιο ισχυρή είναι η συναπτική ζεύξη τόσο πιο πολύ έντονα συμμετέχει το συγκεκριμένο φορτίο εισόδου στο συνολικό άθροισμα. Αν το άθροισμα του φορτίου ξεπερνάει κάποιο κατώφλι τότε ο άξονας του Α αρχίζει να παράγει ηλεκτρικούς παλμούς με μεγάλη συχνότητα οπότε λέμε ότι ο νευρώνας πυροβολεί (fires). Αν όμως το φορτίο δεν περνάει το συγκεκριμένο αυτό όριο τότε ο νευρώνας παράγει πολύ αραιά παλμούς σε τυχαίες στιγμές οπότε λέμε ότι ο νευρώνας είναι αδρανής. Κάθε παλμός έχει συγκεκριμένο χρονικό πλάτος t_p και μετά από κάθε παλμό ο νευρώνας χρειάζεται ένα ελάχιστο χρόνο ανάπαυσης t_r. Έτσι ο μέγιστος ρυθμός των παλμών δεν ξεπερνάει το όριο

$$\text{Firing frequency} < 1 / (t_p + t_r) \qquad (3.1)$$

Τελικά οι παλμοί που παράγονται ταξιδεύουν κατά μήκος του άξονα και τροφοδοτούν τους άλλους νευρώνες με τους οποίους συνδέεται ο Α.



Εικόνα 12 : Δομή βιολογικού νευρώνα. Στην εικόνα απεικονίζεται η μεταφορά ηλεκτρικών παλμών μεταξύ δύο νευρώνων με την βοήθεια των συναπτικών απολήξεων [46].

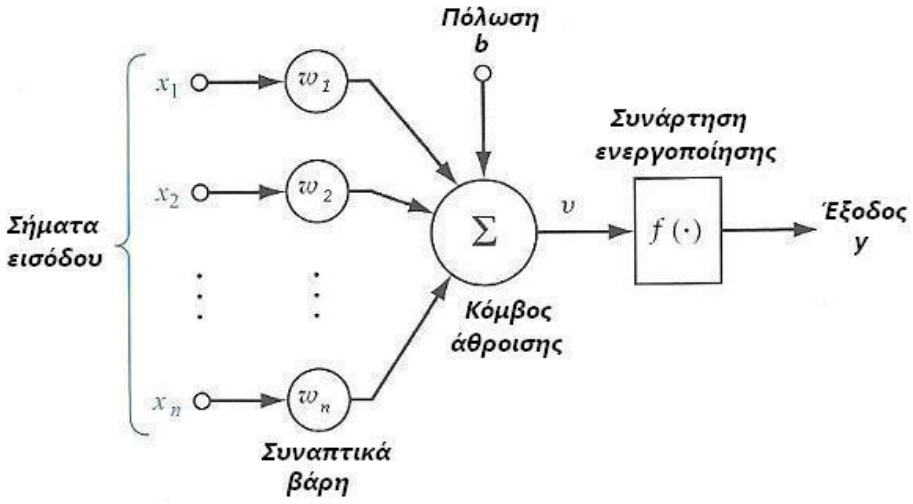
3.3 Τεχνητά Νευρωνικά Δίκτυα

3.3.1 Μοντέλα Νευρώνων

Ένας νευρώνας είναι μια μονάδα επεξεργασίας πληροφορίας, η οποία είναι θεμελιώδης για την λειτουργία ενός νευρωνικού δικτύου. Η Εικόνα 13 παρουσιάζει το μοντέλο ενός νευρώνα που αποτελεί την βάση για την σχεδίαση μιας μεγάλης οικογένειας νευρωνικών δικτύων. Τα τρία βασικά στοιχεία [45] αυτού του μοντέλου είναι:

1. Ένα σύνολο *συνάψεων*, κάθε μια εκ των οποίων χαρακτηρίζεται από το δικό της βάρος. Συγκεκριμένα, ένα σήμα x_i στην είσοδο της σύναψης i που συνδέεται με τον νευρώνα πολλαπλασιάζεται επί το συναπτικό βάρος w_i (συναπτικά βάρη - synaptic weights). Τα συναπτικά βάρη ενός τεχνητού νευρώνα είναι πραγματικοί αριθμοί, θετικοί για τις ενισχυτικές συνάψεις και αρνητικοί για τις ανασταλτικές συνάψεις.
2. Ένας *αθροιστής (adder)* για την άθροιση των σημάτων εισόδου, σταθμισμένων από τα αντίστοιχα συναπτικά βάρη του νευρώνα.
3. Μια *συνάρτηση ενεργοποίησης (activation function)* ***f*()** για τον περιορισμό του πλάτους του σήματος εξόδου ενός νευρώνα.

Παρατηρούμε ότι το μοντέλο νευρώνα της Εικόνας 13 περιλαμβάνει επίσης μια εξωτερικά εφαρμοζόμενη πόλωση (bias) b, που έχει επίδραση στην αύξηση ή μείωση της εισόδου u στην εφαρμοζόμενη συνάρτηση ενεργοποίησης, ανάλογα με το αν είναι θετική ή αρνητική αντίστοιχα.



Εικόνα 13: Μη γραμμικό μοντέλο νευρώνα [45].

Αν x_1, x_2,…, x_n είναι τα σήματα εισόδου, w_1, w_2,…, w_n τα αντίστοιχα συναπτικά βάρη, u η έξοδος του αθροιστή, b η πόλωση, *f()* η συνάρτηση ενεργοποίησης και *y* το σήμα εξόδου του νευρώνα, μπορούμε να περιγράψουμε το μοντέλο της Εικόνας 13 γράφοντας το ακόλουθο ζεύγος εξισώσεων:

$$u = \sum_{i=1}^{n} w_i x_i \tag{3.2}$$

$$y = f(u + b) \tag{3.3}$$

Η πόλωση b είναι ένας πραγματικός αριθμός (θετικός ή αρνητικός) όπως άλλωστε και τα συναπτικά βάρη w_1, w_2,…, w_n. Κατ' αυτή την έννοια η πόλωση b μπορεί να θεωρηθεί ως ένα επί πλέον συναπτικό βάρος συνδεδεμένο με μια σταθερή είσοδο x_0 η οποία έχει πάντα την τιμή 1. Έτσι θα μπορούσαμε να γράψουμε

$$u = \sum_{i=1}^{n} w_i x_i + b = \sum_{i=0}^{n} w_i x_i \tag{3.4}$$

$$y = f(u) \tag{3.5}$$

όπου w_0=b και x_0=1.

Το μοντέλο McCulloch-Pitts

Την δεκαετία του 1940 υπήρξε μια εντονότατη δραστηριότητα προς την κατεύθυνση της μελέτης των βιολογικών νευρωνικών δικτύων και της μαθηματικής μοντελοποίησης τους. Πρωτοπόροι στον τομέα αυτό οι Αμερικανοί επιστήμονες McCulloch και Pitts [47]. Το απλό μοντέλο της δραστηριότητας του νευρώνα που περιέγραψαν έχει τις παρακάτω ιδιότητες:

Χρησιμοποιεί ως συνάρτηση ενεργοποίησης του νευρώνα την λεγόμενη βηματική συνάρτηση

Βηματική συνάρτηση 0/1 (step function 0/1):

$$f(u) = \begin{cases} 0, & u \leq 0 \\ 1, & u > 0 \end{cases} \tag{3.6}$$

και αντίστοιχα η έξοδος του νευρώνα εκφράζεται ως

$$y = f(u) = \begin{cases} 0, & u \leq 0 \\ 1, & u > 0 \end{cases} \quad (3.7)$$

όπου u άθροισμα του φορτίου που δέχεται ο νευρώνας

$$u = \sum_{i=1}^{n} w_i x_i + b \quad (3.8)$$

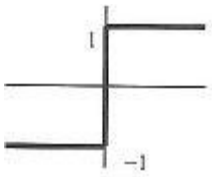
Σχηματικά το παραπάνω μαθηματικό μοντέλο παριστάνεται από έναν αθροιστή ακολουθούμενο από ένα μη-γραμμικό μετασχηματιστή f (Εικόνα 13) .

Άλλα διαδεδομένα μοντέλα

Υπάρχουν πολλές διαφορετικές μοντελοποιήσεις [44] του νευρώνα που αποκλίνουν από το απλό μοντέλο McCulloch-Pitts. Η πιο σημαντική διαφορά είναι στη μορφή της μη γραμμικής συνάρτησης *f*() που χρησιμοποιείται στην έξοδο. Η συνάρτηση αυτή (που καλείται και συνάρτηση ενεργοποίησης του νευρώνα (neuron activation function)) μπορεί να πάρει εναλλακτικά τις παρακάτω μορφές:

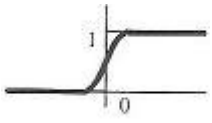
- *Βηματική συνάρτηση -1/1 (step function -1/1):*

$$f(u) = \begin{cases} -1, & u \leq 0 \\ 1, & u > 0 \end{cases} \quad (3.9)$$



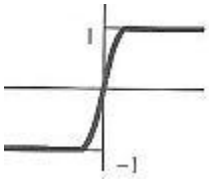
- *Σιγμοειδής (sigmoid):*

$$f(u) = 1/(1 + e^{-u}) \quad (3.10)$$



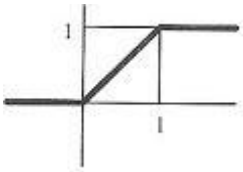
- *Υπερβολική εφαπτομένη (hyperbolic tangent):*

$$f(u) = \tanh(u) = (1 - e^{-u})/(1 + e^{-u}) \quad (3.11)$$



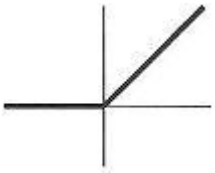
- *Συνάρτηση κατωφλίου (threshold function):*

$$f(u) = \begin{cases} 0, & u \leq 0 \\ u, & 0 < u < 1 \\ 1, & u \geq 1 \end{cases} \qquad (3.12)$$



- *Συνάρτηση ράμπας (ramp function):*

$$f(u) = \begin{cases} 0, & u \leq 0 \\ u, & u > 0 \end{cases} \qquad (3.13)$$



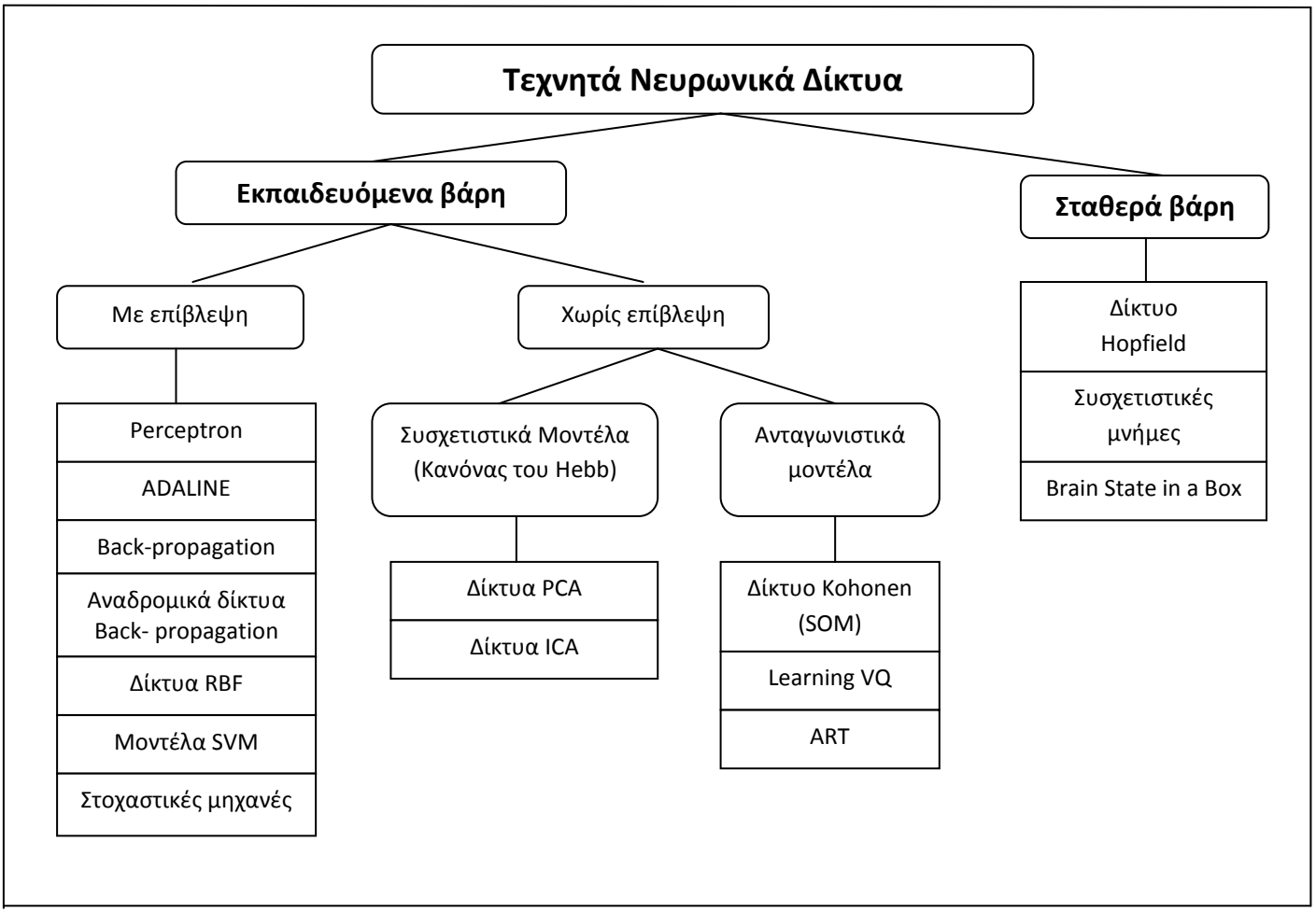
- *Γραμμική (linear):*

$$f(u) = u \qquad (3.14)$$

3.3.2 Ταξινόμηση Νευρωνικών Αλγορίθμων

Μια από τις πιο βασικές ιδιότητες των Νευρωνικών Δικτύων είναι η ικανότητά τους για εκπαίδευση (ή αλλιώς **μάθηση**). Η εκπαίδευση αυτή επιτυγχάνεται μέσω της ανταλλαγής τιμών και βαρών, που αποσκοπεί στη βαθμιαία σύλληψη της πληροφορίας η οποία στη συνέχεια θα είναι διαθέσιμη προς ανάκτηση. Υπό μια ευρεία έννοια, μπορούμε να κατηγοριοποιήσουμε τις διαδικασίες μάθησης μέσω των οποίων λειτουργούν τα νευρωνικά δίκτυα ως εξής: **μάθηση με επίβλεψη** (supervised learning) και **μάθηση χωρίς επίβλεψη** (unsupervised learning) [45],[48].

Στην Εικόνα 14 που ακολουθεί παρουσιάζονται συνοπτικά οι νευρωνικοί αλγόριθμοι:



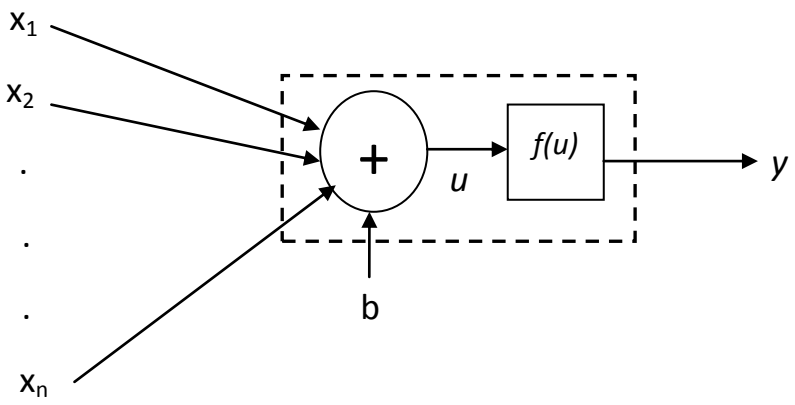
Εικόνα 14 : Αλγόριθμοι νευρωνικών δικτύων ταξινομημένοι ανάλογα με το περιβάλλον εκπαίδευσης[44].

3.3.3 Το δίκτυο Perceptron

Το μοντέλο perceptron (“αισθητήρας”) καταλαμβάνει μια ειδική θέση στην ιστορική εξέλιξη των νευρωνικών δικτύων: ήταν το πρώτο νευρωνικό που μπορούσε να περιγραφεί αλγοριθμικά. Επινοήθηκε από τον Rosenblatt (1958) [49] και βασίζεται σε ένα μη γραμμικό νευρώνα – συγκεκριμένα στο μοντέλο ενός νευρώνα των *McCulloch – Pitts*. Γνωρίζουμε ότι ένα τέτοιο νευρωνικό μοντέλο αποτελείται από ένα γραμμικό συνδυαστή (αθροιστής) ο οποίος ακολουθείται από έναν απότομο περιοριστή (βηματική συνάρτηση 0/1 ή βηματική συνάρτηση -1/1) (Εικόνα 15). Ο κόμβος άθροισης του νευρωνικού μοντέλου υπολογίζει ένα γραμμικό συνδυασμό των εισόδων x_1, x_2,…, x_n που εφαρμόζονται στις συνάψεις του, και ενσωματώνει επίσης μια εξωτερικά εφαρμοζόμενη πολώση b(bias):

$$u = \sum_{i=1}^{n} w_i x_i + b = \sum_{i=0}^{n} w_i x_i \qquad (3.15)$$

όπου w_0=b και x_0=1.



Εικόνα 15 : Μοντέλο Perceptron [44].

Στην Εξίσωση (3.15) θεωρήσαμε την πόλωση b ως ένα επί πλέον συναπτικό βάρος (w_0) το οποίο πολλαπλασιάζεται με μια σταθερή είσοδο x_0=1. Κατά αυτό τον τρόπο αυξάνουμε τις διαστάσεις του διανύσματος εισόδου και του διανύσματος συναπτικών βαρών κατά 1. Έτσι τώρα οι διαστάσεις του επαυξημένου διανύσματος εισόδου

$$\mathbf{x} = [x_0, x_1, x_2, \ldots, x_n]^T$$

και του διανύσματος βαρών

$$\mathbf{w} = [w_0, w_1, w_2, \ldots, w_n]^T$$

είναι n+1 αντί για n.

Ο απότομος περιοριστής ή συνάρτηση ενεργοποίησης ***f***(.) τροφοδοτείται από το προκύπτον άθροισμα **u** και δίδει την έξοδο ***y=f(u)*** του νευρώνα. Η συνάρτηση ενεργοποίησης είναι μη γραμμική και ειδικά στο perceptron παίρνει μια από τις παρακάτω δύο μορφές:

$$f(u) = \begin{cases} 0, & u \leq 0 \\ 1, & u > 0 \end{cases} \tag{3.16}$$

$$f(u) = \begin{cases} -1, & u \leq 0 \\ 1, & u > 0 \end{cases} \tag{3.17}$$

Η έξοδος y είναι λοιπόν ένας δυαδικός αριθμός είτε με την κλασική μορφή (1/0) είτε με τη λεγόμενη διπολική μορφή (1/-1). Η επιλογή ωστόσο μεταξύ κλασικής και διπολικής μορφής είναι ήσσονος σημασίας. Η παράμετρος που ουσιαστικά ρυθμίζει την συμπεριφορά του νευρώνα είναι το διάνυσμα συναπτικών βαρών $\mathbf{w} = [w_0, w_1, w_2, \ldots, w_n]^T$.

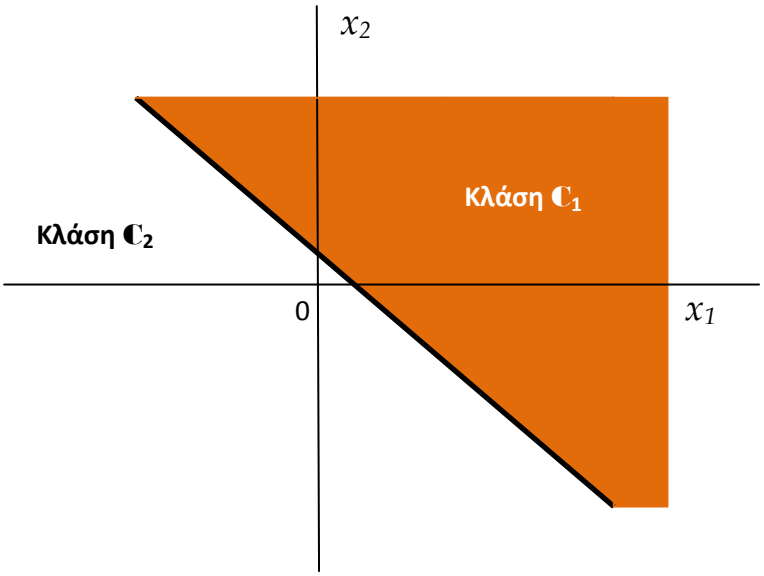
Το perceptron που βασίζεται σ' ένα μεμονωμένο νευρώνα περιορίζεται στην ταξινόμηση προτύπων που ανήκουν σε δύο μόνο κλάσεις. Στόχος του είναι να ταξινομήσει σωστά το σύνολο των εξωτερικά εφαρμοζόμενων διεγέρσεων $x_1, x_2, \ldots, x_n$ σε μια από τις δύο κλάσεις, $\mathbf{C}_1$ ή $\mathbf{C}_2$. Έστω, για παράδειγμα ότι χρησιμοποιείται η δυαδική συνάρτηση (1/0). Τότε *y*=1 αν u>0 και *y*=0 αν u≤0. Η εξίσωση

$$u = \sum_{i=1}^{n} w_i x_i + b = 0 \tag{3.18}$$

αντιστοιχεί σε ένα υπερεπίπεδο στο χώρο των n διαστάσεων (R^n) το οποίο διαχωρίζει τις δύο περιοχές απόφασης $\mathbf{C}_1$ και $\mathbf{C}_2$. Έτσι τα σημεία **x** που αντιστοιχούν σε θετικές τιμές u>0 βρίσκονται από την μια πλευρά του υπερεπιπέδου (έστω $\mathbf{C}_1$) ενώ τα σημεία **x** που αντιστοιχούν σε αρνητικές τιμές u<0 βρίσκονται στην απέναντι πλευρά του υπερεπιπέδου (έστω $\mathbf{C}_2$). Τα σημεία **x** που αντιστοιχούν σε u=0 βρίσκονται πάνω στο υπερεπίπεδο.

Την κατάσταση που προκύπτει μπορούμε να την οπτικοποιήσουμε (Εικόνα 16) καλύτερα στις δύο διαστάσεις. Στο χώρο R^2 η εξίσωση $u = w_1 x_1 + w_2 x_2 + b = 0$ ορίζει μια ευθεία γραμμή η οποία είναι κάθετη στο διάνυσμα των συναπτικών βαρών $\mathbf{w} = [w_1, w_2]^T$. Η ευθεία αυτή χωρίζει τα επίπεδο σε δύο τμήματα: (1) την κλάση $\mathbf{C}_1$ που βρίσκεται πάνω από την γραμμή και περιέχει τα σημεία **x** για τα οποία u>0, και (2) την κλάση $\mathbf{C}_2$ που βρίσκεται κάτω από την γραμμή και περιέχει τα σημεία **x** για τα οποία u<0. Θα πρέπει επίσης να

αναφέρουμε ότι η επίδραση της πόλωσης *b* έγκειται στο να μετατοπίζει απλώς το όριο απόφασης μακριά από το σημείο αρχής των αξόνων. Τα συναπτικά βάρη w_1, w_2,…, w_n του perceptron μπορούν να προσαρμόζονται μέσω μιας επαναληπτικής διαδικασίας. Για την προσαρμογή, μπορούμε να χρησιμοποιήσουμε ένα κανόνα διόρθωσης σφαλμάτων, γνωστό και ως κανόνα εκπαίδευσης perceptron.



Εικόνα 16: Το υπερεπίπεδο (σε αυτό το παράδειγμα μια ευθεία γραμμή) ως όριο απόφασης για ένα πρόβλημα ταξινόμησης προτύπων σε δύο κλάσεις [45].

Κανόνας Εκπαίδευσης Perceptron

Το ζητούμενο σε ένα νευρωνικό δίκτυο όπως το perceptron είναι η αυτόματη εκμάθηση των παραμέτρων του συστήματος ώστε να επιτυγχάνεται ο επιθυμητός στόχος. Το δίκτυο εκπαιδεύεται με επίβλεψη, δηλαδή υπάρχει ένας «δάσκαλος» που μας δίνει την τιμή του στόχου $d^{(P)}$ για κάθε πρότυπο εκπαίδευσης p. Το δίκτυο μαθαίνει προσαρμόζοντας τις παραμέτρους w_0, w_1, w_2,…, w_n λαμβάνοντας υπ' όψη του τα επαυξημένα πρότυπα εκπαίδευσης $\mathbf{x}^{(1)}$,…, $\mathbf{x}^{(P)}$ και τους στόχους $d^{(1)}$,…, $d^{(P)}$ των προτύπων αυτών χρησιμοποιώντας τον επαναληπτικό αλγόριθμο που περιγράφεται στο Πίνακα 2:

Πίνακας 2: Αλγόριθμος εκπαίδευσης δικτύου Perceptron [44]

Είσοδοι
$\mathbf{x}^{(1)},\ldots, \mathbf{x}^{(P)}$ = επαυξημένα διανύσματα εισόδων
$\mathbf{d}^{(1)},\ldots, \mathbf{d}^{(P)}$ = επιθυμητοί στόχοι
η= παράμετρος ρυθμού μάθησης, μια θετική σταθερά μικρότερη από την μονάδα

Έξοδος
$\mathbf{w}=[w_0, w_1, w_2,\ldots, w_n]$= τα εκπαιδευόμενα συναπτικά βάρη

Μέθοδος
1.Αρχικοποίηση:
Θέσε το επαυξημένο διάνυσμα βαρών **w**(0) σε τυχαίες τιμές
k=1
Για κάθε εποχή e=1,…, MAXEPOCHS{

2.Υπολογισμός της Πραγματικής Απόκρισης
Για κάθε πρότυπο p=1,…P{

$y= f(\mathbf{w}(k-1)^T \mathbf{x}^{(P)})$

3.Προσαρμογή του Διανύσματος Βαρών
Αν $y\neq d^{(P)}$ τότε εκπαίδευσε τα βάρη σύμφωνα με τον τύπο:
$\mathbf{w}(k)= \mathbf{w}(k-1) +\eta (d^{(P)} - y) \mathbf{x}^{(P)}$
αλλιώς άφησε τα βάρη όπως είναι.
k=k+1
}
4.Τερματισμός ή συνέχιση
Τερμάτισε αν δεν έγινε καμία αλλαγή βαρών στην εποχή αυτή,
ή αν έχει συμπληρωθεί ο μέγιστος αριθμός εποχών
}

Όπως φαίνεται και στο παραπάνω πίνακα, ο αλγόριθμος τροποποιεί το επαυξημένο διάνυσμα των συναπτικών βαρών **w** μόνο όταν υπάρχει σφάλμα ταξινόμησης, δηλαδή όταν ο στόχος $d^{(P)}$ για το πρότυπο p διαφέρει από την έξοδο *y* του δικτύου. Έτσι, όταν υπάρχει σφάλμα η διόρθωση των βαρών γίνεται προσθέτοντας ή αφαιρώντας ένα ποσοστό του προτύπου $\mathbf{x}^{(P)}$. Η παράμετρος $\boldsymbol{\eta}$ η οποία καλείται ρυθμός μάθησης ή εκπαίδευσης ρυθμίζει το μέγεθος της διόρθωσης αυτής. Αποδεικνύεται ότι η εκπαίδευση του **w** γίνεται με τέτοιο τρόπο ώστε το πρότυπο που ταξινομήθηκε τώρα εσφαλμένα, την επόμενη φορά είτε θα ταξινομηθεί σωστά είτε θα πλησιάζει περισσότερο στο να ταξινομηθεί σωστά.

Στο σημείο αυτό θα πρέπει να επισημάνουμε ότι το μοντέλο perceptron συγκλίνει σε μια λύση η οποία ταξινομεί σωστά όλα τα πρότυπα αρκεί να υπάρχει μια τέτοια λύση, δηλαδή αρκεί το πρόβλημα να είναι γραμμικά διαχωρίσιμο. Με άλλα λόγια θα πρέπει τα

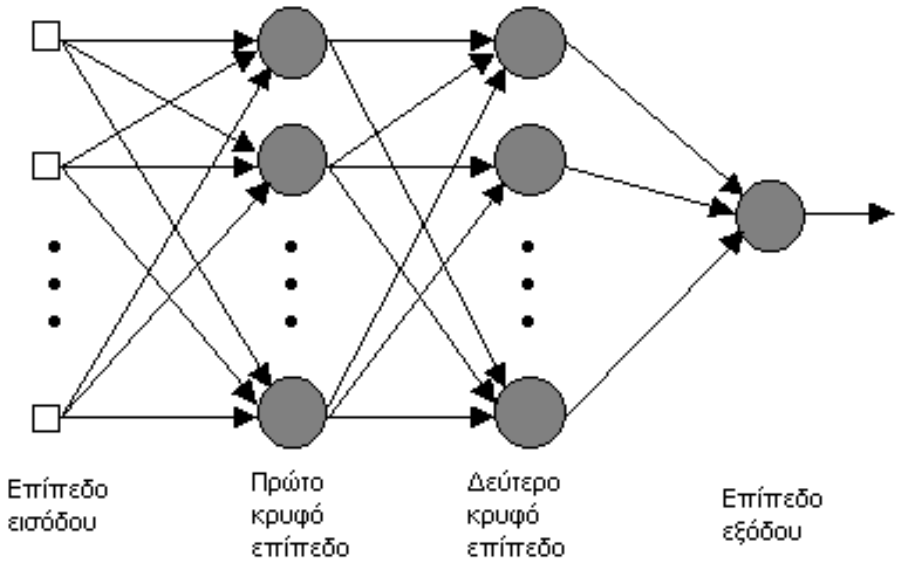
προς ταξινόμηση πρότυπα να είναι επαρκώς διαχωρισμένα το ένα από το άλλο για να μπορεί να σχεδιαστεί ένα υπερεπίπεδο (π.χ μια ευθεία όπως στην Εικόνα 16) ως όριο απόφασης. Σε διαφορετική περίπτωση (μη γραμμικά διαχωρίσιμου προβλήματος) το δίκτυο perceptron αδυνατεί να δώσει λύση.

3.3.4 Perceptron Πολλαπλών Επιπέδων (Multilayer Perceptron-MLP)

Στο προηγούμενο κεφάλαιο μελετήσαμε το perceptron του Rosenblatt, το οποίο είναι ουσιαστικά ένα νευρωνικό δίκτυο ενός επιπέδου. Εκεί δείξαμε ότι οι δυνατότητες αυτού του δικτύου περιορίζονται στην ταξινόμηση γραμμικά διαχωρίσιμων προτύπων. Το πρόβλημα αυτό μπορεί να λυθεί με την χρήση ενός νευρωνικού δικτύου που αποτελείται από περισσότερους νευρώνες. Δίκτυα τέτοιου τύπου καλούνται δίκτυα Perceptron πολλών επιπέδων [45]. Οι ακόλουθες παρατηρήσεις περιγράφουν τα βασικά χαρακτηριστικά των δικτύων perceptron πολλών επιπέδων:

- Το μοντέλο κάθε νευρώνα στο δίκτυο περιλαμβάνει μια μη γραμμική συνάρτηση ενεργοποίησης, η οποία είναι **διαφορίσιμη** .
- Το δίκτυο περιέχει ένα ή περισσότερα επίπεδα τα οποία παραμένουν **κρυφά** για τους κόμβους των επιπέδων εισόδου και εξόδου.
- Το δίκτυο επιδεικνύει μεγάλη **διασυνδεσιμότητα**, ο βαθμός της οποίας καθορίζεται από τα συναπτικά βάρη του δικτύου.

Στην Εικόνα 17 παρουσιάζεται ένα δίκτυο MLP με ένα επίπεδο εισόδου, δύο κρυφά επίπεδα και ένα επίπεδο εξόδου. Όπως παρατηρούμε το δίκτυο είναι πλήρως συνδεδεμένο. Αυτό σημαίνει ότι ένας νευρώνας σε οποιοδήποτε επίπεδο του δικτύου συνδέεται με όλους τους νευρώνες (κόμβους) του προηγούμενου επιπέδου. Η ροή σήματος διαμέσου του δικτύου προχωρά με κατεύθυνση προς τα εμπρός, από τα αριστερά προς τα δεξιά και επίπεδο προς επίπεδο.



Εικόνα 17: Δίκτυο perceptron πολλών επιπέδων με δύο κρυφά επίπεδα [50].

Εκπαίδευση Δικτύων MLP

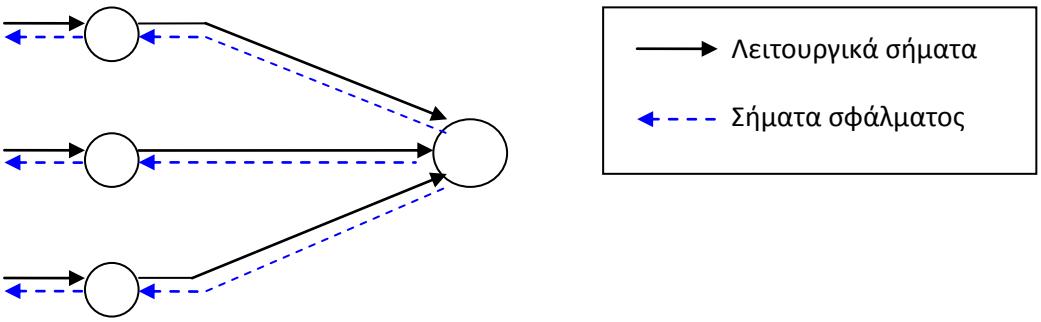
Η εκπαίδευση ενός δικτύου πολλών επιπέδων είναι η διαδικασία ρύθμισης των συναπτικών βαρών του έτσι ώστε να ικανοποιείται κάποιο κριτήριο καταλληλότητας. Αυτό που κάνει την εκπαίδευση ενός δικτύου MLP πολύ πιο ενδιαφέρουσα είναι το εξής: αν έχουμε το κατάλληλο σε μέγεθος δίκτυο μπορούμε να το εκπαιδεύσουμε να μάθει οποιαδήποτε συνάρτηση εμείς επιθυμούμε με οποιαδήποτε ποιότητα προσέγγισης εμείς επιθυμούμε. Αυτό αιτιολογεί και την δημοτικότητα των αλγορίθμων εκπαίδευσης των MLP με κυριότερο εκπρόσωπο τον αλγόριθμο Back – Propagation (Αλγόριθμος Οπισθοδιάδοσης Σφάλματος) [44], [45]. Βασικό χαρακτηριστικό της μεθόδου αυτής είναι η ύπαρξη στόχων, όπως ακριβώς και στο απλό δίκτυο Perceptron. Συνεπώς το μοντέλο αυτό ανήκει στην κατηγορία των δικτύων που εκπαιδεύονται με επίβλεψη.

Η εκπαίδευση με του MLP δικτύου σύμφωνα με τον αλγόριθμο Back Propagation λαμβάνει χώρα σε δύο φάσεις :

1. Στην φάση που εξελίσσεται προς τα εμπρός (forward propagation), όπου τα συναπτικά βάρη του δικτύου είναι σταθερά και το σήμα εισόδου (λειτουργικό

σήμα) διαδίδεται διαμέσου του δικτύου, επίπεδο προς επίπεδο, μέχρι να φτάσει στην έξοδο (Εικόνα 18 – μαύρα βελάκια).

2. Στην φάση που εξελίσσεται προς τα πίσω (back propagation), όπου παράγεται ένα σήμα σφάλματος (error signal) μέσω της σύγκρισης της εξόδου του δικτύου με μια επιθυμητή απόκριση. Το προκύπτον σήμα σφάλματος διαδίδεται διαμέσου του δικτύου, ξανά επίπεδο προς επίπεδο, αλλά αυτή τη φορά η διάδοση γίνεται με κατεύθυνση προς τα πίσω. Σε αυτή τη φάση γίνονται διαδοχικές προσαρμογές στα συναπτικά βάρη του δικτύου (Εικόνα 18 – μπλε βελάκια).



Εικόνα 18 :Οι κατευθύνσεις ροής των δύο βασικών σημάτων σε ένα MLP: διάδοση των λειτουργικών σημάτων προς τα εμπρός και διάδοση των σημάτων σφάλματος προς τα πίσω [45].

Προκειμένου να περιγράψουμε τους υπολογισμούς που λαμβάνουν χώρα κατά την εκπαίδευση του MLP, θα χρειαστεί αρχικά να καθορίσουμε το συμβολισμό των διάφορων μεγεθών. Συγκεκριμένα ορίζουμε:

- L το πλήθος των επιπέδων του δικτύου, εκτός από το επίπεδο εισόδου το οποίο θεωρείται το μηδενικό επίπεδο
- N(0), N(1), …, N(L) το πλήθος των νευρώνων των επιπέδων 0, 1, …, L. Άρα n= N(0) είναι το πλήθος των εισόδων και m= N(L) είναι το πλήθος των εξόδων
- $u_i(l)$ την συνολική είσοδο του νευρώνα i του επιπέδου l
- $y_i(l)$ την έξοδο του νευρώνα i του επιπέδου l
- $w_{ij}(l)$ το συναπτικό βάρος που συνδέει τον νευρώνα j του επιπέδου l-1 με τον νευρώνα i του επιπέδου l
- $w_{i0}(l)$ την πόλωση που εφαρμόζεται στο νευρώνα i του επιπέδου l
- $f()$ την συνάρτηση ενεργοποίησης των νευρώνων

- $x_i = y_i(0)$ τις εισόδους του δικτύου
- $o_i = y_i(L)$ τις εξόδους του δικτύου

Forward Propagation

Όπως αναφέραμε και παραπάνω στην πρώτη φάση εκπαίδευσης ενός δικτύου MLP, σύμφωνα με τον αλγόριθμο Back Propagation, υπολογίζουμε τις τιμές εξόδου για όλους τους νευρώνες του δικτύου με δεδομένες τις τιμές των εισόδων $x_1, x_2,..., x_{N(0)}= x_1, x_2,..., x_n$. Κατά συνέπεια, οι τιμές εξόδου για τα διάφορα επίπεδα του νευρωνικού δικτύου δίνονται από τις παρακάτω σχέσεις.

- Επίπεδο εισόδου:

Για λόγους ευκολίας καλούμε το επίπεδο εισόδου, καταχρηστικά μηδενικό επίπεδο. Συνεπώς οι έξοδοι του μηδενικού επιπέδου θα είναι ίσες με τις αντίστοιχες τιμές εισόδων

$$y_i(0)= x_i \qquad \text{με } i=1,\ldots,n \tag{3.19}$$

Το ίδιο ισχύει φυσικά και για την πόλωση b που εφαρμόζεται στο νευρωνικό δίκτυο και έχει τιμή εξόδου ίση με

$$y_0(0)= x_0=1 \tag{3.20}$$

- Κρυμμένα επίπεδα και επίπεδο εξόδου:

Για τα υπόλοιπα επίπεδα του νευρωνικού δικτύου έχουμε ότι η έξοδος του νευρώνα i του επιπέδου l δίνεται από την σχέση

$$y_i(l)= f(u_i(l)) \qquad \text{με } l=1,..,L,\ i=1,\ldots,N(l), \tag{3.21}$$

και

$$y_0(l)=1 \tag{3.22}$$

όπου:

$$u_i(l)= \sum_{j=1}^{N(l-1)} w_{ij}(l)\, y_j(l-1) + w_{i0}(l) = \sum_{j=0}^{N(l-1)} w_{ij}(l)\, y_j(l-1) \quad \text{με } l=1,..,L \text{ και } i=1,\ldots,N(l) \tag{3.23}$$

είναι η συνολική είσοδος στην συνάρτηση ενεργοποίησης του νευρώνα i και είναι ίση με το άθροισμα των σημάτων εξόδου των νευρώνων του προηγούμενου

στρώματος συνδυασμένων με τα συναπτικά βάρη $w_{ij}(l)$. Το συναπτικό βάρος $w_{i0}(l)$ αντιστοιχεί στην πόλωση του νευρώνα i.

- Έξοδος του δικτύου:

Για την έξοδο του δικτύου έχουμε

$$o_i = y_i(L) \qquad \text{με } i=1,\ldots,m \tag{3.24}$$

Back propagation

Έχοντας υπολογίσει την έξοδο του δικτύου (Εξ.(3.24)), εφαρμόζεται εν συνεχεία η οπισθοδιάδοση του σφάλματος (δεύτερη φάση εκπαίδευσης ενός δικτύου MLP). Σκοπός της εκπαίδευσης του νευρωνικού δικτύου είναι δοθέντος μιας σειράς από P διανύσματα εισόδων, οι έξοδοι να επιτύχουν τιμές που δίνονται από αντίστοιχα P διανύσματα στόχων. Πιο αναλυτικά, έστω

$\mathbf{x}^{(p)} = [x_1^{(p)}, \ldots, x_n^{(p)}]^T$ το p-οστό διάνυσμα εισόδου

$\mathbf{o}^{(p)} = [o_1^{(p)}, \ldots, o_m^{(p)}]^T$ το p-οστό διάνυσμα εξόδου

$\mathbf{d}^{(p)} = [d_1^{(p)}, \ldots, d_m^{(p)}]^T$ το p-οστό διάνυσμα στόχων

Τα δεδομένα που απαιτούνται για να εκπαιδευτεί το δίκτυο είναι τα P ζεύγη διανυσμάτων εισόδων- στόχων $\{\mathbf{x}^{(1)}, \mathbf{d}^{(1)}\}$, $\{\mathbf{x}^{(2)}, \mathbf{d}^{(2)}\}$, ..., $\{\mathbf{x}^{(P)}, \mathbf{d}^{(P)}\}$. Θα ήταν ιδανικό να έχουμε τέλεια ταύτιση εξόδων και στόχων για κάθε πρότυπο εισόδου, δηλαδή $\mathbf{o}^{(1)}= \mathbf{d}^{(1)}$, $\mathbf{o}^{(2)}= \mathbf{d}^{(2)}$, …, $\mathbf{o}^{(P)}= \mathbf{d}^{(P)}$. Ωστόσο αυτό μπορεί να μην είναι απολύτως εφικτό, οπότε επιζητούμε την βέλτιστη προσέγγιση της επιθυμητής κατάστασης χρησιμοποιώντας ένα κριτήριο κόστους. Το μέσο τετραγωνικό σφάλμα

$$J=\frac{1}{P}\sum_{p=1}^{P} ||\mathbf{d}^{(p)} - \mathbf{o}^{(p)}||^2 = \frac{1}{P}\sum_{p=1}^{P}\sum_{i=1}^{m} ||\mathbf{d_i}^{(p)} - \mathbf{o_i}^{(p)}||^2 \tag{3.25}$$

είναι ένα κλασικό κριτήριο κόστους που χρησιμοποιείται ευρέως σε πολλά προβλήματα. Έχει το πλεονέκτημα ότι η ελαχιστοποίησή του σημαίνει την ελαχιστοποίηση της τετραγωνικής απόστασης μεταξύ των διανυσμάτων $\mathbf{o}^{(i)}$, $\mathbf{d}^{(i)}$ και επιπλέον παραγωγίζεται εύκολα οπότε μπορεί να χρησιμοποιηθεί σε μεθόδους κατάβασης δυναμικού (gradient descent).

Τα συναπτικά βάρη w_{ij} είναι οι παράμετροι που πρέπει να διορθωθούν ώστε να ελαχιστοποιηθεί το σφάλμα J, καθώς τόσο οι είσοδοι $\mathbf{x}^{(p)}$ όσο και οι στόχοι $\mathbf{d}^{(p)}$ είναι δεδομένοι και σταθεροί. Σύμφωνα με την μέθοδο κατάβασης δυναμικού, η μεταβολή της παραμέτρου w_{ij} ως προς τον χρόνο t γίνεται χρησιμοποιώντας την παράγωγο του J ως προς w_{ij}.

$$\frac{dw_{\mathrm{ij}}}{dt} = -\frac{\partial J}{dw_{\mathrm{ij}}} \tag{3.26}$$

η οποία ονομάζεται κλίση του J ως προς w_{ij}. Προσθέτοντας και την έννοια του διακριτού χρόνου *k* στην Εξίσωση (3.26) έχουμε

$$w_{ij}(l,k+1) - w_{ij}(l,k) = -\eta \frac{\partial J}{\partial w_{ij}(l,k)} \tag{3.27}$$

όπου η είναι η παράμετρος ρυθμού μάθησης του αλγορίθμου Back propagation και είναι ένας μικρός θετικός αριθμός. Η χρήση του αρνητικού πρόσημου στις εξισώσεις (3.26) και (3.27) σηματοδοτεί βαθμωτή κατάβαση στον χώρο των βαρών (δηλαδή αναζήτηση μιας κατεύθυνσης για την μεταβολή των βαρών η οποία θα μειώνει την τιμή του J)

Για να διευκολυνθούμε στον υπολογισμό της παραγώγου $\partial J/(\partial w_{ij}\ (l,k))$ θα ονομάσουμε

$$\delta_i^{(k)}(l) = -\frac{\partial J}{\partial u_i^{(k)}(l)} \tag{3.28}$$

την παράγωγο του κόστους ως προς τη δικτυακή διέγερση του νευρώνα *i*. Χρησιμοποιώντας το συμβολισμό αυτό μπορούμε να γράψουμε

$$\frac{\partial J}{dw_{\mathrm{ij}}(l,k)} = \frac{\partial J}{\partial u_i^{(k)}(l)} \frac{\partial u_i^{(k)}(l)}{\partial w_{ij}(l,k)} = -\delta_i^{(k)}(l) \frac{\partial u_i^{(k)}(l)}{\partial w_{ij}(l,k)} = -\delta_i^{(k)}(l)\ y_j^{(k)}(l-1) \tag{3.29}$$

με l =1,.., L και j=1,…,*N(l-1)*

Με τη βοήθεια της εξίσωσης (3.29), η εξίσωση διόρθωσης του συναπτικού βάρους που συνδέει το νευρώνα *i* με το νευρώνα *j* (εξίσωση(3.27)) μπορεί να γραφεί ως

$$w_{ij}(l,k+1) = w_{ij}(l,k) + \eta \delta_i^{(k)}(l)\ y_j^{(k)}(l-1) \tag{3.30}$$

Το δεύτερο στάδιο εκπαίδευσης του MLP δικτύου ξεκινά στο επίπεδο εξόδου , στέλνοντας τα σήματα σφάλματος προς τα αριστερά, σε όλα τα επίπεδα του δικτύου, επίπεδο προς επίπεδο, και υπολογίζοντας αναδρομικά το δ (την τοπική κλίση) (εξίσωση 3.28) για κάθε νευρώνα. Αυτή η αναδρομική διαδικασία επιτρέπει στα συναπτικά βάρη του δικτύου να υφίστανται μεταβολές σύμφωνα με τον τύπο της εξίσωσης (3.30) .

Συνεπώς τα βήματα εκπαίδευσης ενός MLP δικτύου ακολουθούν τα παρακάτω βήματα:

- Υπολογισμός των σφαλμάτων στο επίπεδο εξόδου (επίπεδο L):

$$\delta_i^{(k)}(L) = -\frac{\partial J}{\partial u_i^{(k)}(L)} = -\frac{\partial J}{\partial y_i^{(k)}(L)}\frac{\partial y_i^{(k)}(L)}{\partial u_i^{(k)}(L)} = -\frac{\partial J}{\partial o_i^{(k)}}\frac{\partial f\left(u_i^{(k)}(L)\right)}{\partial u_i^{(k)}(L)} =$$

$$= f'(u_i^{(k)}(L))\left(d_i^{(k)} - o_i^{(k)}\right) \quad (3.31)$$

- Διάδοση των σφαλμάτων προς τα πίσω στο δίκτυο μέχρι τους νευρώνες του πρώτου κρυμμένου επιπέδου (επίπεδο l =1,…, L- 1):

$$\delta_i^{(k)}(l) = -\frac{\partial J}{\partial u_i^{(k)}(l)} = -\sum_{\mu=1}^{N(l+1)}\frac{\partial J}{\partial u_\mu^{(k)}(l+1)}\frac{\partial u_\mu^{(k)}(l+1)}{\partial y_i^{(k)}(l)}\frac{\partial y_i^{(k)}(l)}{\partial u_i^{(k)}(l)} =$$

$$= f'(u_i^{(k)}(l))\sum_{\mu=1}^{N(l+1)} w_{\mu i}^{(k)}(l+1)\delta_\mu^{(k)}(l+1) \quad (3.32)$$

- Ενημέρωση των βαρών όλων των επιπέδων (χρήση εξίσωσης (30))

$$w_{ij}(l, k+1) = w_{ij}(l,k) + \eta\delta_i^{(k)}(l)\ y_j^{(k)}(l-1) \quad (3.33)$$

Η συνάρτηση ενεργοποίησης

Ο υπολογισμός του σφάλματος *δ* για κάθε νευρώνα του perceptron πολλών επιπέδων απαιτεί γνώση της παραγώγου της συνάρτησης ενεργοποίησης *f()* που σχετίζεται με αυτό το νευρώνα. Για να υπάρχει αυτή η παράγωγος, πρέπει η συνάρτηση *f()* να είναι συνεχής. Με άλλα λόγια, η διαφορισιμότητα είναι η μόνη απαίτηση που πρέπει να ικανοποιεί μια συνάρτηση ενεργοποίησης. Ένα παράδειγμα συνεχώς διαφορίσιμης, μη γραμμικής

συνάρτησης ενεργοποίησης που χρησιμοποιείται ευρέως στα perceptron πολλών επιπέδων είναι η σιγμοειδής συνάρτηση:

$$f(u) = \frac{1}{1+e^{-u}} \tag{3.34}$$

Με την χρήση της συγκεκριμένης συνάρτησης οι εξισώσεις (3.31) και (3.32) για τον υπολογισμό του σφάλματος μπορούν να γραφούν ως

Επίπεδο L:

$$\delta_i^{(k)}(L) = f'\left(u_i^{(k)}(L)\right)\left(d_i^{(k)} - o_i^{(k)}\right) = f\left(u_i^{(k)}(L)\right)\left(1 - f\left(u_i^{(k)}(L)\right)\right)\left(d_i^{(k)} - o_i^{(k)}\right) =$$
$$= o_i^{(k)}\left(1 - o_i^{(k)}\right)\left(d_i^{(k)} - o_i^{(k)}\right) \tag{3.35}$$

Επίπεδο l =1,…, L- 1:

$$\delta_i^{(k)}(l) = f'\left(u_i^{(k)}(l)\right)\sum_{\mu=1}^{N(l+1)} w_{\mu i}^{(k)}(l+1)\delta_\mu^{(k)}(l+1) =$$

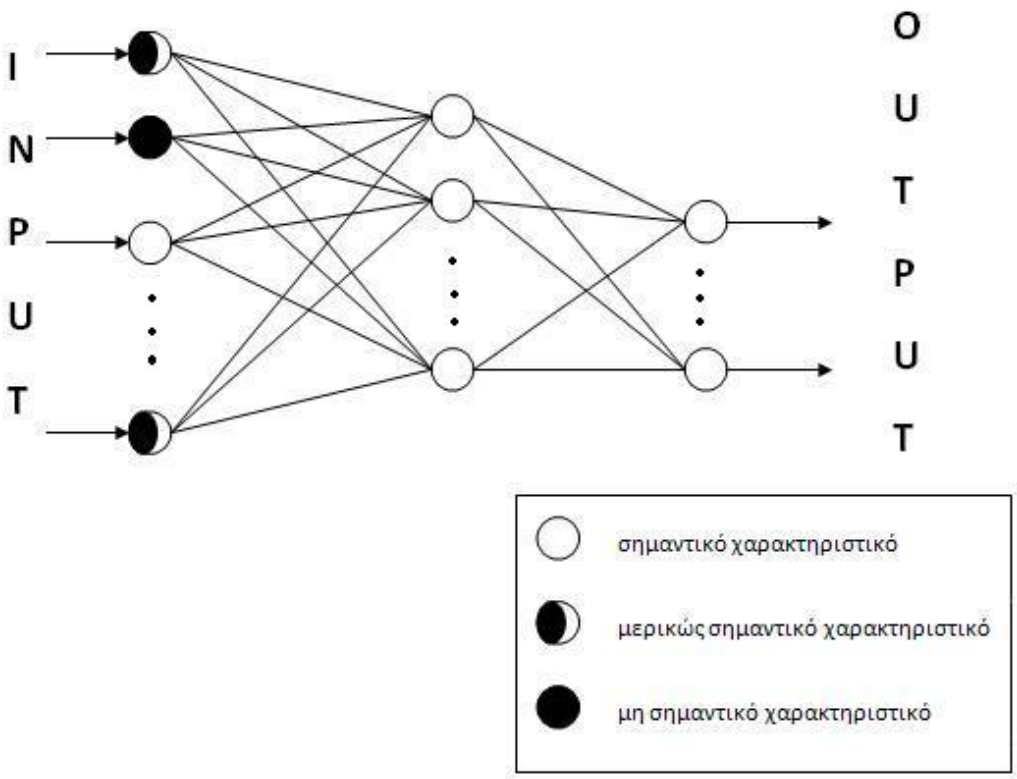
$$= f\left(u_i^{(k)}(l)\right)\left(1 - f\left(u_i^{(k)}(l)\right)\right)\sum_{\mu=1}^{N(l+1)} w_{\mu i}^{(k)}(l+1)\delta_\mu^{(k)}(l+1) =$$

$$= \left(y_i^{(k)}(l)\right)\left(1 - y_i^{(k)}(l)\right)\sum_{\mu=1}^{N(l+1)} w_{\mu i}^{(k)}(l+1)\delta_\mu^{(k)}(l+1) \tag{3.36}$$

3.3.5 Το Feature Selection Multilayer Perceptron (FSMLP)

Στην παρούσα διπλωματική εργασία χρησιμοποιήσαμε ένα νευρωνικό δίκτυο το οποίο προτάθηκε από τους Pal και Chintalapudi [51] και διαθέτει την αρχιτεκτονική ενός δικτύου Perceptron πολλών επιπέδων (MLP) με κάποιες βασικές διαφοροποίησεις. Το συγκεκριμένο μοντέλο βασίζεται στην δυναμική επιλογή ενός κατάλληλου (βέλτιστου) υποσυνόλου χαρακτηριστικών από ένα σύνολο δεδομένων. Η επιλογή αυτή πραγματοποιείται κατά τη διάρκεια εκπαίδευσης του νευρωνικού δικτύου με σκοπό την ταξινόμηση των χαρακτηριστικών που διαθέτει το σύνολο δεδομένων.

Σε ένα FSMLP δίκτυο (Εικόνα 19) κάθε κόμβος εισόδου συνδέεται με ένα μηχανισμό ο οποίος καθορίζει το βαθμό συμμετοχής του σήματος εισόδου στην εκπαίδευση του δικτύου. Αν το χαρακτηριστικό, λοιπόν, που εισέρχεται θεωρείται σημαντικό, τότε ο μηχανισμός διαδίδει την τιμή του χαρακτηριστικού σχεδόν ανέπαφη στο δίκτυο. Σε αντίθετη περίπτωση, ο μηχανισμός δεν επιτρέπει την είσοδο της τιμής του χαρακτηριστικού σε υψηλότερα επίπεδα του δικτύου. Στο τέλος, επιλέγονται τα χαρακτηριστικά εκείνα που παρουσίασαν το μεγαλύτερο βαθμό συμμετοχής σε όλη την διάρκεια εκπαίδευσης του δικτύου.



Εικόνα 19: Το FSMLP νευρωνικό δίκτυο [51].

Ο συγκεκριμένος μηχανισμός μπορεί να υλοποιηθεί ως μια συνάρτηση εξασθένισης (attenuation function) η οποία πολλαπλασιάζεται με κάθε κόμβο στο επίπεδο εισόδου του δικτύου και παράγει υψηλές τιμές για τα σημαντικά χαρακτηριστικά και σχεδόν μηδενικές τιμές για τα μη σημαντικά. Η συνάρτηση εξασθένισης F θα πρέπει να είναι συνεχής στο διάστημα [0,1] έτσι ώστε να μπορεί να προσδιορίζει με ακρίβεια το βαθμό συμμετοχής κάθε χαρακτηριστικού. Επιπλέον θα πρέπει να διαθέτει και μια παράμετρο m η οποία θα εκπαιδεύεται παράλληλα με τα συναπτικά βάρη του δικτύου σύμφωνα με τη μέθοδο κατάβασης δυναμικού (gradient descent method).

Η σιγμοειδής συνάρτηση $F(m) = \frac{1}{1+e^{-m}}$ μπορεί να χρησιμοποιηθεί ως συνάρτηση εξασθένισης αφού είναι μια συνεχής και παραγωγίσιμη συνάρτηση στο διάστημα [0,1] .

Όπως αναφέραμε και στην αρχή, το δίκτυο FSMLP διατηρεί τα χαρακτηριστικά ενός MLP δικτύου. Συνεπώς η εκπαίδευση του θα γίνεται σύμφωνα με τον αλγόριθμο Back – Propagation.

Έστω λοιπόν

- F: η συνάρτηση εξασθένισης
- F'_i : η παράγωγος της συνάρτησης εξασθένισης που σχετίζεται με τον i –οστό κόμβο εισόδου
- m_i : η παράμετρος της συνάρτησης εξασθένισης που σχετίζεται με τον i –οστό κόμβο εισόδου
- μ :ο ρυθμός μάθησης (learning rate) της παραμέτρου εξασθένισης (attenuator)
- w_{ji}^1 :το συναπτικό βάρος που συνδέει τον j –οστό κόμβο του πρώτου κρυφού επιπέδου τον i –οστό κόμβο του επιπέδου εισόδου
- η : ο ρυθμός μάθησης των συναπτικών βαρών
- q : ο αριθμός των κόμβων του πρώτου κρυφού επιπέδου
- δ_j^1:το σφάλμα του j –οστού κόμβου του πρώτου κρυφού επιπέδου

Forward Propagation

Στην πρώτη φάση εκπαίδευσης του δικτύου FSMLP έχουμε :

- Επίπεδο εισόδου:

  Οι έξοδοι του μηδενικού επιπέδου θα είναι ίσες με το γινόμενο της εισόδου επί την συνάρτηση εξασθένισης

$$y_i^0 = x_i F(m_i) \tag{3.37}$$

- Για τα υπόλοιπα επίπεδα του δικτύου:

  Οι τιμές των εξόδων των υπόλοιπων επιπέδων του δικτύου υπολογίζονται σύμφωνα με τις εξισώσεις (3.21)-(3.24).

Back propagation

Στην δεύτερη φάση εκπαίδευσης του δικτύου FSMLP έχουμε:

- Για τα κρυφά επίπεδα η διάδοση των σφαλμάτων πραγματοποιείται σύμφωνα με τις εξισώσεις (3.31),(3.32).

- Μέχρι και το δεύτερο κρυφό επίπεδο η ανανέωση των συναπτικών βαρών γίνεται σύμφωνα με την εξίσωση (3.33)

- Για το πρώτο κρυφό επίπεδο ισχύει:

  Ανανέωση συναπτικών βαρών

$$w_{ji,new}^1 = w_{ji,old}^1 + \eta \delta_j^1(l) y_i^0 = w_{ji,old}^1 + \eta \delta_j^1(l) x_i F(m_i) \tag{3.38}$$

- Για το επίπεδο εισόδου ισχύει:

  Ανανέωση παραμέτρων εξασθένισης

$$m_{i,new} = m_{i,old} + \mu x_i F'(m_i) \sum_{j=1}^{q} w_{ji,new}^1 \delta_j^1 \tag{3.39}$$

Από την σχέση (3.37) παρατηρούμε ότι όταν η τιμή της συνάρτησης εξασθένισης $F(m_i)$ είναι κοντά στο μηδέν τότε η i-οστή τιμή εξόδου $x_i F(m_i)$ του επιπέδου εισόδου (μηδενικού επιπέδου) θα πλησιάζει και αυτή το μηδέν με αποτέλεσμα το i-οστό χαρακτηριστικό να μην εισέρχεται στο δίκτυο. Σε αντίθετη περίπτωση, όταν η τιμή της συνάρτησης $F(m_i)$ πλησιάσει την μονάδα η i-οστή τιμή εξόδου $x_i F(m_i)$ θα είναι περίπου ίση με x_i με αποτέλεσμα το i-οστό χαρακτηριστικό να εισέρχεται σχεδόν ανεπηρέαστο στο δίκτυο. Η εκπαίδευση του δικτύου ξεκινά με την συνάρτηση εξασθένισης να είναι μηδέν για όλα τα χαρακτηριστικά. Κατά συνέπεια, στην αρχή της εκπαίδευσης κανένα από τα χαρακτηριστικά δεν περνάει στο νευρωνικό. Καθώς η εκπαίδευση προχωρά, μόνο τα σημαντικά χαρακτηριστικά ενεργοποιούνται. Αυτό συμβαίνει με την αύξηση της παραμέτρου εξασθένισης m_i όπως υπαγορεύεται από τη μέθοδο κατάβασης δυναμικού (εξίσωση(3.39)). Με άλλα λόγια τα χαρακτηριστικά εκείνα που ελαχιστοποιούν ταχύτερα το σφάλμα δ εισέρχονται και ταχύτερα στο δίκτυο. Η εκπαίδευση του δικτύου σταματά όταν το σφάλμα φτάσει σε μια χαμηλή τιμή (μηδέν ή κοντά στο μηδέν) ή όταν ολοκληρωθεί ένας συγκεκριμένος αριθμός επαναλήψεων. Στο τέλος της εκπαίδευσης επιλέγουμε τα χαρακτηριστικά εκείνα με τις υψηλότερες τιμές της παραμέτρου εξασθένισης.

3.4 Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines- SVM)

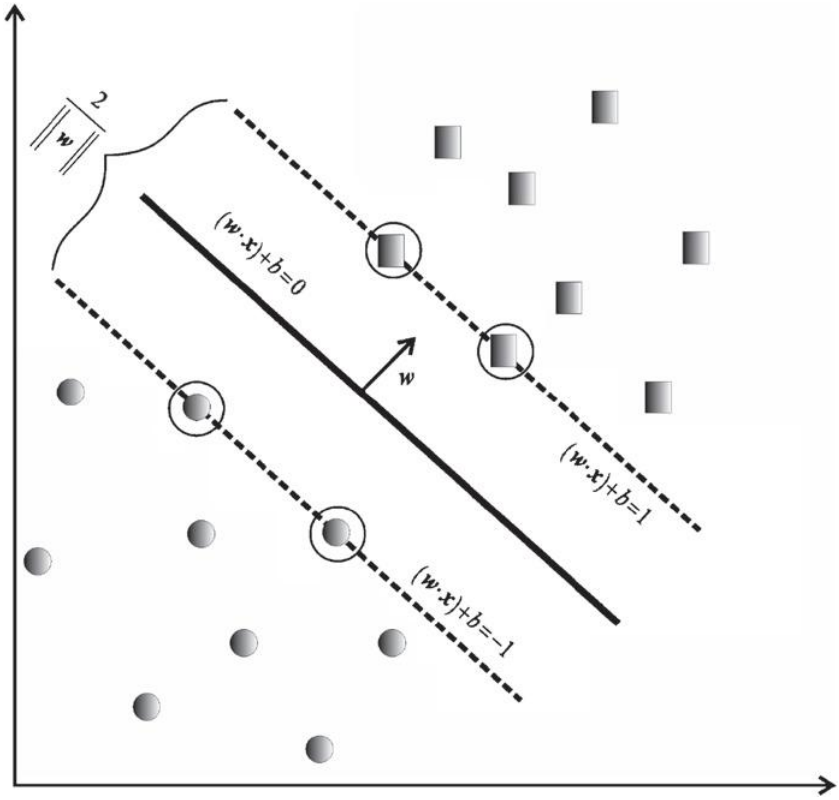
Οι μηχανές διανυσμάτων υποστήριξης (SVMs), οι οποίες παρουσιάστηκαν το 1992 από τους Vapnik, Boser και Guyon [52] , χαρακτηρίζονται ως μοντέλα επιβλεπόμενης μάθησης τα οποία μπορούν να παρέχουν εξαιρετική απόδοση και να επιδεικνύουν εξαιρετική ευρωστία όσον αναφορά την επίλυση προβλημάτων ταξινόμησης προτύπων και προβλημάτων παλινδρόμησης. Η κεντρική ιδέα στην οποία βασίζεται κάθε SVM είναι:

Δοθέντος ενός δείγματος εκπαίδευσης, η μηχανή διανυσμάτων υποστήριξης κατασκευάζει ένα υπερεπίπεδο ως επιφάνεια απόφασης με τρόπο ώστε το περιθώριο διαχωρισμού μεταξύ θετικών και αρνητικών παραδειγμάτων να μεγιστοποιείται [45].

3.4.1 Γραμμικό SVM

Διαχωρίσιμη Περίπτωση

Αρχικά θα παρουσιάσουμε την περίπτωση των γραμμικών μηχανών (linear SVM) [36],[53] που εκπαιδεύονται με διαχωρίσιμα δεδομένα. Θεωρούμε το σύνολο δεδομένων εκπαίδευσης X = {$\mathbf{x}_i$, y_i }, i=1, …, n, y_i ∈ {-1, 1}, $\mathbf{x}_i \in \boldsymbol{R}^d$, όπου $\mathbf{x}_i$ τα είναι τα διανύσματα χαρακτηριστικών του συνόλου εκπαίδευσης. Κάθε ένα από αυτά ανήκει σε κάποια από τις δύο κλάσεις y_i =+1, y_i =-1, οι οποίες είναι γραμμικά διαχωρίσιμες. Ο αλγόριθμος διανυσμάτων υποστήριξης προσπαθεί να βρει το καλύτερο διαχωριστικό υπερεπίπεδο , το οποίο να διαχωρίζει τα δείγματα εκπαίδευσης $\mathbf{x}_i$ που ανήκουν στην θετική κλάση (+1) από εκείνα που ανήκουν στην αρνητική κλάση (-1). Αυτό επιτυγχάνεται με την μεγιστοποίηση της απόστασης $\frac{2}{||w||}$ μεταξύ των δύο παράλληλων γραμμών $H_1: (\boldsymbol{w} \cdot \boldsymbol{x}) + b = 1$ και $H_2: (\boldsymbol{w} \cdot \boldsymbol{x}) + b = -1$, οι οποίες σχηματίζουν το περιθώριο (margin) διαχωρισμού των δύο κλάσεων όπως φαίνεται στην Εικόνα 20.



Εικόνα 20: Επεξήγηση του προβλήματος της δυαδικής ταξινόμησης, που δείχνει το διαχωριστικό περιθώριο μεταξύ των δύο κλάσεων. Τα σημεία που είναι κυκλωμένα πάνω στις διακεκομμένες γραμμές αντιπροσωπεύουν τα support vectors [36].

Το τελικό διαχωριστικό υπερεπίπεδο περνά από το μέσο του περιθωρίου με εξίσωση $(\boldsymbol{w} \cdot \boldsymbol{x}) + b = 0$. Στο σημείο αυτό πρέπει να σημειώσουμε ότι κανένα σημείο εκπαίδευσης δεν πρέπει να βρίσκεται ανάμεσά στα παράλληλα επίπεδα H_1 και H_2. Έτσι όλα τα δεδομένα εκπαίδευσης θα πρέπει να ικανοποιούν τους παρακάτω περιορισμούς:

$$(\boldsymbol{w} \cdot \boldsymbol{x_i}) + b \geq +1\ ,\ \forall \boldsymbol{x} \in \theta\varepsilon\tau\iota\kappa\acute{\eta}\ \kappa\lambda\acute{\alpha}\sigma\eta\ (y_i = +1) \qquad (3.40)$$

$$(\boldsymbol{w} \cdot \boldsymbol{x_i}) + b \leq -1\ ,\ \forall \boldsymbol{x} \in \alpha\rho\nu\eta\tau\iota\kappa\acute{\eta}\ \kappa\lambda\acute{\alpha}\sigma\eta\ (y_i = -1) \qquad (3.41)$$

Αυτές οι δύο σχέσεις μπορούν να συνδυαστούν σε ένα σύνολο ανισοτήτων:

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x_i}) + b \geq 1,\ \forall \boldsymbol{i} \qquad (3.42)$$

Το SVM πρόβλημα μπορεί ισοδύναμα να διατυπωθεί ως: Υπολόγισε τις παραμέτρους $\boldsymbol{w}$, b του υπερεπιπέδου έτσι ώστε:

$$\text{να ελαχιστοποιείται η συνάρτηση } J(\boldsymbol{w}, b) \equiv \frac{1}{2}\|w\|^2 \tag{3.43}$$

$$\text{υπό τους περιορισμούς } y_i(\boldsymbol{w} \cdot \boldsymbol{x_i} + b) \geq 1, \quad i = 1, \dots, n \tag{3.44}$$

Προφανώς η ελαχιστοποίηση της νόρμας οδηγεί σε μεγιστοποίηση του περιθωρίου. Σύμφωνα με τη θεωρεία της δυικότητας (duality theory), το πρόβλημα μπορεί να μετατραπεί στο ακόλουθο πρόβλημα μεγιστοποίησης, όπου το λ αντιπροσωπεύει το διάνυσμα των πολλαπλασιαστών Lagrange και το y_i αναπαριστά την ετικέτα (label) του i-οστού δείγματος:

$$\text{μεγιστοποίηση της ποσότητας} \quad \sum_{i=1}^{n} \lambda_i - \frac{1}{2}\sum_{i,j=1}^{n} \lambda_i \lambda_j y_i y_j \left(\boldsymbol{x}_i \cdot \boldsymbol{x}_j\right), \quad \lambda \epsilon R^n \tag{3.45}$$

$$\text{υπό τους περιορισμούς} \quad \begin{cases} \sum_{i=1}^{n} \lambda_i y_i = 0 \\ \lambda_i \geq 0, \quad i = 1, \dots, n \end{cases} \tag{3.46}$$

Έχοντας υπολογίσει τους βέλτιστους πολλαπλασιαστές Lagrange λ_i , καταλήγουμε στην ακόλουθη έκφραση για το βέλτιστο διάνυσμα κατεύθυνσης **w**:

$$\boldsymbol{w} = \sum_{i=1}^{n} \lambda_i y_i \boldsymbol{x}_i \tag{3.47}$$

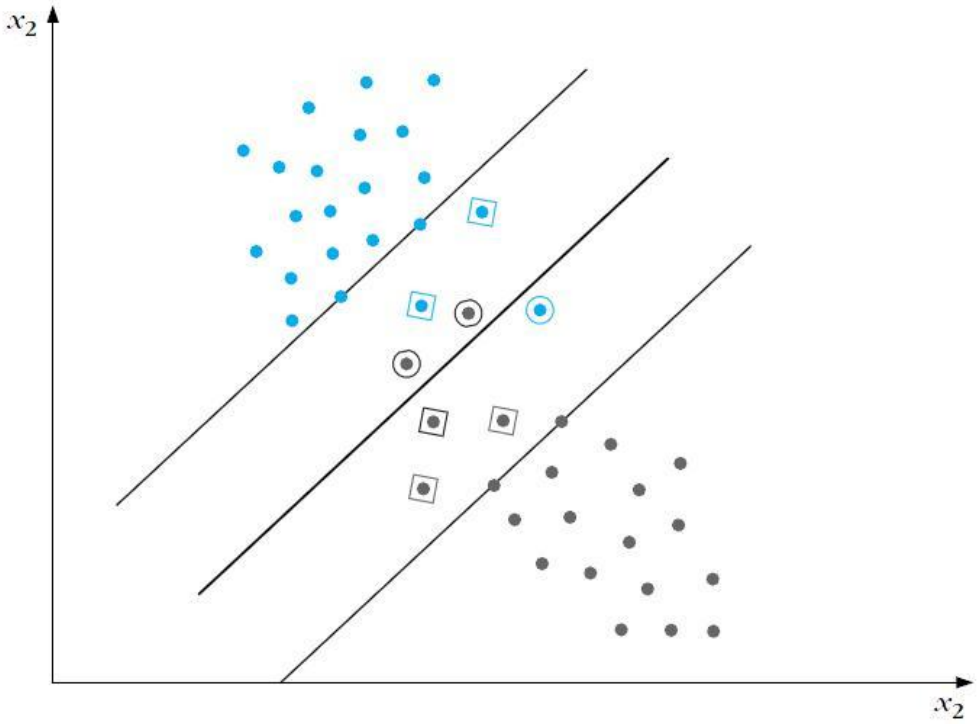
όπου n είναι ο αριθμός των διανυσμάτων υποστήριξης για τα οποία οι πολλαπλασιαστές Lagrange λ_i είναι όλοι μη μηδενικοί. Αποδεικνύεται ότι τα διανύσματα υποστήριξης βρίσκονται στα όρια των κλάσεων (όπως δείχνει η Εικόνα 20) και μπορούν να χρησιμοποιηθούν για τον υπολογισμό του b αντικαθιστώντας ένα από τα support vectors στην ακόλουθη εξίσωση:

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x_i} + b) = 1 \tag{3.48}$$

Μη Διαχωρίσιμη Περίπτωση

Η μέθοδος SVM που παρουσιάσαμε παραπάνω είναι κατάλληλη για επίλυση γραμμικά διαχωρίσιμων προβλημάτων. Βασίζεται στην υπόθεση ότι το πρόβλημα είναι γραμμικά διαχωρίσιμο και με αυτό τον τρόπο ορίζονται τα διαχωριστικό υπερεπίπεδο και τα διανύσματα υποστήριξης. Πολλά προβλήματα ωστόσο δεν είναι διαχωρίσιμα.

Στην Εικόνα 21 απεικονίζεται η περίπτωση στην οποία δύο κλάσεις δεν είναι διαχωρίσιμες [53], αφού όπως παρατηρούμε κάποια πρότυπα της μιας κλάσης πέφτουν μέσα στην περιοχή της άλλης. Κατά συνέπεια, κάθε προσπάθεια να σχεδιαστεί ένα υπερεπίπεδο δεν θα καταλήξει ποτέ σε μια ζώνη διαχωρισμού κλάσεων χωρίς σημεία στο εσωτερικό της, όπως συνέβαινε στην περίπτωση των γραμμικά διαχωρίσιμων κλάσεων. Όπως παρατηρούμε και από την Εικόνα 21 τα διανύσματα χαρακτηριστικών εκπαίδευσης τώρα ανήκουν σε μία από τις ακόλουθες τρεις κατηγορίες:



Εικόνα 21: Στην περίπτωση των μη διαχωρίσιμων κλάσεων ορισμένα σημεία βρίσκονται μέσα στο περιθώριο διαχωρισμού των δύο κλάσεων [53].

- Διανύσματα που βρίσκονται εκτός της ζώνης και είναι σωστά ταξινομημένα.

- Διανύσματα που βρίσκονται εντός της ζώνης και είναι σωστά ταξινομημένα. Αυτά είναι τα σημεία που έχουν τοποθετηθεί σε τετράγωνα στην Εικόνα 21 και ικανοποιούν την ανισότητα

$$0 \leq y_i(\boldsymbol{w} \cdot \boldsymbol{x_i} + b) < 1 \tag{3.49}$$

- Διανύσματα που ταξινομούνται εσφαλμένα. Αυτά περικλείονται από κύκλους και ικανοποιούν την ανισότητα

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x_i} + b) < 0 \tag{3.50}$$

Και οι τρεις περιπτώσεις μπορούν να αντιμετωπιστούν εισάγοντας ένα νέο σύνολο μεταβλητών, και συγκεκριμένα

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x_i} + b) \geq 1 - \xi_i,\ \xi_i \geq 0, \quad i = 1, \dots, n \tag{3.51}$$

Η πρώτη κατηγορία δεδομένων αντιστοιχεί σε $\xi_i = 0$, η δεύτερη σε $0 < \xi_i \leq 1$, και η τρίτη σε $\xi_i > 1$. Οι μεταβλητές ξ_i είναι γνωστές ως μεταβλητές χαλαρότητας (slack variables).

Το πρόβλημα βελτιστοποίησης στην περίπτωση μη διαχωρίσιμων κλάσεων γίνεται πιο περίπλοκο. Ο στόχος τώρα είναι να κάνουμε το περιθώριο διαχωρισμού όσο το δυνατόν μεγαλύτερο αλλά παράλληλα να κρατήσουμε το πλήθος των σημείων για τα οποία ισχύει $\xi > 0$, όσο το δυνατόν πιο μικρό. Οπότε η συνάρτηση κόστους θα πάρει την εξής μορφή:

$$\text{ελαχιστοποίηση της συνάρτησης } J(\boldsymbol{w}, b, \xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \tag{3.52}$$

$$\text{υπό τους περιορισμούς } y_i(\boldsymbol{w} \cdot \boldsymbol{x_i} + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, i = 1, \dots, n \tag{3.53}$$

Η ελαχιστοποίηση του πρώτου όρου της Εξ.(3.52) σχετίζεται με τη μηχανή διανυσμάτων υποστήριξης. Ο δεύτερος όρος $\sum_{i=1}^{n}\xi_i$ αποτελεί ένα άνω φράγμα στον αριθμό των σφαλμάτων ελέγχου. Η παράμετρος C ελέγχει τον συμβιβασμό μεταξύ της πολυπλοκότητας της μηχανής και του αριθμού των μη διαχωρίσιμων σημείων.

Όπως και στην περίπτωση των γραμμικά διαχωρίσιμων κλάσεων, χρησιμοποιώντας την μέθοδο των πολλαπλασιαστών Lagrange μπορούμε να διατυπώσουμε το παραπάνω πρόβλημα (Εξ.((3.52)) ως εξής:

$$\text{μεγιστοποίηση της ποσότητας } \sum_{i=1}^{n}\lambda_i - \frac{1}{2}\sum_{i,j=1}^{n}\lambda_i\lambda_j y_i y_j \left(\boldsymbol{x}_i \cdot \boldsymbol{x}_j\right), \quad \lambda \in R^n \tag{3.54}$$

$$\text{υπό τους περιορισμούς } \begin{cases} \sum_{i=1}^{n}\lambda_i y_i = 0 \\ 0 \leq \lambda_i \leq C, \quad i = 1, \dots, n \end{cases} \tag{3.55}$$

Παρατηρούμε ότι οι μεταβλητές χαλαρότητας ξ_i δεν εμφανίζονται στην Εξ.(3.54), με αποτέλεσμα η συνάρτηση που πρέπει να μεγιστοποιηθεί να είναι ίδια τόσο για την περίπτωση των μη-διαχωρίσιμων όσο και για την περίπτωση των γραμμικά διαχωρίσιμων προτύπων. Η μόνη διαφορά βρίσκεται στο ότι ο περιορισμός $\lambda_i \geq 0$ της Εξ.(3.46) αντικαθίσταται από τον πιο αυστηρό περιορισμό $0 \leq \lambda_i \leq C$ στην μη διαχωρίσιμη περίπτωση. Τέλος οι υπολογισμοί των βέλτιστων τιμών του διανύσματος βαρών $\boldsymbol{w}$ και της πόλωσης b διεξάγονται με τον ίδιο τρόπο όπως ορίζεται από τις Εξ.(3.47) και Εξ.(3.48) αντίστοιχα.

Ταξινόμηση Δεδομένων

Μετά την εκπαίδευση του γραμμικού SVM, κάθε νέο δείγμα ελέγχου x_{test} καθορίζεται σε ποια πλευρά του διαχωριστικού υπερεπιπέδου ανήκει και ταξινομείται στην αντίστοιχη κλάση σύμφωνα με την παρακάτω συνάρτηση απόφασης:

$$f(\boldsymbol{x_{test}}) = sgn\big((\boldsymbol{w} \cdot \boldsymbol{x_{test}}) + b\big) \tag{3.56}$$

όπου το **w** αναπαριστά το διάνυσμα κατεύθυνσης του υπερεπιπέδου. Το πρόσημο της τιμής που επιστρέφεται από την Εξίσωση (3.56) δείχνει την προβλεπόμενη κλάση που σχετίζεται με το δείγμα $\boldsymbol{x_{test}}$, ενώ το $|f(\boldsymbol{x_{test}})|$ δείχνει το βαθμό εμπιστοσύνης στην απόφαση που προέκυψε.

Χρησιμοποιώντας την Εξ.(3.47) για τον υπολογισμό των βέλτιστων τιμών του διανύσματος βαρών $\boldsymbol{w}$, η Εξ.(3.56) μετατρέπεται ως εξής:

$$f(\boldsymbol{x_{test}}) = sgn\big((\textstyle\sum_{i=1}^{n} \lambda_i y_i \boldsymbol{x}_i \cdot \boldsymbol{x_{test}}) + b\big) \tag{3.57}$$

3.4.2 Χρήση Συναρτήσεων Πυρήνα

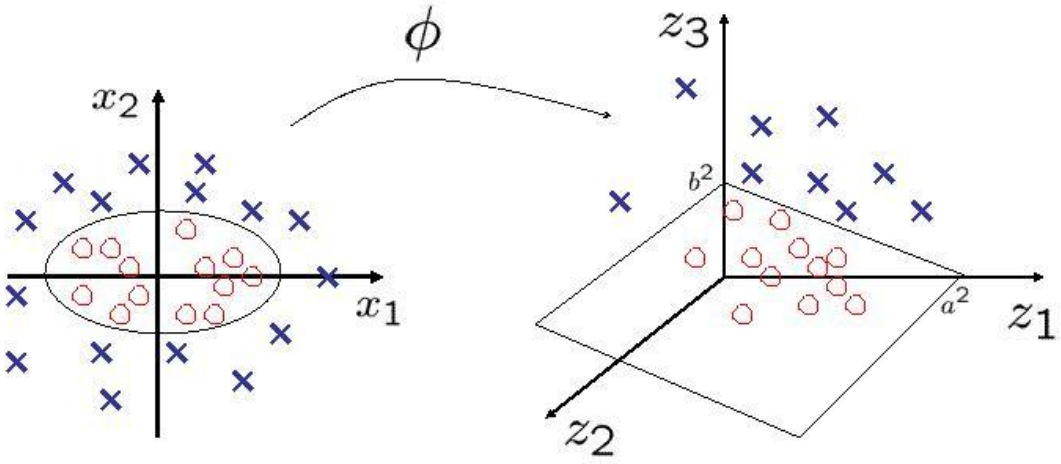
Πολλές φορές τα δεδομένα που έχουμε στην διάθεση μας δεν μπορούν να διαχωριστούν από μία γραμμική επιφάνεια (Εικόνα 22). Το πρόβλημα αυτό μπορεί να λυθεί χάρη σε μια κομψή ιδιότητα στην μεθοδολογία των SVM [53]. Η ιδιότητα αυτή απεικονίζει τα χαρακτηριστικά του χώρου εισόδου $\boldsymbol{x}$ σε ένα ανώτερης διάστασης χώρο χαρακτηριστικών H, όπου τα δεδομένα είναι γραμμικά διαχωρίσιμα, χρησιμοποιώντας μια συνάρτηση $\Phi(\boldsymbol{x})$.

$$\boldsymbol{x} \rightarrow \boldsymbol{\Phi}(\boldsymbol{x}) \in H \tag{3.58}$$

Συνεπώς το πρόβλημα βελτιστοποίησης στην Εξ.(3.54) μετασχηματίζεται ως εξής:

μεγιστοποίηση της ποσότητας $\sum_{i=1}^{n} \lambda_i - \frac{1}{2}\sum_{i,j=1}^{n} \lambda_i \lambda_j y_i y_j \, \Phi(\boldsymbol{x}_i) \cdot \Phi\big(\boldsymbol{x}_j\big), \quad \lambda \in R^n$ (3.59)

υπό τους περιορισμούς $\begin{cases} \sum_{i=1}^{n} \lambda_i y_i = 0 \\ 0 \le \lambda_i \le C, \quad i = 1, \dots, n \end{cases}$ (3.60)



Εικόνα 22: Παράδειγμα μη γραμμικού SVM ταξινομητή για την περίπτωση δύο μη γραμμικά διαχωρίσιμων κλάσεων. Τα δεδομένα εισόδου(αριστερό σχήμα) απεικονίζονται με την βοήθεια της Φ σε έναν υψηλότερης διάστασης χώρο χαρακτηριστικών (δεξί σχήμα) [54].

Το εσωτερικό γινόμενο $\Phi(\boldsymbol{x}_i) \cdot \Phi(\boldsymbol{x}_j)$ καλείται συνάρτηση πυρήνα (*kernel function*) και συμβολίζεται ως $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \Phi(\boldsymbol{x}_i) \cdot \Phi(\boldsymbol{x}_j)$. Έτσι εφαρμόζοντας κάποιον πυρήνα διαφορετικό του γραμμικού η Εξ.(3.59) μετασχηματίζεται στην

μεγιστοποίηση της ποσότητας $\sum_{i=1}^{n} \lambda_i - \frac{1}{2}\sum_{i,j=1}^{n} \lambda_i \lambda_j y_i y_j \, k(\boldsymbol{x}_i, \boldsymbol{x}_j), \quad \lambda \in R^n$ (3.61)

δεδομένου ότι $\begin{cases} \sum_{i=1}^{n} \lambda_i y_i = 0 \\ 0 \le \lambda_i \le C, \quad i = 1, \dots, n \end{cases}$ (3.62)

Εκτός από τον γραμμικό πυρήνα , κάποια τυπικά παραδείγματα πυρήνων που χρησιμοποιούνται στις εφαρμογές αναγνώρισης προτύπων είναι τα ακόλουθα:

Πολυώνυμα

$$k(\boldsymbol{x}, \boldsymbol{y}) = (1 + (\boldsymbol{x} \cdot \boldsymbol{y}))^d \tag{3.63}$$

Συναρτήσεις Ακτινωτής Βάσης (RBF)

$$k(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\gamma ||\boldsymbol{x} - \boldsymbol{y}||^2\right) \tag{3.64}$$

3.5 Αναδρομική Εξάλειψη Χαρακτηριστικών (Recursive Feature Elimination-RFE)

Το πρόβλημα στην επιλογή δεικτών (marker selection) στην ανάλυση μικροσυστοιχιών DNA αντιμετωπίζεται με δύο βασικούς τύπους μεθόδων, τις filter και wrapper μεθόδους. Οι wrapper μέθοδοι λειτουργούν με έναν αναδρομικό τρόπο όπου τα βάρη των χαρακτηριστικών (γονιδίων) επανεκτιμώνται και αλλάζουν δυναμικά από επανάληψη σε επανάληψη. Αντίθετα στις filter μεθόδους τα βάρη παραμένουν σταθερά.

Η μέθοδος RFE-SVM [37] αποτελεί μια wrapper μέθοδο επιλογής γονιδίων η οποία χρησιμοποίει ένα γραμμικό SVM ταξινομητή για την εκτίμηση του διανύσματος βαρών των χαρακτηριστικών. Από την άλλη μεριά, η μέθοδος RFE-LNW [36], που έχουμε ακολουθήσει, αποτελεί μια παραλλαγή της μεθόδου RFE-SVM, εφαρμόζοντας filter κριτήρια (π.χ. μετρική Fisher) με έναν επαναληπτικό (wrapper) τρόπο, όπου τα βάρη ρυθμίζονται από κύκλο σε κύκλο αν αυτό κριθεί απαραίτητο. Η ικανότητα γενίκευσης του αλγορίθμου RFE-LNW σε ανεξάρτητα σύνολα ελέγχου παρουσιάζει αξιοσημείωτα επίπεδα απόδοσης, γεγονός που χαρακτηρίζει τον συγκεκριμένο αλγόριθμο ως ένα χρήσιμο εργαλείο για την επιλογή δεικτών.

3.5.1 Η μέθοδος RFE-SVM

Η RFE-SVM μέθοδος [37], [55] βασίζεται στα SVMs και στην ιδέα της κατάταξης των χαρακτηριστικών (γονιδίων) σύμφωνα με την απόλυτη τιμή των συνιστωσών του διανύσματος κατεύθυνσης $\boldsymbol{w}$ (Εξίσωση (3.47)). Όπως φαίνεται στην Εξίσωση (3.56), κάθε συνιστώσα του $\boldsymbol{w}$ συσχετίζεται με μια συνιστώσα του διανύσματος $\boldsymbol{x_{test}}$, η οποία περιλαμβάνει το επίπεδο έκφρασης ενός χαρακτηριστικού. Με αυτό τον τρόπο, κάθε χαρακτηριστικό (γονίδιο) πολλαπλασιάζεται με ένα βάρος. Όσο μεγαλύτερη είναι η απόλυτη τιμή του βάρους ενός χαρακτηριστικού, τόσο σημαντικότερο είναι αυτό το χαρακτηριστικό σύμφωνα με το αλγόριθμο RFE-SVM, με την έννοια ότι συνεισφέρει περισσότερο στη συνάρτηση απόφασης της Εξίσωσης (3.56). Κατά συνέπεια η κατάταξη των γονιδίων μπορεί να πραγματοποιηθεί σύμφωνα με την απόλυτη τιμή των συνιστωσών του $\boldsymbol{w}$. Σε κάθε βήμα της αναδρομικής διαδικασίας εξάλειψης της μεθόδου RFE-SVM διαγράφεται το λιγότερο σημαντικό χαρακτηριστικό. Λιγότερο σημαντικό θεωρείται το χαρακτηριστικό του οποίου η διαγραφή θα προκαλούσε τη μικρότερη μείωση του

διαχωριστικού περιθωρίου (βλέπε Εικόνα 20). Ο αλγόριθμος ξεκινά εκπαιδεύοντας τον γραμμικό ταξινομητή SVM με όλα τα χαρακτηριστικά. Από την εκπαίδευση υπολογίζεται το διάνυσμα βαρών $\boldsymbol{w}$. Το χαρακτηριστικό που αντιστοιχεί στην ελάχιστη, κατά απόλυτη τιμή, συνιστώσα του διανύσματος $\boldsymbol{w}$ διαγράφεται και ο SVM ταξινομητής εκπαιδεύεται λαμβάνοντας υπόψη τα εναπομείναντα χαρακτηριστικά. Με τον ίδιο τρόπο, η παραπάνω διαδικασία επαναλαμβάνεται μέχρις ότου απομείνει ένας προκαθορισμένος αριθμός χαρακτηριστικών. Ο αλγόριθμος της μεθόδου RFE-SVM παρουσιάζεται συνοπτικά στον Πίνακα 3 που ακολουθεί.

Πίνακας 3 :Η αλγοριθμική παρουσίαση της μεθόδου RFE-SVM [55].

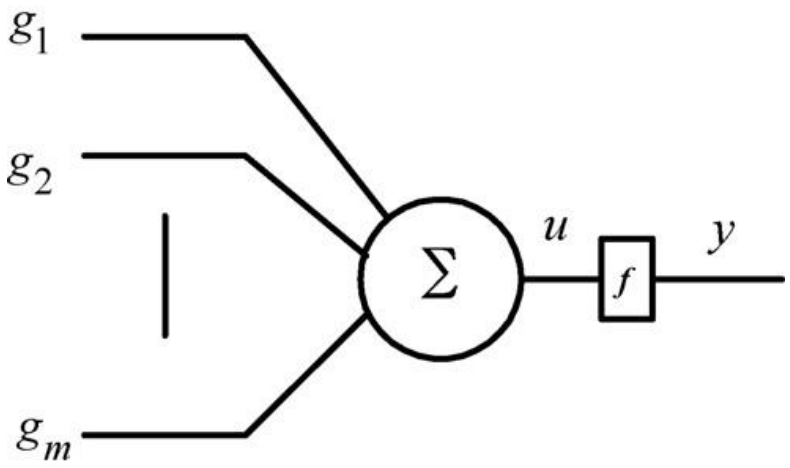
Έστω m ο αρχικός αριθμός των χαρακτηριστικών (γονιδίων)
Όσο (m ≥0)
Υπολόγισε το διάνυσμα κατεύθυνσης $\boldsymbol{w}$ του διαχωριστικού υπερεπιπέδου χρησιμοποιώντας γραμμικό SVM.
Κατάταξε τα χαρακτηριστικά σύμφωνα με τις συνιστώσες του $\boldsymbol{\lvert w \rvert}$**.**
Αφαίρεσε το χαρακτηριστικό με την μικρότερη απόλυτη τιμή βάρους (m←m-1). Περισσότερα από ένα χαρακτηριστικά μπορούν να διαγραφούν σε κάθε επανάληψη.
Εκτίμησε την ακρίβεια ταξινόμησης των m επιζώντων χαρακτηριστικών χρησιμοποιώντας ένα γραμμικό SVM ταξινομητή.
Τέλος Όσο
Βγάλε στην έξοδο ως γονίδια-δείκτες (marker genes) το σύνολο των επιζώντων χαρακτηριστικών που επιτυγχάνουν την μέγιστη ακρίβεια ταξινόμησης.

3.5.2 Η μέθοδος RFE-LNW

Οι περισσότερες τεχνικές επιλογής δεικτών (marker selection) που εφαρμόζονται στο πεδίο των DNA μικροσυστοιχιών, λόγω των πολλών διαστάσεων των δεδομένων, χρησιμοποιούν γραμμικά εργαλεία για να αξιολογήσουν το πρόβλημα. Το RFE-SVM [37], που αναλύθηκε στη προηγούμενη ενότητα, είναι μία τέτοια μέθοδος, όπου ένας γραμμικός πυρήνας χρησιμοποιείται για να εκτιμήσει το διάνυσμα βαρών του διαχωριστικού υπερεπιπέδου, η απόλυτη τιμή του οποίου χρησιμοποιείται στην συνέχεια ως κριτήριο κατάταξης (ranking) των γονιδίων.

Από την άλλη πλευρά, λόγω του σχεδιασμού του (ένας γραμμικός συνδυασμός των εισόδων), ένας γραμμικός νευρώνας (linear neuron-LN) μπορεί να προσεγγίσει οποιαδήποτε γραμμική συνάρτηση. Κατά συνέπεια, μπορούμε να χρησιμοποιήσουμε ένα γραμμικό νευρώνα για την προσέγγιση του διαχωριστικού επιπέδου μεταξύ των θετικών και αρνητικών κλάσεων. Μια τέτοια ανοιχτή αρχιτεκτονική μας δίνει την δυνατότητα να επιλέξουμε ανάμεσα από μια πληθώρα σχημάτων εκμάθησης ή ακόμα να ενσωματώσουμε μια νέα διαδικασία μάθησης κατάλληλα προσαρμοσμένη στο πρόβλημά μας.

Κάτι τέτοιο επιτυγχάνεται με την χρήση ενός δικτύου που αποτελείται από ένα νευρώνα (single neuron network) με m εισόδους (Εικόνα 23), όπου το m αντιστοιχεί στον αριθμό των γονιδίων. Έχοντας δύο πιθανά αποτελέσματα στο επίπεδο εξόδου, την τιμή εξόδου 0 για την αρνητική κλάση και την τιμή 1 για την θετική κλάση, μπορούμε να χρησιμοποιήσουμε ένα τέτοιο γραμμικό νευρώνα για να προσεγγίσουμε το διαχωριστικό υπερεπίπεδο το οποίο διακρίνει τις δύο κλάσεις ενδιαφέροντος.



Εικόνα 23: Ένας μόνο νευρώνας προσαρμοσμένος στο πρόβλημα επιλογής δεικτών [36].

Πιο συγκεκριμένα, χρησιμοποιώντας τη σιγμοειδή συνάρτηση $f(u)$ έχουμε ότι:

$$y = \frac{1}{1 + e^{u}} = f(u) \tag{3.65}$$

$$u = \sum_{i=1}^{m} w_i \, g_i \tag{3.66}$$

$$f'(u) = y(1 - y) \tag{3.67}$$

Όπου $f'(u) \geq 0$ αφού το y κυμαίνεται μεταξύ 0 και 1.

3.5.3 Διαφορικά εκφρασμένα Γονίδια

Η βασική ιδέα πίσω από την ανάπτυξη της μεθόδου [36] RFE-LNW, που χρησιμοποιήσαμε, για την αναδρομική εξάλειψη χαρακτηριστικών (γονιδίων) είναι η αναγνώριση και τελικά η επιλογή των διαφορικά εκφρασμένων γονιδίων. Η ιδέα αυτή δεν είναι καινούρια στο πεδίο επιλογής δεικτών. Σε πολλές μελέτες έχουν χρησιμοποιηθεί παραλλαγές του συντελεστή Fisher ο οποίος δίνεται από την ακόλουθη εξίσωση:

$$f_1(g_i) = \frac{(\mu_+(g_i) - \mu_-(g_i))^2}{\sigma_+(g_i)^2 + \sigma_-(g_i)^2} \tag{3.68}$$

Μια παραλλαγή του συντελεστή Fisher μπορεί να εκφραστεί ως:

$$f_2(g_i) = \frac{\sum_{j=1}^{n} |g_{ij} - c(g_i)|}{\sigma_+(g_i) + \sigma_-(g_i)} \tag{3.69}$$

όπου $\mu_+(g_i)$, $\mu_-(g_i)$, $\sigma_+(g_i)$, και $\sigma_-(g_i)$ είναι οι μέσες τιμές και οι τυπικές αποκλίσεις των εκφράσεων του γονιδίου g_i στην θετική και αρνητική κλάση αντίστοιχα, n ο αριθμός των δειγμάτων και

$$c(g_i) = \frac{(\mu_+(g_i) - \mu_-(g_i))}{2} \tag{3.70}$$

Μία άλλη παραλλαγή με χαμηλότερο υπολογιστικό κόστος δίνεται από την εξίσωση

$$f_3(g_i) = \frac{|\mu_+(g_i) - \mu_-(g_i)|}{\sigma_+(g_i) + \sigma_-(g_i)} \tag{3.71}$$

Κάποιος μπορεί εύκολα να διαπιστώσει πως οι εξισώσεις (3.68), (3.69), (3.71) εκφράζουν στην ουσία την ίδια ιδέα. Όταν χρησιμοποιούνται οι εξισώσεις αυτές για την εκχώρηση βαρών σε ένα σύνολο δοθέντων γονιδίων, είναι προφανές ότι στα γονίδια των οποίων η έκφραση διαφοροποιείται περισσότερο στις δύο καταστάσεις (κλάσεις στις οποίες χωρίζονται τα δεδομένα) εκχωρούνται και υψηλότερες τιμές βαρών σε σχέση με τα γονίδια εκείνα που διαφοροποιούνται λιγότερο ανάμεσα στις δύο κλάσεις. Τέλος, για τα γονίδια τα οποία εκφράζονται με ακριβώς τον ίδιο τρόπο και στις δύο καταστάσεις (έχουν την ίδια έκφραση τόσο στην παθολογική όσο και στην φυσιολογική κατάσταση) εκχωρείται το μικρότερο δυνατό βάρος που είναι ίσο με μηδέν.

3.5.4 Εκπαίδευση του RFE-LNW

Στην ενότητα αυτή παραθέτουμε το βασικό μαθηματικό υπόβαθρο της διαδικασίας μάθησης που χρησιμοποιείται για την ανανέωση των βαρών του γραμμικού νευρώνα. Η συνάρτηση σφάλματος (error function) ενός νευρώνα που πρέπει να ελαχιστοποιηθεί δίνεται από την σχέση:

$$E = \frac{1}{2}\sum_{j=1}^{n}(d_j - y_j)^2 \tag{3.72}$$

όπου το n αντιστοιχεί στον αριθμό των δειγμάτων, το d_j αναπαριστά την επιθυμητή έξοδο του νευρώνα που σχετίζεται με το j-οστό δείγμα και y_j είναι η πραγματική έξοδος που παράγεται από το νευρώνα για το συγκεκριμένο δείγμα. Με την βοήθεια της μεθόδου κατάβασης δυναμικού (gradient descent method), για την ελαχιστοποίηση της εξίσωσης (3.72), ενημερώνουμε το βάρος w_i που σχετίζεται με το γονίδιο g_i όπως φαίνεται παρακάτω:

$$w_i(t+1) = w_i(t) - \left(\mu\frac{\partial E}{\partial w_i}\right) = w_i(t) - \mu\sum_{j=1}^{n}\left(\frac{\partial E}{\partial y_j}\frac{\partial y_j}{\partial u}\frac{\partial u}{\partial w_i}\right) \tag{3.73}$$

Με τη χρήση της μετρικής του Fisher (Εξ.(3.69)) η εξίσωση ανανέωσης βαρών (Εξ.(3.73)) γίνεται:

$$w_i(t+1) = w_i(t) - \frac{\mu}{2}\sum_{j=1}^{n}\left[\left(\frac{\partial E}{\partial y_j}\frac{\partial y_j}{\partial u}\frac{\partial u}{\partial w_i}\right)\frac{\left|g_{ij} - c(g_i)\right|}{\sigma_+(g_i) + \sigma_-(g_i)}\right]$$

$$= w_i(t) - \frac{\mu}{2}\sum_{j=1}^{n}\left(-2\left(d_j - y_j\right)y_j\left(1 - y_j\right)g_{ij}\right)f_2(g_i)$$

$$= w_i(t) + \mu\sum_{j=1}^{n}\left(d_j - y_j\right)y_j\left(1 - y_j\right)g_{ij}f_2(g_i)$$

$$= w_i(t) + \mu\sum_{j=1}^{n}\left(d_j - y_j\right)f'\left(u_j\right)g_{ij}f_2(g_i)$$

Τελικά έχουμε,

$$w_i(t+1) = w_i(t) + \mu\sum_{j=1}^{n} e_j f'\left(u_j\right)g_{ij}\frac{|g_{ij} - c(g_i)|}{\sigma_+(g_i) + \sigma_-(g_i)} \tag{3.74}$$

όπου το t αναπαριστά την τρέχουσα επανάληψη, μ είναι ο βαθμός εκπαίδευσης και

$$e_j = \left(d_j - y_j\right) \tag{3.75}$$

Δουλεύοντας με πρόσημα , η οποία είναι μια ιδέα που παρουσιάζεται στο ελαστικό (resilient) back propagation learning, η Εξ.(3.74) μετατρέπεται σε:

Στην περίπτωση όπου $f_2(g_i) = 1$, το οποίο είναι παρόμοιο με τη πρότυπη back propagation διαδικασία, έχουμε:

$$w_i(t+1) = w_i(t) + \mu\sum_{j=1}^{n} e_j f'\left(u_j\right)g_{ij} \tag{3.76}$$

$$w_i(t+1) = w_i(t) + \mu\sum_{j=1}^{n} sign\left(e_j f'\left(u_j\right)\right)sign\left(g_{ij}\right) \tag{3.77}$$

ή γενικά έχουμε:

$$w_i(t+1) = w_i(t) + \mu\sum_{j=1}^{n} sign\left(e_j f'\left(u_j\right)\right) \text{x } sign\left(g_{ij}\right)f_2(g_i) \tag{3.78}$$

$$w_i(t+1) = w_i(t) + \mu\sum_{j=1}^{n} \left|d_j - y_j\right| sign\left(e_j f'\left(u_j\right)\right) \text{x } sign\left(g_{ij}\right)f_2(g_i) \tag{3.79}$$

Η εξίσωση (3.76) είναι ο βασικός gradient descent αλγόριθμος εκπαίδευσης, ο οποίος ελαχιστοποιεί τη συνάρτηση σφάλματος (Εξ. (3.72)). Επίσης, η εξίσωση (3.77) συγκλίνει,

αφού διατηρώντας το πρόσημο του gradient πλησιάζουμε προς την κατεύθυνση του ελαχίστου, στο οποίο τελικά θα φτάσουμε (εκτός από τις περιπτώσεις όπου παγιδευόμαστε σε ένα τοπικό ελάχιστο) χρησιμοποιώντας το κατάλληλο ρυθμό εκπαίδευσης μ. Στην πραγματικότητα περιμένουμε η Εξ.(3.77) να συγκλίνει ταχύτερα από την Εξ.(3.76), αφού τα e_j και $f'(u_j)$ στην Εξ.(3.76) μπορούν να πάρουν πολύ μικρές τιμές με αποτέλεσμα μικρές τροποποιήσεις στα βάρη w. Αυτό με την σειρά του συνεπάγεται χαμηλές μεταβολές στη συνάρτηση σφάλματος και επιβράδυνση της σύγκλισης. Χρησιμοποιώντας μόνο το πρόσημο του gradient στην Εξ.(3.77) πλησιάζουμε προς τη κατεύθυνση του ελαχίστου κάνοντας «μεγαλύτερα βήματα» και επιταχύνοντας τη σύγκλιση τουλάχιστον όταν η διαδικασία είναι μακριά από το ελάχιστο. Η εξέλιξη αυτή είναι πολύ χρήσιμη στην εφαρμογή της κατά τη διαδικασία επιλογής δεικτών, αφού αναγκάζει τον αλγόριθμο να συγκλίνει πολύ γρήγορα, ειδικότερα στα πρώτα βήματα της διαδικασίας όπου ο αριθμός των χαρακτηριστικών (γονιδίων) είναι εξαιρετικά μεγάλος.

Η Εξ.(3.78) διαφέρει από την (3.77) μόνο στο συντελεστή $f_2(\cdot)$. Η Εξ.(3.78) μπορεί να αποδειχθεί ότι συγκλίνει σε ένα ελάχιστο, αλλά επιπλέον μέσω του όρου άθροισης, λαμβάνει υπόψη της και τελικά υπολογίζει μία προσέγγιση της μετρικής του Fisher. Ωστόσο, αρκετές φορές οι λίγες διαστάσεις των δεδομένων μπορεί να καθυστερήσουν τη σύγκλιση. Καθώς η διαδικασία προχωρά και το πρόβλημα των διαστάσεων μειώνεται σημαντικά, ο λόγος των δειγμάτων (ο αριθμός των οποίων παραμένει σταθερός κατά τη διαδικασία εξάλειψης των χαρακτηριστικών) προς τα χαρακτηριστικά αυξάνεται και το πρόβλημα της εκτίμησης του διαχωριστικού υπερεπιπέδου γίνεται πιο δύσκολο, καθυστερώντας τη σύγκλιση. Σε αυτή τη περίπτωση η αύξηση του αριθμού των επαναλήψεων (εποχών) ή του ρυθμού μάθησης κρίνεται απαραίτητη. Η Εξ.(3.79) μπορεί να χρησιμοποιηθεί στα τελευταία αυτά βήματα για να επιταχύνει τη σύγκλιση διαθέτοντας ένα παραλλαγμένο ρυθμό μάθησης $|d_j - y_j|$. Εύκολα μπορεί να παρατηρήσει κάποιος πως όσο είμαστε μακριά από τον επιθυμητό στόχο, η ποσότητα $|d_j - y_j|$ θα παίρνει μια μεγάλη τιμή επιταχύνοντας τη σύγκλιση, ενώ αντίθετα όσο πλησιάζουμε το στόχο, το $|d_j - y_j|$ θα αρχίζει να παίρνει χαμηλότερες τιμές, επιβραδύνοντας έτσι τη σύγκλιση. Με άλλα λόγια, στα τελευταία βήματα της διαδικασίας επιλογής των χαρακτηριστικών προχωράμε με γρήγορο ρυθμό προς το στόχο όταν βρισκόμαστε μακριά από αυτόν, αλλά καθυστερούμε όταν τον πλησιάσουμε για να δημιουργήσουμε με μεγαλύτερη ακρίβεια το διαχωριστικό υπερεπίπεδο. Οι κανόνες που χρησιμοποιούμε τελικά για την ανανέωση των βαρών είναι η Εξ.(3.78) σε συνδυασμό με την Εξ.(3.79). Οι Εξ.(3.76) και (3.77) χρησιμοποιήθηκαν κυρίως για σκοπούς αιτιολόγησης.

Κλείνοντας την ενότητα για την εκπαίδευση του RFE-LNW, θα θέλαμε να επισημάνουμε ότι η εκπαίδευση ενός μόνο νευρώνα με μια κατάλληλη διαδικασία μάθησης, μπορεί τελικά να εφαρμόσει ένα filter κριτήριο, όπως το Fisher's ratio, με έναν wrapper τρόπο.

3.5.5 Επιλογή των διαφορικά εκφρασμένων γονιδίων

Η επιλογή των διαφορικά εκφρασμένων γονιδίων είναι ένας επιθυμητός στόχος σε οποιαδήποτε προσέγγιση επιλογής δεικτών. Έτσι για την μέθοδο RFE-LNW, πρέπει να δείξουμε ότι ανάμεσα στον εξαιρετικά μεγάλο αριθμό γονιδίων που την τροφοδοτεί, θα επιλεχθούν τελικά δείκτες που εκφράζονται με διαφορετικό τρόπο ανάμεσα στις δύο καταστάσεις (κλάσεις) ενδιαφέροντος.

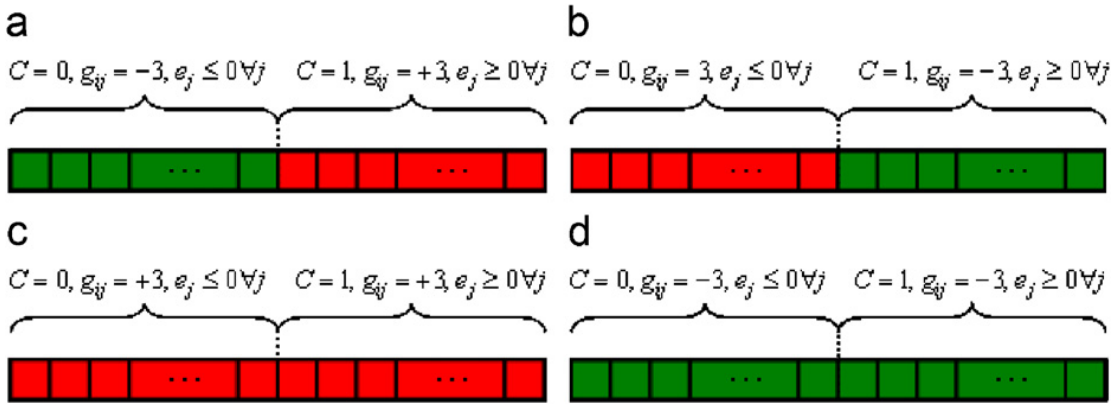
Η Εικόνα 24 δείχνει το επίπεδο έκφρασης ενός υποθετικού γονιδίου g_i στην αρνητική ($C = 0$) και θετική ($C = 1$) κλάση, αντίστοιχα. Στις περιπτώσεις (a) και (b) το υποθετικό γονίδιο εκφράζεται διαφορετικά στις δύο κλάσεις ενδιαφέροντος, με πράσινο χρώμα (αρνητικές τιμές) για την αρνητική κλάση και με κόκκινο (θετικές τιμές) στη θετική κλάση ή αντίστροφα. Από την άλλη μεριά, οι περιπτώσεις (c) και (d) δε δείχνουν καμία διαφοροποίηση στο επίπεδο έκφρασης του συγκεκριμένου γονιδίου στις δύο καταστάσεις ενδιαφέροντος.

Ας πάρουμε, για παράδειγμα, την περίπτωση (a) σε συνδυασμό με την Εξ.(3.77) και ας επικεντρωθούμε στην αρνητική κλάση (πράσινο τμήμα). Παρατηρούμε ότι ο όρος $sign\left(e_j \cdot f'(u_j)\right) \cdot sign(g_{ij}) \geq 0$ διατηρείται. Πράγματι, $e_j \leq 0$ (αφού $d_j = 0$ από την Εξ.(3.75) και $y_j \in [0 \dots 1]$), $f'(u_j)$ από την Εξ.(3.67) είναι θετικό και $g_{ij} = -3$. Τώρα ας επικεντρωθούμε στη θετική κλάση (κόκκινο τμήμα) της Εικόνας 24(a). Χρησιμοποιώντας την ίδια λογική, παρατηρούμε ξανά ότι $sign\left(e_j \cdot f'(u_j)\right) \cdot sign(g_{ij}) \geq 0$. Αφού ο όρος e_j στην πραγματικότητα είναι συχνά μη μηδενικός, ο όρος άθροισης της Εξ.(3.77) παράγει ένα θετικό αποτέλεσμα.

Ακολουθώντας περίπου την ίδια λογική στην περίπτωση (b) της Εικόνας 24 , μπορούμε να καταλάβουμε ότι η Εξ.(3.77) παράγει ένα αρνητικό αποτέλεσμα, ενώ ο όρος άθροισης στις περιπτώσεις (c) και (d) παράγει αποτελέσματα κοντά στο μηδέν αφού οι τιμές των όρων στις δύο κλάσεις αλληλοαναιρούνται.

Εάν τώρα εξετάσουμε τις απόλυτες τιμές των εκχωρημένων βαρών, θα διαπιστώσουμε ότι τα διαφορικά εκφρασμένα γονίδια (περιπτώσεις (a) και (b)) επιτυγχάνουν υψηλότερες τιμές σε σχέση με τα γονίδια τα οποία δεν διαφοροποιούν τις εκφράσεις τους (περιπτώσεις (c) και (d)) στις δύο καταστάσεις ενδιαφέροντος. Αντίθετα, παρατηρούμε ότι η Εξ.(3.76) δεν μπορεί να παράγει το ίδιο αποτέλεσμα, αφού εξαρτάται από τη τιμή και όχι από το πρόσημο του όρου e_j, το οποίο τελικά μπορεί να γίνει πολύ μικρό και να μειώσει το αναμενόμενο αποτέλεσμα.

Από την άλλη μεριά, η Εξ.(3.77) δεν αντιμετωπίζει με δίκαιο τρόπο τα διαφορικά εκφρασμένα γονίδια, αφού χρησιμοποιεί μόνο την τιμή του πρόσημου, με αποτέλεσμα τα γονίδια που εκφράζονται πιο διαφορετικά να λαμβάνουν την ίδια τιμή βάρους με εκείνα που εκφράζονται λιγότερο διαφορικά. Μια πιο δίκαιη λύση στην εκχώρηση βαρών επιτυγχάνεται με την χρήση της Εξ.(3.78) η οποία χρησιμοποιεί τον όρο $f_2(g_i)$.



Εικόνα 24: Διαφορικά εκφρασμένα έναντι μη διαφορικά εκφρασμένων γονιδίων [36].

3.5.6 Βαθμιαία Μάθηση έναντι Ομαδικής Μάθησης (Incremental Vs Batch Learning)

Οι Εξισώσεις (3.76)-(3.79) ανανεώνουν το βάρος $w(t+1)$ μετά την εμφάνιση όλων των δειγμάτων στο δίκτυο, μετά δηλαδή από τον υπολογισμό των όρων των αθροισμάτων. Στην θεωρία των νευρωνικών δικτύων αυτό αναφέρεται ως **ομαδική μάθηση** (batch training). Ένας εναλλακτικός τρόπος ανανέωσης των βαρών είναι μέσω της **βαθμιαίας μάθησης** (incremental learning), όπου τα βάρη ανανεώνονται σταδιακά, εξετάζοντας ένα δείγμα τη φορά. Ακολουθώντας τη συγκεκριμένη στρατηγική, οι όροι άθροισης από των Εξ.(3.76)-(3.79) παραλείπονται και οι εξισώσεις μετασχηματίζονται ως εξής:

$$w_i(t+1) = w(t) + \mu e f'(u) g_i \tag{3.80}$$

$$w_i(t+1) = w(t) + \mu \cdot sign\big(e \cdot f'(u)\big) \cdot sign(g_i) \tag{3.81}$$

$$w_i(t+1) = w(t) + \mu \cdot sign\big(e \cdot f'(u)\big) \cdot sign(g_i) \cdot f_2(g_i) \tag{3.82}$$

$$w_i(t+1) = w(t) + |d-y| \cdot sign\big(e \cdot f'(u)\big) \cdot sign(g_i) \cdot f_2(g_i) \tag{3.83}$$

Στην παρούσα διπλωματική εργασία, για την ανανέωση της τιμής των βαρών στον RFE-LNW αλγόριθμο επιλέξαμε τις Εξ.(3.82) και (3.83), αφού έχει αποδειχθεί ότι στο συγκεκριμένο τομέα η βαθμιαία μάθηση παράγει καλύτερα αποτελέσματα σε σχέση με την ομαδική μάθηση. Τέλος, πρέπει να σημειώσουμε ότι η Εξ.(3.83) χρησιμοποιείται από το σημείο των τελευταίων 100 επιζώντων γονιδίων και κάτω. Για όλα τα υπόλοιπα χρησιμοποιείται η Εξ.(3.82).

3.5.7 Αλγοριθμική Παρουσίαση του RFE-LNW

Στον Πίνακα 4 παρουσιάζεται ο αλγόριθμος της μεθόδου RFE-LNW, όπως περιγράφεται στα προηγούμενα κεφάλαια. Η συγκεκριμένη μέθοδος, εκτός από το πλεονέκτημα της χρήσης ενός μόνου νευρώνα, καταφέρνει να εφαρμόζει filter κριτήρια με ένα wrapper τρόπο, όπου τα βάρη επανεκτιμώνται και προσαρμόζονται από επανάληψη σε επανάληψη. Πράγματι μειώνοντας τον αριθμό των γονιδίων μειώνουμε στην ουσία και τις διαστάσεις του προβλήματος. Σε πρακτικές εφαρμογές είναι σημαντικό τα βάρη των γονιδίων να αλλάζουν από επανάληψη σε επανάληψη, αφού ένας μεγάλος χώρος χαρακτηριστικών με πολλά μη σημαντικά γονίδια, μπορεί να επισκιάσει την επιρροή των πραγματικά σημαντικών. Το συγκεκριμένο γεγονός γίνεται περισσότερο εμφανές καθώς οι διαστάσεις του προβλήματος μειώνονται.

Πίνακας 4 : Η αλγοριθμική παρουσίαση της μεθόδου RFE-LNW [36].

Έστω m ο αρχικός αριθμός των χαρακτηριστικών (γονιδίων)
Όσο (m ≥0)
Ανανέωσε το διάνυσμα βαρών **w** χρησιμοποιώντας τις Εξισώσεις (3.82) και (3.83).
Κατάταξε τα γονίδια σύμφωνα με τις απόλυτες τιμές του διανύσματος **w.**
Αφαίρεσε το χαρακτηριστικό με την μικρότερη απόλυτη τιμή βάρους (m←m-1). Περισσότερα από ένα χαρακτηριστικά μπορούν να διαγραφούν σε κάθε επανάληψη.
Εκτίμησε την ακρίβεια ταξινόμησης των m επιζώντων χαρακτηριστικών χρησιμοποιώντας ένα γραμμικό SVM ταξινομητή.
Τέλος Όσο
Βγάλε στην έξοδο ως γονίδια-δείκτες (marker genes) το σύνολο των επιζώντων χαρακτηριστικών που επιτυγχάνουν την καλύτερη ακρίβεια ταξινόμησης.

3.6 Παλινδρόμηση (Regression)

Η παλινδρόμηση (regression) [45] είναι μια ειδική μορφή προσέγγισης συναρτήσεων η οποία δοθέντος ενός συνόλου τυχαίων μεταβλητών προσπαθεί να βρει τις σχέσεις που μπορεί να υπάρχουν μεταξύ των μεταβλητών. Ένα μοντέλο παλινδρόμησης διαθέτει :

- μια τυχαία μεταβλητή, η οποία θεωρείται «ιδιαίτερου ενδιαφέροντος» και αναφέρεται ως εξαρτώμενη μεταβλητή (dependent variable) ή απόκριση (response).
- ένα σύνολο από τυχαίες μεταβλητές οι οποίες αποκαλούνται ανεξάρτητες μεταβλητές (independent variables) ή παλινδρομητές (regressors). Ο ρόλος τους είναι να εξηγούν ή να προβλέπουν τη στατιστική συμπεριφορά της απόκρισης.
- ένα προσθετικό όρο σφάλματος μέσω του οποίου συνυπολογίζονται οι αβεβαιότητες στον τρόπο με τον οποίο διατυπώνεται η εξάρτηση της απόκρισης από τους παλινδρομητές. Ο όρος σφάλματος καλείται προσδοκώμενο σφάλμα (expectational error) ή εξηγητικό σφάλμα (explanational error).

Υπάρχουν δύο κατηγορίες μοντέλων παλινδρόμησης : γραμμικά και μη γραμμικά. Στα μοντέλα γραμμικής παλινδρόμησης, η εξάρτηση της απόκρισης από τους παλινδρομητές ορίζεται από μια γραμμική συνάρτηση. Αντίθετα στα μοντέλα μη γραμμικής παλινδρόμησης, αυτή η εξάρτηση ορίζεται από μια μη γραμμική συνάρτηση.

3.6.1 Μοντέλο Γραμμικής Παλινδρόμησης (Linear Regression Model)

Ένα μοντέλο γραμμικής παλινδρόμησης [56] μπορεί να περιγραφεί από τη σχέση

$$\hat{Y} = f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \tag{3.84}$$

όπου $X^T = (X_1, X_2, \dots, X_p)$ είναι το διάνυσμα εισόδου και $\hat{Y}$ μια εκτίμηση της πραγματικής εξόδου. Ο όρος β_0 αναπαριστά το προσδοκώμενο σφάλμα του μοντέλου και είναι γνωστός ως πόλωση. Συχνά βολεύει να συμπεριλαμβάνουμε τη σταθερή μεταβλητή 1 στο διάνυσμα X καθώς και το β_0 στο διάνυσμα των συντελεστών β, αυξάνοντας έτσι τη διάσταση των διανυσμάτων X και β από (p) σε $(p+1)$.

Η εκτίμηση των παραμέτρων β πραγματοποιείται με τη βοήθεια των δεδομένων εκπαίδευσης $(x_1, y_1), \dots, (x_N, y_N)$. Κάθε $x_i = \left(x_{i1}, x_{i2}, \dots, x_{ip}\right)^T$ αποτελεί ένα διάνυσμα με τις μετρήσεις των χαρακτηριστικών για την i-οστή περίπτωση. Η πιο γνωστή μέθοδος εκτίμησης των παραμέτρων β είναι η μέθοδος ελαχίστων τετραγώνων (least squares), σύμφωνα με την οποία διαλέγουμε τους συντελεστές $\beta = \left(\beta_0, \beta_1, \dots, \beta_p\right)^T$ που ελαχιστοποιούν το τετραγωνικό αθροιστικό σφάλμα (residual sum of squares):

$$RSS(\beta) = \sum_{i=1}^{N}\left(y_i - f(x_i)\right)^2$$

$$= \sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2. \qquad (3.85)$$

Ορίζουμε ως $\boldsymbol{X}$ το N x $(p+1)$ διάνυσμα το οποίο σε κάθε γραμμή του διαθέτει ένα διάνυσμα εισόδου X διάστασης $(p+1)$. Ομοίως ορίζουμε το διάνυσμα $\boldsymbol{y}$ διάστασης (N x 1), το οποίο αντιστοιχεί στις τιμές εξόδων. Συνεπώς μπορούμε να γράψουμε το τετραγωνικό αθροιστικό σφάλμα (Εξ.(3.85)) ως εξής:

$$RSS(\beta) = (\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta) \qquad (3.86)$$

Διαφορίζοντας τις Εξ.(3.86), (3.87) ως προς β έχουμε:

$$\frac{\partial RSS}{\partial \beta} = -2\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\beta) \qquad (3.87)$$

$$\frac{\partial^2 RSS}{\partial \beta \partial \beta^T} = 2\boldsymbol{X}^T\boldsymbol{X} \qquad (3.88)$$

Υποθέτοντας ότι ο $\boldsymbol{X}$ είναι πλήρους βαθμού στηλών, και επομένως ότι ο $\boldsymbol{X}^T\boldsymbol{X}$ είναι θετικά ορισμένος, θέτουμε την σχέση (3.87) ίση με το μηδέν. Οπότε έχουμε:

$$\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\beta) = 0 \qquad (3.89)$$

Και τελικά καταλήγουμε μέσω της (3.89) στη μοναδική λύση για τους συντελεστές β:

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \qquad (3.90)$$

3.6.2 Μέθοδοι Συρρίκνωσης (Shrinkage Methods)

Για τη βελτίωση του μοντέλου γραμμικής παλινδρόμησης που παρουσιάσαμε προηγουμένως χρησιμοποιούνται συχνά τεχνικές, γνωστές ως μέθοδοι συρρίκνωσης [57]. Οι μέθοδοι συρρίκνωσης βασίζονται στην ιδέα ελάττωσης ή απόρριψης ορισμένων τιμών των παλινδρομητών (regressors) x_i. Αυτό επιτυγχάνεται με την προσθήκη ενός περιορισμού (penalty) στη συνάρτηση του τετραγωνικού αθροιστικού σφάλματος (Εξ.(3.85)), το οποίο έχει ως αποτέλεσμα ορισμένοι συντελεστές β να πλησιάζουν ή να γίνονται ίσοι με μηδέν. Το μικρότερο υποσύνολο παλινδρομητών που παράγεται μας οδηγεί σε πιο ερμηνεύσιμα εννοιολογικά μοντέλα, τα οποία συχνά επιτυγχάνουν υψηλότερη ακρίβεια πρόβλεψης σε σχέση με εκείνα που προκύπτουν μέσω της γραμμικής παλινδρόμησης.

Στη συνέχεια παρουσιάζουμε δύο μεθόδους συρρίκνωσης: το LASSO και το Ridge Regression.

3.6.3 LASSO

Το LASSO (Least Absolute Shrinkage and Selection Operator) είναι μια μέθοδος για εκτίμηση σε γραμμικά μοντέλα. Η συγκεκριμένη μέθοδος, η οποία παρουσιάστηκε από τον Tibshirani [57], ελαχιστοποιεί το τετραγωνικό αθροιστικό σφάλμα υπό την προϋπόθεση ότι το άθροισμα της απόλυτης τιμής των συντελεστών β είναι μικρότερο από μια σταθερά. Επομένως η εκτίμηση για τους συντελεστές $\hat{\beta}^{lasso}$ καθορίζεται από τη σχέση:

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

$$\text{δεδομένου ότι} \quad \sum_{j=1}^{p} |\beta_j| \leq t \tag{3.91}$$

Το t είναι μια ρυθμιζόμενη παράμετρος μεγαλύτερη ή ίση του μηδενός ($t \geq 0$), η οποία καθορίζει το βαθμό συρρίκνωσης των παραμέτρων $\hat{\beta}^{lasso} = [\hat{\beta}_1, \dots, \hat{\beta}_p]$. Επιλέγοντας μια κατάλληλη τιμή για το t κάποιοι από τους συντελεστές β_j μηδενίζονται. Αναλυτικότερα, έστω $\hat{\beta}_j^{ls}$ η εκτίμηση για τις παραμέτρους β που προέκυψε μέσω της μεθόδου των ελαχίστων τετραγώνων του γραμμικού μοντέλου (Εξ.(3.90)). Αν η τιμή του t που θα επιλεγεί

είναι μεγαλύτερη από το $t_0 = \sum_1^p |\hat{\beta}_j^{ls}|$, οι εκτιμητές $\hat{\beta}_j^{lasso}$ θα είναι ίδιοι με εκείνους του γραμμικού μοντέλου. Αντίθετα, αν επιλέξουμε τιμές του $t \leq t_0$ πολλές από τις παραμέτρους $\hat{\beta}_j^{lasso}$ θα πλησιάσουν ή θα γίνουν ίσες με το μηδέν.

Μπορούμε επίσης να διατυπώσουμε το πρόβλημα LASSO (Εξ.(3.91)) στην Lagrangian μορφή του:

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{3.92}$$

Το $\lambda \geq 0$ είναι η ρυθμιστική παράμετρος (regularization parameter), η οποία ελέγχει το βαθμό συρρίκνωσης των συντελεστών β_j.

Συνεπώς, με τη βοήθεια της μεθόδου LASSO, χρησιμοποιώντας είτε την Εξ.(3.91) είτε την (3.92), από ένα μεγάλο σύνολο από παλινδρομητές x_i καταλήγουμε σε ένα μικρότερο υποσύνολο που περιέχει εκείνα τα x_i που εμφανίζουν τα ισχυρότερα χαρακτηριστικά (εκείνα δηλαδή με μη μηδενικούς συντελεστές β_j).

3.6.4 Ridge Regression

Το Ridge Regression [56] αποτελεί μια μέθοδο συρρίκνωσης, όπως το LASSO, με βασική διαφορά τον περιορισμό που επιβάλλεται στη συνάρτηση του τετραγωνικού αθροιστικού σφάλματος. Στην περίπτωση του Ridge Regression ως penalty χρησιμοποιείται τα άθροισμα των τετραγώνων των παραμέτρων β. Οπότε το ridge πρόβλημα μπορεί να περιγραφεί από τη σχέση:

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{3.93}$$

Ένας εναλλακτικός τρόπος να περιγράψουμε το ridge regression είναι:

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2$$

$$\text{δεδομένου ότι} \quad \sum_{j=1}^{p} \beta_j^2 \leq t \tag{3.94}$$

3.6.5 Γεωμετρική σύγκριση LASSO και Ridge Regression

Στην ενότητα 3.9.2.1 περιγράψαμε το λόγο για τον οποίο το LASSO συχνά παράγει συντελεστές οι οποίοι είναι ίσοι με μηδέν. Αντίθετα, το Ridge Regression συρρικνώνει τους συντελεστές χωρίς όμως να τους μηδενίζει [56],[57]. Γνωρίζουμε ότι η διαφορά των δύο μεθόδων βρίσκεται στο penalty που εφαρμόζεται στη συνάρτηση τετραγωνικού αθροιστικού σφάλματος. Στη περίπτωση του LASSO έχουμε τις απόλυτες τιμές των συντελεστών $\sum|\beta_j| \leq t$, ενώ στο Ridge regression έχουμε τα τετράγωνα των συντελεστών $\sum \beta_j^2 \leq t$.

Ας δούμε όμως τι συμβαίνει στην περίπτωση των δύο διαστάσεων όπου p=2 για να κατανοήσουμε καλύτερα το λόγο για τον οποίο παράγεται ή όχι ένας μηδενικός συντελεστής β_j:

Το κοινό κριτήριο και των δύο μεθόδων $\sum_{i=1}^{N}\left(y_i - \sum_j \beta_j x_{ij}\right)^2$ ισοδυναμεί με τη τετραγωνική συνάρτηση:

$$\left(\beta - \hat{\beta}\right)^T \boldsymbol{X}^T \boldsymbol{X}\left(\beta - \hat{\beta}\right) \tag{3.95}$$

Η Εξ.(3.95) απεικονίζεται με ελλειπτικές καμπύλες, οι οποίες είναι κεντραρισμένες στις εκτιμήσεις των ελαχίστων τετραγώνων του $\hat{\beta}$, όπως φαίνεται στην Εικόνα 25.

Για το penalty στην περίπτωση του LASSO έχουμε ότι:

$$|\beta_1| + |\beta_2| \leq t \tag{3.96}$$

το οποίο απεικονίζεται ως ένα μοναδιαίο ορθογώνιο παραλληλόγραμμο. (βλέπε Εικόνα 25)

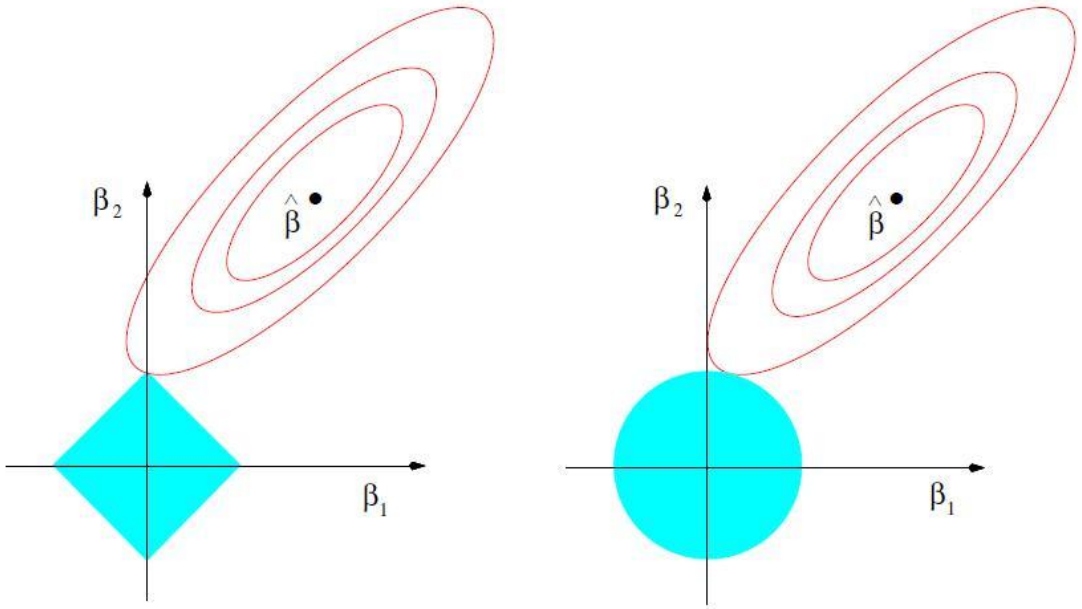
Αντίστοιχα για το penalty στην περίπτωση του Ridge Regression έχουμε ότι:

$$\beta_1^2 + \beta_2^2 \leq t \tag{3.97}$$

το οποίο απεικονίζεται ως ένας μοναδιαίος κύκλος. (βλέπε Εικόνα 25)

Όπως παρατηρούμε στην Εικόνα 25, οι λύσεις για τις δύο μεθόδους βρίσκονται στα σημεία που οι καμπύλες τέμνουν τις περιοχές περιορισμού. Έτσι, η λύση για το LASSO βρίσκεται στο σημείο το οποίο οι καμπύλες τέμνουν το ορθογώνιο. Αν αυτό το σημείο τομής βρίσκεται σε γωνία, όπως συμβαίνει στην περίπτωση που απεικονίζεται, τότε ο συντελεστής β_j που παράγεται είναι ίσος με μηδέν. Αντίστοιχα για το Ridge Regression

παρατηρούμε ότι οι καμπύλες τέμνουν το μοναδιαίο κύκλο σε τιμές κοντά στο μηδέν με αποτέλεσμα οι συντελεστές β_j που παράγονται να πλησιάζουν αλλά να μην παίρνουν ποτέ μηδενικές τιμές.



Εικόνα 25 : Απεικόνιση του Lasso (αριστερά) και του Ridge Regression (δεξιά). Με μπλε χρώμα σημειώνονται οι περιοχές περιορισμού $|\beta_1| + |\beta_2| \leq t$ και $\beta_1^2 + \beta_2^2 \leq t$, αντίστοιχα, ενώ οι κόκκινες ελλείψεις απεικονίζουν τις καμπύλες της συνάρτησης τετραγωνικού αθροιστικού σφάλματος [56].

3.6.6 Εφαρμογή μοντέλου παλινδρόμησης

Στη παρούσα διπλωματική εργασία χρησιμοποιήσαμε το παρακάτω γραμμικό μοντέλο παλινδρόμησης [58] :

$$y = Ax \tag{3.98}$$

Όπου: $A \in \boldsymbol{R}^{m\,x\,n}$ ο πίνακας που περιέχει το διαθέσιμο σύνολο δεδομένων (m δείγματα Χ n γονίδια), $y \in \boldsymbol{R}^m$ το διάνυσμα των εξόδων (η κλάση στην οποία ανήκει κάθε δείγμα) και $x \in \boldsymbol{R}^n$ το διάνυσμα των άγνωστων παραμέτρων .

Η εκτίμηση των παραμέτρων x, από τις οποίες τελικά θα προκύψει το υποσύνολο των επικρατέστερων χαρακτηριστικών (γονιδίων), υπολογίζεται με τη βοήθεια της μεθόδου LASSO. Γνωρίζουμε ότι οι εκτιμητές $\hat{\beta}^{lasso}$ προκύπτουν μέσω της ελαχιστοποίησης της ποσότητας:

$$\sum_{i=1}^{N}\left(y_i - \sum_j \beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p}|\beta_j| = \sum_{i=1}^{N}\left(\sum_j x_{ij}\beta_j - y_i\right)^2 + \lambda \sum_{j=1}^{p}|\beta_j| \quad (3.99)$$

Οπότε αντικαθιστώντας κατάλληλα στην Εξ.(3.99) τις παραμέτρους A, y, x του μοντέλου παλινδρόμησης (Εξ.(3.98)) που χρησιμοποιήσαμε προκύπτει ότι οι παράμετροι x στις οποίες καταλήγουμε είναι αυτές που ελαχιστοποιούν την ποσότητα :

$$\|Ax - y\|_2^2 + \lambda\|x\|_1 \quad (3.100)$$

όπου $\|x\|_1 = \sum_i^n |x_i|$ είναι η l_1 νόρμα του x και λ>0 είναι η ρυθμιστική παράμετρος (regularization parameter).

3.7 Αξιολόγηση του Ταξινομητή

Μετά από την εκπαίδευση ενός ταξινομητή, ακολουθεί η εκτίμηση της απόδοσής του στη ταξινόμηση νέων αθέατων καταστάσεων [59], [60] με τη βοήθεια διάφορων μέτρων αξιολόγησης. Η διαδικασία αξιολόγησης ενός ταξινομητή παίζει σημαντικό ρόλο τόσο στην εκτίμηση της χρησιμότητας του συστήματος που έχει κατασκευαστεί όσο και στη ρύθμιση διαφόρων παραμέτρων του συστήματος με σκοπό να βελτιστοποιηθεί η απόδοσή του. Σε πολλές περιπτώσεις, η ακρίβεια στη ταξινόμηση νέων δεδομένων κρίνεται απαραίτητη και ιδιαίτερα στις βιοϊατρικές εφαρμογές, όπου η ικανότητα πρόβλεψης ενός αποτελέσματος είναι υψίστης σημασίας. Συχνά όμως στον χώρο της ιατρικής, δεν έχουμε στην διάθεσή μας ανεξάρτητα σύνολα δεδομένων εκπαίδευσης και ελέγχου. Στην περίπτωση αυτή για την δημιουργία ενός αξιόπιστου ταξινομητή χρησιμοποιούνται διάφορες στατιστικές τεχνικές. Το Cross Validation, όπως περιγράφηκε στο Κεφάλαιο 2.3, αποτελεί μια γνωστή στατιστική μέθοδο για τον διαχωρισμό ενός ενιαίου συνόλου δεδομένων σε δύο ανεξάρτητα σύνολα δεδομένων για την εκπαίδευση και έλεγχο του ταξινομητή.

3.7.1 Μέτρα Αξιολόγησης

Στην περίπτωση της δυαδικής ταξινόμησης [59] η πρόβλεψη για ένα δείγμα μπορεί να πάρει μια από τις τέσσερεις πιθανές τιμές οι οποίες εμφανίζονται στον Πίνακα 5, true positive(TP), true negative(TN), false positive(FP), false negative(FN). Στην παρούσα εργασία χρησιμοποιήθηκε η δυαδική ταξινόμηση με κλάσεις ενδιαφέροντος την OA (osteoarthritis) και NC (normal control). Ως θετικό θεωρήθηκε κάθε δείγμα του συνόλου δεδομένων που ανήκει στη κλάση OA και ως αρνητικό εκείνο που ανήκει στη κλάση NC. Κατά συνέπεια, ένα true positive (TP) δείγμα είναι ένα δείγμα της κλάσης OA το οποίο ταξινομήθηκε σωστά ενώ αντίθετα ένα false positive (FP) δείγμα είναι ένα δείγμα που ανήκει στη κλάση OA και ταξινομήθηκε λανθασμένα στη κλάση NC. Παρομοίως, ένα true negative(TN) δείγμα είναι ένα δείγμα της κλάσης NC που ταξινομήθηκε σωστά ενώ ένα false negative(FN) είναι ένα δείγμα της κλάσης NC που ταξινομήθηκε λανθασμένα στη κλάση OA.

Πίνακας 5: Πιθανά αποτελέσματα δυαδικής ταξινόμησης [59].

		ΠΡΟΒΛΕΠΟΜΕΝΗ ΚΛΑΣΗ	
		ΟΑ	**NC**
ΠΡΑΓΜΑΤΙΚΗ ΚΛΑΣΗ	**ΟΑ**	TP	FN
	NC	FP	TN

Τα πιο συχνά χρησιμοποιούμενα μέτρα αξιολόγησης [60] , που υπολογίζονται με την βοήθεια των ποσοτήτων TP, TN, FP και FN, είναι η ευαισθησία (Sensitivity-Se), η ιδιαιτερότητα (Specificity-Sp) και η ακρίβεια (Accuracy). Η ευαισθησία σχετίζεται με την ικανότητα του ταξινομητή να προβλέπει σωστά την θετική κλάση. Αντίθετα, η ιδιαιτερότητα σχετίζεται με την ικανότητα του ταξινομητή να προβλέπει σωστά την αρνητική κλάση. Τέλος η ακρίβεια προσδιορίζει την ικανότητα του ταξινομητή στην σωστή κατάταξη ενός δείγματος. Οι σχέσεις που υπολογίζουν τα παραπάνω μέτρα αξιολόγησης είναι:

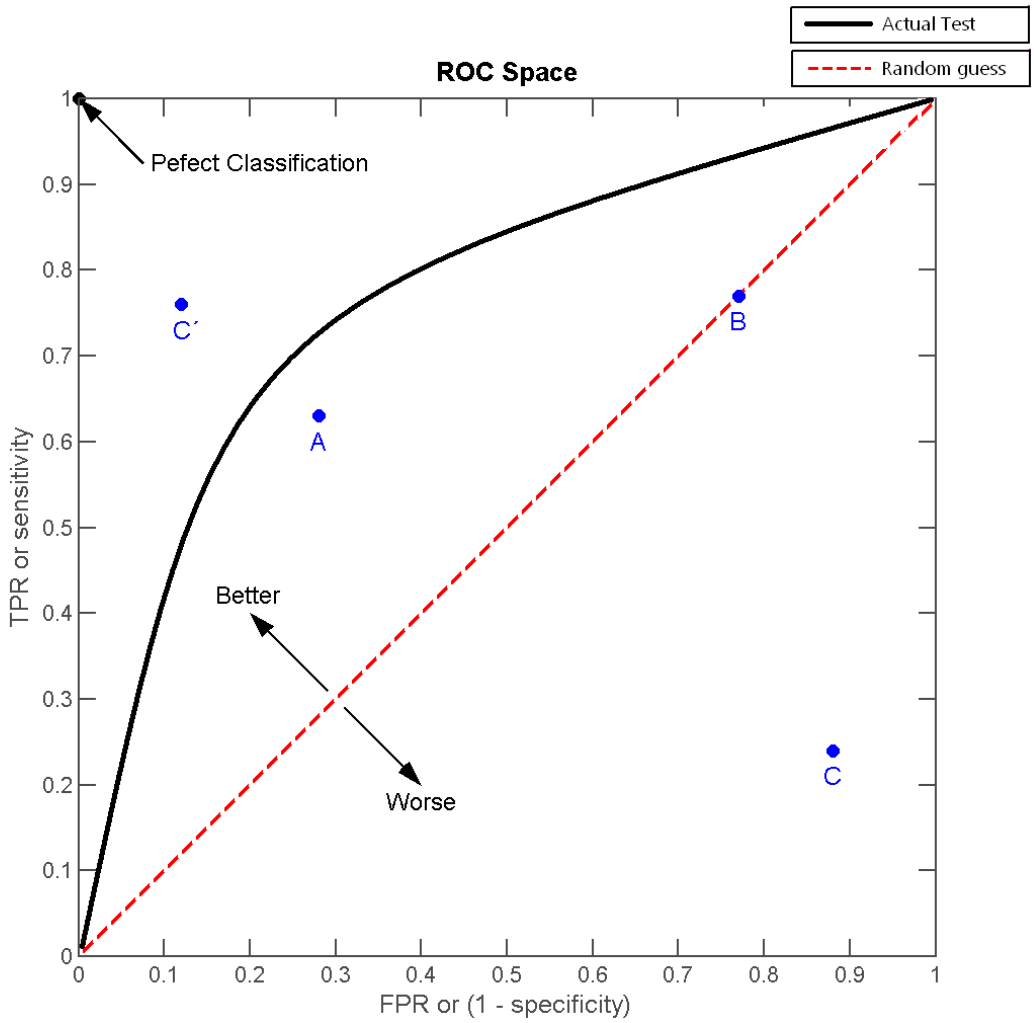
$$\text{Sensitivity} = \frac{TP}{TP+FN} \tag{3.101}$$

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{3.102}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.103}$$

Μια χρήσιμη τεχνική για την απεικόνιση της απόδοσης ενός δυαδικού ταξινομητή αποτελούν οι ROC (Receiver Operating Characteristic) καμπύλες ή αλλιώς καμπύλες λειτουργικού χαρακτηριστικού δείκτη [61]. Για το σχηματισμό μιας καμπύλης ROC χρησιμοποιούνται τα ποσοστά των μετρικών Sensitivity (True Positive Rate, TPR) και 1-Specificity (False Positive Rate, FPR) για διάφορες τιμές κατωφλίου. Αυτά τα ζεύγη τιμών σημειώνονται σε ένα γράφημα όπως φαίνεται στην Εικόνα 26 όπου ο άξονας x αντιστοιχεί στις τιμές του FPR και ο άξονας y στις τιμές του TPR. Η απόδοση κάθε ταξινομητή αναπαρίσταται ως ένα σημείο στην καμπύλη ROC. Η καλύτερη δυνατή μέθοδος πρόβλεψης (τέλεια ταξινόμηση) αποδίδει ένα σημείο στην επάνω αριστερή γωνία (συντεταγμένες (0,1))

του χώρου ROC, αναπαριστώντας με αυτό τον τρόπο 100% Sensitivity (κανένα FN δείγμα) και 100% Specificity (κανένα FP δείγμα). Μια τελείως τυχαία πρόβλεψη δίνει ένα σημείο στη διαγώνια γραμμή που συνδέει τη κάτω αριστερή γωνία με τη πάνω δεξιά του ROC space. Τα σημεία που βρίσκονται πάνω από τη διαγώνιο αντιπροσωπεύουν καλά αποτελέσματα ταξινόμησης (καλύτερα από εκείνα της τυχαίας πρόβλεψης), ενώ αντίστοιχα τα σημεία κάτω από τη διαγώνιο δείχνουν μη αποδοτικά αποτελέσματα ταξινόμησης (χειρότερα από τα τυχαία).



Εικόνα 26 : Καμπύλη ROC [61].

ΚΕΦΑΛΑΙΟ 4 : ΠΡΟΤΕΙΝΟΜΕΝΗ ΜΕΘΟΔΟΛΟΓΙΑ

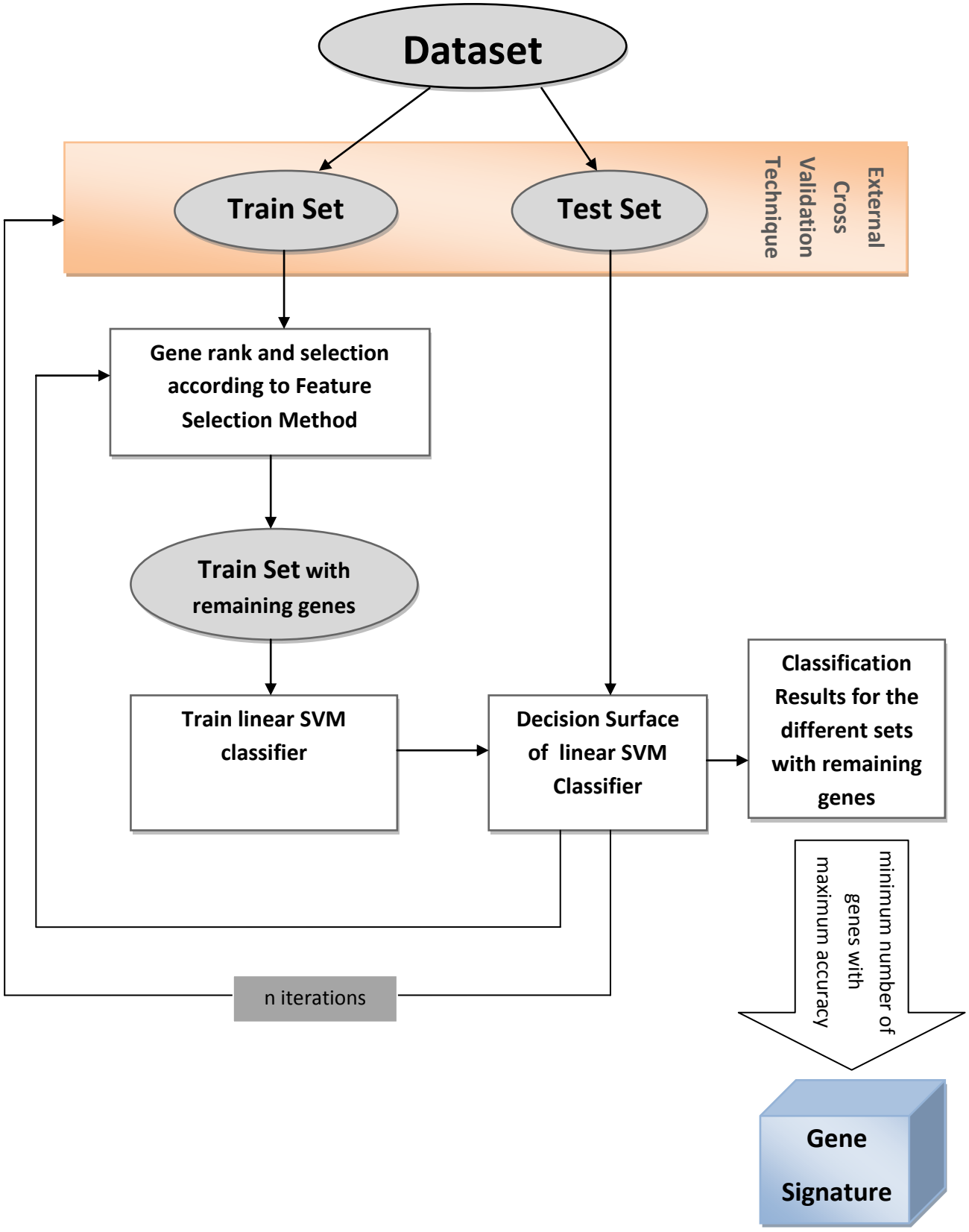
Η ταξινόμηση δεδομένων γονιδιακής έκφρασης [1],[62], τα οποία προέρχονται από DNA μικροσυστοιχίες και περιλαμβάνουν την έκφραση χιλιάδων γονιδίων, είναι γνωστό ότι αποτελεί μια δύσκολη και χρονοβόρα διαδικασία, αφού απαιτεί την επεξεργασία χιλιάδων γονιδιακών τιμών. Το συγκεκριμένο πρόβλημα μπορεί να αντιμετωπιστεί με τη χρήση μεθόδων επιλογής χαρακτηριστικών ή επιλογής γονιδίων. Έτσι προκύπτει ένα διαχειρίσιμο υποσύνολο δεδομένων που είναι μικρότερο από το αρχικό σύνολο και περιέχει τα πιο σημαντικά γονίδια, εκείνα δηλαδή που η έκφρασή τους διαφοροποιείται περισσότερο ανάμεσα στην φυσιολογική (Normal Control-NC) και μη φυσιολογική κατάσταση (Osteoarthritis-OA). Ένα τέτοιο σύνολο γονιδίων, με επαρκή διαχωριστική ικανότητα ανάμεσα στις δύο καταστάσεις-κλάσεις ενδιαφέροντος, το οποίο συχνά καλείται γονιδιακή υπογραφή (gene signature), μπορεί να χρησιμοποιηθεί για την κατασκευή ενός συστήματος πρόβλεψης με σκοπό την ορθή ταξινόμηση νέων αθέατων καταστάσεων (νέων δειγμάτων). Κατανοούμε λοιπόν πως η ανάπτυξη ευέλικτων και εύρωστων μεθόδων επιλογής χαρακτηριστικών οδηγεί σε μείωση του χρόνου επεξεργασίας των δεδομένων καθώς και σε υψηλότερη ακρίβεια ταξινόμησης.

Στην παρούσα διπλωματική εργασία για την επιλογή των επικρατέστερων γονιδίων χρησιμοποιήθηκαν οι μέθοδοι:

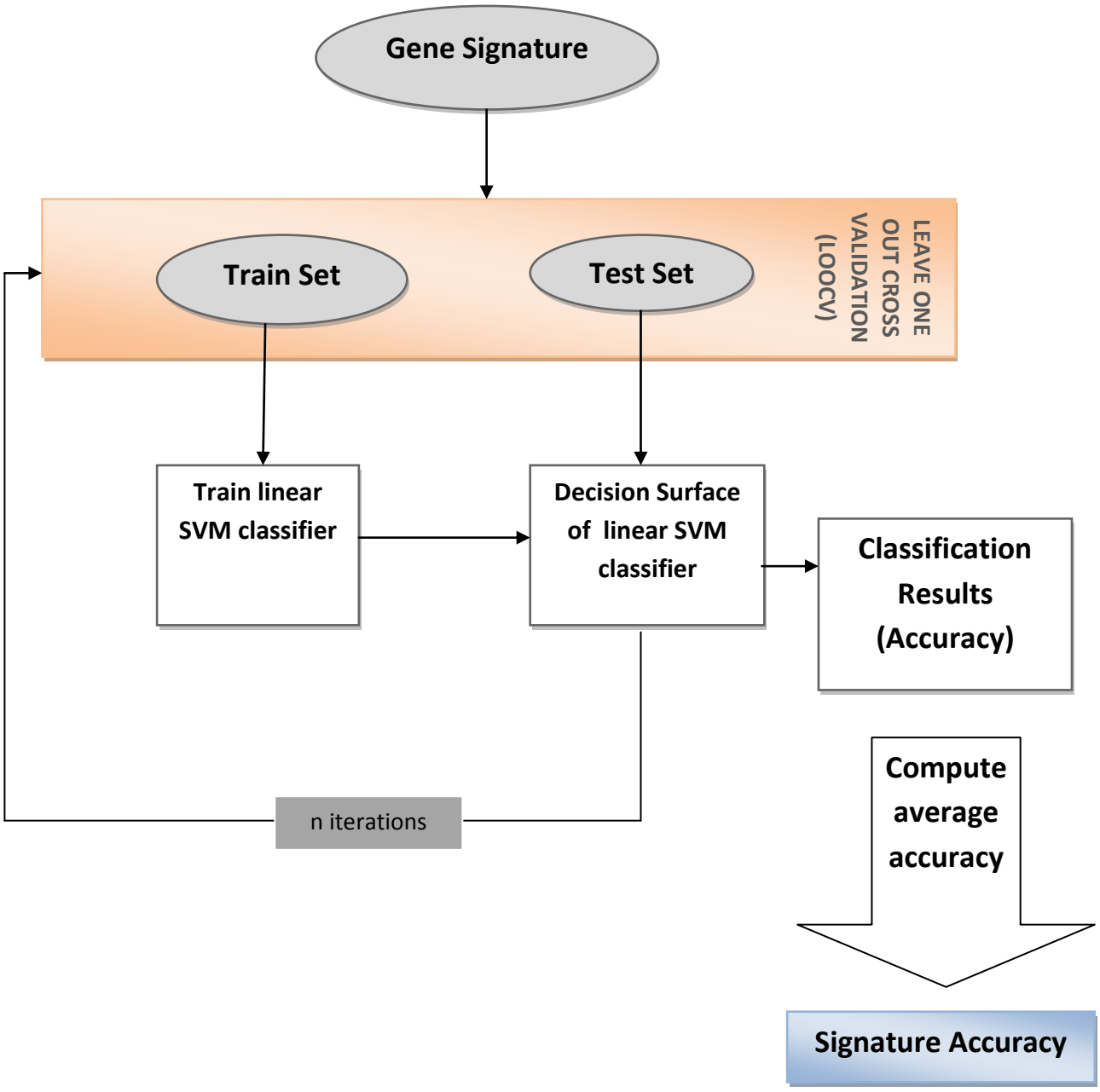
- RFE-LNW (Κεφάλαιο 3.5)
- LASSO (Κεφάλαιο 3.6)
- συνδυασμός RFE-LNW και FSMLP (Κεφάλαιο 3.3.5)

Τα γονίδια που προκύψαν από τις παραπάνω μεθόδους χρησιμοποιήθηκαν για την εκπαίδευση ενός γραμμικού SVM ταξινομητή. Η ικανότητα γενίκευσης του συγκεκριμένου ταξινομητή σε νέες αθέατες καταστάσεις εκτιμήθηκε με την βοήθεια τεχνικών External Cross Validation.

Στα Κεφάλαια που ακολουθούν γίνεται αναλυτική περιγραφή της μεθοδολογίας που ακολουθήσαμε για την εξαγωγή γονιδιακών υπογραφών από σύνολα δεδομένων που αφορούν την ασθένεια της οστεοαρθρίτιδας (ΟΑ). Ένα συνοπτικό περίγραμμα της μεθοδολογίας παρουσιάζεται στις Εικόνες 27 και 28.



Εικόνα 27: Συνοπτική απεικόνιση της μεθοδολογίας για την εξαγωγή της Γονιδιακής Υπογραφής.



Εικόνα 28: Απεικόνιση της μεθοδολογίας για την αξιολόγηση της Γονιδιακής Υπογραφής.

4.1 Διαχωρισμός και Επεξεργασία του Συνόλου Δεδομένων

Η μεθοδολογία της εργασίας εφαρμόζεται σε μια βάση δεδομένων [5] που περιέχει γονίδια, η έκφραση των οποίων μετράται σε 10 δείγματα αρθρικού ιστού από δότες με οστεοαρθρίτιδα (OA) καθώς και σε 9 δείγματα αρθρικού ιστού από υγιείς δότες (NC). Πιο συγκεκριμένα, η βάση δεδομένων χωρίζεται σε δυο επιμέρους σύνολα. Το πρώτο σύνολο, στο οποίο θα αναφερόμαστε ως **Dataset A**, αποτελείται από 22283 γονίδια τα οποία εκφράζονται και στα 19 δείγματα του αρθρικού ιστού (10 OA και 9 NC). Ενώ το δεύτερο σύνολο, το οποίο ονομάζουμε **Dataset B**, περιέχει την έκφραση 22645 γονιδίων για τα 14 από τα 19 δείγματα του αρθρικού ιστού (10 OA και 4 NC). Όπως παρατηρούμε, οι ετικέτες OA και NC συνιστούν τις κλάσεις ενδιαφέροντος για την δυαδική ταξινόμηση του dataset. Έτσι κάθε δείγμα OA έχει ως target τη τιμή +1, ενώ αντίστοιχα κάθε δείγμα NC έχει ως target τη τιμή -1.

Λόγω του μικρού αριθμού δειγμάτων (19 για το Dataset A και 14 για το Dataset B) που έχουμε στη διάθεσή μας και της έλλειψης ανεξάρτητου συνόλου δεδομένων για έλεγχο της απόδοσης του ταξινομητή που κατασκευάζουμε, χρησιμοποιούμε όπως φαίνεται και στην Εικόνα 27 τη τεχνική ELOOCV (Κεφάλαιο 2.3). Σύμφωνα με τη συγκεκριμένη τεχνική χωρίζουμε το ενιαίο σύνολο δεδομένων σε ανεξάρτητα σύνολα δεδομένων εκπαίδευσης (train set) και σύνολα δεδομένων ελέγχου (test set). Η διαδικασία διαχωρισμού επαναλαμβάνεται τόσες φορές όσες και ο αριθμός των δειγμάτων του συνόλου δεδομένων (n=19 φορές για το Dataset A και n=14 φορές για το Dataset B). Σε κάθε επανάληψη χρησιμοποιείται ένα διαφορετικό δείγμα από το σύνολο δεδομένων για έλεγχο και τα υπόλοιπα n-1 δείγματα για εκμάθηση. Όπως αναφέραμε και στο Κεφάλαιο 2.3, στο ELOOCV το δείγμα που χρησιμοποιείται ως test set δεν περιλαμβάνεται στην διαδικασία επιλογής γονιδίων, καθιστώντας κατ' αυτό το τρόπο το ELOOCV μια αμερόληπτη μέθοδο εκτίμησης της απόδοσης του ταξινομητή που κατασκευάζεται. Το test set που χρησιμοποιείται σε κάθε επανάληψη ταξινομείται είτε σωστά (επιτυχία ταξινόμησης = 1) είτε λανθασμένα (αποτυχία ταξινόμησης = 0) από τον ταξινομητή που εκπαιδεύτηκε με το train set. Μετά την ολοκλήρωση και των n επαναλήψεων του ELOOCV προκύπτει ο πίνακας $Save\_all$ διάστασης *m x n* (m =αριθμός επαναλήψεων αλγορίθμου επιλογής γονιδίων , n= αριθμός επαναλήψεων ELOOCV) με στοιχεία $Save\_All_{ij}$ που αντιπροσωπεύουν το αποτέλεσμα της ταξινόμησης του i-οστού υποσυνόλου δεδομένων ελέγχου στην j-οστή επανάληψη του ELOOCV. Στη συνέχεια με τη βοήθεια του πίνακα *Save_all* υπολογίζεται το διάνυσμα *A* :

$$A_i = \frac{1}{n}\sum_{j=1}^{n} Save\_All_{ij} \qquad i = 1, \dots, m \ , j = 1, \dots, n \qquad (4.1)$$

το οποίο περιέχει το μέσο ποσοστό ακρίβειας του ταξινομητή που κατασκευάστηκε από το i-οστό υποσύνολο δεδομένων εκπαίδευσης. Χρησιμοποιώντας το διάνυσμα Α εντοπίζουμε το μικρότερο υποσύνολο γονιδίων στο οποίο σημειώνεται η μεγαλύτερη μέση ακρίβεια ταξινόμησης. Προφανώς, θα υπάρχουν και μεγαλύτερα υποσύνολα γονιδίων στα οποία επιτυγχάνεται η ίδια τιμή μέγιστης ακρίβειας. Όμως εμείς ενδιαφερόμαστε για τον ελάχιστο αριθμό γονιδίων (γονιδιακή υπογραφή) στον οποίον μπορεί να επιτευχθεί αυτή η μέγιστη τιμή.

Τέλος, όπως φαίνεται και στην Εικόνα 28 προσπαθήσαμε να επαληθεύσουμε ότι τα γονίδια στα οποία καταλήξαμε και συνιστούν τη γονιδιακή υπογραφή ($gene\_signature$) είναι ικανά να κατασκευάσουν έναν αποδοτικό ταξινομητή. Για το λόγο αυτό εκπαιδεύουμε ένα γραμμικό SVM ταξινομητή με τα γονίδια της υπογραφής και εκτιμάμε την ικανότητα γενίκευσής του μέσω της LOOCV τεχνικής. Σε κάθε κύκλο επανάληψης της μεθόδου LOOCV υπολογίζουμε την επιτυχία ($S_j = 1$) ή αποτυχία ($S_j = 0$) της ταξινόμησης του j-οστού δείγματος που χρησιμοποιείται για τον έλεγχο της απόδοσης του ταξινομητή. Μετά την ολοκλήρωση των n επαναλήψεων υπολογίζουμε το μέσο ποσοστό ακρίβειας που σημείωσε ο SVM classifier, το οποίο είναι ίσο με:

$$Signature\_Accuracy = \frac{1}{n}\sum_{j=1}^{n} S_j \qquad j = 1, \dots, n \qquad (4.2)$$

4.2 Πρώτη Γονιδιακή Υπογραφή

Για τον υπολογισμό της πρώτης γονιδιακής υπογραφής τόσο στο Dataset A όσο και στο Dataset B ακολουθήσαμε τα παρακάτω βήματα.

Βήμα 1 - Διαχωρισμός Δεδομένων : Το σύνολο δεδομένων μας που αποτελείται από m γονίδια και n δείγματα χωρίζεται σε train και test set σύμφωνα με τη τεχνική ELOOCV, όπως αυτή περιγράφηκε στο προηγούμενο Κεφάλαιο (4.1). Τα νέα ανεξάρτητα σύνολα δεδομένων για την εκπαίδευση και τον έλεγχο του γραμμικού SVM ταξινομητή αποτελούνται αρχικά από m γονίδια.

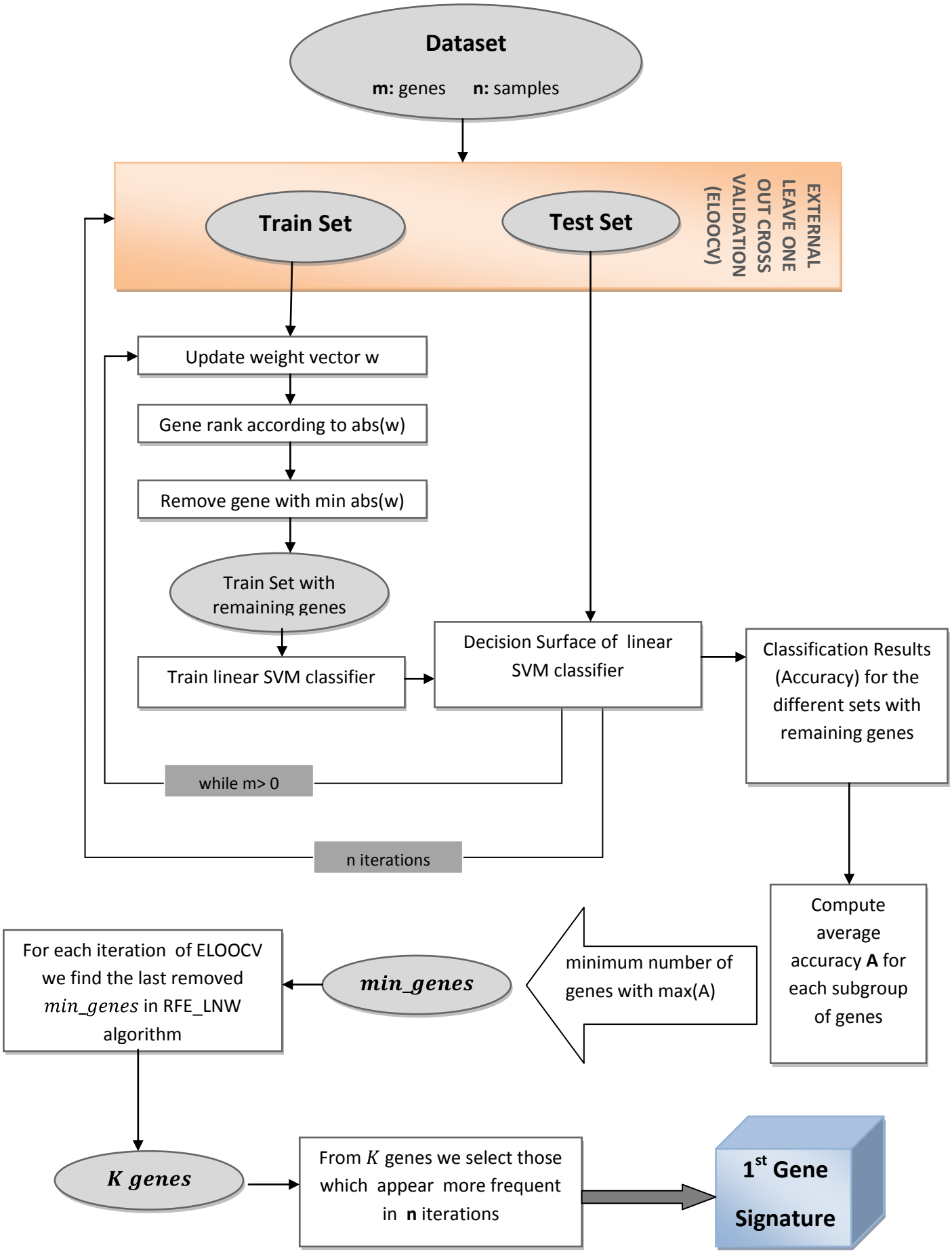
Βήμα 2 - Δημιουργία υποσυνόλων γονιδίων: Μετά από τον διαχωρισμό του διαθέσιμου συνόλου δεδομένων προχωράμε στην διαδικασία επιλογής γονιδίων. Σε κάθε επανάληψη του ELOOCV χρησιμοποιούμε τον αλγόριθμο Recursive Feature Elimination based on Linear Neuron Weights (RFE-LNW). Αρχικά, ο αλγόριθμος RFE-LNW αποδίδει βάρη στα m γονίδια του train set (Εξισώσεις (3.82), (3.83)). Έπειτα τα γονίδια κατατάσσονται σύμφωνα με την απόλυτη τιμή του βάρους τους. Το ή τα γονίδια εκείνα με την μικρότερη απόλυτη τιμή αφαιρούνται τόσο από το train set όσο και από το test set. Οι ανανεωμένοι πίνακες των train set και test set χρησιμοποιούνται για την εκπαίδευση και την εκτίμηση της ακρίβειας ταξινόμησης του γραμμικού SVM classifier. Ο αλγόριθμος RFE-LNW επαναλαμβάνεται μέχρις ότου αφαιρεθούν όλα τα γονίδια. Σε κάθε στάδιο του αλγορίθμου RFE-LNW το train set, πίνακας ο οποίος όπως αναφέραμε περιλαμβάνει τα εναπομείναντα γονίδια για εκπαίδευση, χρησιμοποιείται για να κατασκευάσει το διαχωριστικό υπερεπίπεδο μεταξύ των (θετικών) δειγμάτων που ανήκουν στην κλάση OA και των (αρνητικών) που ανήκουν στη κλάση NC. Έπειτα το δείγμα του test set δοκιμάζεται στο υπερεπίπεδο που έχει παραχθεί και προκύπτει είτε επιτυχία ($Save\_All_{ij}$= 1) είτε αποτυχία ($Save\_All_{ij}$ =0) της ταξινόμησής του.

Βήμα 3 - Επιλογή γονιδίων : Μετά την ολοκλήρωση των n επαναλήψεων του ELOOCV υπολογίζουμε το μέσο ποσοστό ακρίβειας του ταξινομητή σύμφωνα με την Εξίσωση (4.1). Από το διάνυσμα A_i βρίσκουμε τον ελάχιστο αριθμό γονιδίων, έστω min_genes, στον οποίον επιτυγχάνεται η μέγιστη ακρίβεια. Για κάθε μια τις n επαναλήψεις του cross validation συγκεντρώνουμε σε ένα πίνακα τα min_genes που αφαιρέθηκαν τελευταία στον αλγόριθμο RFE-LNW (έστω K genes). Υπολογίζουμε τη συχνότητα εμφάνισης τους μέσα σε αυτές τις n επαναλήψεις και τελικά επιλέγουμε εκείνα

που παρουσιάζουν τις υψηλότερες συχνότητες. Η επιλογή αυτή πραγματοποιείται με τη βοήθεια ενός διαγράμματος συχνοτήτων εμφάνισης.

Βήμα 4 - Αξιολόγηση της γονιδιακής υπογραφής : Τα γονίδια που τελικά επιλέγονται μέσα από το διάγραμμα συχνοτήτων αποτελούν τη γονιδιακή υπογραφή ($gene\_signature$). Η αξιολόγηση της υπογραφής ως προς την ικανότητά της να κατασκευάζει έναν αποδοτικό ταξινομητή πραγματοποιείται σύμφωνα με τη διαδικασία που απεικονίζεται στο διάγραμμα της Εικόνας 28 και περιγράφηκε στη προηγούμενη ενότητα (4.1).

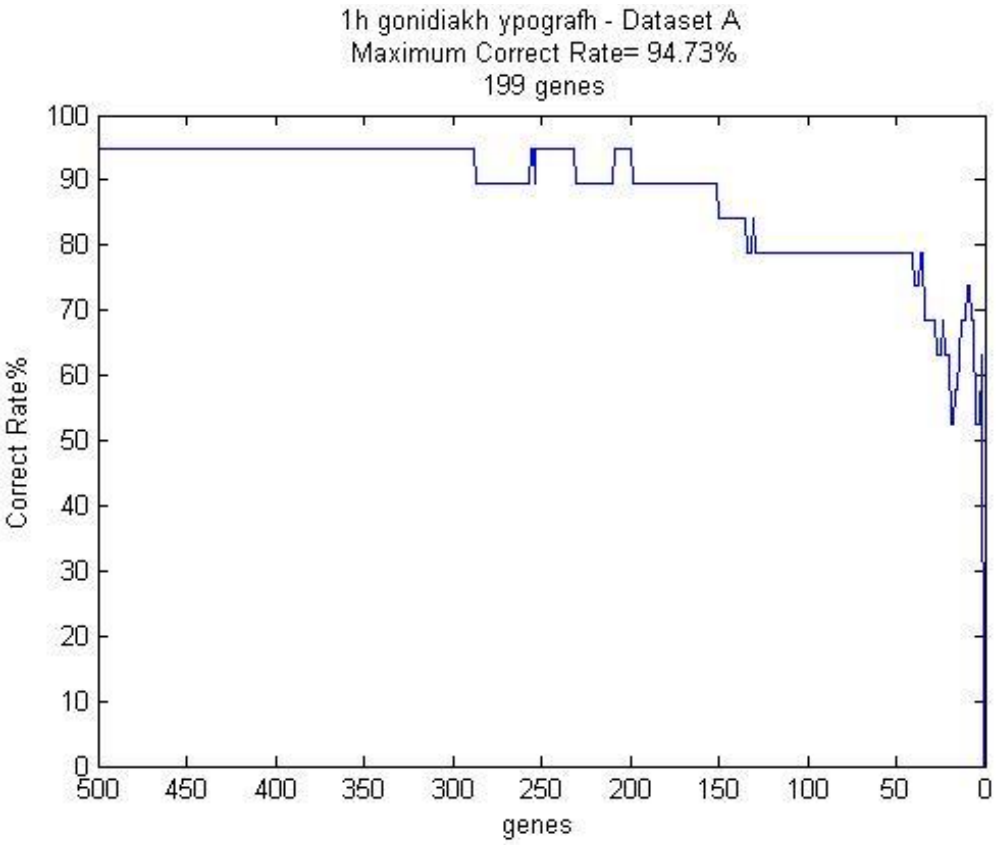
Στην Εικόνα 29 απεικονίζεται το διάγραμμα ροής της μεθοδολογίας που ακολουθήσαμε για τον υπολογισμό της $1^{ης}$ γονιδιακής υπογραφής (Βήμα 1-3).



Εικόνα 29: Αναλυτική απεικόνιση της μεθοδολογίας (Βήματα 1-3) για την εξαγωγή της 1ης Γονιδιακής Υπογραφής.

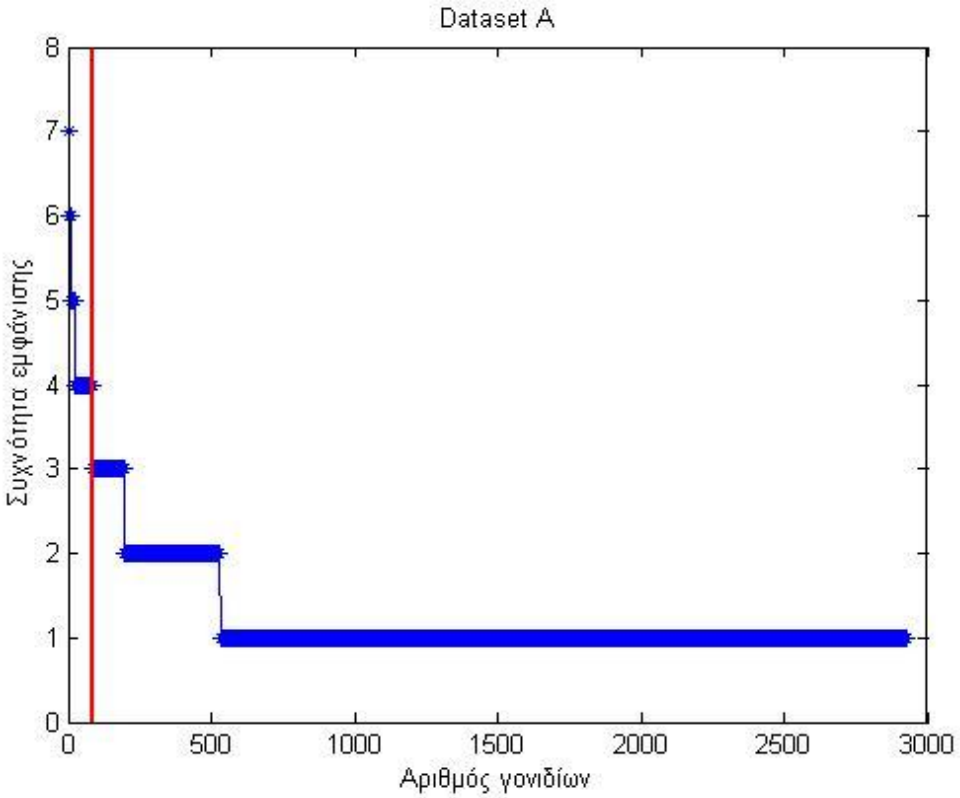
Dataset A

Για το Dataset A (m= 22283 γονίδια , n=19 δείγματα) η γραφική παράσταση του μέσου ποσοστού ακρίβειας A του SVM ταξινομητή για τα τελευταία 500 γονίδια φαίνεται στην Εικόνα 30.



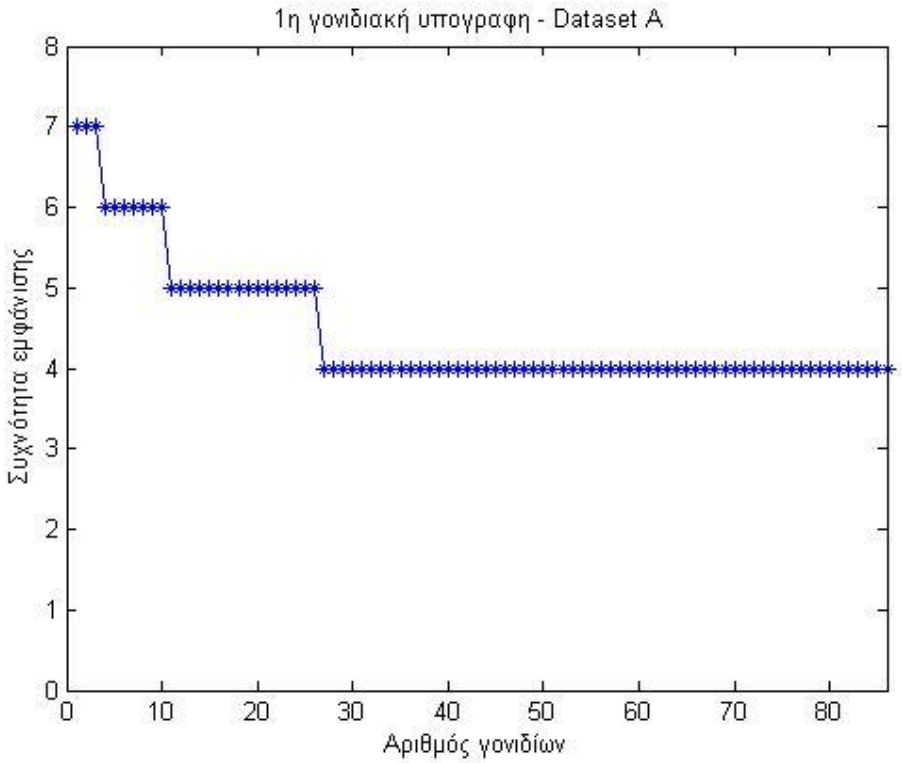
Εικόνα 30 : Γραφική παράσταση της ακρίβειας που σημείωσε ο liner SVM classifier καθώς ελαττώνουμε τα τελευταία 500 γονίδια του Dataset A σύμφωνα με τη μέθοδο RFE-LNW.

Παρατηρούμε ότι ο ταξινομητής επιτυγχάνει στα τελευταία 199 γονίδια ένα αρκετά υψηλό ποσοστό ακρίβειας, το οποίο είναι ίσο με 94.73%. Στη συνέχεια, σύμφωνα με το **Βήμα 3** της μεθοδολογίας που ακολουθήσαμε, συγκεντρώνουμε τα 199 γονίδια που αφαιρέθηκαν τελευταία από τον αλγόριθμο RFE-LNW για κάθε μια από τις 19 επαναλήψεις του ELOOCV. Κατ' αυτό τον τρόπο προκύπτουν 2927 γονίδια με τις συχνότητές εμφάνισης τους, οι οποίες παρουσιάζονται στην Εικόνα 31. Στο σημείο αυτό θα πρέπει να τονίσουμε ότι η μέγιστη τιμή συχνότητας ενός γονιδίου είναι ίση με 19, όσες δηλαδή και οι επαναλήψεις cross validation που πραγματοποιούνται.



Εικόνα 31 : Συχνότητα εμφάνισης των 2927 γονιδίων του Dataset A μέσα στις 19 επαναλήψεις του ELOOCV.

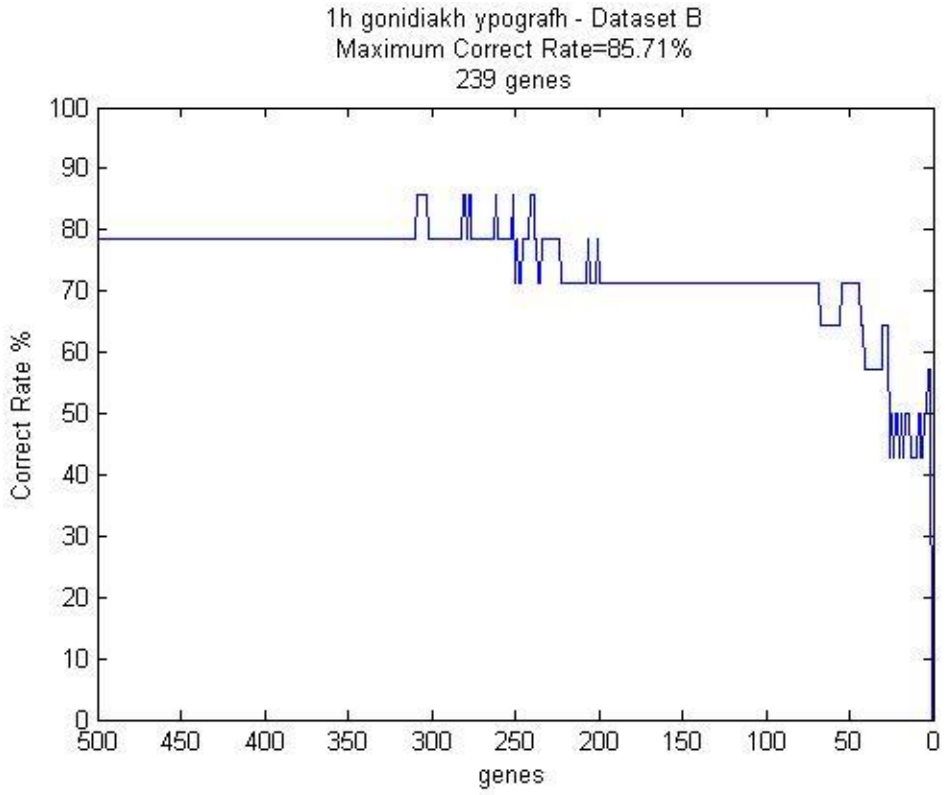
Όπως παρατηρούμε από τη γραφική παράσταση της Εικόνας 31 επιλέγουμε ως γονιδιακή υπογραφή τα 86 γονίδια (από τη κόκκινη γραμμή και αριστερά) με τις υψηλότερες συχνότητες. Τα γονίδια αυτά, τα οποία παρουσιάζονται με τις συχνότητές τους στην Εικόνα 32 σε μια νέα γραφική παράσταση, χρησιμοποιούνται, σύμφωνα με το **Βήμα 4** της μεθοδολογίας που ακολουθήσαμε, για την εκπαίδευση και έλεγχο ενός γραμμικού SVM ταξινομητή. Η ακρίβεια $Signature\_Accuracy$ του ταξινομητή, που κατασκευάστηκε από τα γονίδια της 1^{ης} Γονιδιακής Υπογραφής, υπολογίστηκε σύμφωνα με την Εξίσωση (4.2) και βρέθηκε ίση με **78.94%** .



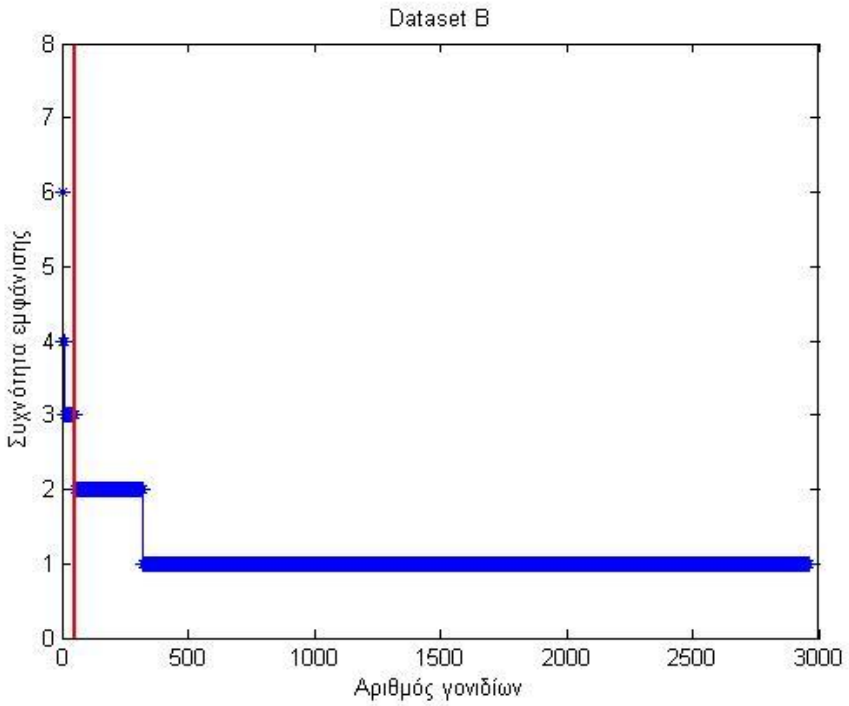
Εικόνα 32 : Τα 86 γονίδια τα οποία αποτελούν την 1η γονιδιακή υπογραφή για το Dataset A, απεικονίζονται με τις συχνότητες εμφάνισης τους.

Dataset B

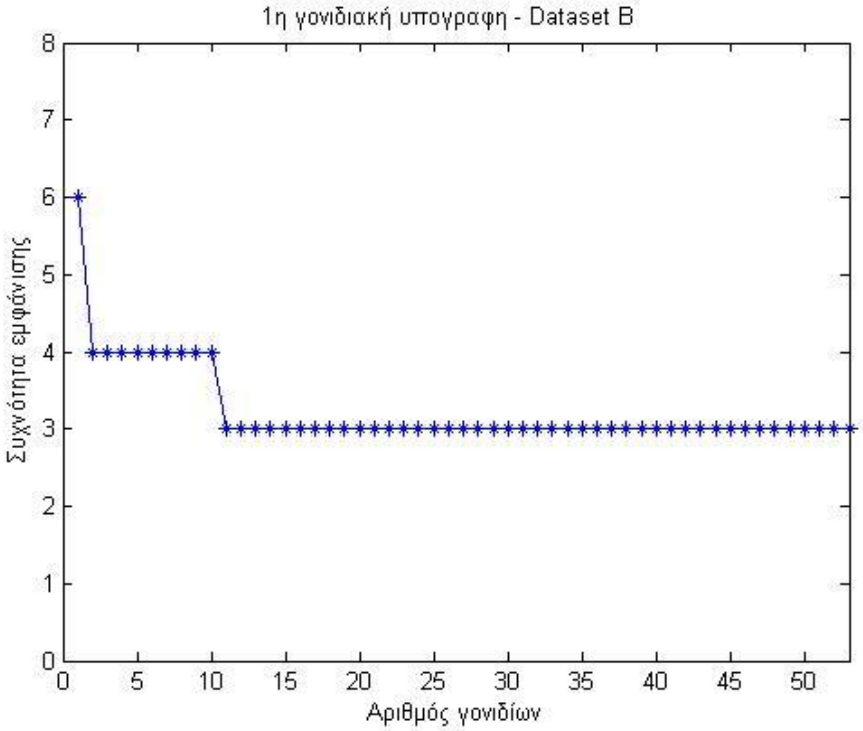
Ακολουθώντας την ίδια διαδικασία για το Dataset B (m=22645 γονίδια n=14 δείγματα) έχουμε αντίστοιχα ότι στα τελευταία 239 γονίδια που αφαιρούνται σύμφωνα με τον RFE-LNW σημειώνεται κατά μέσο όρο 85.71% ακρίβεια ταξινόμησης (Εικόνα 33). Στη συνέχεια, συγκεντρώνουμε τα 239 γονίδια που αφαιρέθηκαν τελευταία στις 14 επαναλήψεις του ELOOCV και καταλήγουμε σε 2963 γονίδια (Εικόνα 34). Στην περίπτωση του Dataset B η μέγιστη τιμή συχνότητας εμφάνισης ενός γονιδίου είναι ίση με 14. Από τα 2963 γονίδια επιλέγουμε πάλι εκείνα με τις υψηλότερες συχνότητες εμφάνισης , οπότε η 1^{η} Γονιδιακή Υπογραφή για το Dataset B αποτελείται από 53 γονίδια (Εικόνα 35). Τέλος ο ταξινομητής που κατασκευάζεται από τα γονίδια της υπογραφής σημειώνει ακρίβεια ταξινόμησης ίση με **64.28%**.



Εικόνα 33 : Γραφική παράσταση της ακρίβειας που σημείωσε ο liner SVM classifier καθώς ελαττώνουμε τα τελευταία 500 γονίδια του Dataset B σύμφωνα με τη μέθοδο RFE-LNW.



Εικόνα 34 : Συχνότητα εμφάνισης των 2963 γονιδίων του Dataset B μέσα στις 14 επαναλήψεις του ELOOCV. Ως γονιδιακή υπογραφή επιλέγουμε τα γονίδια με τις υψηλότερες συχνότητες (από τη κόκκινη γραμμή και αριστερά).



Εικόνα 35 : Τα 53 γονίδια τα οποία αποτελούν την 1^{η} γονιδιακή υπογραφή για το Dataset B, απεικονίζονται με τις συχνότητες εμφάνισης τους.

4.3 Δεύτερη Γονιδιακή Υπογραφή

Για τον υπολογισμό της δεύτερης γονιδιακής υπογραφής ακολουθήσαμε παρόμοια βήματα με τα αντίστοιχα για την πρώτη υπογραφή. Η σημαντική διαφορά βρίσκεται στην χρήση διαφορετικού αλγορίθμου επιλογής των επικρατέστερων γονιδίων. Στη συγκεκριμένη περίπτωση χρησιμοποιήσαμε τη μέθοδο LASSO, όπως αυτή περιγράφηκε στο Κεφάλαιο 3.6.6. Οπότε μπορούμε να πούμε ότι τα Βήματα 1 και 4 που αφορούν το διαχωρισμό του συνόλου δεδομένων και την αξιολόγηση των τελικών γονιδίων που συνιστούν τη $2^{η}$ γονιδιακή υπογραφή, εφαρμόζονται με τον ίδιο τρόπο όπως και προηγουμένως. Η βασική διαφορά παρατηρείται στα Βήματα 2 και 3 τα οποία σχετίζονται με τη δημιουργία υποσυνόλων επικρατέστερων γονιδίων και την επιλογή του κατάλληλου υποσυνόλου ως γονιδιακή υπογραφή αντίστοιχα. Συνεπώς τα Βήματα 2 και 3 διαμορφώνονται ως εξής:

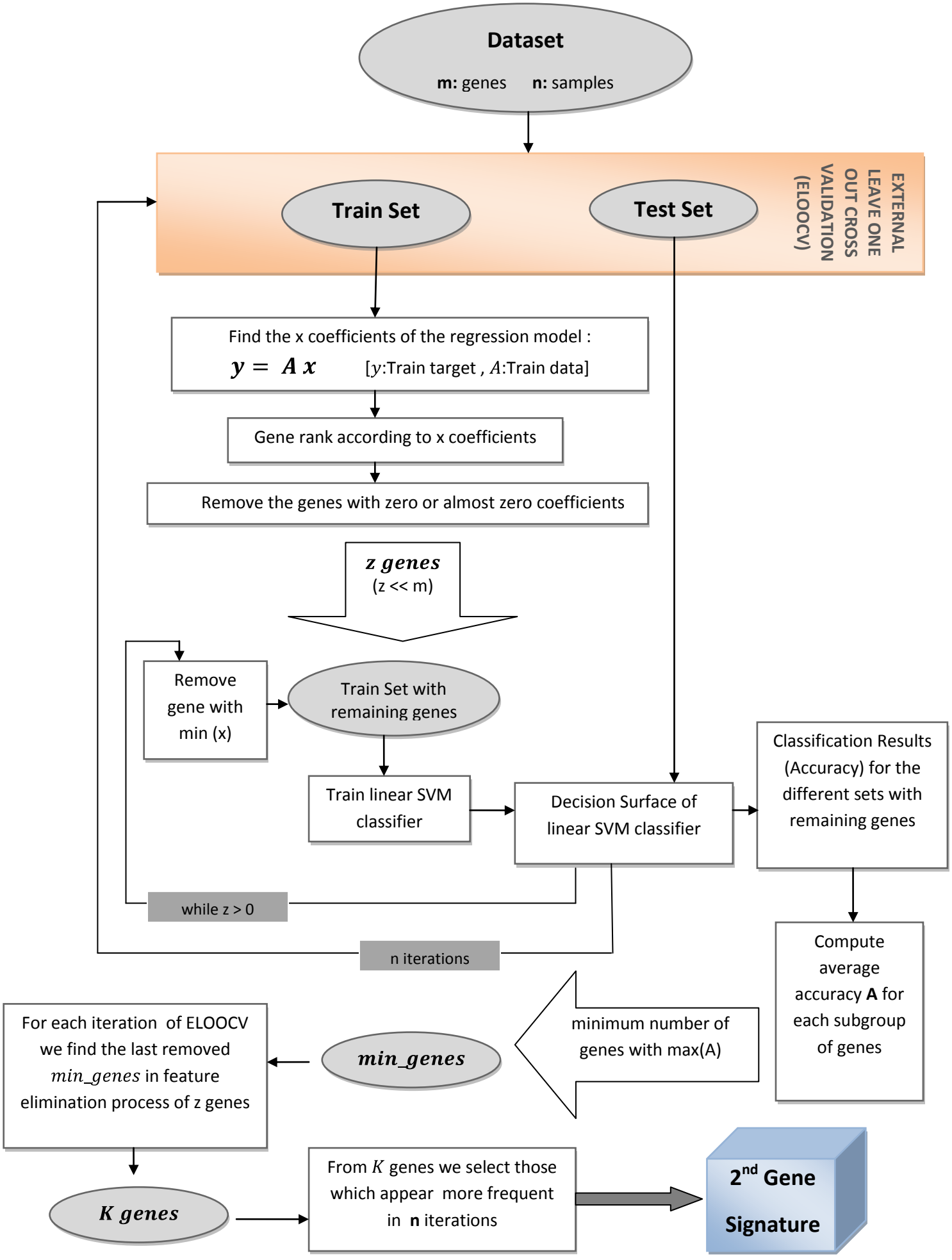
Βήμα 2 - Δημιουργία υποσυνόλων γονιδίων: Όπως έχουμε αναφέρει σε προηγούμενα Κεφάλαια, ο αλγόριθμος LASSO προσπαθεί να λύσει το l_1 regularized least squares πρόβλημα (LSP) : $minimize\ ||Ax - y||_2^2 + \lambda||x||_1$, με σκοπό να βρει τα γονίδια εκείνα για τα οποία παράγονται μη μηδενικοί συντελεστές x. Στην εργασία μας για την λύση του παραπάνω προβλήματος βελτιστοποίησης χρησιμοποιήσαμε ένα υπάρχον matlab solver [63] αντικαθιστώντας όπου A το σύνολο δεδομένων εκπαίδευσης (train set) που προέκυψε από το ELOOCV, και όπου y τα targets (+1,-1) των δειγμάτων που περιλαμβάνονται στο train set. Όσο αναφορά τη ρυθμιστική παράμετρο λ, έγιναν πολλές δοκιμές ώστε να καταλήξουμε σε μια τιμή η οποία θα δίνει μια συγκρίσιμη γονιδιακή υπογραφή σε σχέση με την $1^{η}$ και την $3^{η}$ τόσο στο μέγεθος όσο και στην ακρίβεια ταξινόμησης. Έτσι χρησιμοποιήσαμε τη ρυθμιστική παράμετρο $\lambda = 0.5$ και για τα δύο Dataset (A,B) που επεξεργαστήκαμε. Έχοντας υπολογίσει, λοιπόν, με τη βοήθεια του matlab solver τις τιμές των συντελεστών x για ένα συγκεκριμένο train set, κατατάσσουμε τα γονίδια σε φθίνουσα σειρά ανάλογα με τη τιμή των συντελεστών (μεγάλη τιμή→μικρή τιμή). Στη συνέχεια τα γονίδια με μηδενικούς ή σχεδόν μηδενικούς συντελεστές x διαγράφονται τόσο από το train set όσο και από το test set. Οι ανανεωμένοι πίνακες των train set και test set (οι οποίοι περιέχουν z γονίδια, έναν αριθμό πολύ μικρότερο από τον αρχικό αριθμό των m γονιδίων) χρησιμοποιούνται για την εκπαίδευση και την εκτίμηση της ακρίβειας ταξινόμησης του γραμμικού SVM classifier. Η διαδικασία εκπαίδευσης και αξιολόγησης του ταξινομητή επαναλαμβάνεται αφαιρώντας κάθε φορά από τα train και test set το γονίδιο εκείνο με το μικρότερο συντελεστή x, μέχρις ότου αφαιρεθούν όλα τα z

γονίδια. Με άλλα λόγια, εκτελούνται τόσες επαναλήψεις όσες και ο αριθμός των γονιδίων με μη μηδενικούς συντελεστές x.

Στο Βήμα 2 η διαδικασία υπολογισμού των συντελεστών x επαναλαμβάνεται n φορές, όσες και ο αριθμός επαναλήψεων του ELOOCV, έτσι ώστε για κάθε καινούριο train set που προκύπτει από τη μέθοδο cross validation να αποδοθούν και οι αντίστοιχοι συντελεστές. Όταν ολοκληρωθούν οι n επαναλήψεις προχωράμε στο Βήμα 3:

Βήμα 3 - Επιλογή γονιδίων : Στο στάδιο αυτό υπολογίζουμε το μέσο ποσοστό ακρίβειας του ταξινομητή σύμφωνα με την Εξίσωση (4.1), όπως ακριβώς και για τη 1η γονιδιακή υπογραφή. Από το διάνυσμα A_i βρίσκουμε τον ελάχιστο αριθμό γονιδίων, έστω min_genes, στον οποίον επιτυγχάνεται η μέγιστη μέση ακρίβεια. Για κάθε μια τις n επαναλήψεις του cross validation συγκεντρώνουμε σε ένα πίνακα τα min_genes που αφαιρέθηκαν τελευταία στην επαναληπτική διαδικασία για την εκπαίδευση και αξιολόγηση του ταξινομητή (K genes). Υπολογίζουμε πόσο συχνά εμφανίζονται μέσα στις n επαναλήψεις και τελικά επιλέγουμε εκείνα που παρουσιάζουν τις υψηλότερες συχνότητες.

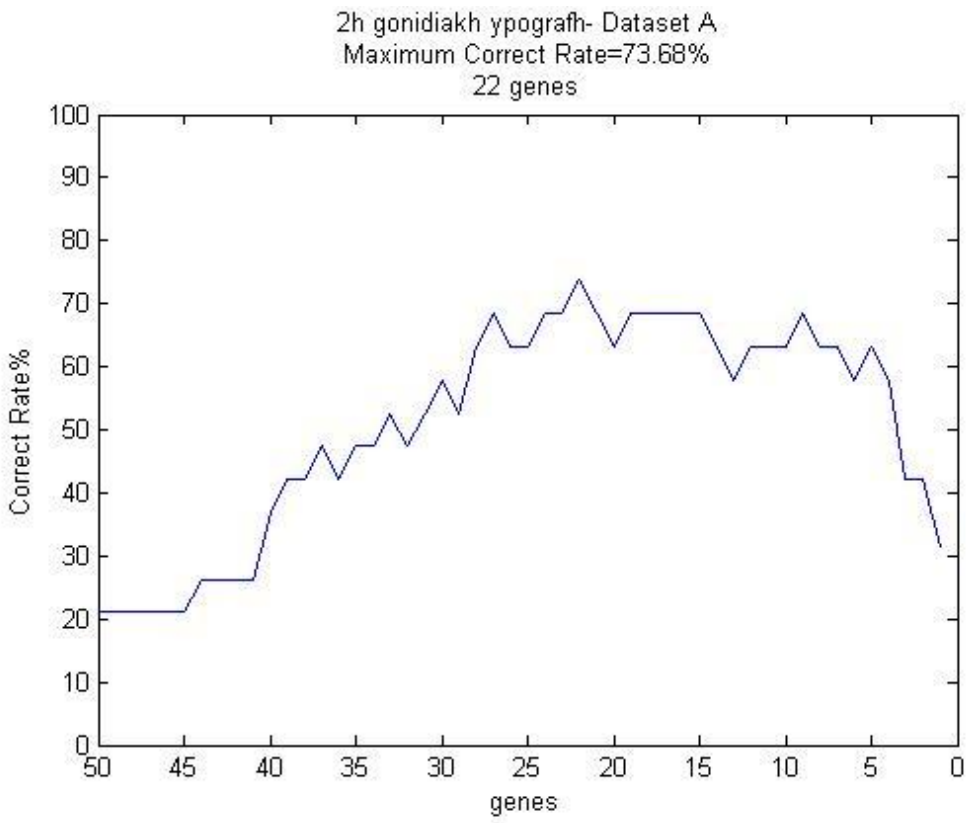
Η μεθοδολογία που ακολουθήθηκε για τον υπολογισμό της 2ης γονιδιακής υπογραφής (Βήματα 1-3) παρουσιάζεται ως διάγραμμα ροής στην Εικόνα 36.



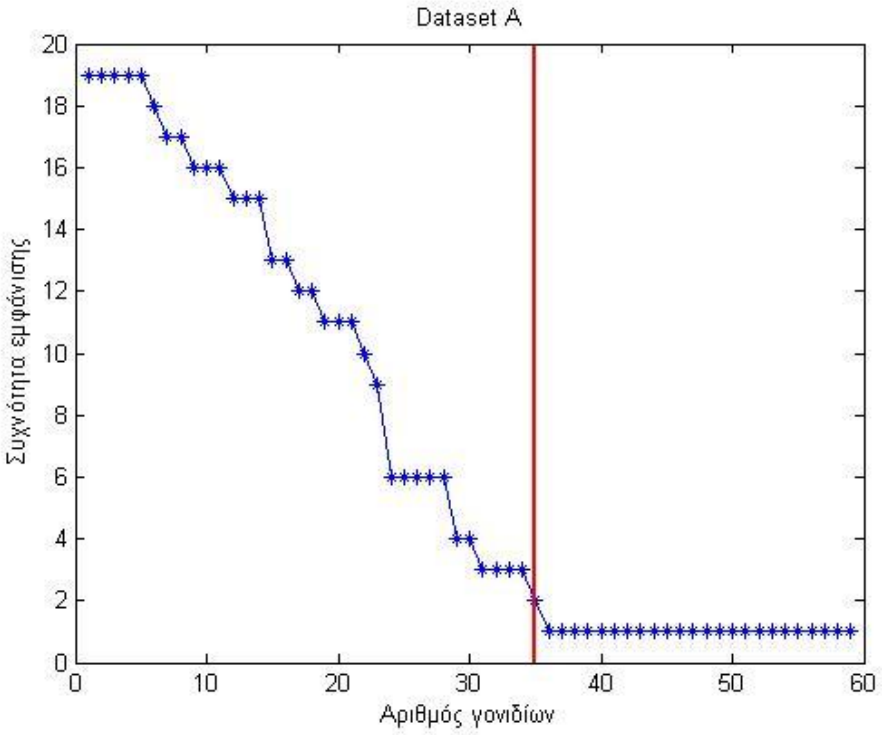
Εικόνα 36 : Αναλυτική απεικόνιση της μεθοδολογίας (Βήματα 1-3) για την εξαγωγή της 2ης Γονιδιακής Υπογραφής.

Dataset A

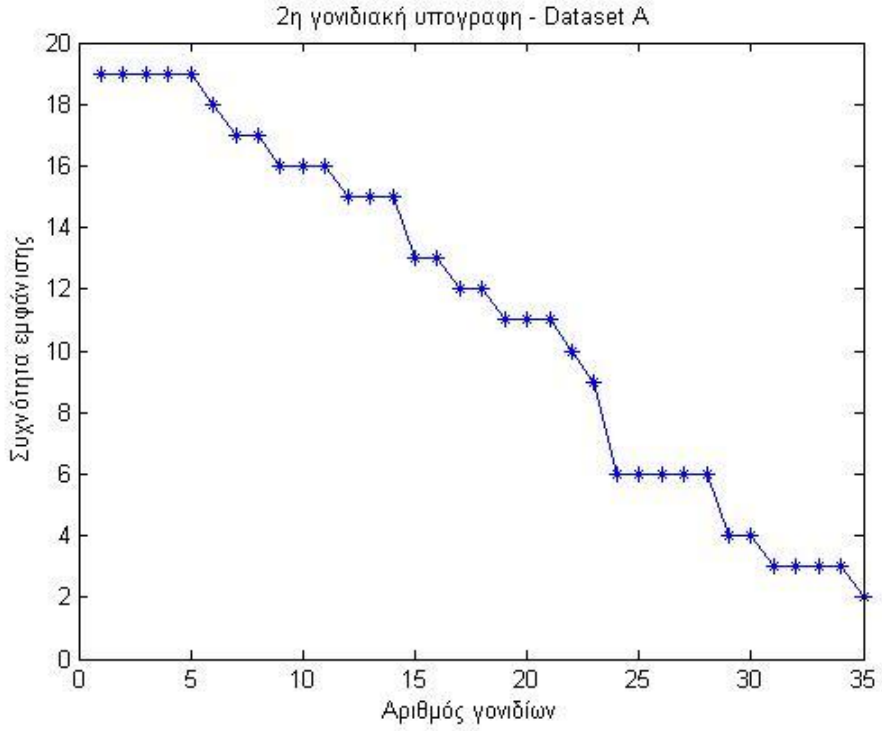
Η γραφική παράσταση του μέσου ποσοστού ακρίβειας Α του SVM ταξινομητή για την περίπτωση της δεύτερης μεθόδου που ακολουθήσαμε φαίνεται στην Εικόνα 37. Παρατηρούμε ότι για 22 γονίδια επιτυγχάνεται η μέγιστη τιμή ακρίβειας, η οποία είναι ίση με 73.68%. Συγκεντρώνοντας τα 22 γονίδια που αφαιρέθηκαν τελευταία και τις συχνότητες εμφάνισης τους μέσα στις 19 επαναλήψεις, προκύπτει το διάγραμμα της Εικόνας 38, το οποίο περιλαμβάνει 59 γονίδια με μέγιστη συχνότητα εμφάνισης τις 19 φορές και ελάχιστη τη μια φορά. Επιλέγουμε ως $2^{η}$ Γονιδιακή υπογραφή για το Dataset A τα 35 γονίδια με τις πιο υψηλές συχνότητες. Τέλος, ο γραμμικός SVM ταξινομητής που κατασκευάζεται από τα 35 γονίδια της υπογραφής (Εικόνα 39) επιτυγχάνει **84.21%** ακρίβεια ταξινόμησης.



Εικόνα 37 : Γραφική παράσταση της μέσης ακρίβειας που σημείωσε ο liner SVM classifier στις 19 επαναλήψεις καθώς ελαττώνουμε τα τελευταία 50 πιο σημαντικά γονίδια του Dataset A.



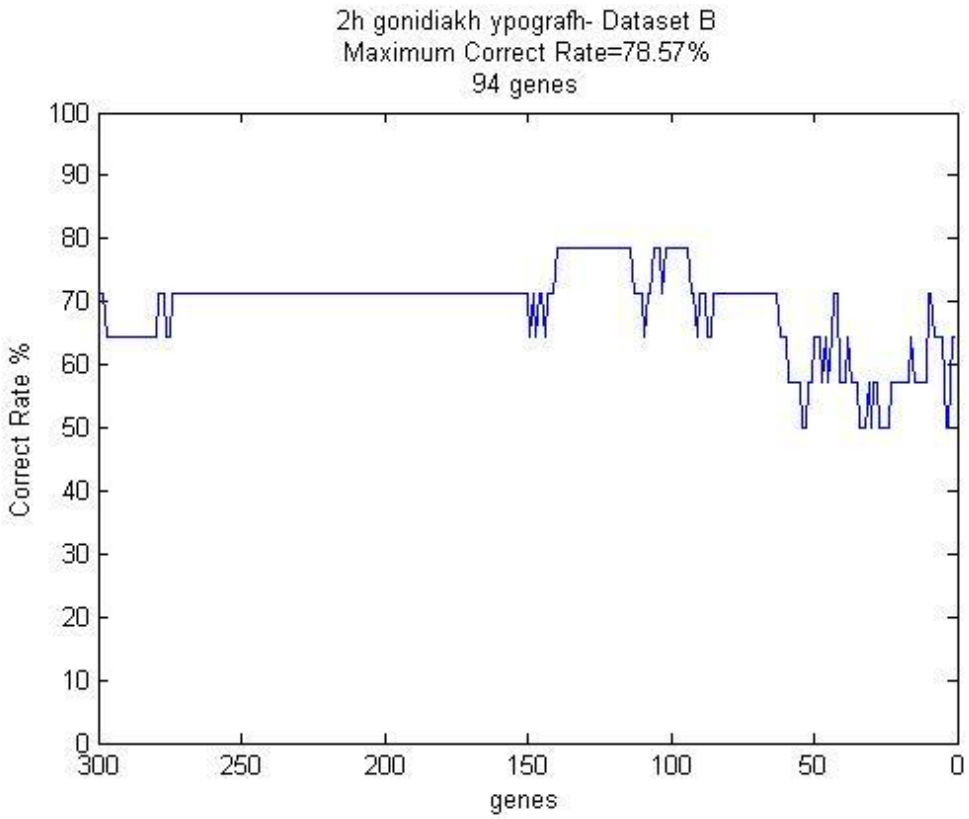
Εικόνα 38 : Συχνότητα εμφάνισης των 59 γονιδίων του Dataset A μέσα στις 19 επαναλήψεις του ELOOCV. Ως γονιδιακή υπογραφή επιλέγουμε τα γονίδια με τις υψηλότερες συχνότητες (από τη κόκκινη γραμμή και αριστερά).



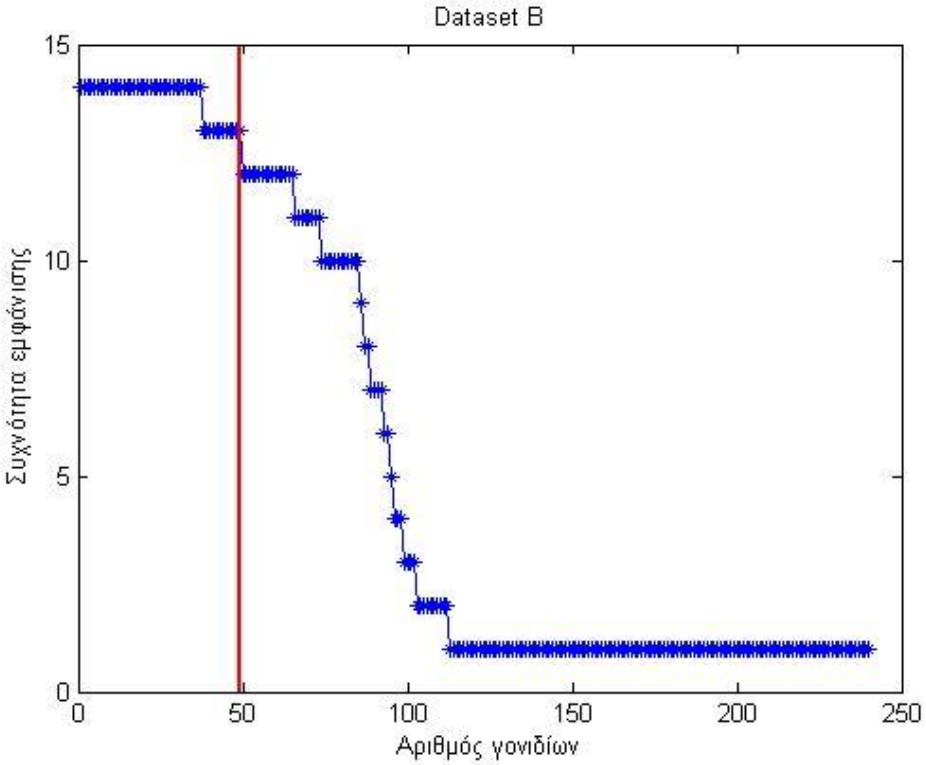
Εικόνα 39 : Τα 35 γονίδια τα οποία αποτελούν την 2^{η} γονιδιακή υπογραφή για το Dataset A, απεικονίζονται με τις συχνότητες εμφάνισης τους.

Dataset B

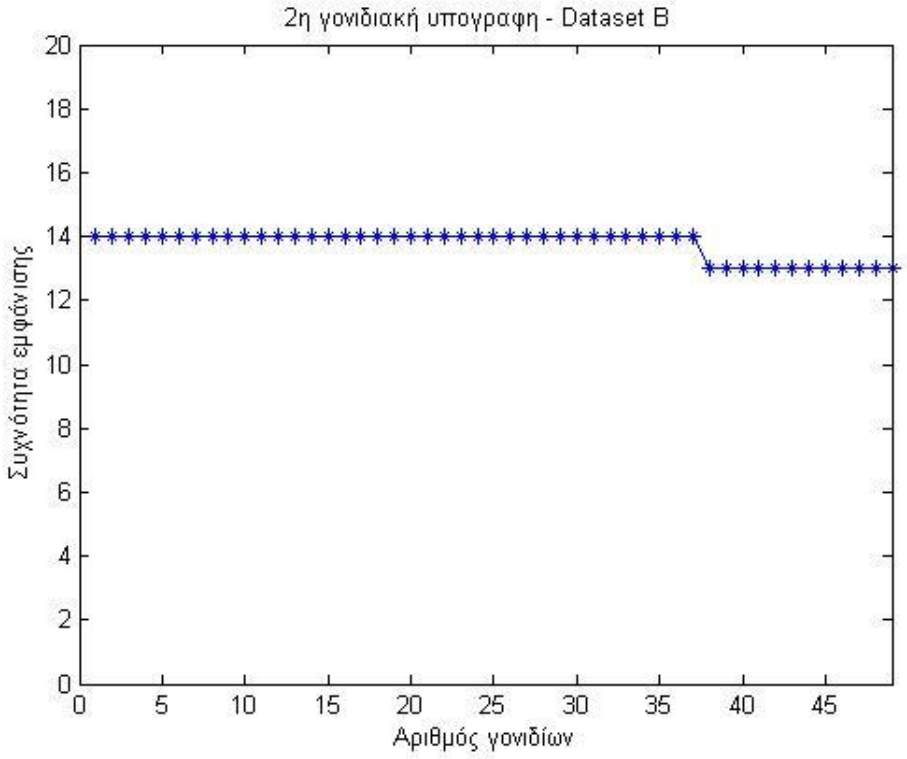
Ομοίως για το Dataset B έχουμε ότι στα 94 γονίδια επιτυγχάνεται ένα παρόμοιο ποσοστό μέγιστης ακρίβειας το οποίο είναι ίσο με 78.57% (Εικόνα 40). Το πλήθος των τελευταίων 94 γονιδίων που εμφανίζονται στις 14 επαναλήψεις του cross validation είναι ίσο με 240 (Εικόνα 41). Από τα 240 γονίδια επιλέγουμε ως $2^{η}$ γονιδιακή υπογραφή για το Dataset B 49 γονίδια (Εικόνα 42), με τα οποία ο ταξινομητής που κατασκευάζεται σημειώνει **85.71%** επιτυχία ταξινόμησης.



Εικόνα 40 : Γραφική παράσταση της μέσης ακρίβειας που σημείωσε ο liner SVM classifier στις 14 επαναλήψεις καθώς ελαττώνουμε τα τελευταία 300 πιο σημαντικά γονίδια του Dataset B.



Εικόνα 41 : Συχνότητα εμφάνισης των 240 γονιδίων του Dataset B μέσα στις 14 επαναλήψεις του ELOOCV. Ως γονιδιακή υπογραφή επιλέγουμε τα γονίδια με τις υψηλότερες συχνότητες (από τη κόκκινη γραμμή και αριστερά).



Εικόνα 42 : Τα 49 γονίδια τα οποία αποτελούν την $2^{η}$ γονιδιακή υπογραφή για το Dataset B, απεικονίζονται με τις συχνότητες εμφάνισης τους.

4.4 Τρίτη Γονιδιακή Υπογραφή

Η τρίτη μεθοδολογία που ακολουθήσαμε για εξαγωγή γονιδιακών υπογραφών χωρίζεται σε 2 στάδια και βασίζεται στο συνδυασμό των αλγορίθμων RFE-LNW και του νευρωνικού FSMLP. Θα πρέπει να τονίσουμε ότι και σε αυτή την περίπτωση η μεθοδολογία που ακολουθείται εφαρμόζεται με τον ίδιο τρόπο τόσο στο Dataset A όσο και στο Dataset B.

Αρχικά, με την βοήθεια του αλγορίθμου RFE-LNW προσπαθούμε να ελαττώσουμε τον αριθμό των m γονιδίων (τάξης των χιλιάδων) που περιέχονται στο διαθέσιμο σύνολο δεδομένων και όχι να επιλέξουμε μια γονιδιακή υπογραφή, όπως στη μεθοδολογία της ενότητας 4.2 ($1^{η}$ γονιδιακή υπογραφή). Στη συνέχεια, το μειωμένο σύνολο γονιδίων που προκύπτει εισέρχεται ως είσοδος σε ένα FSMLP νευρωνικό δίκτυο (Κεφάλαιο 3.3.5), το οποίο εκπαιδεύεται σύμφωνα με τον αλγόριθμο Back – Propagation και το σύνολο γονιδίων που προκύπτει μετά την εκπαίδευση εξετάζεται ως προς την ικανότητα του να κατασκευάσει έναν αποδοτικό SVM ταξινομητή.

Στο σημείο αυτό πρέπει να επισημάνουμε ότι η χρήση του αλγορίθμου RFE-LNW ως ένα ενδιάμεσο στάδιο επεξεργασίας των δεδομένων, μας επιτρέπει να μειώσουμε το χρόνο που απαιτείται για την εκπαίδευση του FSMLP δικτύου, αφού ελαττώνοντας τον αρχικό αριθμό των m γονιδίων μειώνεται σημαντικά και η χρονική καθυστέρηση της επεξεργασίας των δεδομένων από το FSMLP δίκτυο.

Παρακάτω περιγράφουμε αναλυτικά τα 2 στάδια επεξεργασίας των δεδομένων με τα επιμέρους βήματα που περιλαμβάνει το κάθε ένα από αυτά.

$1^{ο}$ Στάδιο

Σκοπός μας σε αυτή τη φάση είναι να δημιουργήσουμε ένα μικρότερο υποσύνολο γονιδίων για περαιτέρω επεξεργασία. Είναι σημαντικό, όμως, το υποσύνολο που θα προκύψει να διαθέτει σημαντικά γονίδια, ικανά να κατασκευάσουν στη συνέχεια (Στάδιο 2) έναν αποδοτικό ταξινομητή. Για αυτό το λόγο το $1^{ο}$ στάδιο επεξεργασίας του συνόλου δεδομένων παρουσιάζει αρκετές ομοιότητες με τα βήματα που ακολουθήσαμε για τον υπολογισμό της $1^{ης}$ γονιδιακής υπογραφής.

Βήμα 1 - Διαχωρισμός Δεδομένων : Για να μειώσουμε το υπολογιστικό κόστος της επεξεργασίας των δεδομένων δεν χωρίσαμε το αρχικό μας Dataset (A ή B) όπως στις προηγούμενες δύο υπογραφές με τη μέθοδο ELOOCV, αλλά με τη βοήθεια της τεχνικής

external 10-fold cross validation (Κεφάλαιο 2.3), σύμφωνα με την οποία τα n δείγματα του Dataset χωρίζονται σε 10 υποσύνολα (folds). Το κάθε υποσύνολο διαθέτει 1 ή 2 δείγματα (τόσο για το Dataset A $(19/10 = 1.9)$ όσο και για το Dataset B $(14/10 = 1.4)$). Εκτελούνται 10 επαναλήψεις του CV και σε κάθε επανάληψη χρησιμοποιείται ακριβώς μια φορά ένα από τα 10 υποσύνολα ως test set και τα υπόλοιπα 9 υποσύνολα ως train set (με 18 ή 17 δείγματα για το Dataset A και 14 ή 13 για το Dataset B).

Βήμα 2 - Δημιουργία υποσυνόλων γονιδίων: Για κάθε μια από τις 10 επαναλήψεις του CV, ο αλγόριθμος RFE-LNW εφαρμόζεται με τον ίδιο τρόπο όπως στην μεθοδολογία της 1ης Γονιδιακής υπογραφής. Θυμίζουμε ότι ο αλγόριθμος επαναλαμβάνεται μέχρις ότου αφαιρεθούν όλα τα m γονίδια από το train και test set ($while\ m \geq 0$). Η ακρίβεια $Save\_All_{ij}$ του γραμμικού SVM ταξινομητή που κατασκευάζεται στο i-οστό βήμα του RFE-LNW από το i-οστό υποσύνολο δεδομένων εκπαίδευσης, εκτιμάται για το i-οστό υποσύνολο δεδομένων ελέγχου από τη σχέση (3.103).

Βήμα 3 - Επιλογή γονιδίων : Μετά την ολοκλήρωση των 10 επαναλήψεων του CV, για κάθε επαναληπτικό κύκλο υπολογίζουμε το μικρότερο i-οστό υποσύνολο γονιδίων στο οποίο επιτυγχάνεται μέγιστη ακρίβεια ταξινόμησης. Οπότε προκύπτουν 10 αριθμοί ελάχιστων γονιδίων, ένας για κάθε fold. Υπολογίζουμε το μέσο όρο των 10 αριθμών, έστω G, και εν συνεχεία εκτελούμε 10 επαναλήψεις των Βημάτων 1, 2. Τώρα ως κριτήριο τερματισμού του αλγορίθμου RFE-LNW χρησιμοποιούμε τη συνθήκη $while\ m > G$, διατηρώντας με αυτό τον τρόπο σε κάθε μια από τις 100 επαναλήψεις του RFE-LNW (= 10 επαναλήψεις x 10 fold CV) τα G τελευταία γονίδια. Τα γονίδια που δεν εξαλείφονται σε κάθε επανάληψη εισάγονται σε έναν πίνακα μαζί με τη συχνότητα εμφάνισης τους μέσα στις 100 επαναλήψεις. Το σημείο εκπαίδευσης και εκτίμησης του SVM ταξινομητή παραλείπεται γιατί βρισκόμαστε σε ένα ενδιάμεσο στάδιο με σκοπό την συγκέντρωση και όχι την αξιολόγηση ενός υποσυνόλου γονιδίων. Μετά την ολοκλήρωση των 100 επαναλήψεων του RFE-LNW αλγορίθμου ελέγχουμε τον πίνακα με τα εναπομείναντα γονίδια (έστω N genes) και επιλέγουμε εκείνα με την μεγαλύτερη συχνότητα εμφάνισης. Τα γονίδια στα οποία καταλήγουμε αποτελούν το μειωμένο σύνολο δεδομένων προς επεξεργασία για το 2ο Στάδιο.

Εφαρμόζοντας το 1ο στάδιο της μεθοδολογίας τόσο στο Dataset A όσο και στο Dataset B καταλήξαμε στα εξής:

- Ο μέσος όρος των γονιδίων που παρουσίασε την μεγαλύτερη ακρίβεια ταξινόμησης προέκυψε ίσος με G = 5 και για τα δύο σύνολα δεδομένων. Οπότε το κριτήριο τερματισμού του αλγορίθμου RFE-LNW στις 100 επαναλήψεις μετατράπηκε στο $while\ m > 5$.
- Το πλήθος των γονιδίων που συγκεντρώθηκε μετά το τέλος των 100 επαναλήψεων ήταν σχετικά πολύ μικρό σε σχέση με το αρχικό σύνολο δεδομένων. Συγκεκριμένα για το Dataset A συγκεντρώσαμε 332 γονίδια ενώ για το Dataset B συγκεντρώσαμε 300 γονίδια. Για το λόγο αυτό προτιμήσαμε να μην εξαλείψουμε κάποιο από τα 332 ή 300 γονίδια ακόμα και αν η συχνότητα εμφάνισής του μέσα στις 100 επαναλήψεις ήταν αρκετά μικρή.

Τα νέα μειωμένα σύνολα δεδομένων που χρησιμοποιούμε ως είσοδο στο νευρωνικό FSMLP είναι :

1) **Reduced_Dataset A**, το οποίο περιλαμβάνει m_{new} =332 γονίδια και n =19 δείγματα
2) **Reduced_Dataset B**, το οποίο περιλαμβάνει m_{new} =300 γονίδια και n =14 δείγματα

2ο Στάδιο

Βήμα 1 - Διαχωρισμός Δεδομένων : Στο δεύτερο αυτό στάδιο επεξεργασίας των δεδομένων διαχωρίσαμε το Reduced_Dataset σε train και test set σύμφωνα με τη τεχνική ELOOCV .

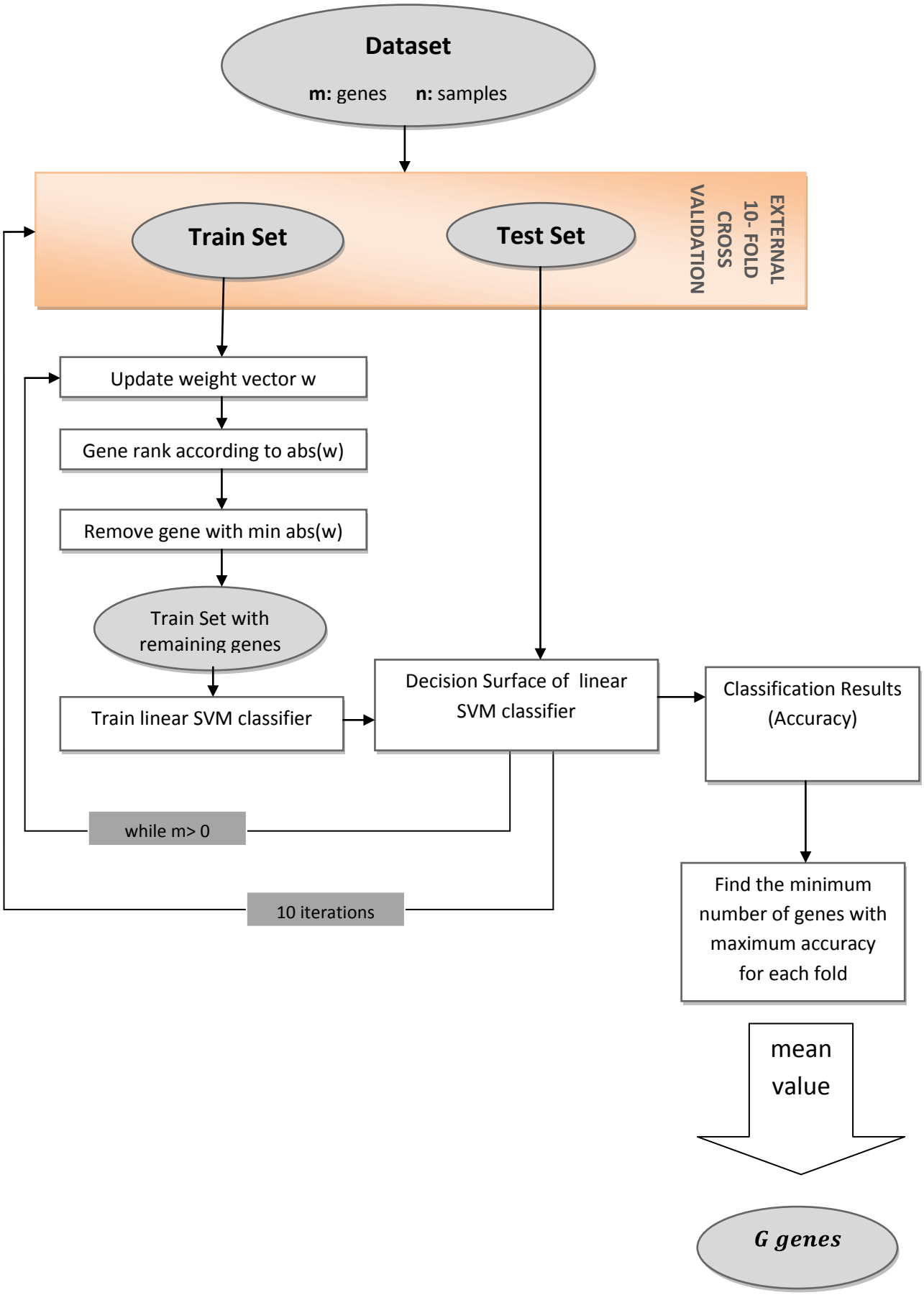
Βήμα 2 - Δημιουργία υποσυνόλων γονιδίων: Ως μέθοδο επιλογής γονιδίων για την 3η γονιδιακή υπογραφή χρησιμοποιήσαμε το νευρωνικό FSMLP. Το συγκεκριμένο MLP δίκτυο συνδέει κάθε γονίδιο στο επίπεδο εισόδου με ένα μηχανισμό ή καλύτερα με μια συνάρτηση πύλης, η οποία καθορίζει το βαθμό διάδοσης του γονιδίου στα υπόλοιπα επίπεδα του δικτύου. Στην αρχή της εκπαίδευσης οι πύλες είναι κλειστές για όλα τα γονίδια. Καθώς η διαδικασία προχωρά οι πύλες ανοίγουν ανάλογα με το αν τα γονίδια με τα οποία συνδέονται μειώνουν το σφάλμα εκπαίδευσης. Επομένως η συνάρτηση πύλης παράγει υψηλές τιμές (κοντά στο 1) για τα σημαντικά γονίδια και χαμηλές (κοντά στο 0) για τα μη σημαντικά.

Το FSMLP δίκτυο που χρησιμοποιήσαμε διαθέτει m_{new} κόμβους εισόδου, 30 κόμβους στο κρυφό επίπεδο και 1 κόμβο εξόδου. Η επιλογή για τον αριθμό των κόμβων του κρυφού επιπέδου πραγματοποιήθηκε μετά από κάποιες δοκιμές. Παρόλα αυτά σκοπός μας δεν είναι να βρούμε ένα βέλτιστο δίκτυο για πρόβλεψη, αλλά ένα σύνολο από χρήσιμα γονίδια. Για το λόγο αυτό θεωρούμε ότι η επιλογή για τον αριθμό των κόμβων του κρυφού επιπέδου δεν αποτελεί κρίσιμο ζήτημα. Η εκπαίδευση του FSMLP δικτύου σταματά όταν το σφάλμα ταξινόμησης φτάσει το μηδέν ή όταν ολοκληρωθεί ο αριθμός των 5000 επαναλήψεων. Μετά το τέλος της εκπαίδευσης του δικτύου κατατάσσουμε τα γονίδια σε φθίνουσα σειρά (σημαντικά→μερικώς σημαντικά→μη σημαντικά) ανάλογα με τη τιμή της συνάρτησης πύλης με την οποία συνδέονται. Το ή τα γονίδια εκείνα με τη μικρότερη τιμή αφαιρούνται τόσο από το train set όσο και από το test set. Οι ανανεωμένοι πίνακες των train set και test set χρησιμοποιούνται για την εκπαίδευση και την εκτίμηση της ακρίβειας ταξινόμησης του γραμμικού SVM classifier. Η διαδικασία εκπαίδευσης και αξιολόγησης του ταξινομητή επαναλαμβάνεται αφαιρώντας κάθε φορά από τα train και test set το ή τα γονίδια εκείνα με τη μικρότερη τιμή της συνάρτησης πύλης, μέχρις ότου αφαιρεθούν όλα τα m_{new} γονίδια.

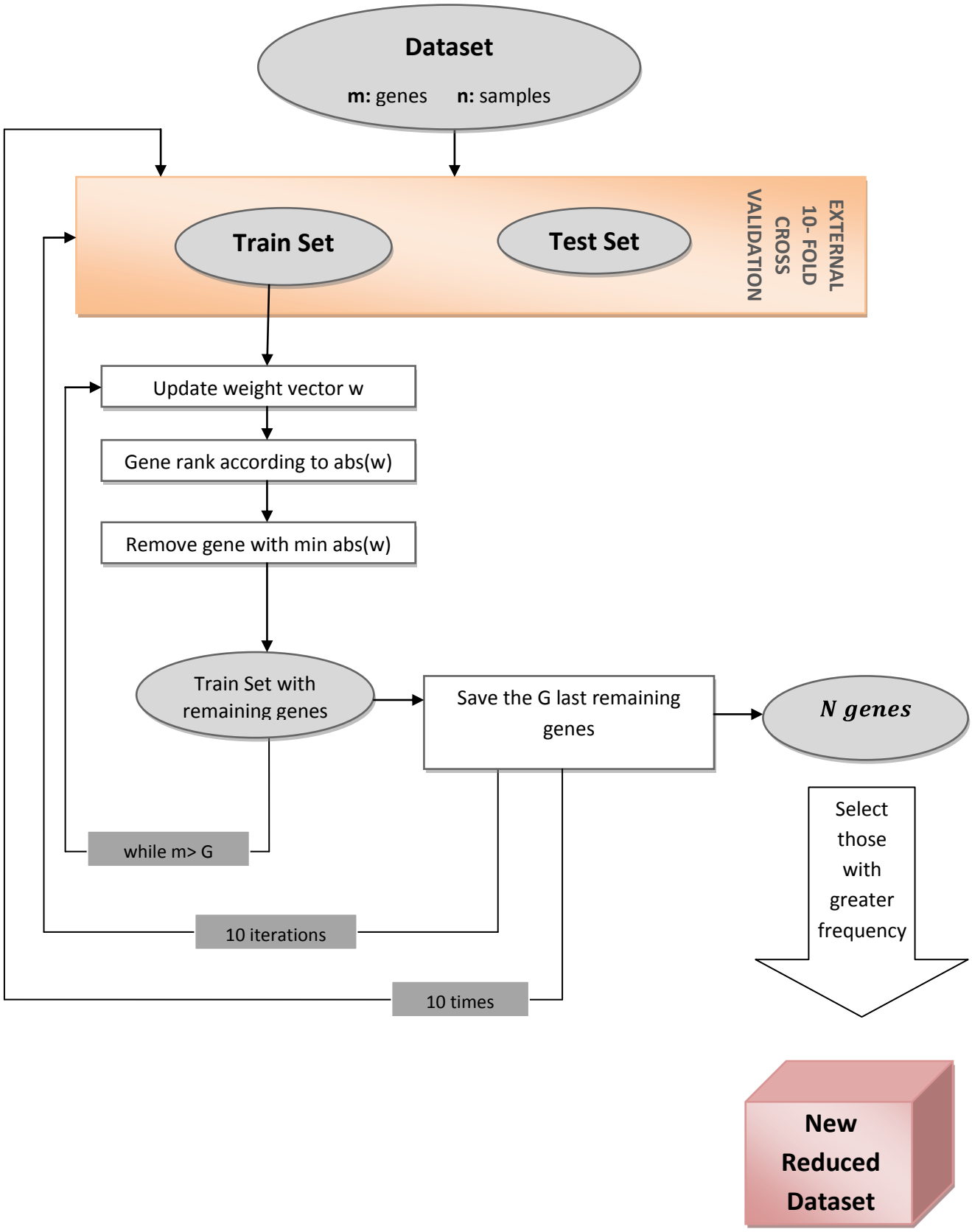
Βήμα 3 - Επιλογή γονιδίων : Μετά την ολοκλήρωση των n επαναλήψεων του ELOOCV υπολογίζουμε το μέσο ποσοστό ακρίβειας του ταξινομητή και τον ελάχιστο αριθμό γονιδίων, min_genes, στον οποίον επιτυγχάνεται η μέση μέγιστη ακρίβεια. Τα min_genes αποτελούν τα γονίδια που αφαιρέθηκαν τελευταία από την επαναληπτική διαδικασία εκπαίδευσης και αξιολόγησης του SVM. Έχοντας κατατάξει τα γονίδιά μας ανάλογα με τη τιμή της συνάρτησης πύλης (Βήμα 2), τα γονίδια που αφαιρούνται τελευταία σε κάθε επανάληψη του CV θεωρούνται και αυτά που διατηρούν αρκετά ανοιχτές τις πύλες τους στο FSMLP. Ως $3^{η}$ Γονιδιακή Υπογραφή επιλέγουμε τα γονίδια εκείνα που εμφανίζονται περισσότερες φορές μέσα στις διάφορες επαναλήψεις του CV.

Βήμα 4 - Αξιολόγηση της γονιδιακής υπογραφής : Τέλος, η $3^{η}$ Γονιδιακή Υπογραφή αξιολογείται , όπως και οι προηγούμενες 2 υπογραφές, ως προς την ικανότητα κατασκευής ενός αξιόπιστου συστήματος ταξινόμησης.

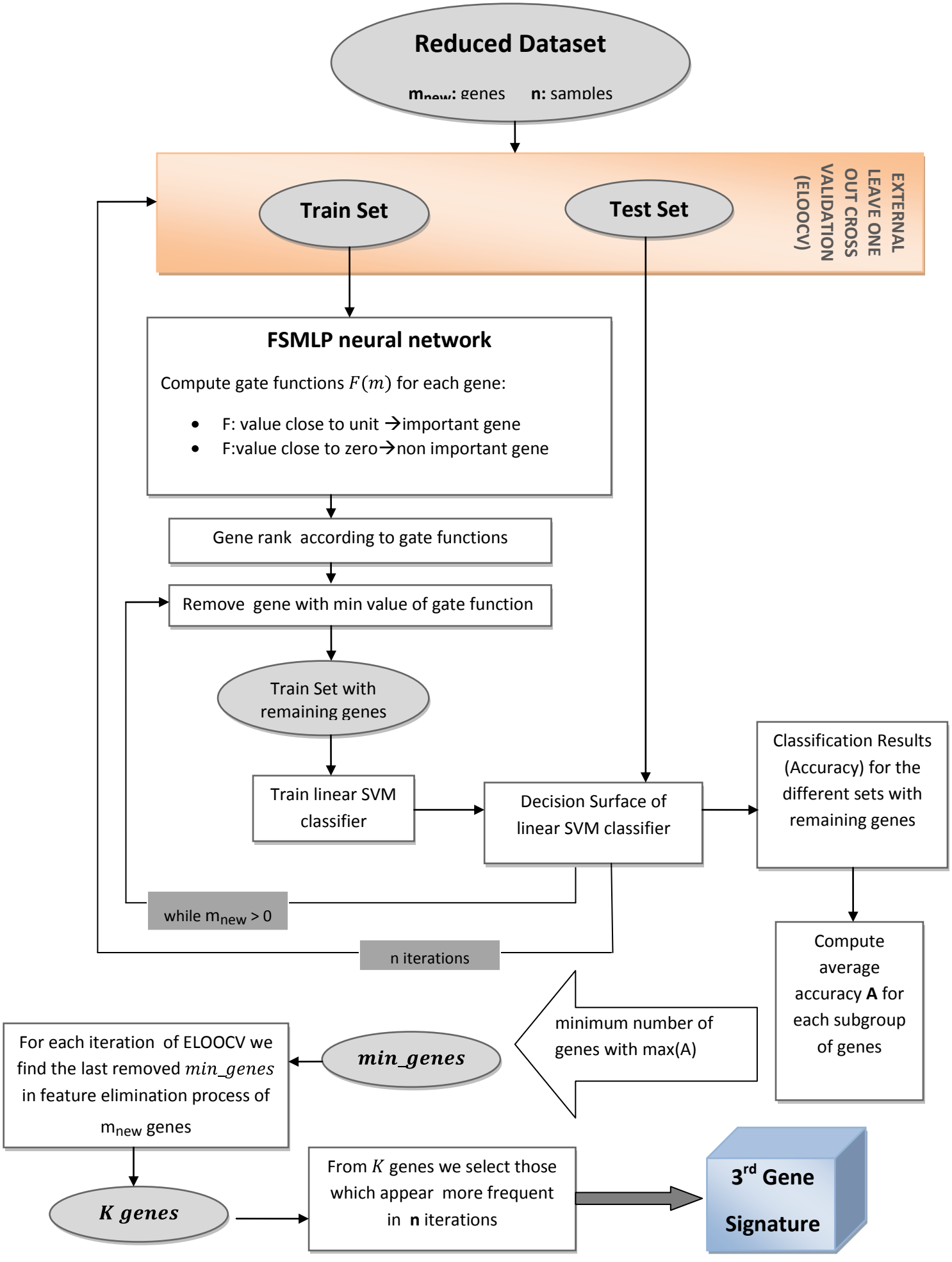
Στη συνέχεια παρουσιάζουμε την σχηματική απεικόνιση της μεθοδολογίας του $1^{ου}$ Σταδίου για τον υπολογισμό των Reduced_Datasets (Εικόνες 43, 44) και του $2^{ου}$ Σταδίου (Εικόνα 45) για τον υπολογισμό της $3^{ης}$ Γονιδιακής Υπογραφής.



Εικόνα 43 : Αναλυτική απεικόνιση της μεθοδολογίας (1^{ο} Στάδιο - Βήματα 1-2) για την εξαγωγή της 3^{ης} Γονιδιακής Υπογραφής.



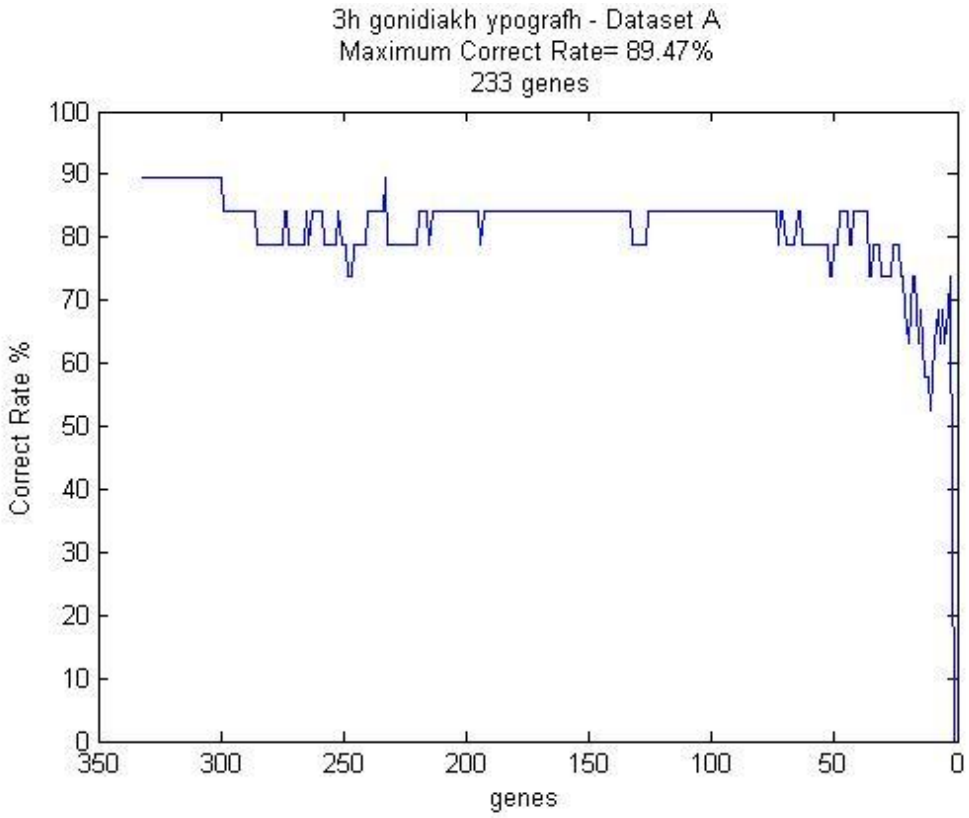
Εικόνα 44 : Αναλυτική απεικόνιση της μεθοδολογίας (1^{ο} Στάδιο - Βήμα 3) για την εξαγωγή της 3^{ης} Γονιδιακής Υπογραφής.



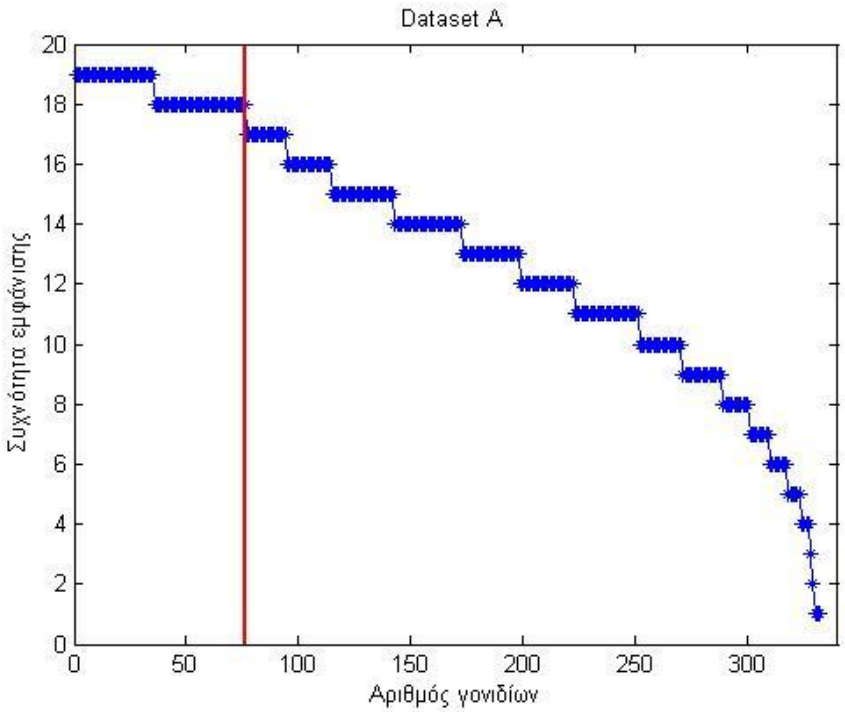
Εικόνα 45 : Αναλυτική απεικόνιση της μεθοδολογίας (2ο Στάδιο – Βήματα 1-3) για την εξαγωγή της 3ης Γονιδιακής Υπογραφής.

Reduced_Dataset A

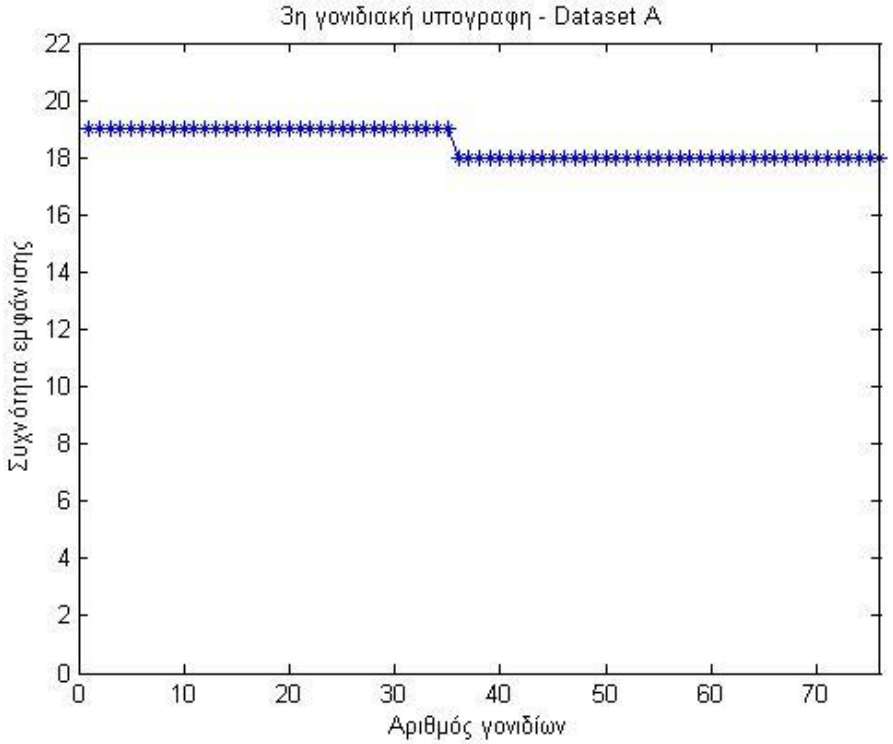
Εφαρμόζοντας το 2ο στάδιο της 3ης μεθοδολογίας στο Reduced_Dataset A καταλήγουμε ότι στα 233 γονίδια εμφανίζεται το μέγιστο ποσοστό επιτυχίας το οποίο είναι ίσο με 89.47%, όπως φαίνεται στην Εικόνα 46. Στην συνέχεια σύμφωνα με το Βήμα 3 συγκεντρώνουμε τα 233 γονίδια που αφαιρέθηκαν τελευταία σε κάθε μια από τις 19 επαναλήψεις του CV και καταλήγουμε σε ένα πλήθος 332 γονιδίων, όσος και ο αρχικός αριθμός γονιδίων του Reduced_Dataset A. Η τυχαιότητα των νευρωνικών δικτύων, ο μεγάλος αριθμός ελαχίστων γονιδίων ($min\_genes = 233$) στον οποίο καταλήξαμε, καθώς και το μικρό σε μέγεθος Reduced_Dataset A οδήγησαν σε αυτό το αποτέλεσμα. Εμείς όμως επιλέγουμε ως 3η Γονιδιακή Υπογραφή τα 76 γονίδια που εμφανίζονται κατά μέσο όρο πάνω από 95% στις 19 επαναλήψεις του CV και των οποίων οι πύλες είναι αρκετά ανοιχτές (Εικόνες 47, 48). Τέλος η ακρίβεια του ταξινομητή που κατασκευάζεται με τη 3η Γονιδιακή Υπογραφή προκύπτει ίση με **84.21%**.



Εικόνα 46 : Γραφική παράσταση της μέσης ακρίβειας που σημείωσε ο liner SVM classifier στις 19 επαναλήψεις καθώς ελαττώνουμε τα 332 γονίδια του Reduced_Dataset A. Καθώς προχωράμε από αριστερά προς τα δεξιά η τιμή της συνάρτησης πύλης των γονιδίων αυξάνεται.



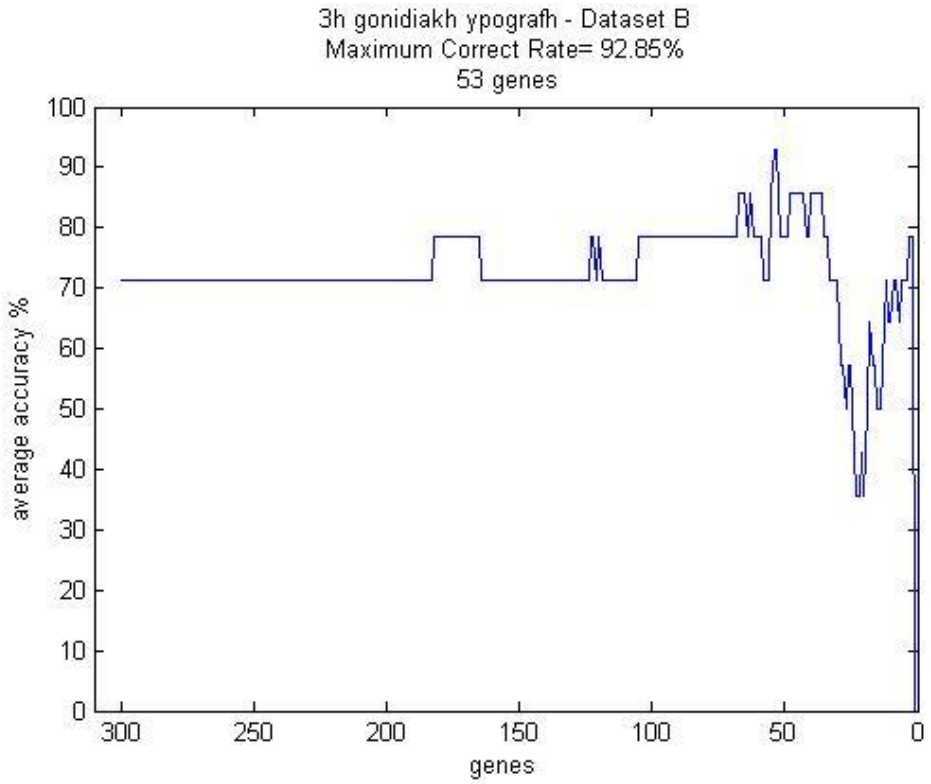
Εικόνα 47 : Συχνότητα εμφάνισης των 332 γονιδίων του Reduced_Dataset A μέσα στις 19 επαναλήψεις του ELOOCV. Ως γονιδιακή υπογραφή επιλέγουμε τα γονίδια με τις υψηλότερες συχνότητες (από τη κόκκινη γραμμή και αριστερά).



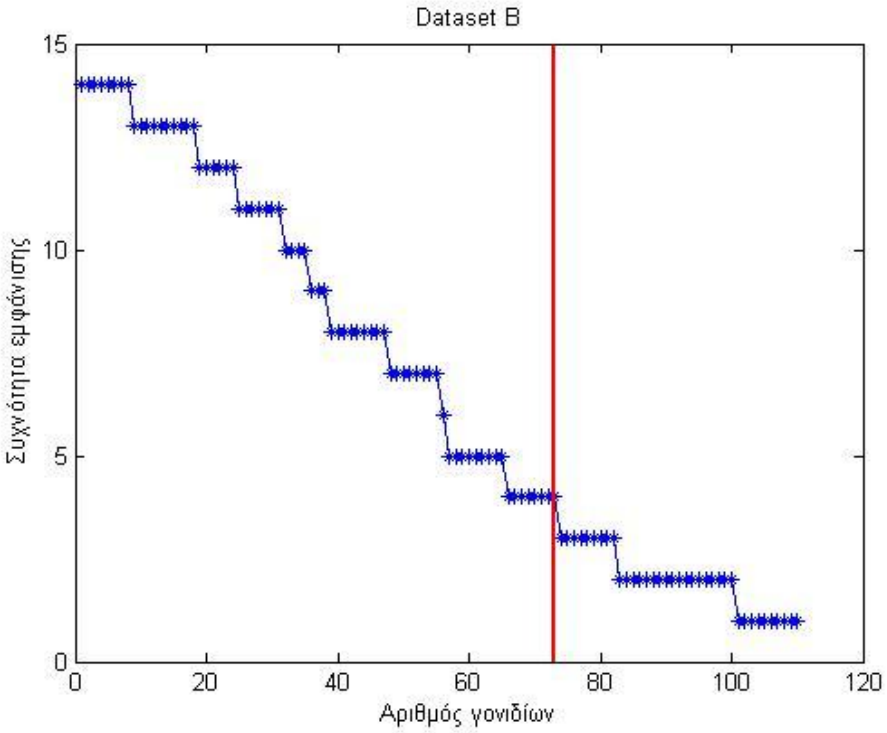
Εικόνα 48 : Τα 76 γονίδια τα οποία αποτελούν την $3^{η}$ γονιδιακή υπογραφή για το Dataset A, απεικονίζονται με τις συχνότητες εμφάνισης τους.

Reduced_Dataset B

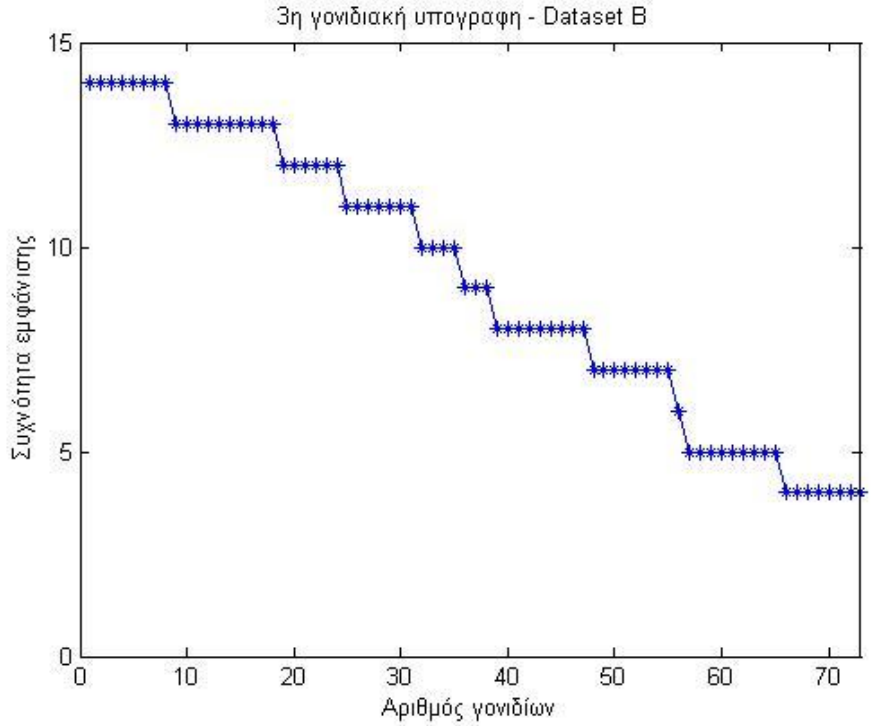
Για το Reduced_Dataset B βρίσκουμε ότι κατά μέσο όρο στα 53 γονίδια που αφαιρούνται τελευταία σε κάθε μία από τις 14 επαναλήψεις του CV επιτυγχάνεται μέγιστη ακρίβεια ταξινόμησης ίση με 95.85% (Εικόνα 49). Ο ελάχιστος αυτός αριθμός γονιδίων είναι κατά πολύ μικρότερος σε σχέση με τον αντίστοιχο για το Reduced_Dataset A (min_genes=233). Αυτό έχει ως αποτέλεσμα ο συνολικός αριθμός των 53 γονιδίων που εμφανίζονται στις 14 επαναλήψεις του cross validation να είναι ίσος με 110 , με το 1/3 δηλαδή του αρχικού αριθμού των m_{new} γονιδίων. Από τα 110 γονίδια επιλέγουμε πάλι εκείνα που εμφανίζονται συχνότερα μέσα στις 14 επαναλήψεις (Εικόνα 50). Καταλήγουμε, λοιπόν, σε μια Γονιδιακή Υπογραφή με 73 γονίδια (Εικόνα 51) και σε ποσοστό επιτυχίας **78.57%** του ταξινομητή που κατασκευάζεται.



Εικόνα 49 : Γραφική παράσταση της μέσης ακρίβειας που σημείωσε ο liner SVM classifier στις 19 επαναλήψεις καθώς ελαττώνουμε τα 300 γονίδια του Reduced_Dataset B. Καθώς προχωράμε από αριστερά προς τα δεξιά η τιμή της συνάρτησης πύλης των γονιδίων αυξάνεται.



Εικόνα 50 : Συχνότητα εμφάνισης των 110 γονιδίων του Reduced_Dataset B μέσα στις 14 επαναλήψεις του ELOOCV. Ως γονιδιακή υπογραφή επιλέγουμε τα γονίδια με τις υψηλότερες συχνότητες (από τη κόκκινη γραμμή και αριστερά).



Εικόνα 51 : Τα 73 γονίδια τα οποία αποτελούν την $3^{η}$ γονιδιακή υπογραφή για το Dataset B, απεικονίζονται με τις συχνότητες εμφάνισης τους.

4.5 Συνδυάζοντας τα Σύνολα Δεδομένων

Οι 2 πρώτες στήλες του Πίνακα 6 που ακολουθεί περιλαμβάνουν για κάθε Dataset (A, B) τις τρείς Γονιδιακές Υπογραφές καθώς και την ακρίβεια του ταξινομητή που κατασκευάστηκε από την αντίστοιχη υπογραφή.

Όπως έχουμε αναφέρει και σε προηγούμενα Κεφάλαια, τα δυο Dataset που επεξεργαστήκαμε δεν περιέχουν κοινά γονίδια και μπορούν να θεωρηθούν ως ανεξάρτητα και πιθανώς συμπληρωματικά σύνολα δεδομένων. Για να εξετάσουμε αν τα 2 σύνολα δεδομένων επιφέρουν καλύτερα αποτελέσματα όταν συνδυάζονται, αθροίσαμε για κάθε μέθοδο (RFE-LNW, LASSO, συνδυασμός RFE-LNW & FSMLP) τις 2 επιμέρους υπογραφές των Dataset, διατηρώντας την έκφρασή των γονιδίων τους μονό ως προς τα 14 κοινά δείγματα (10 OA και 4 NC). Με τη νέα (συνολική) Γονιδιακή Υπογραφή που προέκυψε, κατασκευάσαμε ένα γραμμικό SVM ταξινομητή και ελέγξαμε την ακρίβειά του με τη LOOCV τεχνική. Όπως φαίνεται και στη τρίτη στήλη του Πίνακα 6, οι ταξινομητές που κατασκευάζονται από την συνολική Γονιδιακή Υπογραφή επιτυγχάνουν μεγαλύτερα ποσοστά ακρίβειας σε σχέση με τους ταξινομητές που κατασκευάστηκαν μόνο από τις υπογραφές των Dataset A ή B .

Πίνακας 6: Οι Γονιδιακές Υπογραφές στις οποίες καταλήξαμε και η ακρίβεια ταξινόμησής τους.

	Dataset A	**Dataset B**	**Dataset A + Dataset B**
1η Γονιδιακή υπογραφή (RFE-LNW)	**86** genes **78.94%**	**53** genes **64.28%**	**139** genes **92.85%**
2η Γονιδιακή υπογραφή (LASSO)	**35** genes **84.21%**	**49** genes **85.71%**	**84** genes **92.85%**
3η Γονιδιακή υπογραφή (RFE-LNW, FSMLP)	**76** genes **84.21%**	**73** genes **78.57%**	**149** genes **85.71%**

ΚΕΦΑΛΑΙΟ 5: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Στο συγκεκριμένο Κεφάλαιο συγκεντρώνουμε και αξιολογούμε τα αποτελέσματα που προέκυψαν από την εφαρμογή της μεθοδολογίας του Κεφαλαίου 4 πάνω στο σύνολο δεδομένων των Huber και συνεργατών το οποίο σχετίζεται με την ασθένεια της οστεοαρθρίτιδας [5]. Αρχικά αξιολογούμε τις τρεις γονιδιακές υπογραφές στις οποίες καταλήξαμε για κάθε Dataset ως προς την ικανότητά τους να κατασκευάσουν έναν ταξινομητή που θα επιτυγχάνει υψηλά ποσοστά ακρίβειας. Στη συνέχεια προχωρούμε στη βιολογική ερμηνεία των αποτελεσμάτων όπου με τη βοήθεια των εργαλείων WebGestalt και Genotator επεξεργαζόμαστε τα γονίδια που περιλαμβάνονται στις υπογραφές και ελέγχουμε τη σημαντικότητά τους σχετικά με την ασθένεια της οστεοαρθρίτιδας.

5.1 Στατιστικά Αποτελέσματα

Στους Πίνακες 7 έως 9 παρουσιάζονται συνοπτικά τα αποτελέσματα που μας οδήγησαν στις τρεις Γονιδιακές Υπογραφές σε κάθε Dataset. Συγκεκριμένα ο Πίνακας 7 περιέχει τα αποτελέσματα εφαρμογής της $1^{ης}$ μεθόδου (Κεφάλαιο 4.2), ο Πίνακας 8 της $2^{ης}$ μεθόδου (Κεφάλαιο 4.3) και ο Πίνακας 9 της $3^{ης}$ μεθόδου (Κεφάλαιο 4.4), πάνω στο Dataset A και το Dataset B. Κάθε πίνακας περιλαμβάνει τη μέση μέγιστη ακρίβεια (Max Accuracy) που σημειώθηκε εφαρμόζοντας την αντίστοιχη μέθοδο επιλογής γονιδίων, τον ελάχιστο αριθμό γονιδίων (Minimum Number of Genes with Max Accuracy) στον οποίο σημειώνεται αυτή η μέγιστη ακρίβεια μέσα στις διάφορες επαναλήψεις του ELOOCV (19 επαναλήψεις για το Dataset A και 14 για το Dataset B), τον αριθμό των γονιδίων (Gene Signature) που περιλαμβάνει κάθε υπογραφή καθώς και το μέσο ποσοστό ακρίβειας του γραμμικού SVM ταξινομητή (Signature Accuracy) που κατασκευάζεται από την αντίστοιχη υπογραφή. Στη τελευταία στήλη κάθε πίνακα περιλαμβάνουμε τη γονιδιακή υπογραφή που προέκυψε από τη συνένωση των υπογραφών των 2 Datasets και την ακρίβεια ταξινόμησης που αυτή επιτυγχάνει.

Πίνακας 7: Τα αποτελέσματα που προέκυψαν εφαρμόζοντας την 1η Μεθοδολογία (RFE-LNW) στα διαθέσιμα σύνολα δεδομένων.

1st Method (RFE-LNW, SVM)	Dataset A	Dataset B	Dataset A+Dataset B
Max Accuracy	94.73%	85.71%	-
Minimum Number of Genes with Max Accuracy	2927	2963	-
1st Gene Signature	86	53	139
Signature Accuracy	78.94%	64.28%	92.85%

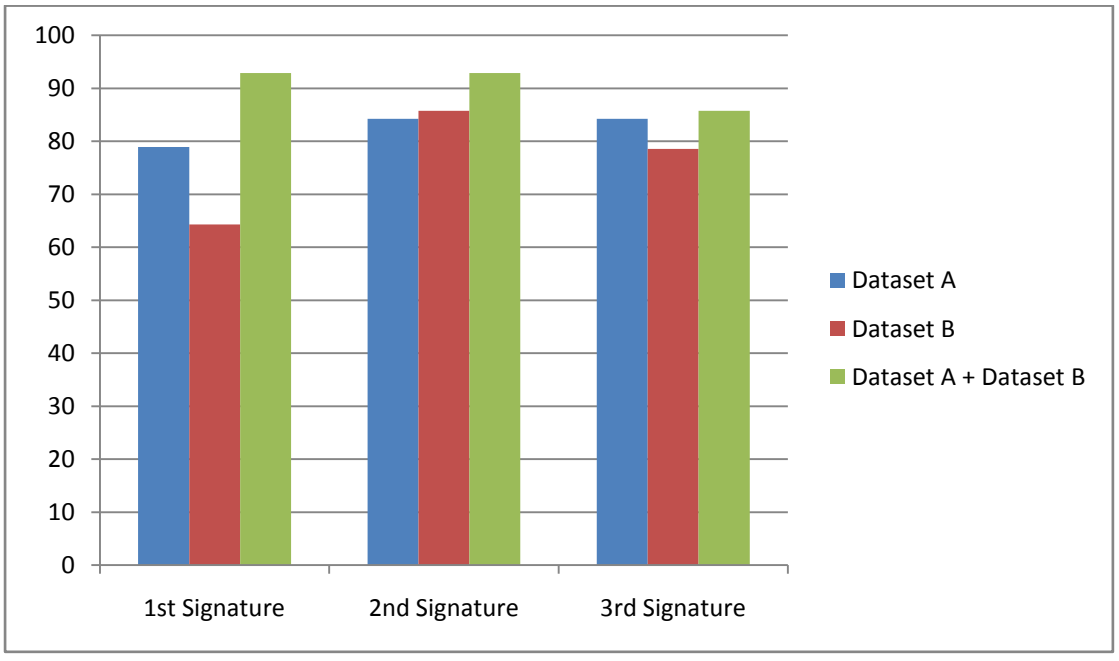
Πίνακας 8: Τα αποτελέσματα που προέκυψαν εφαρμόζοντας την 2η Μεθοδολογία (LASSO) στα διαθέσιμα σύνολα δεδομένων.

2nd Method (LASSO, SVM)	Dataset A	Dataset B	Dataset A+Dataset B
Max Accuracy	73.68%	78.57%	-
Minimum Number of Genes with Max Accuracy	59	240	-
2nd Gene Signature	35	49	84
Signature Accuracy	84.21%	85.71%	92.85%

Πίνακας 9: Τα αποτελέσματα που προέκυψαν εφαρμόζοντας την 3η Μεθοδολογία (RFE-LNW, FSMLP) στα διαθέσιμα σύνολα δεδομένων.

3rd Method (RFE-LNW, FSMLP, SVM)	Dataset A	Dataset B	Dataset A+Dataset B
Max Accuracy	89.47%	92.85%	-
Minimum Number of Genes with Max Accuracy	332	110	-
3rd Gene Signature	76	73	149
Signature Accuracy	84.21%	78.57%	85.71%

Το ραβδόγραμμα εκατοστιαίας βάσης που ακολουθεί παρουσιάζει την μέση ακρίβεια (Signature Accuracy) του γραμμικού SVM ταξινομητή για κάθε γονιδιακή υπογραφή που αφορά το Dataset A, το Dataset B καθώς και τη συνένωση των δύο Datasets.



Εικόνα 52: Η μέση ακρίβεια του γραμμικού SVM ταξινομητή για τις τρεις Γονιδιακές Υπογραφές.

Μελετώντας τους Πίνακες 7 έως 9 και την Εικόνα 52 και συγκρίνοντας τα αποτελέσματα καταλήγουμε στις εξής παρατηρήσεις:

- Τα γονίδια των Υπογραφών που προκύπτουν από την 1η Μέθοδο (RFE-LNW, SVM) κατασκευάζουν ταξινομητές με ακρίβεια που κυμαίνεται από 64% έως και 93%. Αντιθέτως, η 2η (LASSO, SVM) και 3η (RFE-LNW, FSMLP, SVM) Μέθοδος επιτυγχάνουν όχι μόνο υψηλά αλλά παράλληλα και πιο ομαλά ποσοστά ακρίβειας, κάτι το οποίο δεν παρατηρούμε στα αποτελέσματα της 1ης μεθόδου. Συγκεκριμένα, οι ταξινομητές της 2ης Γονιδιακής Υπογραφής σημειώνουν 85% - 93% ακρίβεια ενώ της 3ης κυμαίνονται σε λίγο μικρότερα ποσοστά μεταξύ 79% - 86%.
- Η 2η Μέθοδος (LASSO, SVM) που ακολουθήσαμε μπορεί να χαρακτηριστεί ως η πιο σημαντική σε σχέση με τις υπόλοιπες δύο. Είναι εκείνη που δίνει τα πιο ακριβή αποτελέσματα αφού παρουσιάζει τα υψηλότερα ποσοστά ακρίβειας στις υπογραφές των Dataset A, Dataset B καθώς και στη συνένωση των 2 υπογραφών. Η 3η μέθοδος ακολουθεί με λίγο χαμηλότερα ποσοστά ακρίβειας σε σχέση με την 2η μέθοδο χωρίς όμως να θεωρήσουμε τα αποτελέσματα που δίνει ήσσονος σημασίας. Όσο αναφορά τη 1η μέθοδο μπορεί να προκύπτει 92.85% ακρίβεια ταξινόμησης στην ένωση των υπογραφών όμως δεν ακολουθείται η ίδια συμπεριφορά στις μεμονωμένες υπογραφές για τα Dataset A και B. Το γεγονός αυτό μας οδηγεί στο να χαρακτηρίσουμε την 1η μέθοδο ως εκείνη που παρέχει τα λιγότερα αξιόπιστα αποτελέσματα σε σχέση με τις υπόλοιπες 2 μεθόδους.
- Ο συνδυασμός των υπογραφών των δύο Dataset παρουσιάζει υψηλότερα ποσοστά ακρίβειας και στις 3 μεθόδους (πράσινες ράβδοι στην Εικόνα 52). Το γεγονός αυτό δείχνει ότι πιθανώς τα γονίδια των δύο Dataset να είναι συμπληρωματικά και σε συνδυασμό περιέχουν πληροφορία ικανή να κατασκευάσει πιο εύρωστα μοντέλα ταξινόμησης.
- Ο μικρός αριθμός γονιδίων που περιλαμβάνουν οι υπογραφές και τα αρκετά υψηλά ποσοστά ακρίβειας, που επιτυγχάνουν οι ταξινομητές, αιτιολογούν τη χρησιμότητα των γονιδίων στα οποία καταλήξαμε. Έτσι, από ένα μεγάλο αριθμό γονιδίων (22283 γονίδια για το Dataset A και 22645 για το Dataset B) επιλέξαμε μικρά και διαχειρίσιμα υποσύνολα (πεδίο Gene Signature στους Πίνακες 7-9) με επαρκή διαχωριστική ικανότητα ανάμεσα στις δύο κλάσεις ενδιαφέροντος.

Ο αριθμός των κοινών γονιδίων που βρέθηκαν ανάμεσα στις γονιδιακές υπογραφές είναι πολύ μικρός. Συγκεκριμένα για το Dataset A παρουσιάζονται δύο κοινά γονίδια μεταξύ της $1^{ης}$ και $2^{ης}$ υπογραφής ($1^{η}$ Στήλη Πίνακα 10) και για το Dataset B τρία κοινά γονίδια μεταξύ $1^{ης}$ και $2^{ης}$ υπογραφής ($2^{η}$ Στήλη Πίνακα 10). Όπως έχουμε αναφέρει και σε προηγούμενο Κεφάλαιο (Κεφ. 1.5) τα δύο Dataset δεν περιλαμβάνουν κοινά γονίδια και για τον λόγο αυτό δεν εξετάστηκαν οι υπογραφές τους ως προς την ύπαρξη κοινών γονιδίων.

Πίνακας 10: Συνοπτικός πίνακας με τον αριθμό των κοινών γονιδίων που εμφανίζονται στις διάφορες γονιδιακές υπογραφές των Dataset A και B.

Κοινά Γονίδια	Dataset A	Dataset B
1^{ST} METHOD - 2^{ND} METHOD	2	3
2^{ND} METHOD - 3^{RD} METHOD	-	-
3^{RD} METHOD - 1^{ST} METHOD	-	-

Οι κωδικοί, τα σύμβολα και η περιγραφή όλων των γονιδίων που απαρτίζουν κάθε υπογραφή σε κάθε Dataset παρατίθεται στο Παράρτημα Α.

5.2 Βιολογικά Αποτελέσματα

Μετά τη στατιστική αξιολόγηση προχωράμε στη βιολογική ερμηνεία των αποτελεσμάτων. Εξετάζουμε τη συμμέτοχή των Γονιδιακών Υπογραφών σε διάφορες βιολογικές διεργασίες καθώς και τη συσχέτιση των γονιδίων που περιλαμβάνουν με την ασθένεια της οστεοαρθρίτιδας. Τα εργαλεία WebGestalt και Genotator μας βοηθούν να εργαστούμε προς αυτή την κατεύθυνση.

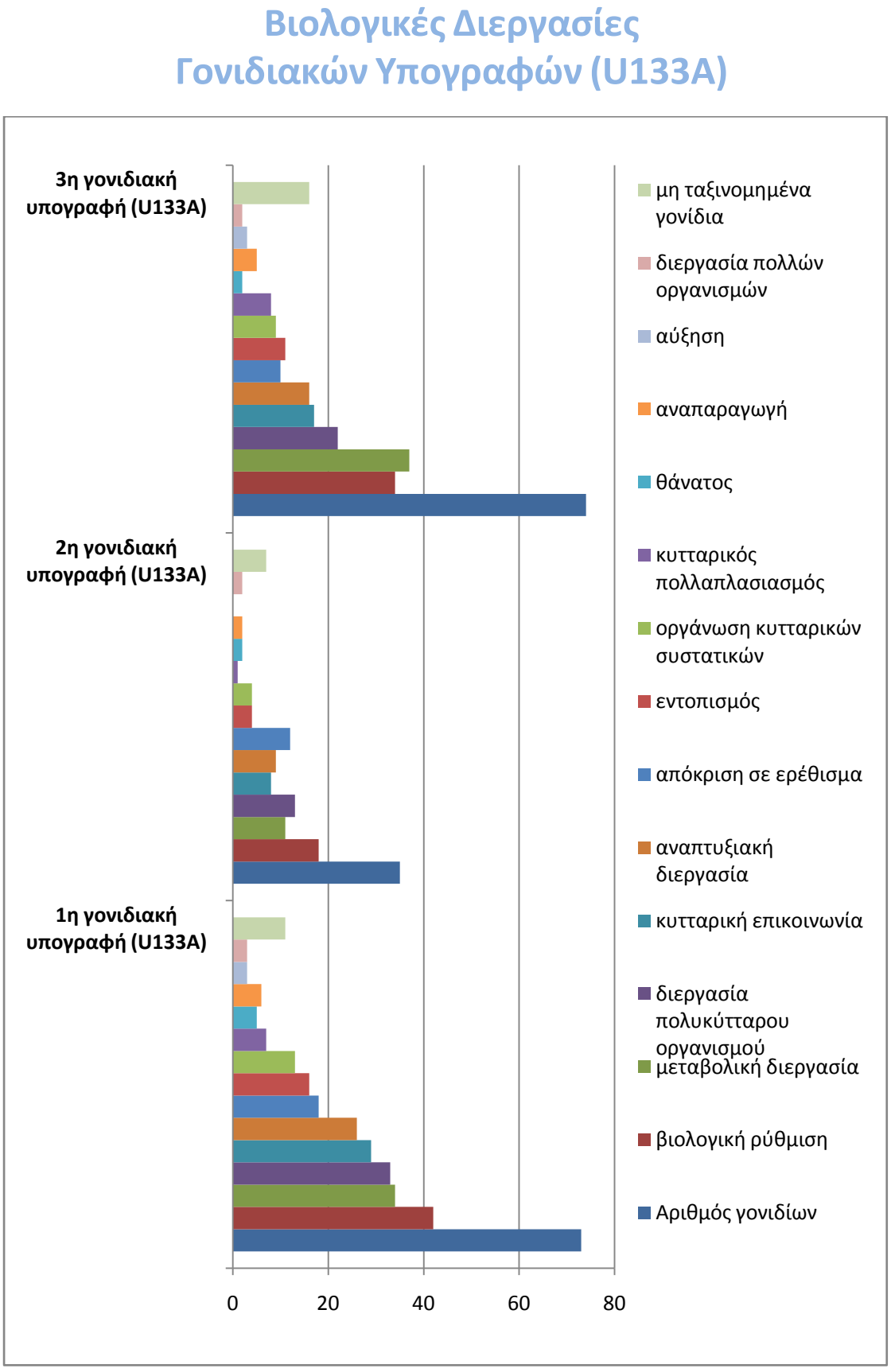
Βιολογικές Διεργασίες

Εξετάζοντας κάθε γονιδιακή υπογραφή με το WebGestalt [26], προσδιορίστηκε αρχικά ένας αριθμός από ενδιαφέρουσες βιολογικές διεργασίες στις οποίες συμμετέχουν τα γονίδια που τις αποτελούν. Οι διεργασίες αυτές παρουσιάζονται στις Εικόνες 53-55. Παρατηρούμε ότι τα γονίδια των υπογραφών, παρόλο που δεν είναι τα ίδια, συμμετέχουν στις ίδιες 13 υπέρ-κατηγορίες βιολογικών διεργασιών.

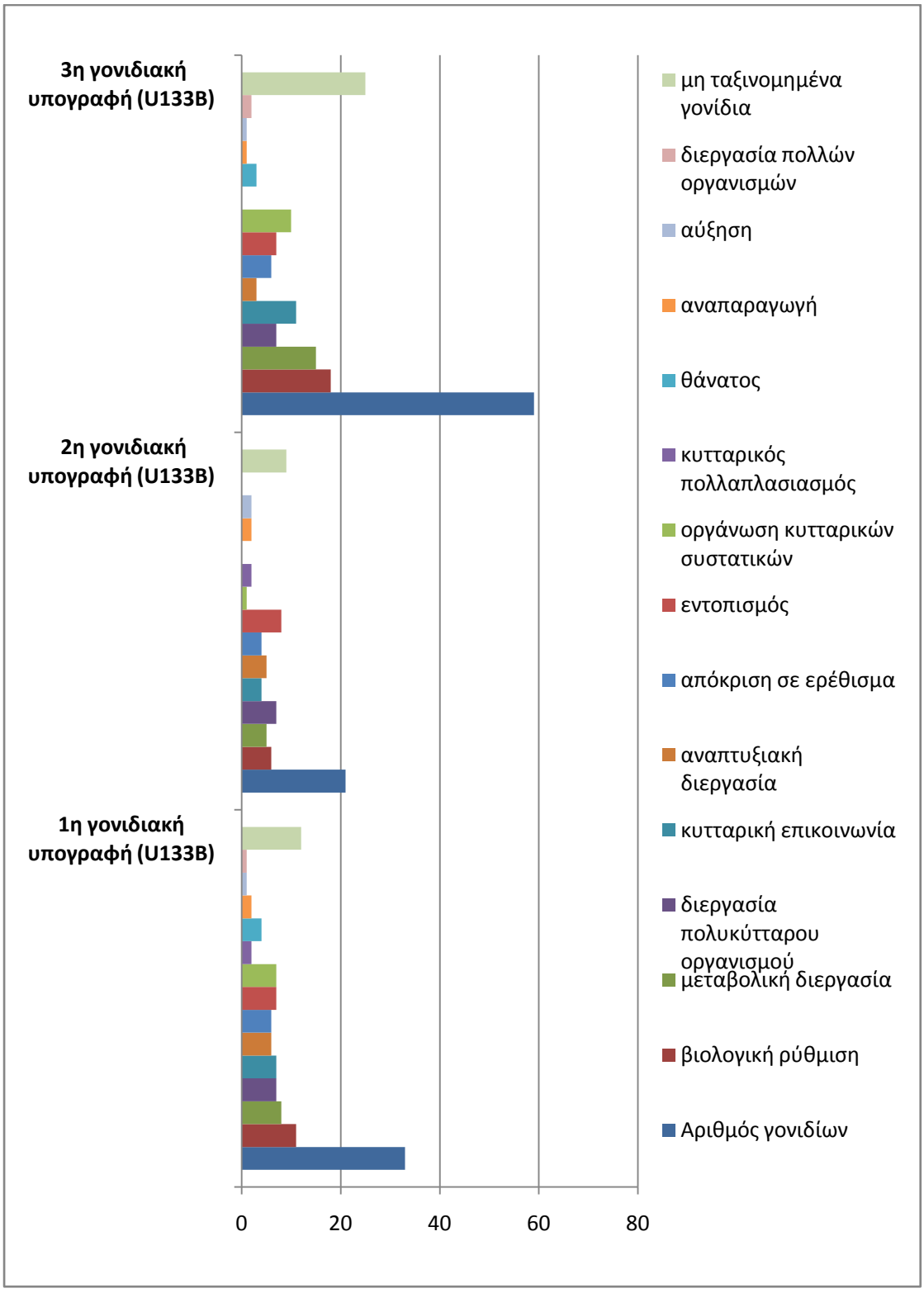
Ανάμεσα στις κυριότερες διεργασίες (όσον αφορά την πλειονότητα των γονιδίων που συμμετέχουν σ' αυτές) της 1^{ης} υπογραφής των Dataset A, B, αλλά και της συνένωσής τους A/B είναι η βιολογική ρύθμιση, ο μεταβολισμός, η διεργασία πολυκύτταρου οργανισμού και η κυτταρική επικοινωνία (Εικόνες 53-55).

Η βιολογική ρύθμιση, ο μεταβολισμός και η κυτταρική επικοινωνία αποτελούν επίσης τις κυριότερες διεργασίες της 3^{ης} υπογραφής των Dataset A, B, και της συνένωσής τους A/B. Ταυτόχρονα, η διεργασία πολυκύτταρου οργανισμού του Dataset A, «αντικαθίσταται» από τη διαδικασία οργάνωσης κυτταρικών συστατικών στο Dataset B, παραμένοντας ωστόσο κύρια κατά την συνένωση A/B των Datasets (Εικόνες 53-55).

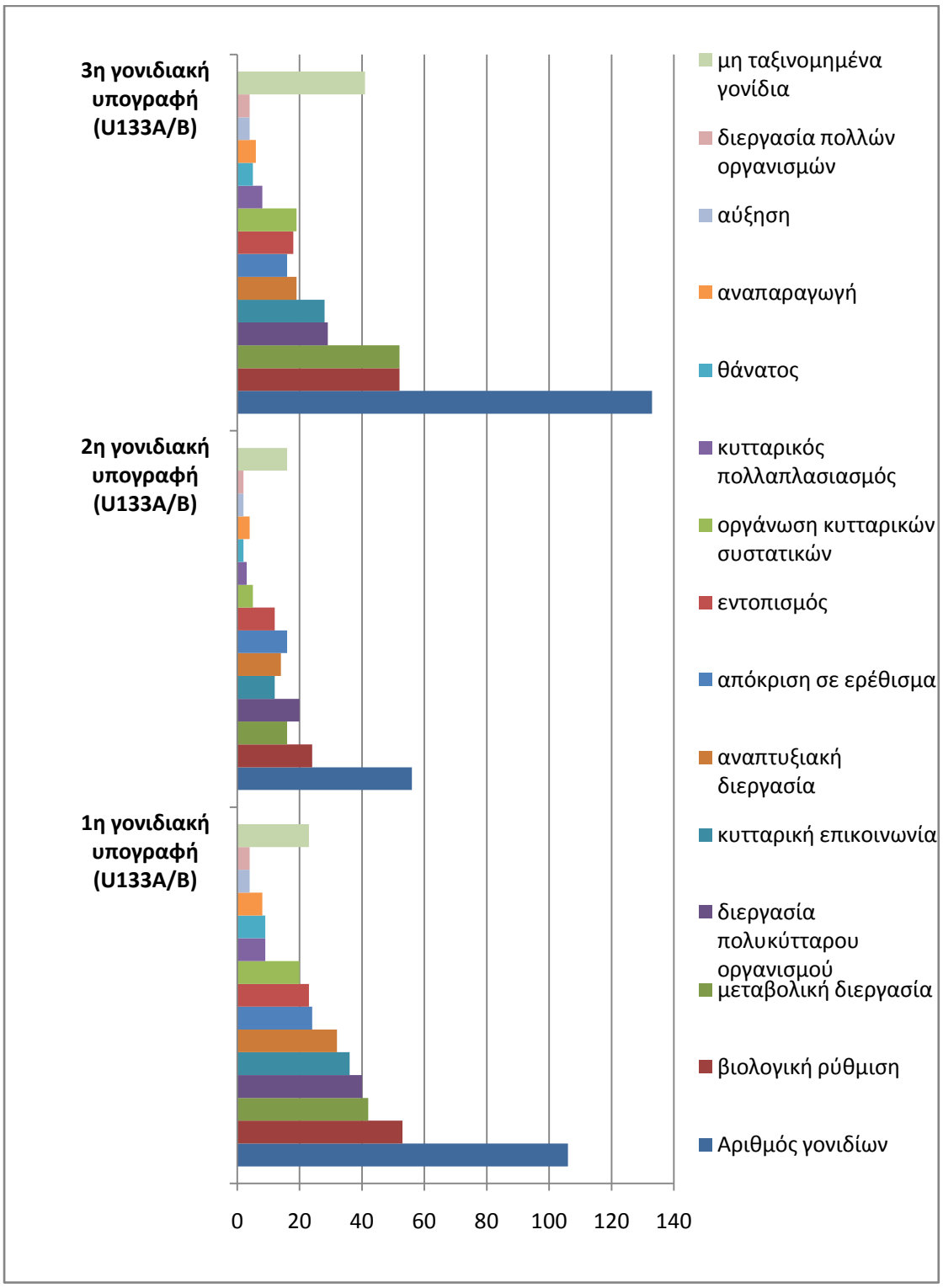
Η 2^{η} υπογραφή διαφοροποιείται σε σχέση με τη 1^{η} και τη 3^{η} γονιδιακή υπογραφή, αφού στις κύριες κατηγορίες εκτός από τη βιολογική ρύθμιση, τη διεργασία πολυκύτταρου οργανισμού, και τον μεταβολισμό, περιλαμβάνονται η απόκριση σε ερέθισμα στο Dataset A, καθώς και η αναπτυξιακή διεργασία και ο εντοπισμός στο Dataset B. Οι κατηγορίες αυτές παραμένουν κύριες κατά την συνένωση A/B των Datasets, με εξαίρεση τον εντοπισμό που εμφανίζεται ως η πρώτη κύρια κατηγορία στο Dataset B και ως δευτερεύουσα κατά τη συνένωση A/B των Datasets (Εικόνες 53-55).



Εικόνα 53 : Συγκριτικά αποτελέσματα για τις 13 βιολογικές διεργασίες στις οποίες συμμετέχουν τα γονίδια που περιλαμβάνονται στις τρεις γονιδιακές υπογραφές του Dataset A.



Εικόνα 54 : Συγκριτικά αποτελέσματα για τις 13 βιολογικές διεργασίες στις οποίες συμμετέχουν τα γονίδια που περιλαμβάνονται στις τρεις γονιδιακές υπογραφές του Dataset B.



Εικόνα 55 : Συγκριτικά αποτελέσματα για τις 13 βιολογικές διεργασίες στις οποίες συμμετέχουν τα γονίδια που περιλαμβάνονται στη συνένωση των τριών γονιδιακών υπογραφών του Dataset A με του Dataset B.

Στην οστεοαρθρίτιδα τόσο ο αρθρικός χόνδρος, όσο ο αρθρικός υμένας (που αφορά στη μελέτη μας) και το υποχόνδριο οστό κατέχουν συγκεκριμένο μερίδιο στον καταρράκτη των κυτταρικών γεγονότων. Σήμερα γνωρίζουμε ότι ο αρθρικός χόνδρος, ο αρθρικός υμένας και το υποχόνδριο οστό, αλληλεπιδρούν μεταξύ τους, και οι διεργασίες που συμβαίνουν στον ένα ιστό, προκαλούνται από γεγονότα που συμβαίνουν στον άλλο και επάγουν με την σειρά τους διεργασίες στους άλλους συμμετέχοντες ιστούς [9].

Γνωρίζοντας ότι η ΟΑ προκαλείται από την αλληλεπίδραση πολλαπλών γονιδίων και περιβαλλοντικών παραγόντων, είναι σκόπιμη η περαιτέρω ανάλυση των μοριακών μονοπατιών ή/και μοριακών δικτύων στα οποία συμμετέχουν πολλαπλά γονίδια [64]. Έτσι, στα πλαίσια αυτά πραγματοποιείται στη παρούσα μελέτη, η ανάλυση των μοριακών μονοπατιών KEGG που θα μπορούσε να είναι αποτελεσματική στην αποκάλυψη των λειτουργιών των γονιδίων για την αιτιολογία μιας τέτοιας πολύπλοκης νόσου.

Μοριακά Μονοπάτια KEGG

Με σκοπό να εμπλουτίσουμε τη βιολογική ανάλυση των αποτελεσμάτων μας, εντοπίσαμε με το WebGestalt τα 10 (“Top 10”) πιο στατιστικά σημαντικά μοριακά μονοπάτια KEGG που σχετίζονται με τα γονίδια των υπογραφών [26]. Η σημαντικότητα ενός μοριακού μονοπατιού για ένα σύνολο γονιδίων καθορίστηκε από τη τιμή της adjust p-value με κατώφλι το 0.05. Ως ελάχιστο αριθμό γονιδίων που συμμετέχουν σε ένα στατιστικά σημαντικό μονοπάτι επιλέξαμε την τιμή 2. Τα αποτελέσματα για τα top 10 μοριακά μονοπάτια που αφορούν τις 3 γονιδιακές υπογραφές των Dataset A, B και της συνένωσης των υπογραφών των 2 Datasets (AB) παρουσιάζονται στις Εικόνες 56-59.

Στις Εικόνες 59-60 παρατηρούμε ότι στις δυο γονιδιακές υπογραφές, $2^{η}$ και $3^{η}$, το σύνολο των μοριακών μονοπατιών εμφανίζεται με στατιστική σημαντικότητα, με εξαίρεση ένα μοριακό μονοπάτι της $3^{ης}$ υπογραφής (κυτταροτοξικότητα φυσικών κυττάρων-φονιάδων), και άρα η $2^{η}$ και ακολούθως η $3^{η}$ υπερέχουν από την $1^{η}$. Παρόλο που η $1^{η}$ γονιδιακή υπογραφή εμφανίζει έναν αριθμό κοινών βιολογικών μονοπατιών τόσο με την $2^{η}$ όσο και την $3^{η}$ υπογραφή, στα περισσότερα μονοπάτια δεν παρατηρείται στατιστική σημαντικότητα. Στην Εικόνα 60 δίνεται μια γραφική απεικόνιση, με τη μορφή ραβδογράμματος, των μοριακών μονοπατιών ($3^{ο}$ επίπεδο κατηγορίας KEGG), όπου τονίζονται οι ομοιότητες και οι διαφορές ανάμεσα στις τρεις γονιδιακές υπογραφές.

ΔΙΑΦΟΡΕΣ ΚΑΙ ΟΜΟΙΟΤΗΤΕΣ ΤΩΝ ΜΟΡΙΑΚΩΝ ΜΟΝΟΠΑΤΙΩΝ KEGG ΑΝΑΜΕΣΑ ΣΤΙΣ 3 ΓΟΝΙΔΙΑΚΕΣ ΥΠΟΓΡΑΦΕΣ ΤΟΥ ΥΠΟΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ Α				
ΜΟΡΙΑΚΑ ΜΟΝΟΠΑΤΙΑ	**U133A**			
	O	**C**	***P***	**adj*P***
Γονιδιακή υπογραφή 1				
[04080] Νευροενεργή αλληλεπίδραση προσδέματος-υποδοχέα	4	256	0,0008	0,004
[04060] Αλληλεπίδραση υποδοχέα κυτοκίνη-κυτοκίνη	*2*	*267*	*0,0691*	*0,1167*
[04010] Μονοπάτι σηματοδότησης MAPK	2	269	0,07	0,1167
[05200] Μονοπάτια στον καρκίνο	*2*	*330*	*0,0992*	*0,124*
[01100] Μεταβολικά μονοπάτια	*2*	*1104*	*0,5335*	*0,5335*
Γονιδιακή υπογραφή 2				
[01100] Μεταβολικά μονοπάτια	*4*	*1104*	*0,01*	*0,015*
[04650] Κυτταροτοξικότητα φυσικών κυττάρων-φονιάδων	2	137	0,0051	0,015
[04060] Αλληλεπίδραση υποδοχέα κυτοκίνη-κυτοκίνη	*2*	*267*	*0,0181*	*0,0181*
Γονιδιακή υπογραφή 3				
[01100] Μεταβολικά μονοπάτια	*8*	*1104*	*0,0004*	*0,004*
[05322] Συστηματικός ερυθηματώδης λύκος	3	140	0,0016	0,008
[00562] Μεταβολισμός φωσφορικής ινοσιτόλης	2	54	0,0036	0,009
[04020] Μονοπάτι σηματοδότησης ασβεστίου	3	178	0,0031	0,009
[00564] Μεταβολισμός γλυκεροφωσφολιπιδίου	2	70	0,0059	0,0115
[04070] Σύστημα σηματοδότησης φωσφατιδυλινοσιτόλης	2	76	0,0069	0,0115
[04666] Fc γάμμα R-μεσολαβούμενη φαγοκυττάρωση	2	97	0,0111	0,0159
[05010] Νόσος Αλτσχάιμερ	2	169	0,0313	0,0391
[05016] Νόσος του Huntington	2	185	0,0369	0,041
[05200] Μονοπάτια στον καρκίνο	*2*	*330*	*0,1015*	*0,1015*

Τα μοριακά μονοπάτια που απαντώνται σε περισσότερες από μια γονιδιακές υπογραφές έχουν σημειωθεί με πλάγια γραφή.
ΑΡΙΘΜΟΣ ΓΟΝΙΔΙΩΝ (O): αναφέρεται στον αριθμό των γονιδίων από την εκάστοτε υπογραφή που συμμετέχουν στα συγκεκριμένα μοριακά μονοπάτια.
ΣΥΝΟΛΟ ΓΟΝΙΔΙΩΝ (C): αναφέρεται στο σύνολο των γονιδίων του ανθρώπινου γονιδιώματος που είναι γνωστά έως σήμερα ότι συμμετέχουν στα συγκεκριμένα μοριακά μονοπάτια.
Συντμήσεις: KEGG (Kyoto Encyclopedia of Genes and Genomes) Εγκυκλοπαίδεια Kyoto των γονιδίων και γονιδιωμάτων, P (power) ισχύς [εφαρμογή του υπεργεωμετρικού τεστ για τον υπολογισμό (της ισχύς) του εμπλουτισμού των όρων των μοριακών μονοπατιών KEGG], adjP (adjustment P) προσαρμογή P [εφαρμογή της μέθοδου Benjamini & Hochberg για την πολλαπλή προσαρμογή δοκιμής].

Εικόνα 56 : Διαφορές και ομοιότητες των μοριακών μονοπατιών KEGG ανάμεσα στις τρεις γονιδιακές υπογραφές του Dataset A.

ΔΙΑΦΟΡΕΣ ΚΑΙ ΟΜΟΙΟΤΗΤΕΣ ΤΩΝ ΜΟΡΙΑΚΩΝ ΜΟΝΟΠΑΤΙΩΝ KEGG ΑΝΑΜΕΣΑ ΣΤΙΣ 3 ΓΟΝΙΔΙΑΚΕΣ ΥΠΟΓΡΑΦΕΣ ΤΟΥ ΥΠΟΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ Β				
ΜΟΡΙΑΚΑ ΜΟΝΟΠΑΤΙΑ	**U133B**			
	O	**C**	***P***	**adj*P***
Γονιδιακή υπογραφή 1				
[04514]Μόρια κυτταρικής προσκόλλησης (ΜΚΠ)	2	134	0,0043	0,0086
[04080]Νευροενεργή αλληλεπίδραση προσδέματος-υποδοχέα	2	256	0,0149	0,0149
Γονιδιακή υπογραφή 2				
[01100] Μεταβολικά μονοπάτια	*2*	*1104*	*0,0917*	*0,0917*
Γονιδιακή υπογραφή 3				
[04530] Στενοσύνδεσμος	3	134	0,0007	0,0021
[05120] Σηματοδότηση επιθηλιακών κυττάρων σε λοίμωξη ελικοβακτηριδίου του πυλωρού	2	68	0,0036	0,0054
[01100] Μεταβολικά μονοπάτια	*2*	*1104*	*0,4227*	*0,4227*

Τα μοριακά μονοπάτια που απαντώνται σε περισσότερες από μια γονιδιακές υπογραφές έχουν σημειωθεί με πλάγια γραφή.
ΑΡΙΘΜΟΣ ΓΟΝΙΔΙΩΝ (O): αναφέρεται στον αριθμό των γονιδίων από την εκάστοτε υπογραφή που συμμετέχουν στα συγκεκριμένα μοριακά μονοπάτια.
ΣΥΝΟΛΟ ΓΟΝΙΔΙΩΝ (C): αναφέρεται στο σύνολο των γονιδίων του ανθρώπινου γονιδιώματος που είναι γνωστά έως σήμερα ότι συμμετέχουν στα συγκεκριμένα μοριακά μονοπάτια.
Συντμήσεις: KEGG (Kyoto Encyclopedia of Genes and Genomes) Εγκυκλοπαίδεια Kyoto των γονιδίων και γονιδιωμάτων, P (power) ισχύς [εφαρμογή του υπεργεωμετρικού τεστ για τον υπολογισμό (της ισχύς) του εμπλουτισμού των όρων των μοριακών μονοπατιών KEGG], adjP (adjustment P) προσαρμογή P [εφαρμογή της μέθοδου Benjamini & Hochberg για την πολλαπλή προσαρμογή δοκιμής].

Εικόνα 57 : Διαφορές και ομοιότητες των μοριακών μονοπατιών KEGG ανάμεσα στις τρεις γονιδιακές υπογραφές του Dataset B.

ΣΥΓΚΡΙΣΗ ΤΩΝ ΜΟΡΙΑΚΩΝ ΜΟΝΟΠΑΤΙΩΝ KEGG ΣΤΙΣ 3 ΓΟΝΙΔΙΑΚΕΣ ΥΠΟΓΡΑΦΕΣ ΤΟΥ ΣΥΝΔΥΑΣΜΟΥ ΤΩΝ ΥΠΟΣΥΝΟΛΩΝ ΔΕΔΟΜΕΝΩΝ Α ΚΑΙ Β ΜΕ ΤΑ ΥΠΟΣΥΝΟΛΑ ΔΕΔΟΜΕΝΩΝ Α ΚΑΙ Β				
ΜΟΡΙΑΚΑ ΜΟΝΟΠΑΤΙΑ	**U133AB**		**U133A**	**U133B**
	O	**adjP**	**adjP**	**adjP**
Γονιδιακή υπογραφή 1				
[04080]Νευροενεργή αλληλεπίδραση προσδέματος-υποδοχέα	6	0,0003	**0,004**	**0,0149**
[04514]Μόρια κυτταρικής προσκόλλησης (ΜΚΠ)	3	0,0156		**0,0086**
[04612] Επεξεργασία και παρουσίαση του αντιγόνου	*2*	*0,0496*		
[04010] Μονοπάτι σηματοδότησης MAPK	3	0,051	**0,1167**	
[04360] Νευρική καθοδήγηση	2	0,059		
[04060] Αλληλεπίδραση υποδοχέα κυτοκίνη-κυτοκίνη	*2*	*0,1725*	***0,1167***	
[05200] Μονοπάτια στον καρκίνο	2	0,2064	***0,124***	
[01100] Μεταβολικά μονοπάτια	*3*	*0,4788*	***0,5335***	
Γονιδιακή υπογραφή 2				
[04672] Εντερικό ανοσοποιητικό δίκτυο για παραγωγή IgA	2	0,0056		
[05320] Αυτοάνοση θυρεοειδική νόσος	2	0,0056		
[01100] Μεταβολικά μονοπάτια	*6*	*0,0056*	***0,015***	***0,0917***
[04612] Επεξεργασία και παρουσίαση του αντιγόνου	*2*	*0,0096*		
[04650] Κυτταροτοξικότητα φυσικών κυττάρων-φονιάδων	*2*	*0,0175*	**0,015**	
[04020] Μονοπάτι σηματοδότησης ασβεστίου	*2*	*0,0239*		
[04060] Αλληλεπίδραση υποδοχέα κυτοκίνη-κυτοκίνη	*2*	*0,0432*	***0,0181***	
Γονιδιακή υπογραφή 3				
[04666] Fc γάμμα R-μεσολαβούμενη φαγοκυττάρωση	3	0,024	**0,0159**	
[01100] Μεταβολικά μονοπάτια	*10*	*0,024*	***0,004***	***0,4227***
[05322] Συστηματικός ερυθηματώδης λύκος	3	0,0328	**0,008**	
[04530] Στενοσύνδεσμος	3	0,0328		**0,0021**
[00562] Μεταβολισμός φωσφορικής ινοσιτόλης	2	0,0355	**0,009**	
[00564] Μεταβολισμός γλυκεροφωσφολιπιδίου	2	0,0362	**0,0115**	
[04020] Μονοπάτι σηματοδότησης ασβεστίου	*3*	*0,0362*	**0,009**	
[05120] Σηματοδότηση επιθηλιακών κυττάρων σε λοίμωξη ελικοβακτηριδίου του πυλωρού	2	0,0362		**0,0054**
[04070] Σύστημα σηματοδότησης φωσφατιδυλινοσιτόλης	2	0,0375	**0,0115**	
[04650] Κυτταροτοξικότητα φυσικών κυττάρων-φονιάδων	*2*	*0,0896*		
[05010] Νόσος Αλτσχάιμερ			0,0391	
[05016] Νόσος του Huntington			0,041	
[05200] Μονοπάτια στον καρκίνο			*0,1015*	

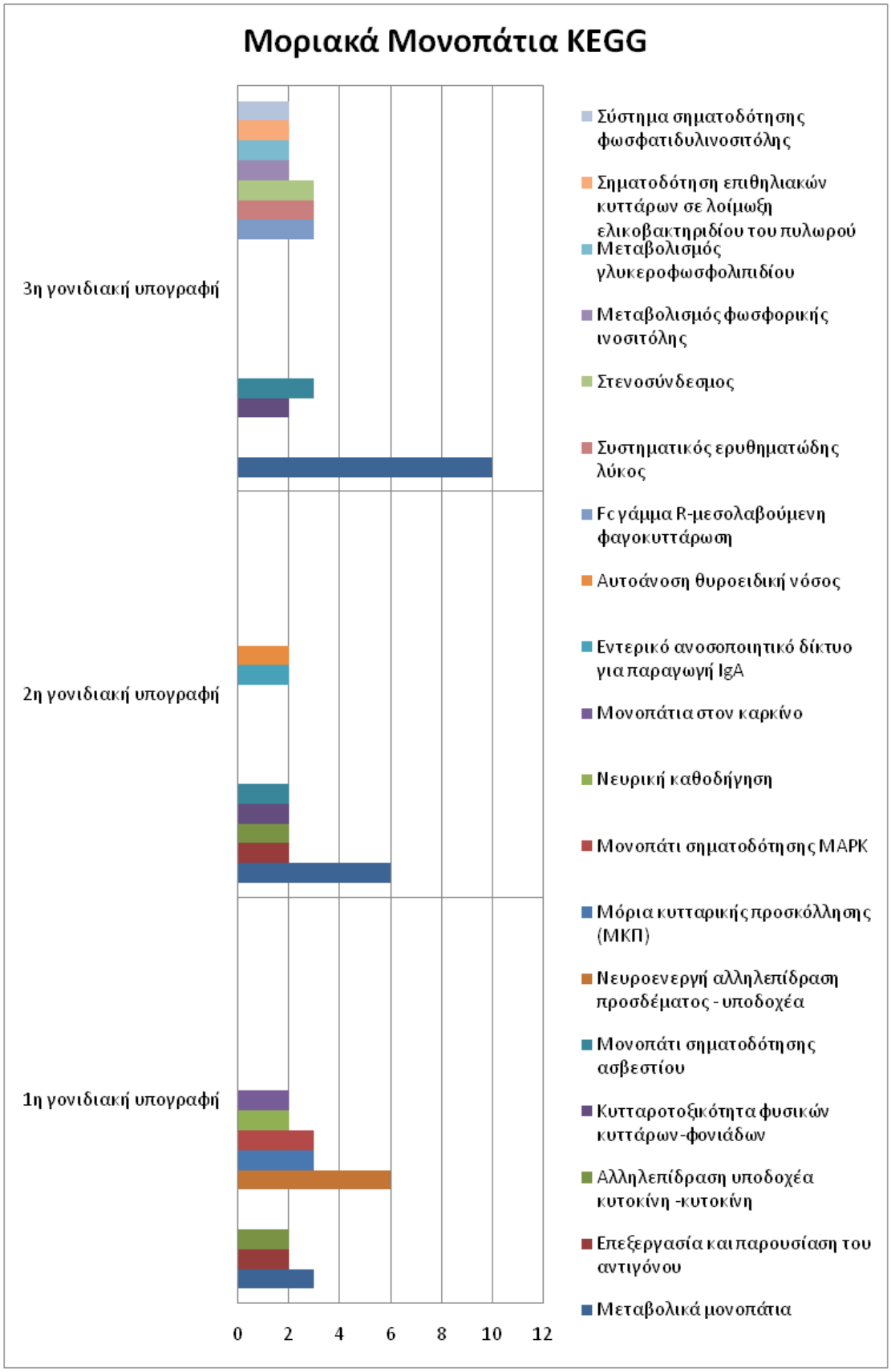
Τα μοριακά μονοπάτια που απαντώνται σε περισσότερες από μια γονιδιακές υπογραφές έχουν σημειωθεί με πλάγια γραφή.
ΑΡΙΘΜΟΣ ΓΟΝΙΔΙΩΝ (O): αναφέρεται στον αριθμό των γονιδίων από την εκάστοτε υπογραφή που συμμετέχουν στα συγκεκριμένα μοριακά μονοπάτια.
ΣΥΝΟΛΟ ΓΟΝΙΔΙΩΝ (C): αναφέρεται στο σύνολο των γονιδίων του ανθρώπινου γονιδιώματος που είναι γνωστά έως σήμερα ότι συμμετέχουν στα συγκεκριμένα μοριακά μονοπάτια.
Συντμήσεις: KEGG (Kyoto Encyclopedia of Genes and Genomes) Εγκυκλοπαίδεια Kyoto των γονιδίων και γονιδιωμάτων, P (power) ισχύς [εφαρμογή του υπεργεωμετρικού τεστ για τον υπολογισμό (της ισχύς) του εμπλουτισμού των όρων των μοριακών μονοπατιών KEGG], adjP (adjustment P) προσαρμογή P [εφαρμογή της μέθοδου Benjamini & Hochberg για την πολλαπλή προσαρμογή δοκιμής].

Εικόνα 58 : Σύγκριση των μοριακών μονοπατιών KEGG ανάμεσα στις τρεις γονιδιακές υπογραφές του Dataset A, του Dataset B και της συνένωσης των υπογραφών των δύο Datasets (AB).

ΟΜΟΙΟΤΗΤΕΣ ΤΩΝ ΜΟΡΙΑΚΩΝ ΜΟΝΟΠΑΤΙΩΝ KEGG ΑΝΑΜΕΣΑ ΣΤΙΣ 3 ΓΟΝΙΔΙΑΚΕΣ ΥΠΟΓΡΑΦΕΣ ΜΕΤΑ ΑΠΟ ΤΟΝ ΣΥΝΔΥΑΣΜΟ ΤΩΝ ΥΠΟΣΥΝΟΛΩΝ ΔΕΔΟΜΕΝΩΝ Α ΚΑΙ Β (U133A/U133B)						
ΒΙΟΛΟΓΙΚΕΣ ΛΕΙΤΟΥΡΓΙΕΣ ΜΟΡΙΑΚΩΝ ΜΟΝΟΠΑΤΙΩΝ	ΜΟΡΙΑΚΑ ΜΟΝΟΠΑΤΙΑ 2ο ΕΠΙΠΕΔΟ	ΜΟΡΙΑΚΑ ΜΟΝΟΠΑΤΙΑ 3ο ΕΠΙΠΕΔΟ ΚΑΤΗΓΟΡΙΑΣ KEGG	ΑΡΙΘΜΟΣ ΓΟΝΙΔΙΩΝ	ΣΥΝΟΛΟ ΓΟΝΙΔΙΩΝ	*P*	adj*P*
Γονιδιακή υπογραφή 1						
Ανοσολογική λειτουργία	Ανοσολογικό (Ανοσοποιητικό) σύστημα	*[04612] Επεξεργασία και παρουσίαση του αντιγόνου*	2	89	0,0186	0,0496
Κυτταρική σηματοδότηση	Μόρια σηματοδότησης και αλληλεπίδραση	[04080] Νευροενεργή αλληλεπίδραση προσδέματος-υποδοχέα	6	256	3,25E-05	0,0003
		[04514] Μόρια κυτταρικής προσκόλλησης (ΜΚΠ)	3	134	0,0039	0,0156
		[04060] Αλληλεπίδραση υποδοχέα κυτοκίνη-κυτοκίνη	2	267	0,1294	0,1725
	Μεταγωγή σήματος	[04010] Μονοπάτι σηματοδότησης MAPK	3	269	0,0255	0,051
Μεταβολισμός	Μεταβολισμός	*[01100] Μεταβολικά μονοπάτια*	3	1104	0,4788	0,4788
Κυτταρική ανάπτυξη	Ανάπτυξη	[04360] Νευρική καθοδήγηση	2	129	0,0369	0,059
Άλλες λειτουργίες	Καρκίνοι	[05200] Μονοπάτια στον καρκίνο	2	330	0,1806	0,2064
Γονιδιακή υπογραφή 2						
Ανοσολογική λειτουργία	Ανοσολογικό (Ανοσοποιητικό) σύστημα	[04672] Εντερικό ανοποιητικό δίκτυο για παραγωγή IgA	2	50	0,0018	0,0056
		[04612] Επεξεργασία και παρουσίαση του αντιγόνου	2	89	0,0055	0,0096
		[04650] Κυτταροτοξικότητα φυσικών κυττάρων-φονιάδων	2	137	0,0125	0,0175
	Ανοσολογικές παθήσεις	[05320] Αυτοάνοση θυρεοειδική νόσος	2	53	0,002	0,0056
Κυτταρική σηματοδότηση	Μόρια σηματοδότησης και αλληλεπίδραση	*[04060] Αλληλεπίδραση υποδοχέα κυτοκίνη-κυτοκίνη*	2	267	0,0432	0,0432
	Μεταγωγή σήματος	*[04020] Μονοπάτι σηματοδότησης ασβεστίου*	2	178	0,0205	0,0239
Μεταβολισμός	Μεταβολισμός	*[01100] Μεταβολικά μονοπάτια*	6	1104	0,0024	0,0056
Γονιδιακή υπογραφή 3						
Ανοσολογική λειτουργία	Ανοσολογικό (Ανοσοποιητικό) σύστημα	[04666] Fc γάμμα R-μεσολαβούμενη φαγοκυττάρωση	3	97	0,003	0,024
		[04650] Κυτταροτοξικότητα φυσικών κυττάρων-φονιάδων	2	137	0,0616	0,0896
	Ανοσολογικές παθήσεις	[05322] Συστηματικός ερυθηματώδης λύκος	3	140	0,0082	0,0328
Κυτταρική σηματοδότηση	Μεταγωγή σήματος	*[04020] Μονοπάτι σηματοδότησης ασβεστίου*	3	178	0,0157	0,0362
		[04070] Σύστημα σηματοδότησης φωσφατιδυλινοσιτόλης	2	76	0,0211	0,0375
Κυτταρική επικοινωνία	Κυτταρική επικοινωνία	[04530] Στενοσύνδεσμος	3	134	0,0073	0,0328
Μεταβολισμός	Μεταβολισμός	*[01100] Μεταβολικά μονοπάτια*	10	1104	0,0016	0,024
	Μεταβολισμός των υδατανθράκων	[00562] Μεταβολισμός φωσφορικής ινοσιτόλης	2	54	0,0111	0,0355
	Μεταβολισμός των λιπιδίων	[00564] Μεταβολισμός γλυκεροφωσφολιπιδίου	2	70	0,0181	0,0362
Άλλες λειτουργίες	Λοιμώδη νοσήματα	**[05120] Σηματοδότηση επιθηλιακών κυττάρων σε λοίμωξη ελικοβακτηριδίου του πυλωρού**	2	68	0,0171	0,0362

Τα μοριακά μονοπάτια που απαντώνται σε περισσότερες από μια γονιδιακές υπογραφές έχουν σημειωθεί με πλάγια γραφή.
ΑΡΙΘΜΟΣ ΓΟΝΙΔΙΩΝ: αναφέρεται στον αριθμό των γονιδίων από την εκάστοτε υπογραφή που συμμετέχουν στα συγκεκριμένα μοριακά μονοπάτια.
ΣΥΝΟΛΟ ΓΟΝΙΔΙΩΝ: αναφέρεται στο σύνολο των γονιδίων του ανθρώπινου γονιδιώματος που είναι γνωστά έως σήμερα ότι συμμετέχουν στα συγκεκριμένα μοριακά μονοπάτια.
Συντμήσεις: P (power) ισχύς [εφαρμογή του υπεργεωμετρικού τεστ για τον υπολογισμό (της ισχύς) του εμπλουτισμού των όρων των μοριακών μονοπατιών KEGG], adjP (adjustment P) προσαρμογή P [εφαρμογή της μέθοδου Benjamini & Hochberg για την πολλαπλή προσαρμογή δοκιμής].

Εικόνα 59 : Ομοιότητες των μοριακών μονοπατιών KEGG ανάμεσα στη συνένωση των τριών γονιδιακών υπογραφών του Dataset A και B.

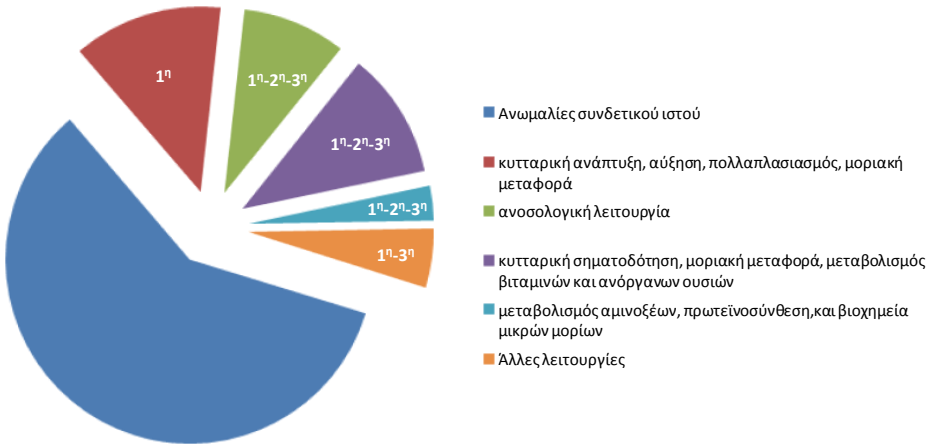


Εικόνα 60: Σχηματική απεικόνιση των συγκριτικών αποτελεσμάτων για τα μοριακά μονοπάτια KEGG (3^{ο} επίπεδο κατηγορίας KEGG) στα οποία συμμετέχουν τα γονίδια που περιλαμβάνονται στη συνένωση των τριών γονιδιακών υπογραφών του Dataset A με του Dataset B.

Τα βιολογικά μονοπάτια που κυριαρχούν στη 2η γονιδιακή υπογραφή και εμφανίζονται στη 3η και σε μικρότερο βαθμό στη 1η γονιδιακή υπογραφή είναι κυρίως τα μονοπάτια που σχετίζονται με την ειδική ανοσολογική απάντηση (επεξεργασία και παρουσίαση του αντιγόνου, εντερικό ανοσοποιητικό δίκτυο για παραγωγή IgA, Fc γάμμα R-μεσολαβούμενη φαγοκυττάρωση) και τη μη-ειδική ανοσολογική απάντηση (αλληλεπίδραση υποδοχέα κυτοκίνη-κυτοκίνη) δηλαδή τη φλεγμονή, ενώ άλλα κοινά μοριακά μονοπάτια και των τριών γονιδιακών υπογραφών αποτελούν τα μεταβολικά μονοπάτια και η κυτταρική σηματοδότηση (Εικόνες 59-60). Τα συγκεκριμένα ευρήματα συμφωνούν με μελέτες γονιδιακής έκφρασης [23] καθώς και πρωτεομικές και μεταβολομικές μελέτες για την ΟΑ. Αναλυτικότερα, μεταβολομικές και πρωτεομικές μελέτες αναφέρονται σε έναν τροποποιημένο ενεργειακό μεταβολισμό και σε μεταβολές μορίων που εμπλέκονται στον μεταβολισμό των υδατανθράκων και των λιπιδίων, γεγονός που συνάδει με τα ευρήματά μας (μεταβολικά μονοπάτια) και ειδικότερα με τη 3η γονιδιακή υπογραφή. Ταυτόχρονα, η κυτταρική σηματοδότηση, όπως για παράδειγμα οι αλλαγές στην έκφραση διαφόρων κυτοκινών, επιβεβαιώνεται από διάφορες μελέτες γονιδιακής έκφρασης [23]. Τέλος, η ανοσολογική λειτουργία που αποτυπώνεται και στις τρεις γονιδιακές υπογραφές ανακλά τις φλεγμονώδεις αλλοιώσεις που επισυμβαίνουν δευτερογενώς (με τα μέχρι στιγμής δεδομένα) στον αρθρικό υμένα [9].

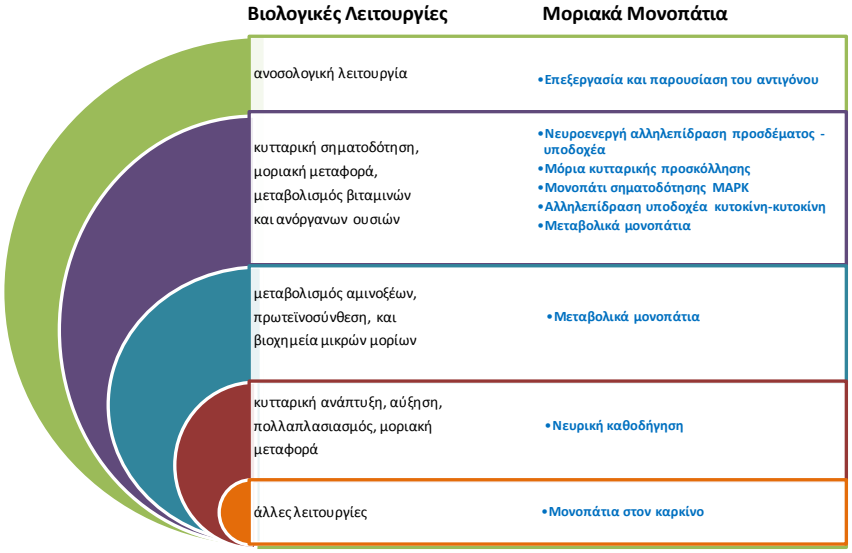
Στην Εικόνα 61, που ακολουθεί, παρατηρούμε την αντιστοίχιση των μοριακών μονοπατιών των τριών γονιδιακών υπογραφών στις γνωστές μοριακές λειτουργίες, οι οποίες έχουν προκύψει από αρκετές μελέτες γονιδιακής έκφρασης στην ΟΑ [23].

Σύνδεση των μοριακών μονοπατιών από τις 3 γονιδιακές υπογραφές με γνωστές βιολογικές λειτουργίες της ΟΑ



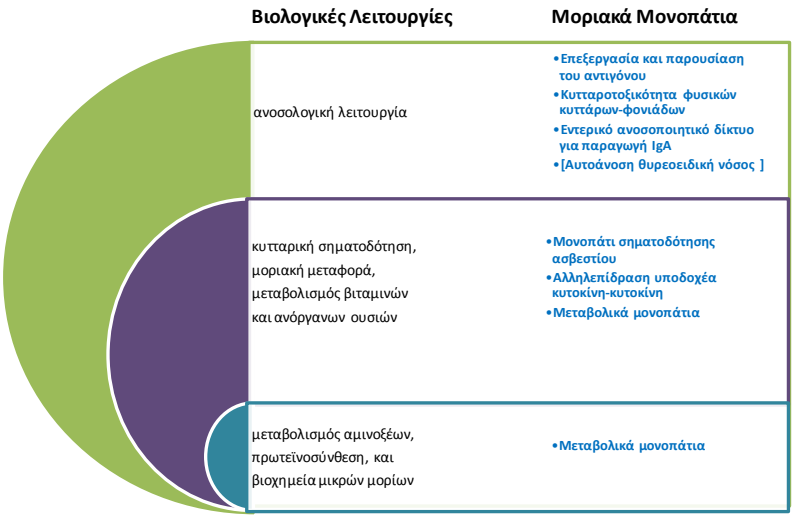
Εικόνα 61: Γράφημα – πίτα που παρουσιάζει τη κατανομή των βιολογικών λειτουργιών στο επίπεδο των μεταγράφων (mRNAs) που έχουν προκύψει από αρκετές έρευνες για την ασθένεια της ΟΑ. Σε κάθε κομμάτι σημειώνεται η γονιδιακή υπογραφή της οποίας τα μονοπάτια βρέθηκαν να σχετίζονται με την αντίστοιχη βιολογική λειτουργία.

Αντιστοίχιση των μοριακών μονοπατιών της 1ης γονιδιακής υπογραφής σε γνωστές βιολογικές λειτουργίες της ΟΑ από μελέτες γονιδιακής έκφρασης



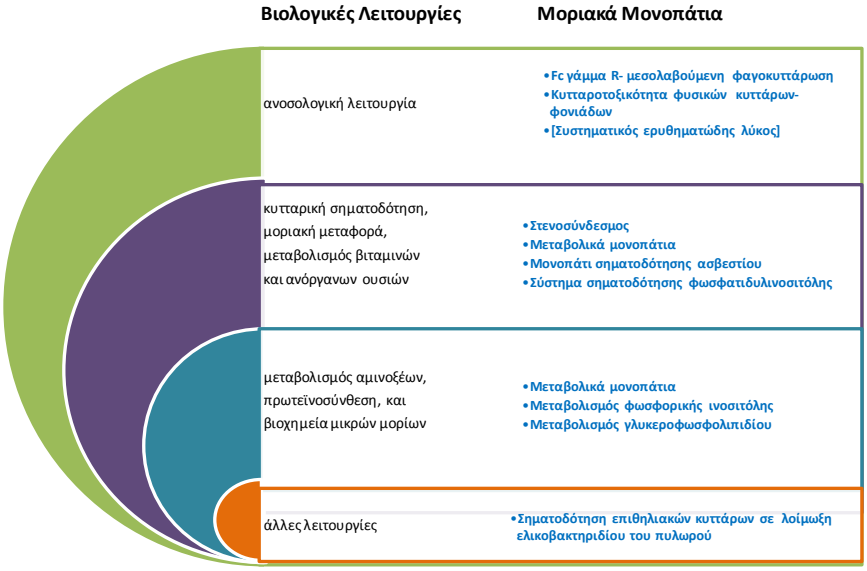
Εικόνα 62: Σχηματικό διάγραμμα που αντιστοιχίζει γνωστές βιολογικές λειτουργίες της ΟΑ που εμφανίζονται στο γράφημα της Εικόνας 61 με τα μοριακά μονοπάτια της 1ης γονιδιακής υπογραφής.

Αντιστοίχιση των μοριακών μονοπατιών της 2ης γονιδιακής υπογραφής σε γνωστές βιολογικές λειτουργίες της ΟΑ από μελέτες γονιδιακής έκφρασης



Εικόνα 63: Σχηματικό διάγραμμα που αντιστοιχίζει γνωστές βιολογικές λειτουργίες της ΟΑ που εμφανίζονται στο γράφημα της Εικόνας 61 με τα μοριακά μονοπάτια της 2ς γονιδιακής υπογραφής.

Αντιστοίχιση των μοριακών μονοπατιών της 3ης γονιδιακής υπογραφής σε γνωστές βιολογικές λειτουργίες της ΟΑ από μελέτες γονιδιακής έκφρασης



Εικόνα 64: Σχηματικό διάγραμμα που αντιστοιχίζει γνωστές βιολογικές λειτουργίες της ΟΑ που εμφανίζονται στο γράφημα της Εικόνας 61 με τα μοριακά μονοπάτια της 3ης γονιδιακής υπογραφής.

Σήμερα, η ΟΑ θεωρείται ως νόσος όλων των ιστών της άρθρωσης, δηλαδή του χόνδρου του αρθρικού υμένα και του υποχόνδριου οστού. Η ΟΑ, σε μεγαλύτερη ή μικρότερη έκταση χαρακτηρίζεται πάντοτε από εκφύλιση του αρθρικού χόνδρου και ταυτόχρονη ανάπτυξη νέου οστού, χόνδρου και συνδετικού ιστού [9]. Στα κύτταρα του ΟΑ χόνδρου, συντελούνται διεργασίες όπως αυτές που αποτυπώνονται στην Εικόνα 61 και περιλαμβάνουν τον κυτταρικό πολλαπλασιασμό, τη βιοσυνθετική και καταβολική δραστηριότητα (μεταβολισμός), και άλλες λειτουργίες (απόπτωση, φαινοτυπικές αλλαγές των χονδροκυττάρων) [9]. Οι ανωμαλίες του συνδετικού ιστού που φαίνεται ότι κατέχουν μια σημαντική θέση στην κατανομή των βιολογικών λειτουργιών (Εικόνα 61) απεικονίζουν τις ανωμαλίες που σημειώνονται κατά τη βλάβη του δικτύου του κολλαγόνου στα κύτταρα του ΟΑ χόνδρου (εξειδικευμένη μορφή συνδετικού ιστού). Η βλάβη του δικτύου του κολλαγόνου προκύπτει από έναν καταρράκτη κυτταρικών γεγονότων, που διαδραματίζουν σημαντικό ρόλο στην παθογένεια της ΟΑ, και έχει ως αποτέλεσμα τη πλήρη καταστροφή της θεμέλιας ουσίας αλλά και των χονδροκυττάρων και κατ' επέκταση του αρθρικού χόνδρου [9]. Επιπρόσθετα, νέες μελέτες που έρχονται στο φως όσον αφορά στη φλεγμονή που αναπτύσσεται στον αρθρικό υμένα, αναδεικνύουν την πιθανότητα η οστεοαρθρίτιδα να είναι μια φλεγμονώδης νόσος, όπως η ρευματοειδής αρθρίτιδα [9].

Σύγκριση βιολογικών αποτελεσμάτων με άλλες μελέτες

Μετά την επεξεργασία των γονιδίων που περιλαμβάνονται στις υπογραφές με το σύστημα ταξινόμησης WebGestalt προχωράμε στη σύγκριση των βιολογικών ευρημάτων μας με αντίστοιχα ευρήματα από τη μελέτη των Huber και συνεργατών [5] καθώς και από την έρευνα των Davis και συνεργατών [64]. Όπως έχουμε αναφέρει στόχος της παρούσας εργασίας είναι η εύρεση ενός συνόλου γονιδίων (γονιδιακή υπογραφή), με επαρκή διαχωριστική ικανότητα ανάμεσα στις καταστάσεις-κλάσεις ενδιαφέροντος, το οποίο μπορεί να χρησιμοποιηθεί για την κατασκευή ενός συστήματος πρόβλεψης με σκοπό την ορθή ταξινόμηση νέων αθέατων καταστάσεων (νέων δειγμάτων). Αντίθετα, η έρευνα των Huber και συνεργατών [5], από την οποία αντλήσαμε τα σύνολα δεδομένων που επεξεργαστήκαμε, χρησιμοποιεί διαφορετικό τρόπο επεξεργασίας των δεδομένων και στοχεύει στην εύρεση μεμονωμένων γονιδίων τα οποία παρουσιάζουν διαφορετικές τιμές διακύμανσης ανάμεσα στις κλάσεις ενδιαφέροντος. Όσο αναφορά τη μελέτη των Davis και συνεργατών [64] η μεθοδολογία που ακολουθείται για επεξεργασία των γονιδιακών

δεδομένων είναι παρόμοια σε σύγκριση με αυτή που εφαρμόσαμε στη παρούσα διπλωματική εργασία. Ωστόσο, η βασική και αρκετά σημαντική διαφορά της εργασίας μας με αυτή του Davis εντοπίζεται στη χρήση ΟΑ δεδομένων που προέρχονται από διαφορετικά τμήματα ιστού (παρούσα εργασία : δείγματα από αρθρική μεμβράνη – εργασία Davis και συνεργατών : δείγματα από αρθρικό χόνδρο).

Για την επίτευξη της σύγκρισης της μελέτης μας με τις παραπάνω εργασίες [5], [65] σε επίπεδο βιολογικών μονοπατιών ακολούθησε η κατάλληλη επεξεργασία των στοιχείων που αντλήθηκαν από την εκάστοτε έρευνα. Έτσι, 1) στη μελέτη των Huber και συνεργατών έγινε αντιστοίχιση των μοριακών μονοπατιών ($3^{ο}$ επίπεδο κατηγορίας KEGG) στις ευρύτερες κατηγορίες τους ($2^{ο}$ επίπεδο κατηγορίας KEGG) [5], ενώ 2) στη μελέτη των Davis και συνεργατών βρέθηκαν με τη βοήθεια του συστήματος ταξινόμησης WebGestalt, τα μοριακά μονοπάτια KEGG ($2^{ο}$ επίπεδο κατηγορίας KEGG) για τα 32 γονίδια που σχετίζονται με την ΟΑ [65]. Η παραπάνω επεξεργασία παρουσιάζεται αναλυτικά στο Παράρτημα Β.

Έπειτα από επεξεργασία και των δικών μας αποτελεσμάτων, όπου τα εξειδικευμένα μοριακά μονοπάτια ($3^{ο}$ επίπεδο κατηγορίας KEGG) αποδόθηκαν στις πιο γενικευμένες κατηγορίες τους ($2^{ο}$ επίπεδο κατηγορίας KEGG) (Εικόνα 59), και προχωρώντας στη σύγκρισή τους με τις παραπάνω μελέτες (Huber και συν., Davis και συν.) παρατηρήσαμε αρκετές ομοιότητες, όπως φαίνεται στον Πίνακα 11.

Για παράδειγμα, η κυτταρική σηματοδότηση και η κυτταρική επικοινωνία ($2^{ο}$ επίπεδο κατηγορίας KEGG) εμφανίζονται τόσο στα δικά μας αποτελέσματα όσο και στις άλλες δυο μελέτες, ενώ η ανοσολογική λειτουργία και ο μεταβολισμός/πρωτεϊνοσύνθεση παρουσιάζονται τόσο στα δικά μας αποτελέσματα όσο και στη μελέτη των Huber και συνεργατών. Επίσης, η περαιτέρω σύγκριση της μελέτης μας με εκείνη του Huber και συν.[5] (στο $3^{ο}$ επίπεδο κατηγορίας KEGG) ανέδειξε ένα κοινό μονοπάτι που είναι το μονοπάτι σηματοδότησης επιθηλιακών κυττάρων σε λοίμωξη ελικοβακτηριδίου του πυλωρού ($3^{ο}$ επίπεδο κατηγορίας KEGG) της $3^{ης}$ γονιδιακής υπογραφής της παρούσας εργασίας. Αξίζει να σημειωθεί, ότι το μονοπάτι της αλληλεπίδρασης υποδοχέα κυτοκίνη-κυτοκίνη της $2^{ης}$ γονιδιακής υπογραφής, αλλά και το μονοπάτι σηματοδότησης MAPK της $1^{ης}$ υπογραφής που σχετίζονται με την ανοσολογική λειτουργία (Εικόνα 59, Πίνακας 11) αναφέρονται στην εργασία των Huber και συνεργατών ως μονοπάτια που απαντώνται μόνο στην ρευματοειδή αρθρίτιδα και δεν εμφανίζονται σε ασθενείς με ΟΑ [5].

Οι ομοιότητες αλλά και οι διαφορές που εντοπίζονται (Πίνακας 11) θα μπορούσαν ενδεχομένως να οφείλονται στη διαφορετική μεθοδολογική προσέγγιση και στα διαφορετικά χαρακτηριστικά των ΟΑ δεδομένων (παρούσα εργασία και μελέτη Huber και συν.: δείγματα από αρθρικό υμένα, εργασία Davis και συν.: δείγματα από αρθρικό χόνδρο) που χρησιμοποίησε κάθε μελέτη. Παράλληλα, όπως ήδη έχει αναφερθεί, η ΟΑ - γνωστή και ως εκφυλιστική νόσος των αρθρώσεων - είναι μια νόσος ολόκληρης της άρθρωσης, με την έννοια ότι όλα τα δομικά της στοιχεία (αρθρικός χόνδρος, αρθρικός υμένας, υποχόνδριο οστό) συμμετέχουν στην παθογένειά της, η οποία προκύπτει από την αλληλεπίδραση πολύπλοκων μηχανικών και βιολογικών διεργασιών. Αυτό έχει ως αποτέλεσμα, οι διάφορες βιολογικές διεργασίες, μοριακά μονοπάτια και παράγοντες (κυτοκίνες, ένζυμα, αυξητικοί παράγοντες) που εμπλέκονται στην παθογένεια της νόσου, να σχετίζονται πολλές φορές μ' έναν ξεχωριστό ιστό (π.χ. αρθρικό υμένα ή αρθρικό χόνδρο) ή να έχουν διαφορετικό ποσοστό και βαρύτητα συμμετοχής σε κάθε ιστό (όσον αφορά την παθογένεια της νόσου) [9].

Πίνακας 11 : Πίνακας που αποτυπώνει τη σύγκριση των μοριακών μονοπατιών που προκύπτουν από τη παρούσα μελέτη με αντίστοιχα μονοπάτια που παρουσιάζονται στις εργασίες των Huber και Davis. Με "+" ή "-" συμβολίζεται αντίστοιχα η εμφάνιση ή η απουσία ενός συγκεκριμένου μονοπατιού στις εργασίες των Huber και Davis.

ΣΥΓΚΡΙΣΗ ΜΟΡΙΑΚΩΝ ΜΟΝΟΠΑΤΙΩΝ [2^{ο} ΕΠΙΠΕΔΟ ΚΑΤΗΓΟΡΙΑΣ KEGG] ΜΕ ΠΑΡΟΜΟΙΕΣ ΜΕΛΕΤΕΣ		
Μοριακά Μονοπάτια στη παρούσα εργασία	**Εργασία Huber (κοινό Dataset- διαφορετική μεθοδολογία)**	**Εργασία Davis (διαφορετικό Dataset- παρόμοια μεθοδολογία)**
Ανοσολογικό σύστημα	+	-
Μεταγωγή σήματος	+	-
Μόρια σηματοδότησης και αλληλεπίδραση	-	+
Μεταβολισμός / Πρωτεϊνοσύνθεση	+	-
Κυτταρική επικοινωνία	+	+
Άλλες λειτουργίες (π.χ. καρκίνοι, λοιμώδη νοσήματα)	+	+

Ταυτόχρονα, τα αποτελέσματά μας (Εικόνες 61-64) συμφωνούν με πρόσφατες έρευνες, οι οποίες καταδεικνύουν μέσω ορολογικών και ιστολογικών αποδείξεων την ύπαρξη φλεγμονής στον αρθρικό υμένα (synovitis) σε αρχικά στάδια της νόσου, και την

κατάρριψη έτσι της άποψης ότι η φλεγμονή εμφανίζεται μόνο σε προχωρημένα στάδια της νόσου [9]. Επίσης, με τη χρήση απεικονιστικής μεθοδολογίας (MRI) έχει δειχθεί ότι, σε ασθενείς με ΟΑ γόνατος παρατηρείται πάχυνση της αρθρικής μεμβράνης που ανέρχεται σ' ένα ποσοστό 73%, και μάλιστα σε αυτούς με σχετικά πρώιμη μορφή της νόσου [9]. Αυτή η πάχυνση, βρέθηκε ότι αντιστοιχούσε σε μετρίου βαθμού χρόνια φλεγμονή του αρθρικού υμένα [9].

Συνοψίζοντας, η αποτύπωση των σημαντικότερων μοριακών μονοπατιών των γονιδιακών υπογραφών, ιδιαίτερα της $2^{ης}$ και ακολούθως της $3^{ης}$, συνηγορούν με την άποψη των πρόσφατων μελετών, που προσδίδουν στη φλεγμονή κύριο ρόλο τόσο για τη συμπτωματολογία της νόσου, όσο και για την εξέλιξή της, γεγονός που πιθανολογεί τον μελλοντικό αποχαρακτηρισμό της οστεοαρθρίτιδας ως μη φλεγμονώδους και τον χαρακτηρισμό της ως φλεγμονώδους αρθροπάθειας, όπως είναι η ρευματοειδής αρθρίτιδα.

Genotator

Εκτός από την παράθεση των βιολογικών διεργασιών και των μοριακών μονοπατιών, θελήσαμε 1) να διερευνήσουμε τον αριθμό των κοινών γονιδίων στις τρείς γονιδιακές υπογραφές, και 2) να αξιολογήσουμε τις υπογραφές μας σύμφωνα με το εργαλείο Genotator, έτσι ώστε να αναδείξουμε σημαντικά γονίδια που σχετίζονται με την ΟΑ.

Όπως αναφέρθηκε, ο αριθμός των κοινών γονιδίων μεταξύ των γονιδιακών υπογραφών τόσο του Dataset A, όσο και του Dataset B (Πίνακας 12) είναι πολύ μικρός και αφορά μόνο την $1^{η}$ και $2^{η}$ γονιδιακή υπογραφή, περιλαμβάνοντας συνολικά πέντε γονίδια, δυο εκ των οποίων δεν έχουν σχολιαστεί ακόμα (άγνωστο σύμβολο, κωδικός γονιδίου και περιγραφή) και άρα δεν γνωρίζουμε τον λειτουργικό τους ρόλο. Επιπλέον, μέχρι σήμερα η περιγραφή των γονιδίων *ASCL2* (μέλος της οικογένειας των μεταγραφικών παραγόντων της δομής βασική έλικα-βρόχος-έλικα (bHLH) που φαίνεται να διαδραματίζουν σημαντικό ρόλο στην ανάπτυξη και την κυτταρική δραστηριότητα), *NRTN* (μέλος της υπό-οικογένειας TRN του τροποποιητικού παράγοντα της ανάπτυξης-β (TGF-β), που αποτελεί έναν τύπο κυτοκίνης και παίζει ρόλο στην ανοσία, τον καρκίνο και άλλες ασθένειες) και ιδιαίτερα του *FAM196A* (C10orf141) δεν είναι επαρκής οδηγώντας μόνο στην έμμεση συσχέτιση αυτών των γονιδίων με την ΟΑ [66]. Ωστόσο, αξίζει να σημειωθεί ότι σύμφωνα με την μελέτη των Huber και συνεργατών ο τροποποιητικός παράγοντας της ανάπτυξης-β (TGF-β), που

συμμετέχει στη φλεγμονώδη απάντηση, εμφανίζεται σε ασθενείς με ρευματοειδή αρθρίτιδα και δεν εμφανίζεται σε ασθενείς με ΟΑ [5].

Πίνακας 12 : Στις πρώτες δύο γραμμές του πίνακα παρουσιάζονται τα 2 κοινά γονίδια μεταξύ της 1ης και 2ης Γονιδιακής Υπογραφής του Dataset A. Στις υπόλοιπες γραμμές παρουσιάζονται τα 3 κοινά γονίδια της 1ης και 2ης Γονιδιακής Υπογραφής του Dataset B.

Dataset	Κωδικός καταχώρησης στην GEO (GEO Accession viewer ID)	Κωδικός γονιδίου (GENE_ID)	Σύμβολο γονιδίου
A	207607_at	430	*ASCL2*
	210683_at	4902	*NRTN*
B	240738_at	A	A
	241109_at	A	A
	244435_at	642938	*FAM196A (C10orf141)*

Το Genotator [29], όπως περιγράψαμε στο Κεφάλαιο 1.5, είναι ένα πολύ χρήσιμο εργαλείο για να εξετάσουμε κατά πόσο τα δεδομένα μας σχετίζονται με κάποια ασθένεια. Συγκρίναμε λοιπόν τα αποτελέσματα των γονιδιακών μας υπογραφών με τη λίστα που μας δίνει το Genotator για τα γονίδια που σχετίζονται με την οστεοαρθρίτιδα. Εξετάσαμε κάθε μια από τις τρεις υπογραφές στο σύνολο της (γονίδια υπογραφής Dataset A + γονίδια υπογραφής Dataset B) και τα αποτελέσματα που προέκυψαν παρουσιάζονται στους Πίνακες 13 και 14.

Πίνακας 13 : Γονίδια στο Genotator.

Γονιδιακή Υπογραφή	Αριθμός Γονιδίων	Σχετιζόμενα με ΟΑ
1	139	5
2	84	2
3	149	1

Πίνακας 14 : Τα γονίδια των υπογραφών που σύμφωνα με τη λίστα του Genotator σχετίζονται με την ΟΑ.

	Κωδικός Γονιδίου (NCBI Gene ID)	Σύμβολο Γονιδίου (Gene Symbol)	Περιγραφή (Gene Name)	Σκορ (Genotator Score)	Σειρά Ταξινόμησης (Genotator)
1η Γονιδιακή Υπογραφή	2100	ESR2	estrogen receptor 2 (ER beta)	4.8	**48**
	4049	LTA	lymphotoxin alpha (TNF superfamily, member 1)	1.1	154
	100132285	KIR2DS2	killer cell immunoglobulin-like receptor, two domains, short cytoplasmic tail, 2	1.1	106
	3806	KIR2DS1	killer cell immunoglobulin-like receptor, two domains, short cytoplasmic tail, 1	1.1	107
	3809	KIR2DS4	killer cell immunoglobulin-like receptor, two domains, short cytoplasmic tail, 4	1.1	108
2η Γονιδιακή Υπογραφή	1301	COL11A1	collagen, type XI, alpha 1	1.1	115
	3117	HLA-DQA1	major histocompatibility complex, class II, DQ alpha 1	7.0	**28**
3η Γονιδιακή Υπογραφή	23245	ASTN2	astrotactin 2	4.9	**44**

Παρόλο που το ποσοστό των γονιδίων που σχετίζονται με την ΟΑ, σύμφωνα με το Genotator, είναι μόλις 2% σε ένα σύνολο 374 γονιδίων, μπορούμε να παρατηρήσουμε στην τελευταία στήλη του παραπάνω πίνακα (Πίνακας 14) ότι τα συγκεκριμένα γονίδια, και ιδιαίτερα το *HLA-DQA1* της 2^{ης} υπογραφής αλλά και τα γονίδια *ASTN2* της 3^{ης} υπογραφής και *ESR2* της 1^{ης} υπογραφής βρίσκονται σε υψηλή κατάταξη. Επίσης, λαμβάνοντας υπόψη ότι 176 από τα 374 γονίδια έχουν σκορ 0.0, προκύπτει ότι αυτά τα 8 γονίδια αποτελούν σημαντικά ευρήματα της παρούσας εργασίας. Αναλυτικότερα, τα γονίδια που παρουσιάζονται στον Πίνακα 14 ενισχύουν τα ευρήματα που προκύπτουν από τη μελέτη των μοριακών μονοπατιών καθώς τα γονίδια *HLA-DQA1, KIR2DS2, KIR2DS1, KIR2DS4,* και *LTA* συσχετίζονται με την ανοσολογική απάντηση [66]. Επίσης, το γονίδιο *COL11A1* της 2^{ης} υπογραφής αποτελεί έναν σημαντικό παράγοντα διαταραχής του συνδετικού ιστού, το γονίδιο *ESR2* της 1^{ης} υπογραφής αποτελεί έναν σημαντικό μεταγραφικό παράγοντα (με άγνωστο λειτουργικό ρόλο, πιθανά στην αύξηση), ενώ το γονίδιο *ASTN2* της 3^{ης} υπογραφής πιθανά παίζει ρόλο στην νευρωνική μετανάστευση [66].

Επιπλέον, μελέτες όπως οι ευρείες γονιδιωματικές μελέτες σύνδεσης (GWAS) [3], οι οποίες αποτελούν την πιο αποτελεσματική προσέγγιση στη μελέτη της γενετικής των ανθρώπινων ασθενειών, αναδεικνύουν τη σημασία των γονιδίων *COL11A1* [67]*, HLA-DQA1* [68] της 2^{ης} υπογραφής, καθώς και του γονιδίου *ASTN2* [69] της 3^{ης} υπογραφής. Ειδικότερα, οι GWAS αναλύουν γενετικές παραλλαγές, και συγκεκριμένα μονονουκλεοτιδικούς πολυμορφισμούς (SNPs) προκειμένου να καθορίσουν τις γενετικές ποικιλομορφίες που σχετίζονται είτε με την εμφάνιση είτε με τη βαρύτητα της ασθένειας της ΟΑ. Συγκεκριμένα, έχουν εντοπιστεί τέσσερις παραλλαγές στο γονίδιο *COL11A1* (γονίδιο κολλαγόνου τύπου *XIα1,* ένα ισχυρό υποψήφιο γονίδιο για την εκφυλιστική μυοσκελετική πάθηση της ΟΑ) της 2^{ης} γονιδιακής υπογραφής, που μπορεί να συμβάλλουν στην παθογένεση της πρώιμης εμφάνιση της νόσου της ΟΑ, η οποία συχνά θεωρείται ότι είναι μια ολιγογονιδιακή ασθένεια [67]. Μια πρόσφατη μελέτη αναφέρεται σε 2 SNPs (rs7775228 and rs10947262) που συνδέονται με ευαισθησία σε ΟΑ του γόνατος [68]. Οι δύο SNPs εντοπίστηκαν σε μια περιοχή που περιέχει γονίδια της τάξης II/III του μείζονος συμπλέγματος ιστοσυμβατότητας HLA (MHC) που περικλείει το γονίδιο *HLA-DQA1* [68] της 2^{ης} υπογραφής, ενώ έχει βρεθεί ένας σημαντικός γενετικός τόπος στο χρωμόσωμα 9 που γειτνιάζει με το γονίδιο *ASTN2* [69] της 3^{ης} υπογραφής. Οι συγκεκριμένες μελέτες αναδεικνύουν τον δυνητικό ρόλο αυτών των γονιδίων, *HLA-DQA1, COL11A1*, και *ASTN2* στον εντοπισμό πρόωρων οστεοαρθριτιδικών αλλοιώσεων.

Όλες αυτές οι μελέτες, όπως αναφέρθηκε, έχουν ως στόχο τον προσδιορισμό γενετικών ποικιλομορφιών που θα προβλέπουν είτε την προδιάθεση για εμφάνιση είτε την έκφραση της σοβαρότητας της οστεοαρθρίτιδας. Επιπλέον, μπορεί να χρησιμεύσουν ως οδηγός για εξατομικευμένες θεραπευτικές παρεμβάσεις, καθώς η ΟΑ αντιμετωπίζεται σήμερα με θεραπευτικές επεμβάσεις, οι οποίες είναι συμπωματικές θεραπείες, και δεν δρουν στην παθογένεια της νόσου [9], [23]. Οι θεραπευτικές επιλογές αποσκοπούν στην εξάλειψη των συμπτωμάτων, ιδίως του πόνου (με την χρήση των απλών αναλγητικών ή των μη στεροειδών αντιφλεγμονωδών φαρμάκων ή ακόμα και με την ενδοαρθρική έγχυση κορτιζόλης στην ΟΑ άρθρωση), ενώ σε προχωρημένα στάδια της νόσου η θεραπευτική παρέμβαση είναι αρκετές φορές συνυφασμένη με την χειρουργική επέμβαση (αρθροπλαστική). Έτσι, σήμερα δίνεται έμφαση στη γονιδιακή μελέτη της ΟΑ ώστε να εντοπιστούν γονίδια που εμπλέκονται όχι μόνο στην προδιάθεση για ΟΑ, αλλά και στην εξέλιξη και στη σοβαρότητα της νόσου. Επιπρόσθετα, ιδιαίτερη βαρύτητα δίνεται στην κατανόηση των διάφορων μοριακών μονοπατιών που συμμετέχουν στην ΟΑ και της αλληλεπίδρασή τους, για την ανάπτυξη νέων θεραπειών που θα έχουν ως στόχο την αιτιοπαθογένεση της ΟΑ, με αποτέλεσμα την οριστική της θεραπεία [9].

Συμπερασματικά, τα αποτελέσματα της παρούσας διπλωματικής εργασίας, τα οποία συνάδουν με τα στατιστικά αποτελέσματα της παρούσας μελέτης αναδεικνύουν την $2^{η}$ γονιδιακή υπογραφή ως την σημαντικότερη υπογραφή, η οποία αποτυπώνει ποσοτικά και ποιοτικά τις βιολογικές διεργασίες και τα βιοχημικά μονοπάτια που λαμβάνουν χώρα σε δείγματα αρθρικού υμένα από ασθενείς με ΟΑ, και αναδεικνύει τη φλεγμονή της αρθρικής μεμβράνης ως έναν σημαντικό αιτιοπαθογενετικό παράγοντα στην εμφάνιση της νόσου της ΟΑ.

ΚΕΦΑΛΑΙΟ 6: ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΕΠΕΚΤΑΣΕΙΣ

6.1 Συμπεράσματα

Στη παρούσα διπλωματική εργασία μελετήσαμε διάφορες μεθοδολογίες με σκοπό τη ταξινόμηση και την επιλογή των κατάλληλων χαρακτηριστικών (γονιδιακές υπογραφές) από σύνολα γονιδιακών δεδομένων. Η κάθε μεθοδολογία επιλογής γονιδίων σε συνδυασμό με ένα γραμμικό SVM ταξινομητή μας παρείχε διαφορετικά αποτελέσματα για το εκάστοτε πρόβλημα τα οποία αξιολογήσαμε στατιστικά αλλά και ερμηνεύσαμε βιολογικά.

Εφαρμόσαμε τη πειραματική διαδικασία σε δύο Datasets (Dataset A, Dataset B) τα οποία σχετίζονται με την ασθένεια της οστεοαρθρίτιδας και καταλήξαμε σε 3 διαφορετικές γονιδιακές υπογραφές για το κάθε σύνολο δεδομένων. Τα αποτελέσματα που λάβαμε ήταν αρκετά ικανοποιητικά τόσο από την άποψη του μικρού αριθμού των γονιδίων που περιλαμβάνονται στις υπογραφές όσο και από την υψηλή προγνωστική ακρίβεια που αυτές επιτυγχάνουν. Συγκεκριμένα:

- Παρατηρούμε ότι και οι τρεις διαφορετικές τεχνικές επιλογής γονιδίων που ακολουθήσαμε δημιούργησαν μικρά και διαχειρίσιμα υποσύνολα από σημαντικά γονίδια τα οποία είναι ικανά να κατασκευάσουν ένα εύρωστο και αποδοτικό σύστημα ταξινόμησης. Η δεύτερη μεθοδολογία (LASSO, SVM) παρουσίασε τα καλύτερα αποτελέσματα με τη τρίτη (RFE-LNW, FSMLP, SVM) και τη πρώτη (RFE-LNW, SVM) να ακολουθούν.
- Για κάθε μεθοδολογία που ακολουθήσαμε, η συνένωση των γονιδιακών υπογραφών των 2 Datasets προσφέρει καλύτερη προγνωστική ακρίβεια σε σχέση με εκείνη που επιτυγχάνει η κάθε υπογραφή μόνη της. Το γεγονός αυτό καταδεικνύει τη συμπληρωματικότητα της πληροφορίας που περιλαμβάνουν οι υπογραφές των 2 Datasets.
- Η εφαρμογή διαφορετικής τεχνικής επιλογής γονιδίων οδηγεί σε ποικίλα μεγέθη γονιδιακής υπογραφής, με διαφορές στα επιλεγμένα γονίδια. Προχωρώντας όμως

στη βιολογική αξιολόγηση των υπογραφών παρατηρούμε ότι τα γονίδιά τους επικαλούνται τις ίδιες βιολογικές διεργασίες και ότι συμμετέχουν σε αρκετά κοινά μοριακά μονοπάτια. Τέλος, ένας αριθμός γονιδίων που περιλαμβάνονται στις υπογραφές βρέθηκε να σχετίζεται με την ασθένεια της οστεοαρθρίτιδας και ειδικότερα με τη φλεγμονώδη παθογένεια της νόσου.

6.2 Μελλοντικές Επεκτάσεις

Ενδιαφέρον θα αποτελούσε η εφαρμογή της προτεινόμενης μεθοδολογίας σε διαφορετικά σύνολα δεδομένων με σκοπό τον έλεγχο της εγκυρότητας των τεχνικών που χρησιμοποιήσαμε.

Επίσης, θα μπορούσαν να εξεταστούν διαφορετικοί τρόποι επιλογής γονιδίων πέραν των RFE-LNW, LASSO, FSMLP που μελετήθηκαν στη παρούσα εργασία.

Τέλος, ως μελλοντική επέκταση συνίσταται η χρήση άλλων μοντέλων μάθησης και πρόβλεψης πέρα από το γραμμικό SVM ταξινομητή (π.χ. SVM με πολυωνυμικό πυρήνα) στην ίδια γονιδιακή βάση δεδομένων με απώτερο στόχο την σύγκριση των αποτελεσμάτων με αυτά που εμείς καταλήξαμε και την ενίσχυση των συμπερασμάτων.

Βιβλιογραφία

[1] MAGLIETTA R., D'ADDABBO A., PIEPOLI A., PERRI F., LIUNI S., PESOLE G., & ANCONA N. (2007). Selection of relevant genes in cancer diagnosis based on their prediction accuracy. *Artificial Intelligence In Medicine.* vol. 40, pp. 29-44.

[2] PAN X., HUANG L., CHEN J., DAI Y., & CHEN X. (2012). Analysis of synovial fluid in knee joint of osteoarthritis:5 proteome patterns of joint inflammation based on matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *International Orthopaedics.* vol. 36, pp. 57-64.

[3] ZHAI G., AREF ESHGHI E., & ZHAI, GUANGJU. (2012). Biomarkers for osteoarthritis: investigation, identification, and prognosis. *Dove Pres*s. vol. 2, pp. 19-28. http:///www.dovepress.com/getfile.php?fileID=13140.

[4] HAYASHI D., ROEMER F., KATUR A., FELSON D., YANG S., ALOMRAN F., & GUERMAZI A. (2011). Imaging of Synovitis in Osteoarthritis: Current Status and Outlook. *Seminars in Arthritis and Rheumatism.* vol. 41, pp. 116-130.

[5] HUBER R., HUMMERT C., GAUSMANN U., POHLERS D., KOCZAN D., GUTHKE R., & KINNE R. W. (2008). Identification of intra-group, inter-individual, and gene-specific variances in mRNA expression profiles in the rheumatoid arthritis synovial membrane. *Arthritis Research & Therapy.* vol. 10, no. 4.

[6] HAN H.-S., LEE S., KIM J.H., SEONG S.C., & LEE M.C. (2010). Changes in chondrogenic phenotype and gene expression profiles associated with the in vitro expansion of human synovium-derived cells. *Journal of Orthopaedic Research.* vol. 28, pp. 1283-1291.

[7] MARSHALL K. W., ZHANG H., & NOSSOVA N. (2006). Chondrocyte genomics: implications for disease modification in osteoarthritis. *Drug Discovery Today.* vol. 11, pp. 825-832.

[8] [Online] http://www.trnres.com/ebook/uploads/pelletier/T_13271279757%20Pelletier.pdf

[9] ΠΑΓΙΩΤΟΠΟΥΛΟΣ Α. Δ. Μεταπτυχιακή Εργασία: Μοριακοί παθογενετικοί μηχανισμοί στην εκφυλιστική νόσο των αρθρώσεων. Πανεπιστήμιο Πατρών – Τμήμα Ιατρικής. [Online] http://nemertes.lis.upatras.gr/jspui/handle/10889/772

[10] XIE T., GUO S., ZHANG J., CHEN Z., & PEAVY G.M. (2006). Determination of characteristics of degenerative joint disease using optical coherence tomography and polarization sensitive optical coherence tomography. *Lasers in Surgery and Medicine.* vol.38, pp. 852-865.

[11] [Online]: http://www.mednutrition.gr/biohimikoi-deiktes-ti-lene-gia-ti-diatrofi-mas

[12] Σκελετική Υγεία. (2009). *Ελληνικό Ίδρυμα Οστεοπόρωσης (ΕΛ.Ι.ΟΣ.).* Τόμος 8ο, Τεύχος 4ο. [Online]: http://heliost.gr/images/SKELETAL_HEALTH/TOMOS_8/T8T4.pdf

[13] Βασικές Αρχές Τεχνικών Μοριακής Διαγνωστικής. 15ο σεμινάριο. (2000). Κείμενα Διαλέξεων. *Ελληνική Εταιρεία Κλινικής Χημείας – Κλινικής Βιοχημείας*

[14] LUBERT L. (1994). Βιοχημεία,Τόμος Πρώτος. *Πανεπιστημιακές Εκδόσεις Κρήτης*

[15] [Online]: http://www.embl.it/training/scienceforschools/teacher_training/teachingbase/microarray_greek/intro_gr.pdf

[16] Φαρμακευτική Βιοτεχνολογία, Γονιδιωματική. (2009) [Online]: http://www.pharmacy.upatras.gr/index.php/en/latest-news/cat_view/130--/53--7/136--

[17] KURELLA M., HSIAO L.L., YOSHIDA T., RANDALL J.D., CHOW G., SARANG S.S., JENSEN R.V., & GULLANS S.R. (2001). DNA microarray analysis of complex biologic processes. *Journal of the American Society of Nephrology : JASN. v*ol. 12, pp. 1072-1078.

[18] DHIMAN N., BONILLA R., O'KANE J, D., & POLAND G. A. (2001). Gene expression microarrays: a 21st century tool for directed vaccine design. *Vaccine.* vol. 20, pp. 22-30.

[19][Online]: http://www.biooptics.ece.ntua.gr/postgrand_notes/2%CE%B7_%CE%B4%CE%B9%CE%B1%CE%BB%CE%B5%CE%BE%CE%B7_%CF%85%CE%BB%CE%B9%CE%BA%CF%8C.pdf

[20] [Online]: http://digitalschool.minedu.gov.gr/modules/ebook/show.php/DSGL-C112/52/390,1507/

[21] TARCA A. L., ROMERO R., & DRAGHICI S. (2006). Analysis of microarray experiments of gene expression profiling. *American Journal of Obstetrics and Gynecology.* vol. 195, pp. 373-388

[22] ΠΑΠΑΕΥΑΓΓΕΛΙΟΥ Δ., ΣΟΛΑΚΙΔΗ Σ., & ΖΟΥΜΠΟΥΡΛΗΣ Β. (2003). Μοριακή Ανάλυση των Νεοπλασμάτων με τη Χρήση Μικροσυστοιχιών DNA. *Ιατρική Επικαιρότητα*, pp. 2092-2095

[23] [Online] https://openaccess.leidenuniv.nl/bitstream/handle/1887/15125/3%20Hoofdstuk%202.pdf?sequence=4

[24] GENE EXPRESSION OMNIBUS-Series GSE12021 [Online]: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE12021

[25] [Online]: http://www.ima.umn.edu/talks/workshops/9-29-10-3.2003/mclachlan/mclachlan.pdf

[26] ZHANG B., KIROV S., & SNODDY J. (2005). WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Research.* vol. 33, W741 - W748.
[Online]: http://bioinfo.vanderbilt.edu/webgestalt/

[27] [Online] http://www.geneontology.org/GO.doc.shtml

[28] [Online] http://www.genome.jp/kegg/pathway.html

[29] WALL D.P., PIVOVAROV R., TONG M., JUNG J.Y., FUSARO V.A., DELUCA T.F., & TONELLATO P.J. (2010). Genotator: a disease-agnostic tool for genetic annotation of disease. *BMC Medical Genomics.* vol. 3. [Online]: http://genotator.hms.harvard.edu/geno/

[30] BOUSQUET O., LUXBURG U. V., RÄTSCH G., & GHAHRAMANI Z. (2004). Advanced Lectures on Machine Learning. *Springer -Verlag*. LNAI 3176, pp. 72-112.
[Online]: http://eprints.pascal-network.org/archive/00000763/

[31] [Online]: http://www.dataminingarticles.com/data-mining-introduction.html

[32] CHAPELLE O., SCHÖLKOPF B., & ZIEN A. (2006). Semi-supervised learning. London, *MIT Press*.

[33] ZHU X. (2012). Semi-Supervised Learning Literature Survey. *University of Wisconsin - Madison Department of Computer Sciences*.
[Online]: http://digital.library.wisc.edu/1793/60444.

[34] ZHU Z., ONG Y.S., & DASH M. (2007). Wrapper-filter feature selection algorithm using a memetic framework. *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics.* vol. 37, pp. 70-76.

[35] INZA I., LARRAÑAGA P., BLANCO R., & CERROLAZA A. J. (2004). Filter versus wrapper gene selection approaches in DNA microarray domains. *Artificial Intelligence in Medicine.* vol. 31, pp. 91-103.

[36] BLAZADONAKIS M., & ZERVAKIS M. (2008). Wrapper filtering criteria via linear neuron and kernel approaches. *Computers in Biology and Medicine.* vol. 38, pp. 894-912.

[37] GUYON I., WESTON J., BARNHILL,S., & VAPNIK V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning.* vol. 46, pp.389-422

[38] KOHAVI R., & JOHN G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence.* vol. 97, pp. 273-324.

[39] REFAEILZAHDEH P., TANG L., LIU H. Cross Validation.
[Online]: http://www.cse.iitb.ac.in/~tarung/smt/papers_ppt/ency-cross-validation.pdf

[40] GUTIEREZZ – OSUNA R. Lecture 13: Validation. *Intelligent Sensor Systems*.
[Online]: http://research.cs.tamu.edu/prism/lectures/iss/iss_l13.pdf

[41] DZIUDA D. M. (2010). Data mining for genomics and proteomics: Analysis of gene and protein expression data. Hoboken*, J. Wiley Publication*.

[42] SADOGHI YAZDI J., KALANTARY F., & SADOGHI YAZDI H. (2012). Prediction of liquefaction potential based on CPT up-sampling. *Computers and Geosciences.* vol. 44, pp. 10-23.

[43] TETKO I., BASKIN I., & VARNEK A. Tutorial on Machine Learning- Part 2: Descriptor Selection Bias. [Online]: http://masterchemoinfo.ustrasbg.fr/Documents/TutoChemo/Descriptor_selection.pdf

[44] ΔΙΑΜΑΝΤΑΡΑΣ Κ. (2007). Τεχνητά Νευρωνικά Δίκτυα, *Εκδόσεις Κλειδάριθμος*

[45] HAYKIN S. (2009). Νευρωνικά Δίκτυα και Μηχανική Μάθηση. *Εκδόσεις Παπασωτηρίου*. Τρίτη Έκδοση

[46] [Online]: http://hplusmagazine.com/2012/10/17/four-statements-about-the-future/

[47] MCCULLOCH W. S., & PITTS W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics.* vol. 5, pp. 115-133.

[48] [Online]: http://el.wikipedia.org/wiki/Νευρωνικό_δίκτυο

[49] ROSENBLATT F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Cornell Aeronautical Laboratory, Psychological Review.* vol. 65, no.6, pp. 386-408.

[50] [Online]: http://www.nd.com/neurosolutions/products/ns/whatisNN.html

[51] PAL N. R., & CHINTALAPUDI K. K. (1997). A connectionist system for feature selection. *Neural, Parallel & Scientific Computations.* vol. 5, pp. 359-382.

[52] BOSER B.E., GUYON I.M., & VAPNIK V.N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of 5th Annual Workshop on Computational Learning Theory.* pp. 144-152.

[53] THEODORIDIS S., & KOUTROUMBAS K. (2012). Αναγνώριση Προτύπων. *Εκδόσεις Π.Χ. Πασχαλίδης.*

[54] [Online]: http://omega.albany.edu:8008/machine-learning-dir/notes-dir/ker1/ker1-l.html

[55] BLAZADONAKIS M.E., & ZERVAKIS M. (2008). The linear neuron as marker selector and clinical predictor in cancer gene analysis. *Computer Methods and Programs in Biomedicine.* vol. 91, pp. 22-35.

[56] HASTIE T., TIBSHIRANI R., & FRIEDMAN J. H. (2001). The elements of statistical learning data mining, inference, and prediction. New York, *Springer*.
[Online]: http://www-stat.stanford.edu/~tibs/ElemStatLearn/

[57] TIBSHIRANI R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological).* vol. 58, no.1, pp. 267-288.

[58] KIM S.J., KOH K., LUSTIG M., BOYD S., & GORINEVSKY D. (2007). An Interior-Point Method for Large-Scale l1-Regularized Least Squares. *IEEE Journal of Selected Topics in Signal Processing.* vol.1, no.4, pp. 606-617.

[59] WITTEN I. H., & FRANK E. (2005). Data mining: practical machine learning tools and techniques. Amsterdam. *Morgan Kaufman.* Second Edition.

[60] HUDSON D. L., & COHEN M. E. (2000). Neural networks and artificial intelligence for biomedical engineering. New York. *IEEE Press*.

[61] Receiver operating characteristic (ROC).
[Online]: http://en.wikipedia.org/wiki/Receiver_operating_characteristic

[62] SEO M., & OH S. (2012). Derivation of an artificial gene to improve classification accuracy upon gene selection. *Computational Biology and Chemistry.* vol. 36, pp. 1-12.

[63] KOH K., KIM S.J., BOYD S. (2008). `l1_ls`: Simple Matlab Solver for l1-regularized Least Squares Problems.
[Online]: http://www.stanford.edu/~boyd/l1_ls/

[64] WANG S., DUAN C., ZHANG F., MA W., & GUO X. (2013). Regulatory gene networks and signaling pathways from primary osteoarthritis and Kashin-Beck disease, an endemic osteoarthritis, identified by three analysis software. *Gene.* vol.512, no.1, pp. 89-96.

[65] DAVIS C.A., GERICK F., HINTERMAIR V., FRIEDEL C.C., FUNDEL K., KÜFFNER R., & ZIMMER R. (2006). Reliable gene signatures for microarray classification: assessment of stability and performance. *Bioinformatics.* vol. 22, no.19, pp. 2356-2363.

[66] [Online] http://www.genecards.org/

[67] JAKKULA E., MELKONIEMI M., KIVIRANTA I., LOHINIVA J., RAINA S.S., PERALA M., WARMAN M.L., AHONEN K., KROGER H., GORING H.H., & ALA-KOKKO L. (2005). The role of sequence variations within the genes encoding collagen II, IX and XI in non-syndromic, early-onset osteoarthritis. *Osteoarthritis and Cartilage.* vol.13, no.6, pp. 497-507.

[68] NAKAJIMA M., TAKAHASHI A., KOU I., RODRIGUEZ-FONTELA C., GOMEZ-REINO J.J., FURUICHI T., DAI J., SUDO A., UCHIDA A., FUKUI N., KUBO M., KAMATANI N., TSUNODA T., MALIZOS K.N., TSEZOU A., GONZALEZ A., NAKAMURA Y., IKEGAWA S. (2010). New sequence variants in HLA class II/III region associated with susceptibility to knee osteoarthritis identified by genome-wide association study. *PLoS One*. vol. 5, no.3.

[69] arcOGEN Consortium; arcOGEN Collaborators. (2012). Identification of new susceptibility loci for osteoarthritis (arcOGEN): a genome-wide association study. *Lancet*. vol. 380, no.9844, pp.815-23

Παράρτημα Α

Ακολουθούν οι κωδικοί, τα σύμβολα και η περιγραφή των γονιδίων που αποτελούν τις γονιδιακές υπογραφές για κάθε Dataset.

 Με πράσινο χρώμα σημειώνονται τα κοινά γονίδια των υπογραφών του Dataset A.

 Με μπλε χρώμα σημειώνονται τα κοινά γονίδια των υπογραφών του Dataset B.

DATASET A (GPL96 U133A)

1η Γονιδιακή Υπογραφή

A/A	Κωδικός καταχώρησης στην GEO (GEO Accession viewer ID)	Αριθμός καταχώρησης στην GenBank (GenBank Accession number)	Κωδικός γονιδίου (GENE_ID)	Σύμβολο γονιδίου	Περιγραφή
1	AFFX-DapX-3_at	A	A	A	A
2	AFFX-PheX-5_at	A	A	A	A
3	203609_s_at	NM_001080	7915	ALDH5A1	aldehyde dehydrogenase 5 family, member A1
4	204322_at	BF002254	A	A	A
5	204660_at	NM_005262	2671	GFER	growth factor, augmenter of liver regeneration
6	205477_s_at	NM_001633	259	AMBP	alpha-1-microglobulin/bikunin precursor
7	205800_at	NM_000341	A	A	A
8	205958_x_at	NM_022579	1444	CSHL1	chorionic somatomammotropin hormone-like 1
9	205979_at	NM_002407	4246	SCGB2A1	secretoglobin, family 2A, member 1
10	206014_at	NM_016188	51412	ACTL6B	actin-like 6B
11	206067_s_at	NM_024426	7490	WT1	Wilms tumor 1
12	206237_s_at	NM_013957	3084	NRG1	neuregulin 1
13	206443_at	NM_006914	6096	RORB	RAR-related orphan receptor B
14	206678_at	NM_000806	2554	GABRA1	gamma-aminobutyric acid (GABA) A receptor, alpha 1
15	206975_at	NM_000595	4049	LTA	lymphotoxin alpha (TNF superfamily, member 1)
16	207245_at	NM_001077	7367	UGT2B17	UDP glucuronosyltransferase 2 family, polypeptide B17
17	207259_at	NM_017928	55018	C17orf73	chromosome 17 open reading frame 73
18	207397_s_at	NM_000523	3239	HOXD13	homeobox D13
19	207607_at	NM_005170	430	ASCL2	achaete-scute complex homolog 2 (Drosophila)
20	207672_at	NM_002920	A	A	A
21	208053_at	NM_001522	2986	GUCY2F	guanylate cyclase 2F, retinal
22	208085_s_at	NM_006125	395	ARHGAP6	Rho GTPase activating protein 6
23	208088_s_at	NM_030787	81494	CFHR5	complement factor H-related 5
24	208356_x_at	NM_022642	1442	CSH1	chorionic somatomammotropin hormone 1 (placental

					lactogen)
25	208483_x_at	NM_004138	3883	KRT33A	keratin 33A
26	208589_at	NM_020389	57113	TRPC7	transient receptor potential cation channel, subfamily C, member 7
27	209173_at	AF088867	10551	AGR2	anterior gradient homolog 2 (Xenopus laevis)
28	209261_s_at	BF000629	2063	NR2F6	nuclear receptor subfamily 2, group F, member 6
29	210122_at	BC005303	5620	PRM2	protamine 2
30	210165_at	M55983	1773	DNASE1	deoxyribonuclease I
31	210245_at	L78207	6833	ABCC8	ATP-binding cassette, sub-family C (CFTR/MRP), member 8
32	210341_at	AB020642	4661	MYT1	myelin transcription factor 1
33	210433_at	BC000582	23509	POFUT1	protein O-fucosyltransferase 1
34	210683_at	AL161995	4902	NRTN	neurturin
35	211083_s_at	Z25428	9175	MAP3K13	mitogen-activated protein kinase kinase kinase 13
36	211119_at	AF060555	2100	ESR2	estrogen receptor 2 (ER beta)
37	211179_at	AY004251	861	RUNX1	runt-related transcription factor 1
38	211237_s_at	AF202063	2264	FGFR4	fibroblast growth factor receptor 4
39	211532_x_at	L76668	100132285 /// 3806 /// 3809	KIR2DS1 /// KIR2DS2 /// KIR2DS4	killer cell immunoglobulin-like receptor, two domains, short cytoplasmic tail, 1 /// killer cell immunoglobulin-like receptor, two domains, short cytoplasmic tail, 2 /// killer cell immunoglobulin-like receptor, two domains, short cytoplasmic tail, 4
40	211586_s_at	M97260	100134286 /// 491	ATP2B2 /// LOC100134286	ATPase, Ca++ transporting, plasma membrane 2 /// similar to ATPase, Ca++ transporting, plasma membrane 2
41	211689_s_at	AF270487	7113	TMPRSS2	transmembrane protease, serine 2
42	211739_x_at	BC005921	115548 /// 1442 /// 1443 /// 2688	CSH1 /// CSH2 /// FCHO2 /// GH1	chorionic somatomammotropin hormone 1 (placental lactogen) /// chorionic somatomammotropin hormone 2 /// FCH domain only 2 /// growth hormone 1
43	211869_at	AF049656	A	A	A
44	213845_at	AL355532	2898	GRIK2	glutamate receptor, ionotropic, kainate 2

45	214207_s_at	AW024347	29775	CARD10	Full length insert cDNA clone ZD53E07
46	214304_x_at	AI077476	23336	SYNM	synemin, intermediate filament protein
47	214346_at	AW026646	10882	C1QL1	Complement component 1, q subcomponent-like 1, mRNA (cDNA clone MGC:3776 IMAGE:3635430)
48	214602_at	D17391	1286	COL4A4	collagen, type IV, alpha 4
49	214832_at	X87870	3172	HNF4A	hepatocyte nuclear factor 4, alpha
50	214978_s_at	AK023365	8497	PPFIA4	protein tyrosine phosphatase, receptor type, f polypeptide (PTPRF), interacting protein (liprin), alpha 4
51	215107_s_at	AI923972	55001	TTC22	tetratricopeptide repeat domain 22
52	215145_s_at	AC005378	26047	CNTNAP2	contactin associated protein-like 2
53	215426_at	AL117532	23174	ZCCHC14	zinc finger, CCHC domain containing 14
54	215514_at	AL080072	A	A	A
55	215790_at	AA835004	55966	AJAP1	adherens junctions associated protein 1
56	215804_at	Z27409	2041	EPHA1	EPH receptor A1
57	216281_at	AK001827	23405	DICER1	dicer 1, ribonuclease type III
58	216311_at	AI206718	728361	LOC728361	hypothetical LOC728361
59	216346_at	AC004832	266629	SEC14L3	SEC14-like 3 (S. cerevisiae)
60	216567_at	L41657	A	A	A
61	216717_at	AK021457	55578	FAM48A	Hypothetical protein
62	216734_s_at	X68829	643	CXCR5	chemokine (C-X-C motif) receptor 5
63	216796_s_at	AK026847	A	A	A
64	217108_at	X63966	A	A	A
65	217265_at	AL020989	51090	PLLP	plasma membrane proteolipid (plasmolipin)
66	217290_at	AL030995	A	A	A
67	217351_at	AL024458	A	A	A
68	217571_at	AV661138	A	A	A

69	219995_s_at	NM_024702	79755	ZNF750	zinc finger protein 750
70	220075_s_at	NM_017717	53841	MUPCDH	mucin-like protocadherin
71	220294_at	NM_014379	27012	KCNV1	potassium channel, subfamily V, member 1
72	220385_at	NM_020433	57158	JPH2	junctophilin 2
73	220479_at	NM_014116	29034		PRO0132 protein
74	220679_s_at	NM_004361	1005	CDH7	cadherin 7, type 2
75	220807_at	NM_005331	3049	HBQ1	hemoglobin, theta 1
76	221089_at	NM_015718	50508	NOX3	NADPH oxidase 3
77	221353_at	NM_002550	4994	OR3A1	olfactory receptor, family 3, subfamily A, member 1
78	221417_x_at	NM_030760	53637	S1PR5	sphingosine-1-phosphate receptor 5
79	221456_at	NM_016943	50831	TAS2R3	taste receptor, type 2, member 3
80	221576_at	BC000529	9518	GDF15	growth differentiation factor 15
81	221721_s_at	AF123656	11178	LZTS1	leucine zipper, putative tumor suppressor 1
82	221977_at	AW303460	6909	TBX2	Hs-TBX2=T-box gene {T-box region} [human, fetal kidney, mRNA Partial, 283 nt]
83	222055_at	AA723370	151313 /// 51011 /// 729234	FAHD2A /// FAHD2B /// LOC729234	fumarylacetoacetate hydrolase domain containing 2A /// fumarylacetoacetate hydrolase domain containing 2B /// fumarylacetoacetate hydrolase domain containing 2 pseudogene
84	222255_at	AB046840	57716	PRX	periaxin
85	222260_at	AK026947	A	A	A
86	222296_at	AI668610	A	A	A

DATASET A (GPL96 U133A)

2η Γονιδιακή Υπογραφή

A/A	Κωδικός καταχώρησης στην GEO (GEO Accession viewer ID)	Αριθμός καταχώρησης στην GenBank (GenBank Accession number)	Κωδικός γονιδίου (GENE_ID)	Σύμβολο γονιδίου	Περιγραφή
1	37892_at	J04177	1301	COL11A1	collagen, type XI, alpha 1
2	203000_at	BF967657	11075	STMN2	stathmin-like 2
3	204320_at	NM_001854	1301	COL11A1	collagen, type XI, alpha 1
4	204704_s_at	BF195998	229	ALDOB	aldolase B, fructose-bisphosphate
5	204712_at	NM_007191	11197	WIF1	WNT inhibitory factor 1
6	204810_s_at	NM_001824	1158	CKM	creatine kinase, muscle
7	205358_at	NM_000826	2891	GRIA2	glutamate receptor, ionotropic, AMPA 2
8	205553_s_at	NM_003476	8048	CSRP3	cysteine and glycine-rich protein 3 (cardiac LIM protein)
9	206089_at	NM_006157	4745	NELL1	NEL-like 1 (chicken)
10	206202_at	NM_005924	4223	MEOX2	mesenchyme homeobox 2
11	206287_s_at	NM_002218	3700	ITIH4	inter-alpha (globulin) inhibitor H4 (plasma Kallikrein-sensitive glycoprotein)
12	206348_s_at	NM_005391	5165	PDK3	pyruvate dehydrogenase kinase, isozyme 3
13	206439_at	NM_004950	1833	EPYC	epiphycan
14	206505_at	NM_021139	7363	UGT2B4	UDP glucuronosyltransferase 2 family, polypeptide B4
15	206627_s_at	NM_005635	6756	SSX1	synovial sarcoma, X breakpoint 1
16	206641_at	NM_001192	608	TNFRSF17	tumor necrosis factor receptor superfamily, member 17
17	207054_at	NM_001563	3617	IMPG1	interphotoreceptor matrix proteoglycan 1
18	207256_at	NM_000242	4153	MBL2	mannose-binding lectin (protein C) 2, soluble (opsonic defect)

19	207607_at	NM_005170	430	ASCL2	achaete-scute complex homolog 2 (Drosophila)
20	207678_s_at	NM_007017	11063	SOX30	SRY (sex determining region Y)-box 30
21	207932_at	NM_002170	3445	IFNA8	interferon, alpha 8
22	209904_at	AF020769	7134	TNNC1	troponin C type 1 (slow)
23	210116_at	AF072930	4068	SH2D1A	SH2 domain protein 1A
24	210121_at	AF288390	8707	B3GALT2	UDP-Gal:betaGlcNAc beta 1,3-galactosyltransferase, polypeptide 2
25	210683_at	AL161995	4902	NRTN	neurturin
26	211187_at	AF118079	A	A	A
27	211644_x_at	L14458	28875 /// 28876 /// 28912 /// 3514 /// 440871 /// 50802	IGK@ /// IGKC /// IGKV3-20 /// IGKV3D-11 /// IGKV3D-15 /// LOC440871	immunoglobulin kappa locus /// immunoglobulin kappa constant /// immunoglobulin kappa variable 3-20 /// immunoglobulin kappa variable 3D-11 /// immunoglobulin kappa variable 3D-15 (gene/pseudogene) /// similar to hCG2043206
28	214586_at	T16257	2861	GPR37	G protein-coupled receptor 37 (endothelin receptor type B-like)
29	216560_x_at	D87021	3535	IGL@	immunoglobulin lambda locus
30	216686_at	AL137717	645784	FLJ40330	hypothetical LOC645784
31	216974_at	S80491	A	A	A
32	217037_at	S83374	A	A	A
33	217320_at	AJ275413	A	A	A
34	220084_at	NM_018168	55195	C14orf105	chromosome 14 open reading frame 105
35	220437_at	NM_018687	55908	LOC55908	hepatocellular carcinoma-associated gene TD26

DATASET A (GPL96 U133A)

3η Γονιδιακή Υπογραφή

Α/Α	Κωδικός καταχώρησης στην GEO (GEO Accession viewer ID)	Αριθμός καταχώρησης στην GenBank (GenBank Accession number)	Κωδικός γονιδίου (GENE_ID)	Σύμβολο γονιδίου	Περιγραφή
1	37586_at	D87073	7701	ZNF142	zinc finger protein 142
2	201094_at	NM_001032	6235	RPS29	ribosomal protein S29
3	201134_x_at	NM_001867	1350	COX7C	cytochrome c oxidase subunit VIIc
4	203561_at	NM_021642	2212	FCGR2A	Fc fragment of IgG, low affinity IIa, receptor (CD32)
5	203838_s_at	AI146308	10188	TNK2	tyrosine kinase, non-receptor, 2
6	203993_x_at	U84569	755	C21orf2	MRNA; candidate gene for APECED
7	204231_s_at	NM_001441	2166	FAAH	fatty acid amide hydrolase
8	204277_s_at	BE895437	7084	TK2	thymidine kinase 2, mitochondrial
9	204362_at	NM_003930	8935	SKAP2	src kinase associated phosphoprotein 2
10	205175_s_at	NM_000221	3795	KHK	ketohexokinase (fructokinase)
11	205188_s_at	NM_005903	4090	SMAD5	SMAD family member 5
12	205264_at	NM_012099	10849	CD3EAP	CD3e molecule, epsilon associated protein
13	205376_at	NM_003866	8821	INPP4B	inositol polyphosphate-4-phosphatase, type II, 105kDa
14	205377_s_at	AI190022	43	ACHE	acetylcholinesterase (Yt blood group)
15	205518_s_at	NM_003570	8418	CMAH	cytidine monophosphate-N-acetylneuraminic acid hydroxylase (CMP-N-acetylneuraminate monooxygenase) pseudogene
16	205989_s_at	NM_002433	4340	MOG	myelin oligodendrocyte glycoprotein
17	206714_at	NM_001141	247	ALOX15B	arachidonate 15-lipoxygenase, type B
18	206763_at	NM_003602	8468	FKBP6	FK506 binding protein 6, 36kDa
19	206880_at	NM_005446	9127	P2RX6	purinergic receptor P2X, ligand-gated ion channel, 6
20	207028_at	NM_006316	100129296 /// 10408	LOC100129296 /// MYCNOS	hypothetical protein LOC100129296 /// v-myc myelocytomatosis viral related oncogene,

					neuroblastoma derived (avian) opposite strand
21	207193_at	NM_001138	181	AGRP	agouti related protein homolog (mouse)
22	207401_at	NM_002763	5629	PROX1	prospero homeobox 1
23	207841_at	NM_019003	54466	SPIN2A	spindlin family, member 2A
24	207877_s_at	NM_002533	4931	NVL	nuclear VCP-like
25	208004_at	NM_021225	58503	PROL1	proline rich, lacrimal 1
26	208101_s_at	NM_030914	81605	URM1	ubiquitin related modifier 1 homolog (S. cerevisiae)
27	208395_s_at	NM_014825	9875	URB1	URB1 ribosome biogenesis 1 homolog (S. cerevisiae)
28	208432_s_at	NM_000721	777	CACNA1E	calcium channel, voltage-dependent, R type, alpha 1E subunit
29	208468_at	NM_007084	11166	SOX21	SRY (sex determining region Y)-box 21
30	208490_x_at	NM_003522	8339 /// 8343 /// 8344 /// 8346 /// 8347	HIST1H2BC /// HIST1H2BE /// HIST1H2BF /// HIST1H2BG /// HIST1H2BI	histone cluster 1, H2bc /// histone cluster 1, H2be /// histone cluster 1, H2bf /// histone cluster 1, H2bg /// histone cluster 1, H2bi
31	208783_s_at	AL570661	4179	CD46	CD46 molecule, complement regulatory protein
32	209060_x_at	AI438999	8202	NCOA3	nuclear receptor coactivator 3
33	209269_s_at	AW450910	6850	SYK	spleen tyrosine kinase
34	209359_x_at	L34598	861	RUNX1	runt-related transcription factor 1
35	209641_s_at	AF009670	8714	ABCC3	ATP-binding cassette, sub-family C (CFTR/MRP), member 3
36	209693_at	AF116574	23245	ASTN2	astrotactin 2
37	210227_at	AF119817	9228	DLGAP2	discs, large (Drosophila) homolog-associated protein 2
38	210263_at	AF029780	3754	KCNF1	potassium voltage-gated channel, subfamily F, member 1
39	210388_at	BC000939	5330	PLCB2	phospholipase C, beta 2

40	210459_at	AB033605	5710	PSMD4	proteasome (prosome, macropain) 26S subunit, non-ATPase, 4
41	210607_at	U03858	2323	FLT3LG	fms-related tyrosine kinase 3 ligand
42	211170_s_at	AF127480	10846	PDE10A	phosphodiesterase 10A
43	211484_s_at	AF023450	1826	DSCAM	Down syndrome cell adhesion molecule
44	211822_s_at	AF229061	22861	NLRP1	NLR family, pyrin domain containing 1
45	211827_s_at	AF187964	3752	KCND3	potassium voltage-gated channel, Shal-related subfamily, member 3
46	212991_at	AL137520	26268	FBXO9	F-box protein 9
47	213303_x_at	AF097916	51341	ZBTB7A	zinc finger and BTB domain containing 7A
48	213478_at	AB028949	23254	RP1-21O18.1	kazrin
49	213713_s_at	R48779	89944	GLB1L2	galactosidase, beta 1-like 2
50	213806_at	BE222739	5813	PURA	Pur alpha extended 3'untranslated region
51	213958_at	AW134823	923	CD6	CD6 molecule
52	214119_s_at	AI936769	2280	FKBP1A	FK506 binding protein 1A, 12kDa
53	214542_x_at	NM_003509	8329 /// 8330 /// 8331 /// 8332 /// 8336 /// 85235 /// 8969	HIST1H2AG /// HIST1H2AH /// HIST1H2AI /// HIST1H2AJ /// HIST1H2AK /// HIST1H2AL /// HIST1H2AM	histone cluster 1, H2ag /// histone cluster 1, H2ah /// histone cluster 1, H2ai /// histone cluster 1, H2aj /// histone cluster 1, H2ak /// histone cluster 1, H2al /// histone cluster 1, H2am
54	214945_at	AW514267	202134 /// 285596 /// 653316	FAM153A /// FAM153B /// FAM153C	family with sequence similarity 153, member A /// family with sequence similarity 153, member B /// family with sequence similarity 153, member C
55	215085_x_at	AL137706	9940	DLEC1	deleted in lung and esophageal cancer 1
56	215302_at	AU150691	257152	LOC257152	hypothetical protein LOC257152
57	215563_s_at	U28055	11223	MSTP9	macrophage stimulating, pseudogene 9
58	215617_at	AU145711	26010	LOC26010	viral DNA polymerase-transactivated protein 6

59	215817_at	BE148534	5275	SERPINB13	serpin peptidase inhibitor, clade B (ovalbumin), member 13
60	216085_at	AL080128	26105	DKFZP434C153	DKFZP434C153 protein
61	216482_x_at	X65232	7633	ZNF79	zinc finger protein 79
62	216562_at	AL121777	A	A	A
63	216653_at	AL137673	1810	DR1	Down-regulator of transcription 1, TBP-binding (negative cofactor 2), mRNA (cDNA clone MGC:29766 IMAGE:4555131)
64	217027_x_at	AC004941	3837	KPNB1	karyopherin (importin) beta 1
65	217137_x_at	K00627	A	A	A
66	217772_s_at	NM_014342	23788	MTCH2	mitochondrial carrier homolog 2 (C. elegans)
67	218038_at	NM_018035	55101	ATP5SL	ATP5S-like
68	218612_s_at	NM_005706	10078	TSSC4	tumor suppressing subtransferable candidate 4
69	219268_at	NM_018208	55224	ETNK2	ethanolamine kinase 2
70	219676_at	NM_025231	80345	ZSCAN16	zinc finger and SCAN domain containing 16
71	219799_s_at	NM_005771	10170	DHRS9	dehydrogenase/reductase (SDR family) member 9
72	220610_s_at	NM_006309	9209	LRRFIP2	leucine rich repeat (in FLII) interacting protein 2
73	220687_at	NM_018175	A	A	A
74	220895_at	NM_020903	57663	USP29	ubiquitin specific peptidase 29
75	221522_at	AL136784	84079	ANKRD27	ankyrin repeat domain 27 (VPS9 domain)
76	221913_at	AI492888	23410	SIRT3	sirtuin (silent mating type information regulation 2 homolog) 3 (S. cerevisiae)

Συντμήσεις: **GEO**: GeneExpressionOmnibus, "ΓενικήΓονιδιακήΈκφραση" (δημόσια αποθήκη λειτουργικών γονιδιωματικών δεδομένων),
GenBank: Γονιδιακή Βιβλιοθήκη (σχολιασμένη συλλογή όλων των δημόσια προσβάσιμων αλληλουχιών DNA),
A: άγνωστο.

DATASET B (GPL97_U133B)

1η Γονιδιακή Υπογραφή

Α/Α	Κωδικός καταχώρησης στην GEO (GEO Accession viewer ID)	Αριθμός καταχώρησης στην GenBank (GenBank Accession number)	Κωδικός γονιδίου (GENE_ID)	Σύμβολο γονιδίου	Περιγραφή
1	224101_x_at	BC001028	57380	MRS2	MRS2 magnesium homeostasis factor homolog (S. cerevisiae)
2	224408_at	AF347063	84539	MCHR2	melanin-concentrating hormone receptor 2
3	224536_s_at	AF152526	26025 /// 56097	PCDHGA12 /// PCDHGC5	protocadherin gamma subfamily A, 12 /// protocadherin gamma subfamily C, 5
4	224547_at	L10404	A	A	A
5	228178_s_at	AI739514	A	A	A
6	228413_s_at	BF057567	6422	SFRP1	Secreted frizzled related protein
7	229093_at	AW663964	4846	NOS3	nitric oxide synthase 3 (endothelial cell)
8	229478_x_at	AW274311	54841	BIVM	basic, immunoglobulin-like variable motif containing
9	229634_at	AI627262	135932	TMEM139	transmembrane protein 139
10	230037_at	AI798655	100133898	LOC100133898	similar to anaphase promoting complex subunit 1
11	230824_at	AI819206	162333	MARCH10	membrane-associated ring finger (C3HC4) 10
12	231009_at	BF939574	84647	PLA2G12B	phospholipase A2, group XIIB
13	231545_at	BE503728	A	A	A
14	231639_at	AW003106	A	A	A
15	232067_at	BC004869	84553	C6orf168	chromosome 6 open reading frame 168
16	232325_at	AA693817	A	A	A
17	233147_at	AI868401	A	A	A
18	233171_at	AL359651	116443	GRIN3A	glutamate receptor, ionotropic, N-methyl-D-aspartate 3A

19	233464_at	AK000127	79370	BCL2L14	BCL2-like 14 (apoptosis facilitator)
20	234060_at	AK026824	A	A	A
21	234365_at	Z68274	A	A	A
22	234663_at	AK026713	A	A	A
23	234798_x_at	AL136532	343629	C20orf66	chromosome 20 open reading frame 66
24	235270_at	BG027325	84307	ZNF397	zinc finger protein 397
25	235950_at	BE676210	146542	ZNF688	zinc finger protein 688
26	236152_at	AW135330	90737	PAGE5	P antigen family, member 5 (prostate associated)
27	236690_at	AW294251	84236	RHBDD1	Rhomboid domain containing 1, mRNA (cDNA clone IMAGE:5228783)
28	236844_at	BF195045	22907	DHX30	DEAH (Asp-Glu-Ala-His) box polypeptide 30, mRNA (cDNA clone MGC:34339 IMAGE:5171702)
29	237060_at	BF590569	A	A	A
30	237191_x_at	AI279615	A	A	A
31	237380_at	BF434708	A	A	A
32	237805_at	AI684717	729296	LOC729296	PREDICTED: Homo sapiens similar to hCG2017976 (LOC729296), mRNA
33	238131_at	AA431100	1912	PHC2	polyhomeotic homolog 2 (Drosophila)
34	240140_s_at	AW293282	26018	LRIG1	leucine-rich repeats and immunoglobulin-like domains 1
35	240345_x_at	BF445961	A	A	A
36	240635_at	BE220436	A	A	A
37	240672_at	AA416829	A	A	A
38	240738_at	AI245924	A	A	A
39	241006_at	AW449100	A	A	A
40	241047_at	AI638532	401237	FLJ22536	hypothetical locus LOC401237
41	241109_at	AW590666	A	A	A

42	241142_at	AA994013	A	A	A
43	241238_at	AI733438	A	A	A
44	241350_at	AL533913	283807	FBXL22	F-box and leucine-rich repeat protein 22
45	241549_at	AI800518	A	A	A
46	242038_at	BG037106	23507	LRRC8B	leucine rich repeat containing 8 family, member B
47	243239_at	AI033500	25813	SAMM50	sorting and assembly machinery component 50 homolog (S. cerevisiae)
48	243562_at	BE326951	A	A	A
49	243809_at	AI627810	113510	HEL308	DNA helicase HEL308
50	244191_at	BF437817	6176	RPLP1	CDNA: FLJ22926 fis, clone KAT06984, highly similar to HUMPPARP1 Human acidic ribosomal phosphoprotein P1 mRNA
51	244345_at	AI627453	23705	CADM1	cell adhesion molecule 1
52	244410_at	BG431652	5314	PKHD1	polycystic kidney and hepatic disease 1 (autosomal recessive)
53	244435_at	AI377320	642938	C10orf141 (FAM196A)	chromosome 10 open reading frame 141

DATASET B (GPL97 U133B)

2η Γονιδιακή Υπογραφή

A/A	Κωδικός καταχώρησης στην GEO (GEO Accession viewer ID)	Αριθμός καταχώρησης στην GenBank (GenBank Accession number)	Κωδικός γονιδίου (GENE_ID)	Σύμβολο γονιδίου	Περιγραφή
1	AFFX-r2-Bs-dap-M_at				
2	223737_x_at	AF239821	83539	CHST9	carbohydrate (N-acetylgalactosamine 4-0) sulfotransferase 9
3	224219_s_at	AF063825	7223	TRPC4	transient receptor potential cation channel, subfamily C, member 4
4	229472_at	AI991240	84826	SFT2D3	SFT2 domain containing 3
5	230344_x_at	AI053890	A	A	A
6	230865_at	N29837	167410	LIX1	Lix1 homolog (chicken)
7	231425_at	AI935040	89869	PLCZ1	phospholipase C, zeta 1
8	231612_at	AW183059	85438	C4orf35	chromosome 4 open reading frame 35
9	231655_x_at	AW238005	A	A	A
10	231898_x_at	AW026426	347689	SOX2OT	SOX2 overlapping transcript (non-protein coding)
11	233040_at	AK026344	54477	PLEKHA5	pleckstrin homology domain containing, family A member 5
12	233616_at	AK022413	A	A	A
13	233821_at	H99386	A	A	A
14	233822_x_at	AW736788	A	A	A
15	234139_s_at	AK023382	A	A	A
16	234407_s_at	AF067628	57113	TRPC7	transient receptor potential cation channel, subfamily C, member 7
17	234632_x_at	AK026267	A	A	A
18	234702_x_at	S64699	1080	CFTR	Cystic fibrosis transmembrane conductance regulator isoform 36 (CFTR)
19	234755_x_at	AF083130	A	A	A
20	236203_at	AI377755	3117	HLA-DQA1	major histocompatibility complex, class II, DQ alpha 1
21	236739_at	AI885627	150622	FLJ30594	CDNA FLJ34044 fis, clone FCBBF2007080
22	237477_at	AW139167	132954	PDCL2	phosducin-like 2
23	237530_at	T77543	A	A	A

24	237648_x_at	H10673	A	A	A
25	237717_x_at	BE176177	A	A	A
26	237937_x_at	AI939541	A	A	A
27	238343_x_at	AW390231	A	A	A
28	238370_x_at	AI252081	6146	RPL22	Full open reading frame cDNA clone RZPDo834F116D for gene RPL22, ribosomal protein L22; complete cds, incl. stopcodon
29	239178_at	AL583692	2254	FGF9	fibroblast growth factor 9 (glia-activating factor)
30	239776_at	AI027091	642987	FLJ43080	hypothetical protein LOC642987
31	240026_x_at	AW517851	A	A	A
32	240724_at	AI668629	A	A	A
33	240734_at	AW510851	A	A	A
34	240738_at	AI245924	A	A	A
35	240876_x_at	AA861839	145645	C15orf43	chromosome 15 open reading frame 43
36	241109_at	AW590666	A	A	A
37	241147_at	AI346849	A	A	A
38	241188_at	BF223340	A	A	A
39	241306_at	AI346649	A	A	A
40	241436_at	AI985987	6340	SCNN1G	sodium channel, nonvoltage-gated 1, gamma
41	241545_x_at	N66591	A	A	A
42	241638_at	AW973235	A	A	A
43	241868_at	AA120882	A	A	A
44	241880_x_at	R39960	A	A	A
45	241979_x_at	AI733283	A	A	A
46	242118_x_at	N80145	A	A	A
47	244435_at	AI377320	642938	C10orf141 (FAM196A)	chromosome 10 open reading frame 141
48	244621_x_at	H05469	A	A	A
49	244742_at	H47984	5053	PAH	phenylalanine hydroxylase

DATASET B (GPL97 U133B)

3η Γονιδιακή Υπογραφή

A/A	Κωδικός καταχώρησης στην GEO (GEO Accession viewer ID)	Αριθμός καταχώρησης στην GenBank (GenBank Accession number)	Κωδικός γονιδίου (GENE_ID)	Σύμβολο γονιδίου	Περιγραφή
1	222824_at	AW237290	A	A	A
2	223132_s_at	AF220034	81603	TRIM8	tripartite motif-containing 8
3	223393_s_at	AL136805	57616	TSHZ3	teashirt zinc finger homeobox 3
4	223971_at	AF327904	393046 /// 401428 /// 441295	OR2A20P /// OR2A5 /// OR2A9P	olfactory receptor, family 2, subfamily A, member 20 pseudogene /// olfactory receptor, family 2, subfamily A, member 5 /// olfactory receptor, family 2, subfamily A, member 9 pseudogene
5	224298_s_at	BC004528	337867	UBAC2	UBA domain containing 2
6	224418_x_at	AY008407	5369	PMCHL1	pro-melanin-concentrating hormone-like 1
7	224609_at	AI264216	57153	SLC44A2	solute carrier family 44, member 2
8	224946_s_at	AL571677	84317	CCDC115	coiled-coil domain containing 115
9	224957_at	AL572206	497661	C18orf32	chromosome 18 open reading frame 32
10	225051_at	AA522435	2035	EPB41	erythrocyte membrane protein band 4.1 (elliptocytosis 1, RH-linked)
11	225111_s_at	AK022817	63908	NAPB	N-ethylmaleimide-sensitive factor attachment protein, beta
12	225222_at	AI243268	64645	HIAT1	hippocampus abundant transcript 1
13	225348_at	AI954700	10772	FUSIP1	FUS interacting protein (serine/arginine-rich) 1
14	225413_at	BG291685	84833	USMG5	up-regulated during skeletal muscle growth 5 homolog (mouse)
15	225450_at	AI433831	154810	AMOTL1	angiomotin like 1
16	225568_at	BE728983	84960 /// 85014	KIAA1984 /// TMEM141	KIAA1984 /// transmembrane protein 141
17	225653_at	AV755269	A	A	A
18	226030_at	BE897866	36	ACADSB	acyl-Coenzyme A dehydrogenase, short/branched chain
19	226056_at	AB033030	57514	CDGAP	Cdc42 GTPase-activating protein

20	226163_at	AW291499	221504 /// 8831	SYNGAP1 /// ZBTB9	synaptic Ras GTPase activating protein 1 homolog (rat) /// zinc finger and BTB domain containing 9
21	226225_at	BE967311	4163	MCC	mutated in colorectal cancers
22	226308_at	AA099118	93323	HICE1	HEC1/NDC80 interacting, centrosome associated 1
23	226415_at	AA156723	57687	VAT1L	vesicle amine transport protein 1 homolog (T. californica)-like
24	226507_at	AU154408	5058	PAK1	p21 protein (Cdc42/Rac)-activated kinase 1
25	226527_at	AI569785	23248	RPRD2	regulation of nuclear pre-mRNA domain containing 2
26	226815_at	BE464367	132001	C3orf31	chromosome 3 open reading frame 31
27	226902_at	BF109140	8975	USP13	Isopeptidase T-3 (ISOT-3)
28	227351_at	H06491	730094	C16orf52	chromosome 16 open reading frame 52
29	227379_at	AI734993	154141	MBOAT1	membrane bound O-acyltransferase domain containing 1
30	227503_at	N26620	A	A	A
31	227699_at	BF511003	112849	C14orf149	chromosome 14 open reading frame 149
32	227908_at	BG236006	57465	TBC1D24	TBC1 domain family, member 24
33	228517_at	AW466905	64769	C1orf149	chromosome 1 open reading frame 149
34	228811_at	AI493276	A	A	A
35	229204_at	BE218428	50809	HP1BP3	Heterochromatin protein 1, binding protein 3, mRNA (cDNA clone IMAGE:5013089)
36	230648_at	AI377398	283663	LOC283663	hypothetical LOC283663
37	230715_at	AI138969	85460	ZNF518B	zinc finger protein 518B
38	230750_at	AI290475	A	A	A
39	230775_s_at	BF590192	646871	RP11-251J8.3	hypothetical LOC646871
40	230784_at	BG498699	84366	PRAC	prostate cancer susceptibility candidate
41	230840_at	BE504634	388588	LOC388588	hypothetical LOC388588
42	231197_at	H46689	89801	PPP1R3F	CDNA FLJ56283 complete cds, highly similar to Homo sapiens protein phosphatase 1, regulatory (inhibitor) subunit 3F (PPP1R3F), mRNA
43	231720_s_at	AF356518	83700	JAM3	junctional adhesion molecule 3

44	231829_at	AB033097	57506	VISA	virus-induced signaling adapter
45	231863_at	AF161419	54556	ING3	inhibitor of growth family, member 3
46	231913_s_at	X64643	79184	BRCC3	BRCA1/BRCA2-containing complex, subunit 3
47	231984_at	BE958291	4507	MTAP	methylthioadenosine phosphorylase
48	232058_at	AU158358	A	A	A
49	233676_at	AF339831	A	A	A
50	234098_at	AK021973	55084	SOBP	CDNA FLJ33560 fis, clone BRAMY2009557
51	235026_at	AI885871	144577	C12orf66	chromosome 12 open reading frame 66
52	235181_at	H12075	129450	C2orf60	chromosome 2 open reading frame 60
53	235483_at	AA858058	A	A	A
54	235984_at	AL036662	A	A	A
55	236408_at	AW367380	A	A	A
56	237192_at	AI435590	A	A	A
57	238395_at	AI254013	A	A	A
58	238790_at	BE738988	374443	LOC374443	CLR pseudogene
59	239026_x_at	H20019	100134082	LOC100134082	hypothetical protein LOC100134082
60	239726_at	AI743588	A	A	A
61	241368_at	AI190693	440503	LSDP5	lipid storage droplet protein 5
62	242057_at	AI301859	A	A	A
63	242301_at	R60224	147381	CBLN2	cerebellin 2 precursor
64	242471_at	AI916641	A	A	A
65	242866_x_at	BF509229	A	A	A
66	243408_at	AI252664	A	A	A
67	243575_at	AI272825	100128443 /// 375449	LOC100128443 /// MAST4	hypothetical protein LOC100128443 /// microtubule associated serine/threonine kinase family member 4
68	243879_at	BG055027	A	A	A
69	243917_at	AW083491	53405	CLIC5	chloride intracellular channel 5

70	244043_at	AI049624	A	A	A
71	244065_at	AW016751	643827	LOC643827	similar to cell recognition molecule CASPR3
72	244151_at	AI078206	285733	LOC285733	hypothetical LOC285733
73	244407_at	AI796334	51302	CYP39A1	cytochrome P450, family 39, subfamily A, polypeptide 1

Συντμήσεις: **GEO**: GeneExpressionOmnibus, "ΓενικήΓονιδιακήΈκφραση" (δημόσια αποθήκη λειτουργικών γονιδιωματικών δεδομένων),
GenBank: Γονιδιακή Βιβλιοθήκη (σχολιασμένη συλλογή όλων των δημόσια προσβάσιμων αλληλουχιών DNA),
A: άγνωστο.

Παράρτημα Β

Ακολουθεί η επεξεργασία που πραγματοποιήθηκε στα δεδομένα της εργασίας των Huber και συνεργατών καθώς και στη μελέτη του Davis και συνεργατών. Όπως αναφέραμε και στο Κεφάλαιο των Βιολογικών αποτελεσμάτων (Κεφ. 5.2) στη μελέτη των Huber και συν. έγινε αντιστοίχιση των μοριακών μονοπατιών (3^{ο} επίπεδο κατηγορίας KEGG) στις ευρύτερες κατηγορίες τους (2^{ο} επίπεδο κατηγορίας KEGG) ενώ αντίστοιχα στη μελέτη των Davis και συνεργατών βρέθηκαν με τη βοήθεια του συστήματος ταξινόμησης WebGestalt, τα μοριακά μονοπάτια KEGG (2^{ο} επίπεδο κατηγορίας KEGG) για τα 32 γονίδια που σχετίζονται με την ΟΑ.

ΣΥΓΚΡΙΤΙΚΕΣ ΜΕΛΕΤΕΣ ΣΕ ΕΠΙΠΕΔΟ ΜΟΡΙΑΚΩΝ ΜΟΝΟΠΑΤΙΩΝ KEGG

1η Μελέτη [Huber και συν.]*

ΒΙΟΛΟΓΙΚΕΣ ΛΕΙΤΟΥΡΓΙΕΣ ΜΟΡΙΑΚΩΝ ΜΟΝΟΠΑΤΙΩΝ	ΜΟΡΙΑΚΑ ΜΟΝΟΠΑΤΙΑ 2ο ΕΠΙΠΕΔΟ ΚΑΤΗΓΟΡΙΑΣ KEGG	ΜΟΡΙΑΚΑ ΜΟΝΟΠΑΤΙΑ 3ο ΕΠΙΠΕΔΟ ΚΑΤΗΓΟΡΙΑΣ KEGG	B	E	*P*	Γονίδια
Κυτταρική σηματοδότηση	Μεταγωγή σήματος	[04310] Μονοπάτι σηματοδότησης Wnt	7	4	0.21	CSNK2A1, SMAD2, PPP3CB, PRKACA, TBL1X, BTRC, RBX1
		[04310] Μονοπάτι σηματοδότησης Wnt (κανονικό υπο-μονοπάτι)	6	3	0.12	CSNK2A1, BTRC, SMAD2, PRKACA, TBL1X, RBX1
Κυτταρική επικοινωνία	Κυτταρική επικοινωνία	[04520] Ζώνη πρόσφυσης	5	2	0.07	CSNK2A1, SMAD2, ACP1, TGFBR2, YES1
Κυτταρική ανάπτυξη	Κυτταρική ανάπτυξη και θάνατος	[04210] Απόπτωση	6	2	0.01	AKT2, IKBKB, PP3CB, PRKACA, RKAR2A, BCL2L
Μεταβολισμός, Πρωτεϊνοσύνθεση, και Βιοχημεία μικρών μορίων	Μετάφραση	[03010] Ριβόσωμα	5	2	0.16	RPL18, RPL35A, RPL38, RPS10, RPL14
		[03010] Ριβόσωμα (μεγάλη υπομονάδα)	4	1	0.04	RPL18, RPL35A, RPL38, RPL14
	Αναδίπλωση, διαλογή και αποικοδόμηση	[04120] Ουβικιτίνη-μεσολαβούμενη πρωτεόλυση	4	2	0.01	ANAPC5, UBE2D2, BTRC, RBX1
Άλλες λειτουργίες	Καρκίνοι	[05212] Καρκίνος του παγκρέατος	5	1	0.04	AKT2, IKBKB, SMAD2, BCL2L1, TGFBR2
	Ασθένειες του νευρικού συστήματος (νευροεκφυλιστικές ασθένειες)	[05050] Οδοντωτο-ερυθρο-ωχρο-λαϊσιανή ατροφία (DRPLA)	3	1	<0.01	ATN1, RERE, MAGI1

1η Μελέτη [Huber και συν.] **

ΒΙΟΛΟΓΙΚΕΣ ΛΕΙΤΟΥΡΓΙΕΣ ΜΟΡΙΑΚΩΝ ΜΟΝΟΠΑΤΙΩΝ	ΜΟΡΙΑΚΑ ΜΟΝΟΠΑΤΙΑ 2ο ΕΠΙΠΕΔΟ ΚΑΤΗΓΟΡΙΑΣ KEGG	ΜΟΡΙΑΚΑ ΜΟΝΟΠΑΤΙΑ 3ο ΕΠΙΠΕΔΟ ΚΑΤΗΓΟΡΙΑΣ KEGG	B	E	*P*	Γονίδια
Ανοσολογική λειτουργία	Ανοσολογικό (Ανοσοποιητικό) σύστημα	[04620] Μονοπάτι σηματοδότησης υποδοχέα τύπου Toll (TLR)	5	2	0.12	AKT2, JUN, NFKBIA, TLR7, STAT1
Κυτταρική σηματοδότηση	Μεταγωγή σήματος	[04310] Μονοπάτι σηματοδότησης Wnt	8	3	0.04	CSNK1A1, DKK2, JUN, MYC, PPP2R1B, PRKACB, WNT5B, FZD1
Άλλες λειτουργίες	Λοιμώδη νοσήματα	[05120] Σηματοδότηση επιθηλιακών κυττάρων σε λοίμωξη ελικοβακτηριδίου του πυλωρού	5	2	0.01	JUN, NFKBIA, ATP6V1C1, ADAM17, ATP6V0D1
	Καρκίνοι	[05211] Καρκίνος του νεφρού	5	2	0.01	AKT2, HGF, JUN, TCEB1, VEGFA

2η Μελέτη [Davis και συν.]

ΒΙΟΛΟΓΙΚΕΣ ΛΕΙΤΟΥΡΓΙΕΣ ΜΟΡΙΑΚΩΝ ΜΟΝΟΠΑΤΙΩΝ	ΜΟΡΙΑΚΑ ΜΟΝΟΠΑΤΙΑ 2ο ΕΠΙΠΕΔΟ ΚΑΤΗΓΟΡΙΑΣ KEGG	ΜΟΡΙΑΚΑ ΜΟΝΟΠΑΤΙΑ 3ο ΕΠΙΠΕΔΟ ΚΑΤΗΓΟΡΙΑΣ KEGG	Σ ΓΟΝΙΔΙΩΝ	ΣΥΝΟΛΟ ΓΟΝΙΔΙΩΝ	*P*	adj*P*
Κυτταρική σηματοδότηση	Μόρια σηματοδότησης και αλληλεπίδραση	[04512] Αλληλεπίδραση ECM-υποδοχέα	11	85	2.99e-23	8.97e-23
Κυτταρική επικοινωνία	Κυτταρική επικοινωνία	[04510] Εστιακή προσκόλληση	11	200	5.20e-19	1.04e-18
Άλλες λειτουργίες	Πεπτικό σύστημα	[04974] Πέψη και απορρόφηση των πρωτεϊνών	11	81	1.70e-23	8.97e-23
	Λοιμώδη νοσήματα	[05146] Αμοιβάδωση	9	106	2.22e-17	3.33e-17
	Καρκίνοι	[05222] Μικροκυτταρικός καρκίνος του πνεύμονα	2	85	0.0015	0.0015
		[05200] Μονοπάτια στον καρκίνο	3	326	0.0014	0.0015

ΑΡΙΘΜΟΣ ΓΟΝΙΔΙΩΝ: αναφέρεται στον αριθμό των γονιδίων από την εκάστοτε υπογραφή που συμμετέχουν στα συγκεκριμένα μοριακά μονοπάτια. **ΣΥΝΟΛΟ ΓΟΝΙΔΙΩΝ:** αναφέρεται στο σύνολο των γονιδίων του ανθρώπινου γονιδιώματος που είναι γνωστά έως σήμερα ότι συμμετέχουν στα συγκεκριμένα μοριακά μονοπάτια.
Συντμήσεις: P (power) ισχύς [εφαρμογή του υπεργεωμετρικού τεστ για τον υπολογισμό (της ισχύς) του εμπλουτισμού των όρων των μοριακών μονοπατιών KEGG], adjP (adjustment P) προσαρμογή P [εφαρμογή της μέθοδου Benjamini & Hochberg για την πολλαπλή προσαρμογή δοκιμής], B (absolute frequency) απόλυτη συχνότητα, E (expected frequency) αναμενόμενη συχνότητα.
ΣΗΜΕΙΩΣΕΙΣ: 1) Για την εύρεση των μοριακών μονοπατιών KEGG από τη μελέτη των Davis και συν., πραγματοποιήθηκε ανάλυση των 33 γονιδίων με το σύστημα ταξινόμησης WebGestalt,
2) Γονιδιακή υπογραφή των Davis και συν.: *LECSF1,COL11A1,COL11A2,COL1A2,COL2A1,COL3A1,COL4A1,COL6A1,EFEMP1,LOC83690,MT1X,MT2A,OGN,SOD2,SPARC, TXNIP,FN1,IGFBP3,RPS2,COL1A1,COL5A1,MT1G,COL9A2,DUSP1,GPX3,COL6A3,PRSS11,ANGPTL2,MT1E,OGN,SFRP4,MMP3,MMP2,* και
3) *βλέπε Πίνακα 4 στην εργασία Huber και συν., ** βλέπε Πίνακα 5 στην εργασία Huber και συν.

Παράρτημα Γ

Παραθέτουμε ένα λεξικό των όρων που χρησιμοποιήθηκαν στη παρούσα εργασία.

Α

αγκρεκάνη = aggrecan

ακρίβεια = accuracy

ακτινογραφία = radiography

άνυδρο πυροφωσφορικό ασβεστίο = calcium pyrophosphate dehydrate

αρθρική παθολογία = synovial pathology

αρθρική χονδρομάτωση = synovial chondromatosis

αρθρικό υγρό = synovial fluid

αρθρικός ινώδης θυλάκας = joint capsule

αρθρικός υμένας = synovium / synovial membrane

αρθροθυλακίτιδα = synovitis

άρθρωση = joint

Β

βιοδείκτης = biomarker

Γ

γονιδιακή έκφραση = gene expression

γονιδιακή υπογραφή = gene signature

γονιδιωματική = genomics

γονιδιωματικές μελέτες σύνδεσης = Genome-Wide Association Studies (GWAS)

Ε

εξόρυξη δεδομένων = data mining

εξωκυττάρια ουσία = extracellular matrix (ECM)

επιλογή χαρακτηριστικών = feature selection

ευαισθησία = sensitivity (Se)

Θ

θεμέλια ουσία του χόνδρου = soluble cartilage matrix

Ι

ιδιαιτερότητα = specificity (Sp)

ιξωδοελαστικών = viscoelastic

Κ

καμπύλες λειτουργικού χαρακτηριστικού δείκτη = Receiver Operating Characteristic (ROC)

κρύσταλλοι υδροξυαπατίτη ασβεστίου = calicium hydroxyapatite crystals

κυτοκίνες = cytokines

Μ

μαγνητική τομογραφία = magnetic resonance imaging (MRI)

μεταβολωμική = metabolomics

μεταγραφή = transcription

μείζων σύμπλεγμα ιστοσυμβατότητας = Major Histocompatibility Complex (MHC)

μηνίσκος = menisci

μηχανή διανυσμάτων υποστήριξης = support vector machine (SVM)

μηχανική μάθηση = machine learning

μικροσυστοιχίες = microarrays

μονοπάτι =pathway

Ν

νευρωνικό δίκτυο = neural network

Ο

ολιγομερή πρωτεΐνη της θεμέλιας ουσίας του χόνδρου = Cartilage Oligomeric Matrix Protein (COMP)

ομαδοποίηση = clustering

ορογόνος θύλακας = bursa

οστεοαρθρίτιδα = osteoarthritis

οστεοχόνδρινα ίχνη = osteochondral spurs

Π

Παγκόσμια Πρωτοβουλία για Βιοδείκτες της ΟΑ = OA Biomakers Global Initiative

παλινδρόμηση = regression

πρωτεάσες = proteases

πρωτεομική = proteomics

Ρ

ρευματοειδή αρθρίτιδα = rheumatoid arthritis (RA)

Σ

στόχος = target

σύνδεσμος = ligament

Τ

ταξινόμηση = classification

τένοντας = tendon

τετραγωνικό αθροιστικό σφάλμα = residual sum of squares (RSS)

Υ

υαλώδης χόνδρος = hyaline cartilage

υμενίτιδα = (βλ. αρθροθυλακίτιδα)

υποχόνδρινο οστό = subchondral bone

Χ

χόνδρος = cartilage