

Τίτλος Διπλωματικής εργασίας:

Σύστημα επεξεργασίας, ανάλυσης και ταξινόμησης εικόνων δισδιάστατης ηλεκτροφόρησης με τεχνικές αναγνώρισης προτύπων

Περίληψη:

Σκοπός της παρούσας εργασίας είναι η ανάπτυξη ενός συστήματος ανάλυσης και ταξινόμησης εικόνων δισδιάστατης ηλεκτροφόρησης από ασθενείς με μυελογενή και λεμφογενή λευχαιμία, με τεχνικές ανάλυσης εικόνας και αναγνώρισης προτύπων.

Ένα σημαντικό πλεονέκτημα των εικόνων, δισδιάστατης ηλεκτροφόρησης (2-d Gel Electrophoresis), είναι η πληθώρα πρωτεϊνών που απεικονίζεται σε καθεμιά απ' αυτές, κάτι που αποτελεί ταυτόχρονα και αντικείμενο ιδιαίτερου ερευνητικού ενδιαφέροντος, καθώς θα πρέπει να εντοπιστούν από το σύνολο των κηλίδων, εκείνες οι οποίες θα μπορούσαν να αποτελέσουν πιθανούς βιοδείκτες είτε μεμονωμένα, είτε σε συνδυασμό.

Για την πρακτική εφαρμογή χρησιμοποιήθηκαν εικόνες από μια διαδικτυακή βάση δεδομένων με εικόνες από σκαναρισμένα gel δισδιάστατης ηλεκτροφόρησης, η LECB 2-D PAGE. Η συγκεκριμένη βάση είχε το πλεονέκτημα ότι οι περιοχές ενδιαφέροντος, δηλαδή, οι κηλίδες, ήταν και αυτές διαθέσιμες, καθώς δίδονταν το κέντρο βάρους κάθε μίας από αυτές.

Η εξαγωγή των περιοχών ενδιαφέροντος, έγινε με δύο τρόπους: χειροκίνητα, με επιλογή από τον χρήστη των κηλίδων που υποδεικνύονταν στη βάση δεδομένων και αυτόματα, ορίζοντας μια περιοχή ενδιαφέροντος για κάθε κηλίδα, μεγέθους 13x13 εικονοστοιχείων με κέντρο το σημείο που έδινε η βάση δεδομένων. Από τις περιοχές ενδιαφέροντος (κυλίδες/spot) υπολογίστηκαν χαρακτηριστικά υφής 1^{ης} και 2^{ης} τάξεως, δημιουργώντας για κάθε κηλίδα ένα διάνυσμα χαρακτηριστικών.

Στη συνέχεια για κάθε σύνολο ξεχωριστά, έγινε μείωση των χαρακτηριστικών πρώτα με βάση τα αποτελέσματα που προέκυψαν από την εφαρμογή ενός Man Whitney test στο οποίο εντοπίστηκαν τα τρία χαρακτηριστικά για κάθε κυλίδα (από σύνολο 22 κυλίδων) που είχαν μεγαλύτερη στατιστικά σημαντική διαφορά ($p < 0.001$) μεταξύ των δυο κατηγοριών (μυελογενή και λεμφογενή λευχαιμία). Στη συνέχεια πραγματοποιήθηκε κατάταξη των χαρακτηριστικών με βάση τη συσχέτιση και το Wilcoxon test. Με αυτόν τον τρόπο για κάθε εικόνα δημιουργήθηκε ένα διάνυσμα μεγέθους 16 χαρακτηριστικών. Με τα διανύσματα αυτά εκπαιδεύτηκε ένα Πιθανοκρατικό Νευρωνικό Δίκτυο (Probabilistic Neural Network PNN), και ο kNN ταξινομητής για τον οποίο δοκιμάστηκαν διάφορες τιμές του k. Οι ταξινομητές εκπαιδεύτηκαν έτσι ώστε να διαχωρίζουν με τη μεγαλύτερη δυνατή ακρίβεια τις δυο κατηγορίες, χρησιμοποιώντας το μικρότερο πλήθος χαρακτηριστικών. Η επιλογή των χαρακτηριστικών έγινε με τη μέθοδο εξαντλητικής αναζήτησης. Ο αξιολόγηση των ταξινομητών έγινε χρησιμοποιώντας τη μέθοδο αποκλεισμού ενός (Leave One Out-LOO). Επίσης, το σύστημα αξιολογήθηκε περαιτέρω, για κάθε σύνολο χωριστά, με τη μέθοδο

εξωτερικής διασταυρούμενης επικύρωσης (External Cross Validation- ECV), με αποκλεισμό από το στάδιο της εκπαίδευσης του 30% του συνόλου των δειγμάτων.

Για τον PNN ταξινομητή και την LOO μέθοδο αξιολόγησης, η συνολική ακρίβεια (accuracy) του συστήματος ταξινόμησης, φάνηκε να φτάνει το 98.1% ενώ με τη μέθοδο ECV η απόδοση του ταξινομητή κυμαίνεται στο $89.4 \pm 4.5\%$. Επιπλέον, για το σύνολο αυτό έγινε και ταξινόμηση με τον kNN ταξινομητή, για τα 16 αυτά χαρακτηριστικά, και με τις δύο μεθόδους LOO, αλλά και ECV. Η ολική ακρίβεια ταξινόμησης στην περίπτωση του kNN βρέθηκε 96.3% για $k=3$ και $k=5$, με τη LOO μέθοδο, ενώ με την ECV, υπολογίστηκε $89.5 \pm 4.9\%$ για $k=3$.

Το πείραμα επαναλήφθηκε και για το δεύτερο σύνολο περιοχών ενδιαφέροντος όπου ο PNN κατάφερε ακρίβεια 96.3%, με την LOO μέθοδο αξιολόγησης και $88.6 \pm 3.6\%$ με μέθοδο αξιολόγησης την ECV. Τα αντίστοιχα ποσοστά του kNN, ταξινομητή είναι 97.2% με την LOO μέθοδο, τόσο για $k=3$, $k=5$ και $k=7$. Με την μέθοδο ECV, ο kNN για $k=3$ είχε απόδοση $91.1 \pm 4.6\%$.

Ιδιαίτερα σημαντικά χαρακτηριστικά με ικανότητα διάκρισης μεταξύ των δύο κατηγοριών, φάνηκε να δίνουν οι κηλίδες {12, 22, 8} που αποτελούσαν τον κορμό σε όλα τα σύνολα που προέκυψαν για όλους τους ταξινομητές και των δύο συνόλων περιοχών ενδιαφέροντος. Επιπλέον και με σειρά συχνότητας εμφάνισης, ιδιαίτερη σημασία φάνηκε να έχουν οι κηλίδες {13, 17 και 5}, χωρίς αυτό να σημαίνει ότι στα σύνολα που προτάθηκαν από τους ταξινομητές, δεν υπήρχαν και άλλες κηλίδες, οι οποίες είχαν συχνότητα εμφάνισης από δύο και λιγότερες φορές στο πλήθος των δέκα επαναλήψεων.

Όσον αφορά τα χαρακτηριστικά υφής τα οποία ήταν πιο συχνά εμφανιζόμενα στα σύνολα, αυτά ήταν ο μέσος και το εύρος της ενέργειας (mean Energy, range Energy) και της αντίθεσης, (mean Contrast, range Contrast) καθώς επίσης και ο μέσος της ομοιογένειας (mean Homogeneity).

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Επεξεργασία Εικόνας

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Δισδιάστατη ηλεκτροφόρηση Μυελογενείς Λευχαιμία, Λεμφογενής Λευχαιμία, Αναγνώριση προτύπων, Βιοδείκτες, Πιθανοκρατικό Νευρωνικό Δίκτυο.

Εξεταστική Επιτροπή:

Δρ. Διονύσης Κάβουρας, Καθηγητής, Τμήμα Τεχνολ. Ιατρικών Οργάνων, Σχολή Τεχνολογικών Εφαρμογών, ΤΕΙ Αθήνας – Επιβλέπων

Δρ. Μανώλης Σαγκριώτης, Αν. Καθηγητής, Τμήμα Πληροφορικής και Τηλεπικοινωνιών, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Δρ. Ερρίκος Βεντούρας, Καθηγητής, Τμήμα Τεχνολ. Ιατρικών Οργάνων, Σχολή Τεχνολογικών Εφαρμογών, ΤΕΙ Αθήνας