NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCE
DEPARTMENT OF INFORMATICS & TELECOMMUNICATIONS

POSTGRADUATE PROGRAM
"INFORMATIONS TECHNOLOGIES IN MEDICINE AND BIOLOGY"

MASTER THESIS

Radiogenomics methods on the relationship between molecular and imaging characteristics to improve breast cancer classification

Georgios-Eleftherios I. Kalykakis

Supervisor: **Georgios Spyrou,** Bioinformatics ERA Chair, Head of the Bioinformatics Group, The Cyprus Institute of Neurology and Genetics

ATHENS

DECEMBER 2017

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΣΤΗΝ ΙΑΤΡΙΚΗ ΚΑΙ ΤΗ ΒΙΟΛΟΓΙΑ»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μέθοδοι Ραδιογενομικής (Radiogenomics) επί της συσχέτισης μοριακών και απεικονιστικών χαρακτηριστικών για την βελτίωση της ταξινόμησης στον καρκίνο του μαστού

Γεώργιος Ελευθέριος Ι. Καλυκάκης

Επιβλέπων: **Γεώργιος Σπύρου,** Ειδικός Λειτουργικός Επιστήμονας (βαθμίδα Α΄)

ΑΘΗΝΑ

ΔΕΚΕΜΒΡΙΟΣ 2017

MASTER THESIS

Radiogenomics methods on the relationship between molecular and imaging characteristics to improve breast cancer classification

Georgios-Eleftherios I. Kalykakis

Student Registration Number: ΠΙΒ0147

Suprvisor: **Georgios Spyrou,** Bioinformatics ERA Chair, Head of the Bioinformatics Group, The Cyprus Institute of Neurology and Genetics

EXAMINING COMMITTEE

- **Dr. Georgios Spyrou,** Bioinformatics ERA Chair, Head of the Bioinformatics Group, The Cyprus Institute of Neurology and Genetics
- **Dr. Evanthia Anastasiadou,** PhD, Investigator - Lecturer Level
- **Dr. George Th. Tsangaris**, PhD, Staff Research Scientist - Professor Level

December 2017

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μέθοδοι Ραδιογενομικής (Radiogenomics) επί της συσχέτισης μοριακών και απεικονιστικών χαρακτηριστικών για την βελτίωση της ταξινόμησης στον καρκίνο του μαστού

Γεώργιος Ελευθέριος Ι. Καλυκάκης

Α.Μ.: ΠΙΒ0147

ΕΠΙΒΛΕΠΩΝ: **Γεώργιος Σπύρου,** Ειδικός Λειτουργικός Επιστήμονας (βαθμίδα Α΄)

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

- **Δρ.Γεώργιος Σπύρου,** Ειδικός Λειτουργικός Επιστήμονας (βαθμίδα Α΄)
- **Δρ. Ευανθία Αναστασιάδου**, Ερευνήτρια Δ΄, Ίδρυμα Ιατροβιολογικών Ερευνών Ακαδημίας Αθηνών
- **Δρ. Γεώργιος Τσάγκαρης,** Ειδικός Λειτουργικός Επιστήμονας (βαθμίδα Α΄), Ίδρυμα Ιατροβιολογικών Ερευνών Ακαδημίας Αθηνών

Δεκέμβριος 2017

ABSTRACT

The purpose of this Master Thesis is to create a mixed features matrix consisting of image and mRNA gene values in patients with breast cancer, using MRI images in order to investigate the classification performance at different stages of cancer.

The development of the methodology for the creation of the mixed feature matrix was implemented in two separate processes. The first was the extraction of features from the images of the patients and the second was the analysis of the mRNA gene data before the final classification. In particular, we worked on the The Cancer Genome Atlas (TCGA) database, and we obtained MRI images from 85 patients. Then, on the same base, we chose the corresponding mRNA gene values for the respective patients.

A Medical Doctor indicated the annotated regions of interest (ROI) from the images and morphological and statistical characteristics were extracted for each of the 85 patients. The values from the mRNA were ranked based on the Absolute Log Fold Change and then the first 100 were selected. Because of the nature of the classifiers, Support Vector Machine and k-Nearest Neighbor, the feature space was reduced by the Wilcoxon statistical test and the sequential forward selection (SFS). For the purpose of the comparison, the data were examined separately and then compared to a combined table.

According to the results, the mean accuracy has a fairly significant improvement in SVM and in some cases with the k-NN classifier. Finally, it is worth mentioning that with this methodology, some genes are more likely to help with the classification which means that they need further investigation into the importance of overexpression or under expression in the different stages of breast cancer.

SUBJECT AREA: Processing of MRI images, processing of mRNA expression gene data

KEYWORDS: magnetic resonance imaging(MRI), computer aided diagnosis system, breast cancer, gene data

ΠΕΡΙΛΗΨΗ

Σκοπός της παρούσας διπλωματικής εργασίας αποτελεί η δημιουργία ενός μικτού πίνακα χαρακτηριστικών, ο οποίος αποτελείται από χαρακτηριστικά τιμών εικόνας μαγνητικής τομογραφίας (MRI) και τιμές mRNA γονιδίων σε ασθενείς με διαγνωσμένο καρκίνο του μαστού, με στόχο την διερεύνηση της απόδοσης της ταξινόμησης στα διαφορετικά στάδια του καρκίνου.

Η ανάπτυξη της μεθοδολογίας για την δημιουργία του μεικτού πίνακα υλοποιήθηκε σε δύο ξεχωριστές διεργασίες. Η πρώτη αφορούσε την εξαγωγή χαρακτηριστικών από τις εικόνες των ασθενών και η δεύτερη την ανάλυση των δεδομένων mRNA γονιδίων πριν την τελική ταξινόμηση. Πιο συγκεκριμένα, εργαστήκαμε πάνω στην βάση The Cancer Genome Atlas (TCGA), και αντλήσαμε εικόνες MRI για 85 ασθενείς. Στην συνέχεια από την ίδια βάση επιλέξαμε για τους αντίστοιχους ασθενείς τις τιμές γονιδιακής έκφρασης (mRNA profiling).

Με την βοήθεια ιατρού επιλέχθηκαν στις εικόνες οι περιοχές ενδιαφέροντος (ROI) και έγινε η εξαγωγή μορφολογικών και στατιστικών χαρακτηριστικών για κάθε ένα από τους 85 ασθενείς. Οι τιμές από το mRNA κατατάχθηκαν με βάση την διαφορική τους έκφραση (Absolute Log Fold Change) και στην συνέχεια επιλέχθηκαν οι πρώτες 100 τιμές. Λόγω της φύσης των ταξινομητών Support Vector Machine (SVM) και k-Nearest Neighbor (k-NN), έγινε μείωση των χαρακτηριστικών με το στατιστικό τεστ του Wilcoxon και με την ακολουθία προσθήκη χαρακτηριστικών Sequential Forward Selection (SFS). Για τον σκοπό της σύγκρισης ελέγχθηκαν ξεχωριστά τα μοριακά και απεικονιστικά δεδομένα μόνα τους ως προς την ακρίβεια της ταξινόμησης και στην συνέχεια σε συνδυασμό.

Σύμφωνα με τα αποτελέσματα, ο μικτός πίνακας χαρακτηριστικών έχει μία αρκετά σημαντική βελτίωση στο ποσοστό ακρίβειας του διαχωρισμού των σταδίων του καρκίνου με βάση τον SVM και σε κάποιες περιπτώσεις με τον k-NN ταξινομητή. Αξίζει τέλος να αναφερθούμε ότι με τη μεθοδολογία αυτή έρχονται στην επιφάνεια γονίδια τα οποία χρήζουν περαιτέρω διερεύνησης για το κατά πόσο παίζει ρόλο η υπερέκφραση ή η υποέκφραση τους στα διαφορετικά στάδια του καρκίνου του μαστού.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Επεξεργασία Εικόνων μαγνητικής τομογραφίας,επεξεργασία δεδομένα έκφρασης γονιδίων mRNA

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: μαγνητική τομογραφία, σύστημα υποβοηθούμενης διάγνωσης,καρκίνος του μαστού,γονιδιακά δεδομένα