ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΦΥΣΙΚΗΣ

ΔΠΜΣ «ΦΥΣΙΚΗ ΚΑΙ ΤΕΧΝΟΛΟΓΙΚΕΣ ΕΦΑΡΜΟΓΕΣ»

ΜΕΛΕΤΗ ΜΑΣΤΙΚΟΥ ΙΣΤΟΥ ΜΕ ΦΑΣΜΑΤΟΣΚΟΠΙΑ RAMAN ΓΙΑ ΤΗ ΔΙΑΓΝΩΣΗ ΚΑΡΚΙΝΟΥ ΤΟΥ ΜΑΣΤΟΥ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Δήμητρα Κάντα

Επιβλέποντες : Ι. Ζεργιώτη Αναπληρώτρια Καθηγήτρια ΕΜΠ

Ι. Ράπτης Καθηγητής ΕΜΠ

Αθήνα, Ιανουάριος 2017

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΈΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΟΜΕΑΣ ΦΥΣΙΚΗΣ
ΔΠΜΣ «ΦΥΣΙΚΗ ΚΑΙ ΤΕΧΝΟΛΟΓΙΚΕς ΕΦΑΡΜΟΓΕΣ»

ΜΕΛΕΤΗ ΜΑΣΤΙΚΟΥ ΙΣΤΟΥ ΜΕ ΦΑΣΜΑΤΟΣΚΟΠΙΑ RAMAN ΓΙΑ ΤΗ ΔΙΑΓΝΩΣΗ ΚΑΡΚΙΝΟΥ ΤΟΥ ΜΑΣΤΟΥ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Δήμητρα Κάντα

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την

.....	.....	.....
Ι. Ζεργιώτη	Γ. Ράπτης	Μ. Μακροπούλου
Αν. Καθηγήτρια ΕΜΠ	Καθηγητής ΕΜΠ	Καθηγήτρια ΕΜΠ

Αθήνα, Ιανουάριος 2017

Ευχαριστίες

Θα ήθελα καταρχάς να ευχαριστήσω τους υπεύθυνους της διπλωματικής μου Δρ. Ιωάννη Ράπτη και Δρ. Ιωάννα Ζεργιώτη της σχολής ΕΜΦΕ του ΕΜΠ. Το γραφείο τους ήταν πάντα ανοιχτό για μένα όταν είχα ερωτήσεις ή προβλήματα σχετικά με τη διπλωματική μου. Με προέτρεψαν και με βοήθησαν ώστε αυτή η διπλωματική να είναι δική μου δουλειά, αλλά πάντα με κατηύθυναν προς τη σωστή κατεύθυνση όποτε το χρειαζόμουν.

Θα ήθελα επίσης να ευχαριστήσω τους εμπειρογνώμονες από το ΙΙΒΕΑΑ που συμμετείχαν στην έρευνα αυτή και μας προμήθευσαν με τα δείγματα, αλλά και για τις συμβουλές τους πάνω στα δείγματα αυτά και πάνω σε βιολογικά θέματα.

Ευχαριστώ πολύ τους διδακτορικούς Μαριάνθη Παναγοπούλου και Σίμο Παπάζογλου για το χρόνο που αφιέρωσαν για να μου δείξουν πώς λειτουργεί η διάταξη, πώς δουλεύουμε στα πειράματα αυτά, αλλά και για τις πολλές απορίες και ζητήματα που μου έλυναν καθ' όλη τη διάρκεια της εργασίας μου.

Τέλος, θα ήθελα να εκφράσω την ευγνωμοσύνη μου στους γονείς μου και στους φίλους μου για την αμέριστη υποστήριξη και συνεχή ενθάρρυνση όλα αυτά τα χρόνια των σπουδών μου και καθ'όλη τη διάρκεια της διπλωματικής μου.

Σας ευχαριστώ πολύ.

Περίληψη

Μελέτη μαστικού ιστού με φασματοσκοπία Raman για τη διάγνωση καρκίνου του μαστού

Συγγραφέας: Δήμητρα Κάντα

Υπεύθυνοι: Δρ. Ιωάννης Ράπτης[1] , Δρ. Ιωάννα Ζεργιώτη[1]

[1]Εθνικό Μετσόβιο Πολυτεχνείο, Αθήνα, Ελλάδα

Η φασματοσκοπία Raman μπορεί να χρησιμοποιηθεί για τη μέτρηση της χημικής σύνθεσης ενός δείγματος, η οποία μπορεί με τη σειρά της να χρησιμοποιηθεί για την εξαγωγή βιολογικών πληροφοριών. Η εφαρμογή της φασματοσκοπίας Raman στη βιολογία αυξάνεται ραγδαία αφού μπορεί να παρέχει χημικές αλλά και δομικές πληροφορίες[1]. Τα κύρια πλεονεκτήματα της πιθανής χρήσης της φασματοσκοπίας Raman είναι: άμεση in vivo διάγνωση, μείωση του αριθμού των βιοψιών, συνδυασμός βιοχημικής και δομικης διάγνωσης [2].Τα φάσματα Raman του κανονικού, πρώιμου και όψιμου καρκινικού ιστού του μαστού των ποντικών μελετήθηκαν Ex-vivo χρησιμοποιώντας ένα 532 nm σύστημα Raman. Ένας συνολικός αριθμός 146 Raman φασμάτων αποκτήθηκαν από φυσιολογικούς (49), πρώιμους καρκινικούς (39) και όψιμους καρκινικούς (58) ιστούς μαστού. Οι διαφορές μεταξύ των φασμάτων Raman φυσιολογικού και καρκινικού ιστού έχουν καταγραφεί και αναλυθεί ως μια μέθοδος για την έγκαιρη ανίχνευση του καρκίνου του μαστού. Συλλέξαμε χαρακτηριστικές κορυφές από τους ιστούς μέσα στην περιοχή μεταξύ 2.550 έως 3.050 cm^{-1}. Μια ευρεία ζώνη κορυφών εντός της περιοχής αυτής παρατηρήθηκε, λόγω των ομάδων (-CH3) και (-CH2-)[3]. Διενεργήσαμε Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis, PCA) για τη μείωση των διαστάσεων των δεδομένων και εκπαιδεύσαμε πολλαπλούς αλγόριθμους με τα δεδομένα μας. Τα αποτελέσματα είναι πολύ ελπιδοφόρα, με την sensitivity του συνόλου δοκιμής να κυμαίνεται από 83% - 100%, ανάλογα με τον αλγόριθμο που χρησιμοποιείται. Το αρκετά μικρό σύνολο δεδομένων μας θα μπορούσε ωστόσο να προκαλέσει overfitting στα δεδομένα μας και ως εκ τούτου για τις μελλοντικές μελέτες θα πρέπει να χρησιμοποιείται ένα πολύ μεγαλύτερο σύνολο δεδομένων.

Abstract

Raman spectroscopy study of breast tissue in the detection of breast cancer

Author: Dimitra Kanta

Supervisors: Dr. Ioannis Raptis[1] , Dr. Ioanna Zergioti[1]

[1]National Technical University of Athens, Athens, Greece

Raman spectroscopy can be used to measure the chemical composition of a sample, which can in turn be used to extract biological information. The application of Raman spectroscopy and microscopy within biology is rapidly increasing since it can provide chemical and compositional information [1].The main advantages of potentially employing Raman spectroscopy are: immediate in vivo diagnosis, reduction of the number of biopsies, combination of biochemical and structural diagnosis [2]. The Raman spectra of normal, early and late cancerous breast tissues of mice were investigated in vitro using a 532 nm Raman system. A total number of 146 Raman spectra were acquired from normal (49), early (39) and late (58) cancerous breast tissues. Differences between Raman spectra of normal, late and early cancer tissues have been recorded and analyzed as a method for the early detection of breast cancer. We collected characteristic peaks from the tissues within the region of 2550-3050. A broad band of peaks within this region was observed, due to the methyl ($-CH_3$) and methylene ($-CH_2-$) group bonds [3]. We conducted Principal Component Analysis (PCA) for data reduction and trained our data with multiple algorithms. The results are quite promising, with a sensitivity of the test set ranging from 83%-100% depending on the algorithm used. The small data set could however cause overfitting to our data and as a result for future studies a much bigger data set should be used.

REFERENCES

[1] Holly J Butler and Lorna Ashton and Benjamin Bird and Gianfelice Cinque and Kelly Curtis and Jennifer Dorney and Karen Esmonde-White and Nigel J Fullwood and Benjamin Gardner and Pierre L Martin-Hirsch and Michael J Walsh and Martin R McAinsh and Nicholas Stone and Francis L Martin, “Using Raman spectroscopy to characterize biological materials”, Nature Protocols, 2016

[2] H. Abramczyk, I. Placek, B. Brozek-Pluska, K. Kurczewski, Z. Morawiecc and M. Tazbir “Human breast tissue cancer diagnosis by Raman spectroscopy”, Spectroscopy, 2008

[3] C.-H. Liu, et.al. “Resonance Raman and Raman Spectroscopy for Breast Cancer Detection”, Technology in Cancer Research and Treatment , ISSN 1533-0346, Volume 12, Number 4, August 2013, pp 371-382.

Table of Contents

Κεφάλαιο 1: Ο καρκίνος και το ερευνητικό μας κίνητρο

1.1 Η ασθένεια του καρκίνου

Ο καρκίνος αποτελεί τη δεύτερη αιτία θνησιμότητας παγκοσμίως (1), λόγω των πολλών διαφορετικών τύπων καρκίνου, αλλά και του ραγδαίου ρυθμού ανάπτυξής του. Η αποτελεσματική θεραπεία του καρκίνου εξαρτάται κατά πολύ από την έγκαιρη διάγνωση σε πρώιμο στάδιο.

Ο όρος «καρκίνος του μαστού» αναφέρεται στην ανάπτυξη κακοήθους όγκου στην ευρύτερη περιοχή του μαστού (2). Οφείλεται στον αφύσικο και ανεξέλεγκτο πολλαπλασιασμό των κυττάρων στους ιστούς του μαστού, προκαλώντας έτσι, τον σχηματισμό κακοήθους όγκου και η επέκτασή τους σε γειτονικές περιοχές πιθανή μετάσταση όγκου.

Ο καρκίνος του μαστού είναι η δεύτερη πιο κοινή μορφή καρκίνου στον κόσμο και με διαφορά, η συχνότερη μορφή καρκίνου μεταξύ των γυναικών, κατ' εκτίμηση 1.670.000 νέες περιπτώσεις καρκίνου διαγνώστηκαν το 2012 (25% όλων των καρκίνων). Είναι η πιο κοινή μορφή καρκίνου στις γυναίκες τόσο στις περισσότερο όσο και στις λιγότερο ανεπτυγμένες περιοχές με ελαφρώς περισσότερες περιπτώσεις στις λιγότερο ανεπτυγμένες (883.000 περιπτώσεις) από ό,τι στις πιο ανεπτυγμένες (794.000) περιοχές. Τα ποσοστά εμφάνισης ποικίλουν σχεδόν τέσσερις φορές κατά μήκος των όλων των περιοχών του κόσμου, με ποσοστά που κυμαίνονται από 27 ανά 100.000 στη Μέση Αφρική και την Ανατολική Ασία και εώς και 92 στη βόρεια Αμερική. (3)

Ο καρκίνος του μαστού κατατάσσεται ως η πέμπτη αιτία θανάτου από καρκίνο συνολικά (522.000 θάνατοι) και ενώ είναι η πιο συχνή αιτία θανάτου από καρκίνο στις γυναίκες στις λιγότερο αναπτυγμένες περιοχές (324.000 θάνατοι, 14,3% του συνόλου), είναι πλέον η δεύτερη αιτία θανάτου από καρκίνο στις πιο ανεπτυγμένες περιοχές (198.000 θάνατοι, 15,4%) μετά τον καρκίνο του πνεύμονα. Το εύρος των ποσοστών θνησιμότητας μεταξύ των περιοχών ανά τον κόσμο είναι μικρότερο από

εκείνο των περιστατικών, λόγω της ευνοϊκότερης επιβίωσης του καρκίνου του μαστού σε ανεπτυγμένες περιοχές (υψηλής συχνότητας περιστατικών καρκίνου του μαστού), με ποσοστά που κυμαίνονται από 6 ανά 100.000 στην Ανατολική Ασία έως 20 ανά 100.000 στη Δυτική Αφρική.

1.2 Η διαγνωστική τεχνολογία σήμερα

Ο πιο γνωστός και πλέον διαδεδομένος τρόπος διάγνωσης του καρκίνου του μαστού είναι η βιοψία, η όποια μας δίνει πληροφορίες για τη χημική σύσταση του όγκου. Μια σημαντική αλλά όχι διαγνωστική μέθοδος είναι η μαστογραφία, μία μέθοδος απεικονιστική η οποία δείχνει μεταβολή στην πυκνότητα του μαστού με τη βοήθεια των ακτίνων χ. Η μέθοδος αυτή δεν μπορεί να είναι ακριβείας μιας και οι μεταβολές στην πυκνότητα μπορέι να έχουν σχέση και με καλοήθεις όγκους (4). Υπάρχουν και άλλες απεικονιστικές μέθοδοι με σημαντικότερη την MRI (Magnetic Resonance imaging). Η MRI, ή απεικόνιση μαγνητικού συντονισμού, είναι μια τεχνολογία που χρησιμοποιεί μαγνήτες και ραδιοκύματα για την παραγωγή λεπτομερών εικόνων εγκάρσιας διατομής του εσωτερικού του σώματος. Η MRI δεν χρησιμοποιεί ακτίνες Χ, γι 'αυτό δεν περιλαμβάνει καμία έκθεση σε ακτινοβολία. Παρόλ’αυτά η αξία της μαγνητικής τομογραφίας μαστού για την ανίχνευση του καρκίνου του μαστού παραμένει αβέβαιη. Μερικοί γιατροί πιστεύουν ότι η MRI μπορεί να διακρίνει τον καρκίνο του μαστού από το φυσιολογικό ιστό του αδένα του μαστού καλύτερα από άλλες τεχνικές. Αλλά η MRI μαστού είναι δαπανηρή και απαιτεί πολύ εξειδικευμένο εξοπλισμό και άρτια καταρτισμένους εμπειρογνώμονες. Υπάρχουν σχετικά λίγα κέντρα MRI μαστού, ιδιαίτερα έξω από τις μεγάλες πόλεις. Και ακόμη και στις καλύτερες περιπτώσεις, η MRI παράγει πολλά αβέβαια συμπεράσματα..

Όσον αφορά στη βιοψία σε όγκους που έχουν ανιχνευθεί μέσω μιας απεικονιστικής μεθόδου προκύπτει ότι ένα μεγάλο μέρος των όγκων αυτών είναι καλοήθεις. Το μεγαλύτερο μειονέκτημα όμως είναι ότι η βιοψία στον μαστό απαιτεί χειρουργική επέμβαση και πολλές φορές πάνω από μία φορά, διαδικασία επίπονη για τον ασθενή. Η βιοψία εξαρτάται από την εμπειρία και τη διακριτική ικανότητα του γιατρού και ως εκ τούτου είναι αρκετά υποκειμενική. Σημαντικό επίσης είναι ότι η όλη διαδικασία

από την αρχή μέχρι και την τελική διάγνωση μπορεί να διαρκέσει μέχρι και αρκετές εβδομάδες με μεγάλη πιθανότητα να χρειαστούν περισσότερες από μία βιοψίες.

Ιδανικά θα θέλαμε τεχνικές μη επεμβατικές, ασφαλείς, γρήγορες, οικονομικές και να παρέχουν χημικές πληροφορίες (να μην είναι μονό απεικονιστικές). Επιπλέον, θα ήταν ιδανικό να μπορούσαμε να κάνουμε διάγνωση προληπτική και να μην χρειαστεί να βρεθεί ο όγκος πρώτα. Οι ανάγκες αυτές οδήγούν την έρευνα στην εύρεση νέων διαγνωστικών. Οι νέες αυτές τεχνικές χρησιμοποιούν πηγή φωτός στο ορατό ή το κοντινό υπέρυθρο για να δώσουν χημικές πληροφορίες και είναι λιγότερο επεμβατικές από μεθόδους όπως η βιοψία. Μία από αυτές τις μεθόδους είναι η φασματοσκοπία Raman. Η φασματοσκοπία αυτή μπορεί να ανιχνεύσει πολλά χημικά στοιχεία στον ιστό και συνεπώς και πολλές χημικές μεταβολές που σχετίζονται με την παρουσία κακοήθους καρκινικού όγκου (5).

1.3 Η φασματοσκοπία Raman ως τεχνική σε βιολογικά δείγματα

Η παρουσία μεγάλου αριθμού ενεργών κατά Raman μορίων μέσα στον ιστό του μαστού είναι ένα επιπλέον πλεονέκτημα της μεθόδου αυτής. Η ενεργός διατομή της σκέδασης Raman είναι πολλές τάξεις μεγέθους ασθενέστερη από τη σκέδασης Rayleigh που σημαίνει ότι πρέπει να ενισχυθεί το σήμα μας με κατάλληλους τρόπους όπως θα αναλύσουμε παρακάτω. Η φασματοσκοπία Raman θα μπορούσε να αποτελεί μία αποτελεσματική και έγκαιρη διαγνωστική μέθοδο, καθώς κάθε μόριο χαρακτηρίζεται από δικούς του χαρακτηριστικούς τρόπους δόνησης τους οποίους και ανιχνεύει η μέθοδος Raman με αποτέλεσμα να έχουμε χημικές πληροφορίες για το δείγμα μας και άρα να έχουμε εφαρμογή σε μια διαδικασία διάγνωσης καρκίνου του μαστού κα όχι μόνο (6).

Η σκέδαση Raman είναι μια ανελαστική σκέδαση, κατά την οποία τα φωτόνια μια πηγής (συνήθως) laser προσπίπτουν στην επιφάνεια ενός δείγματος και σκεδάζονται με ενέργεια μικρότερη ή μεγαλύτερη της αρχικής τους. Η διαφορά ενέργειας αντιστοιχεί σε συγκεκριμμένες δονήσεις ή περιστροφές μορίων και γι’ αυτό το αποτύπωμα Raman για κάθε μόριο είναι μοναδικό. Η τεχνική αυτή είναι επίσης ικανή να χρησιμοποιηθεί in vivo, αφού τα μήκη κύματος που χρησιμοποιούνται αλλά και η

ενέργεια του laser είναι ακίνδυνη για τους ιστούς και επιπλέον έχουν αρκετά μεγάλο βάθος διείσδυσης.

Το νερό είναι κακός σκεδαστής του φωτός και συνεπώς όταν έχουμε φασματοσκοπία Raman υδατικού διαλύματος, δεν έχουμε σκέδαση του νερού που μπορεί να αλλοιώνει το αποτέλεσμά μας. Το γεγονός αυτό μας δίνει τη δυνατότητα να μπορούμε να μελετήσουμε φασματοσκοπικά κατά Raman βιολογικά δείγματα που έχουν μεγάλη περιεκτικότητα σε νερό (7).

Θα πρέπει όμως να είμαστε ιδιαίτερα προσεκτικοί να μη δημιουργήσουμε τοπική υπερθέρμανση του δείγματος για να μην αλλοιωθεί το υλικό μας. Επιπλέον πρέπει να λάβουμε υπόψην μας την ύπαρξη ακτινοβολίας φθορισμού που είναι η κύρια πηγή υποβάθρου στη φασματοσκοπία Raman.

1.4 Πλεονεκτήματα φασματοσκοπίας έναντι βιοψίας

Ίσως αναρωτηθεί κανείς γιατί θα μας ήταν χρήσιμο να αντικαταστήσουμε μια μέθοδο παγιωμένη και κοινώς αποδεκτή από την ιατρική κοινότητα (βιοψία), με μία μέθοδο καινούργια όπως είναι η φασματοσκοπία Raman. Η βιοψία από τη φύση της παρουσιάζει έναν αριθμό μειονεκτημάτων. Αρχικά, το δείγμα που λαμβανεται από τον ασθενή πρέπει να περάσει μια διαδικασία και επεξεργασία πριν μπορέσει να αξιολογηθεί η οποία κραταέι μέρες. Αντίθετα, στη φασματοσκοπία Raman απαιτούνται μερικά λεπτά (ή και δευτερόλεπτα αν χρησιμοποιηθεί και τεχνική ενίσχυσης σήματος) για να ληφθεί το τελικό φάσμα του ιστού. Στη βιοψία, το κόστος είναι πιο μεγάλο, καθώς απαιτείται επέμβαση και εξειδικευμένοι γιατροί για να ανάλυση του δείγματος. Σε αντίθεση με τη βιοψία, στη φασματοσκοπία Raman μπορεί να υπάρχει μόνο μια συσκευή που θα χειρίζεται ο γιατρός και θα ελέγχει τα δείγματα. Όπως προαναφέραμε ένα μεγάλο μειονέκτημα της βιοψίας είναι ότι είναι επεμβατική μέθοδος. Αντίθετα, η φασματοσκοπία Raman μπορεί να λειτουργήσει με τη βοήθεια οπτικών ινών που εισέρχονται μέσα στο σώμα και κατευθύνονται στον ιστό που μας ενδιαφέρει. Η βιοψία βασίζεται σε μορφολογική ανάλυση οπότε η διάγνωση εξαρτάται σε μεγάλο βαθμό στην εμπειρία του παθολόγου (8). Στη

φασματοσκοπία Raman η διάγνωση μπορεί να βασιστεί σε αλγορίθμους και να δώσει αντικειμενικά αποτελέσματα.

Ένα μειονέκτημα της φασματοσκοπίας Raman είναι ότι η μέτρηση σε ιστό in-vivo δεν μας δίνει εύκολη πρόσβαση σε πολλά σημεία και πρέπει εκ των προτέρων να γνωρίζουμε πού βρίσκεται το πρόβλημα. Γι' αυτό το λόγο γίνονται πολλές μελέτες ώστε να αναπτυχθεί η τεχνική αυτή για μελέτη βιολογικών υγρών όπως το αίμα ή τα ούρα (9).

Κεφάλαιο 2 :Φασματοσκοπία Raman: Θεωρία

2.1 Βασικές Αρχές

Φασματοσκοπία είναι η μελέτη της αλληλεπίδρασης της ακτινοβολίας, σε όλες της τις μορφές, με την ύλη. Η αλληλεπίδραση ενδέχεται να προκαλέσει ηλεκτρονικές διεγέρσεις, (π.χ. UV), μοριακές δονήσεις (π.χ. Raman) ή προσανατολισμούς στο πυρηνικό σπιν (π.χ. NMR). Η **φασματοσκοπία** και φασματογραφία είναι όροι που χρησιμοποιούνται για να αναφερθούμε στην μέτρηση της έντασης της ακτινοβολίας ως συνάρτηση του μήκους κύματος και συχνά χρησιμοποιούνται για να περιγράψουν πειραματικές φασματοσκοπικές μεθόδους.

Όταν μια δέσμη λευκού φωτός χτυπά ένα τριγωνικό πρίσμα, το φως χωρίζεται στις διάφορες συνιστώσες του (ROYGBIV: red, orange, yellow, green, blue, indigo and violet). Αυτό είναι γνωστό ως **φάσμα**. Το οπτικό σύστημα το οποίο επιτρέπει την παραγωγή και την προβολή του φάσματος ονομάζεται **φασματόμετρο**. Υπάρχουν πολλές άλλες μορφές του φωτός οι οποίες δεν είναι ορατές στο ανθρώπινο μάτι και φασματοσκοπία επεκτείνεται για να τις καλύψει (10).

Τα φάσματα μπορούν να ταξινομηθούν ανάλογα με τη φύση της προέλευσής τους, τον τρόπο δηλαδή με τον τρόπο που αλληλεπίδρασαν με την ύλη. Οι κύριοι τρόποι περιλαμβάνουν:

- Απορρόφηση: Ενέργεια από την πηγή ακτινοβολίας απορροφάται από το υλικό. Σε ένα φάσμα απορρόφησης, τμήματα ενός συνεχούς φάσματος λείπουν, επειδή έχουν απορροφηθεί από το μέσο από το οποίο έχει περάσει το φως.
- Εκπομπή: Υποδεικνύει ότι ακτινοβολούσα ενέργεια απελευθερώνεται από το υλικό. Για παράδειγμα, υο φάσμα μέλανος σώματος ενός υλικού είναι ένα αυθόρμητο φάσμα εκπομπής που καθορίζεται από τη θερμοκρασία του.
- Ελαστική σκέδαση: Προσπίπτουσα ακτινοβολία ανακλάται ή σκεδάζεται από ένα υλικό. (Παράδειγμα: Rayleigh scattering)

- Ανελαστική σκέδαση: Περιλαμβάνει την ανταλλαγή ενέργειας μεταξύ της αλληλεπίδρασης της ακτινοβολίας και της ύλης κατά την οποία μετατοπίζεται το μήκος κύματος της σκεδαζόμενης ακτινοβολίας. (Raman, Compton)
- Σύμφωνη φασματοσκοπία ή φασματοσκοπία συντονισμού: Παράδειγμα NMR.

Η φασματοσκοπία Raman βασίζεται στην ανελαστική σκέδαση. Όταν κάποιο υλικό σύστημα (στερεό, υγρό, αέριο, άμορφο ή κρυσταλλικό), ακτινοβολείται με μονοχρωματική ακτινοβολία, τότε η σκεδαζόμενη ακτινοβολία εκτός από την αρχική συχνότητα (Rayleigh) περιέχει και νέες φασματικές περιοχές. Οι περιοχές αυτές αντιστοιχούν σε μοναδικές για το κάθε μόριο συχνότητες με αποτέλεσμα ένα «δαχτυλικό αποτύπωμα» του κάθε μορίου.

Ανάλογα με το αν η σκεδαζόμενη ακτινοβολία είναι μικρότερου ή μεγαλύτερου μήκους κύματος – ενέργειας από την προσπίπτουσα έχουμε αντίστοιχα σκέδαση Stokes ή Anti-Stokes.

Η σκέδαση Stokes λαμβάνει χώρα όταν το μόριο προσλαμβάνει ενέργεια από το προσπίπτον φωτόνιο και μεταβαίνει από μία κατάσταση χαμηλότερης ενέργειας σε μία διεγερμένη δονητική κατάσταση. Αντίστοιχα στην Anti-Stokes το μόριο μπορεί να βρίσκεται ήδη σε διεγερμένη δονητική κατάσταση πριν την αλληλεπίδραση με το προσπίπτον φωτόνιο και το μόριο να απελευθερώσει ενέργεια και να μεταβεί σε στάθμη χαμηλότερης ενέργειας (11).

Όταν αναλυθεί η σκεδαζόμενη συχνότητα παρατηρούμε την ελαστική σκέδαση Rayleigh του φαινομένου (την οποία τα όργανά μας αφαιρούνε όπως θα δούμε παρακάτω) και ένα πολύ μικρό ποσοστό της ακτινοβολίας της τάξης του 10^{-7}, που σκεδάζεται σε διαφορετικά μήκη κύματος. Αυτό το μικρό ποσοστό είναι που αποτελεί τη σκεδαζόμενη ακτινοβολία Raman (12).

Το φαινόμενο σκέδασης Raman οφείλει το όνομά του στον Ινδό φυσικό Sir C.V. Raman, ο οποίος στα πλαίσια της έρευνάς του για τη μοριακή σκέδαση φωτός, το απέδειξε πειραματικά το 1928 και τιμήθηκε με το Nobel Φυσικής το 1930.

2.2 Μακροσκοπική ερμηνεία φαινομένου Raman και κλασσική προσέγγιση

Στην κλασσική προσέγγιση του φαινομένου το προσπίπτον φως πολώνει τα μόρια του μέσου με το οποίο αλληλεπιδρά επάγοντας ταλαντούμενα ηλεκτρικά δίπολα τα οποία ακτινοβολούν. Η σκέδαση της ηλεκτρομαγνητικής ακτινοβολίας προέρχεται από την αλληλεπίδραση των φωτονίων με το ηλεκτρονιακό νέφος δηλαδή το φως σκεδάζεται από τα ηλεκτρόνια του (13).Η σκέδαση Raman αποδίδεται στην σύζευξη των κινήσεων των ηλεκτρονίων και των πυρήνων.

Σε ένα υλικό, που βρίσκεται υπό την επίδραση ενός ταλαντευόμενου ηλεκτρικού πεδίου

$$\mathbf{E} = \boldsymbol{E_0} \cos(2\pi f_{laser} t) \qquad [1]$$

αναπτύσσεται μία επαγόμενη ταλαντούμενη πόλωση:

$$\boldsymbol{P} = \tilde{\alpha} \cdot \boldsymbol{E} \qquad [2]$$

Το μέγεθος $\tilde{\alpha}$ ονομάζεται μοριακή πολωσιμότητα και είναι η ευκολία με την οποία η πυκνότητα της ηλεκτρονιακής κατανομής μπορεί να παραμορφωθεί από ένα ηλεκτρικό πεδίο $\boldsymbol{E}$. Η πόλωση ως εκ τούτου γράφεται ως:

$$\boldsymbol{P} = a\, \boldsymbol{E_0} \cos(2\pi f_{laser} t) \qquad [3]$$

Η πολωσιμότητα εξαρτάται από τη γεωμετρία του μορίου, καθώς το μόριο δονείται εκείνη αλλάζει. Μπορούμε λοιπόν να γράψουμε την πολωσιμότητα σαν ανάπτυγμα σειράς Taylor γύρω από τη θέση ισορροπίας ως προς τη γενικευμένη συντεταγμένη της δονητικής μετατόπισης q_i του i-οστού τρόπου ταλάντωσης:

$$Qi = Q_i^0 \, cos(2\pi f_{vibr} t) \qquad [4]$$

$$\text{Taylor ανάπτυγμα: } a = a_0 + \frac{\partial a}{\partial Q} Q + \cdots \qquad [5]$$

Αντικαθιστώντας την [5] στην [4] και αναπτύσσοντας το γινόμενο των τριγωνομετρικών συναρτήσεων σε άθροισμα, προκύπτει:

$$p = a_0 E_0 \cos(2\pi f_{laser} t) + Q_i^0 \frac{\partial a}{\partial Qi} \frac{Q_i^0 E_0}{2} [\cos(2\pi(f_{laser} - f_{vibr})t) + \cos(2\pi(f_{laser} + f_{vibr})t)] \quad [6]$$

Αυτό το νέο Η/Μ κύμα αποτελείται από τον πρώτο όρο, που με συχνότητα f_{iaser} αντιστοιχεί στην ελαστική σκέδαση Rayleigh και άλλους δύο όρους με νέες συχνότητες $\omega_i \pm \omega_{i0}$, που αντιστοιχούν στις μη ελαστικές σκεδάσεις Stokes (2ος όρος) και Anti-Stokes (3ος όρος). Η πολωσιμότητα στη θέση ισορροπίας είναι α_0 και $(\frac{\partial \alpha}{\partial Q_i})$ είναι ο ρυθμός μεταβολής της πολωσιμότητας ως προς μια μετατόπιση του πυρήνα.

Από τη σχέση [6] συμπεραίνουμε ότι αν η μεταβολή της πολωσιμότητας ως προς την κανονική συνταταγμένη (Q_i) είναι μη μηδενική γύρω από το σημείο ισορροπίας, τότε

$$\frac{\partial a}{\partial Qi}\Big|_{Q_i=0} \neq 0 \quad [7]$$

αφού για:

$$\frac{\partial a}{\partial Qi}\Big|_{q_i=0} = 0 \quad [8]$$

μηδενίζονται όλοι οι όροι εκτός από τον πρώτο της ελαστικής σκέδασης.

2.3 Κβαντική προσέγγιση φαινομένου Raman

Για να περιγράψουμε πλήρως το φαινόμενο Raman πρέπει να το μελετήσουμε κβαντομηχανικά. Η σκέδαση Raman πρώτης τάξης περιγράφεται ως μια διαδικασία όπου το φως αλληλεπιδρά έμμεσα με τα ηλεκτρόνια σθένους του υλικού.

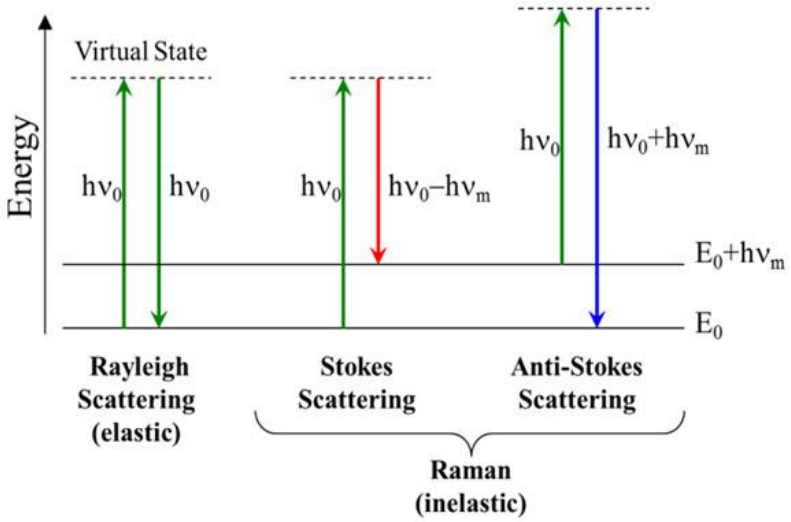
Για την κβαντική ερμηνεία, η ηλεκτρομαγνητική ερμηνεία περιγράφεται ως ροή φωτονίων ενέργειας $\hbar\omega_i$ και ορμής $\hbar \boldsymbol{k}_i$. Η ενέργεια και η ορμή διατηρούνται μεταξύ της αρχικής και τελικής κατάστασης του συστήματος. Στη σκέδαση Rayleigh έχουμε $\omega_0 = \omega_s$ και $k_0 = k_s$. Για τη σκέδαση Raman η διατήρηση της ενέργειας και της ορμής είναι:

$$\omega_0 = \omega_s \pm \omega_j(q) \text{ και}$$

$$k_0 = k_s \pm q \tag{9}$$

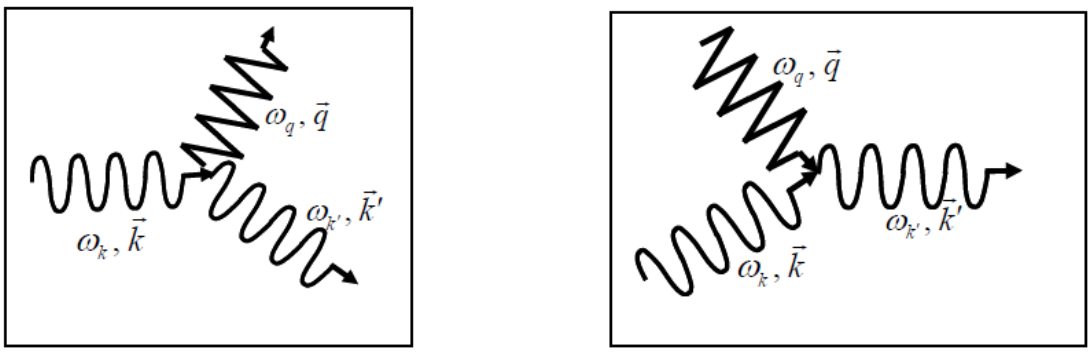
όπου το πρόσημο (+) αντιστοιχεί στη σκέδαση Stokes, καθώς έχουμε τη δημιουργία ενός φωνονίου $\omega_j(q)$, ενώ το πρόσημο (-) αντιστοιχεί στην anti-Stokes όπου έχουμε την καταστροφή ενός $\omega_j(q)$.

Διαγραμματικά:



Εικόνα 2.3.1

Συγκεκριμένα η διαδικασία Stokes και anti-Stokes (αντίστοιχα) σε διαγράμματα απεικονίζεται ως εξής (14):



Εικόνα 2.3.2

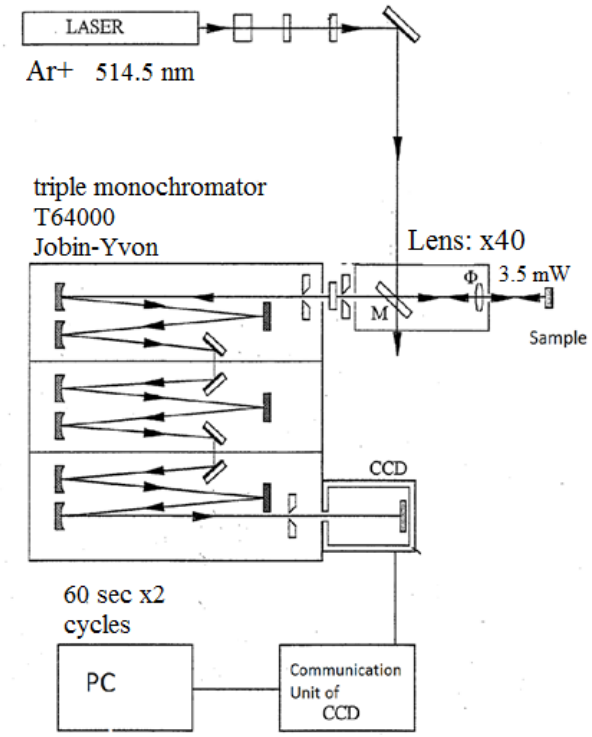
Κεφάλαιο 3: Πειραματική διαδικασία

3.1 Επιστημονικό ερώτημα

Τα δύο κύρια είδη μελέτης που μπορεί να κάνει κανείς όταν μελετάει τον καρκίνο με φασματοσκοπία Raman είναι η διαγνωστική και η διερευνητική μελέτη. Στη διερευνητική μελέτη αυτό που μας απασχολεί είναι τι είδους χημικές αλλαγές συμβαίνουν μέσα βιολογικό υλικό που μελετάμε. Αυτό που μας ενδιαφέρει είναι μια διαγνωστική είδους μελέτη. Αυτό που μας ενδιαφέρει είναι η παρουσία των χημικών ενώσεων και των αλλαγών τους ώστε να βγάλουμε συμπεράσματα για το δείγμα μας. Γι' αυτό το λόγο είναι απαραίτητο ένα μεγάλο σύνολο δεδομένων (τα φάσματά μας) το οποίο θα επεξεργαστεί ώστε να βγει συμπέρασμα μέσω στατιστικής ανάλυσης.

3.2 Πειραματική διάταξη

Η πειραματική διάταξη micro - Raman που χρησιμοποιήσαμε έχει στηθεί στο Εθνικό Μετσόβιο Πολυτέχνειο, στο κτήριο φυσικής με φασματόμετρο ένα τριπλό μονοχρωμάτορα T64000 της Jobin-Yvon. Σχηματικά:



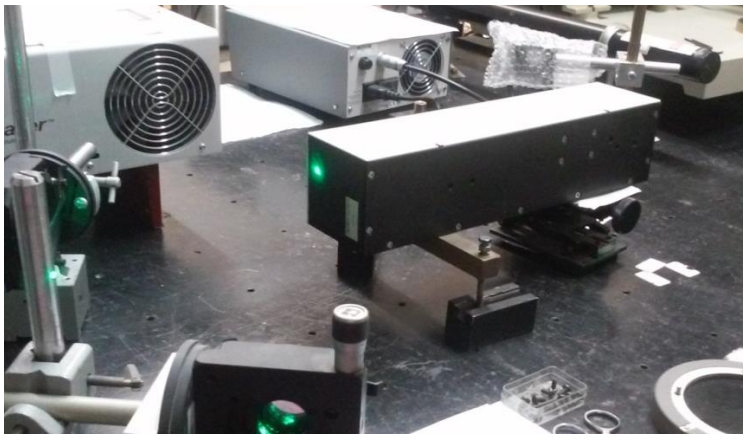
(Εικόνα 3.2.1)

Η διάταξη μας είναι η παρακάτω:

Εικόνα 3.2.2

Για τη διέγερση των υλικών, χρησιμοποιήθηκε λέιζερ (laser) αερίου Ar+ (ιόντα Αργού) που εκπέμπει μονοχρωματικό και πολωμένο ορατό φως στα 514.5 nm. Ο φακός του μικροσκοπίου ήταν x40.

Για την απομάκρυνση γραμμών πλάσματος αλλά και ενός ασθενούς υποβάθρου, υπάρχει στην αρχή της διάταξής μας ένα ρυθμιζόμενο **φίλτρο διέγερσης - μονοχρωμάτορας** (SPEX 1450 Tunable Excitation Filter monochromator). Στην εικόνα είναι το μάυρο κουτί που περιέχει την κατάλληλη οπτική διάταξη για αυτή τη δουλειά:

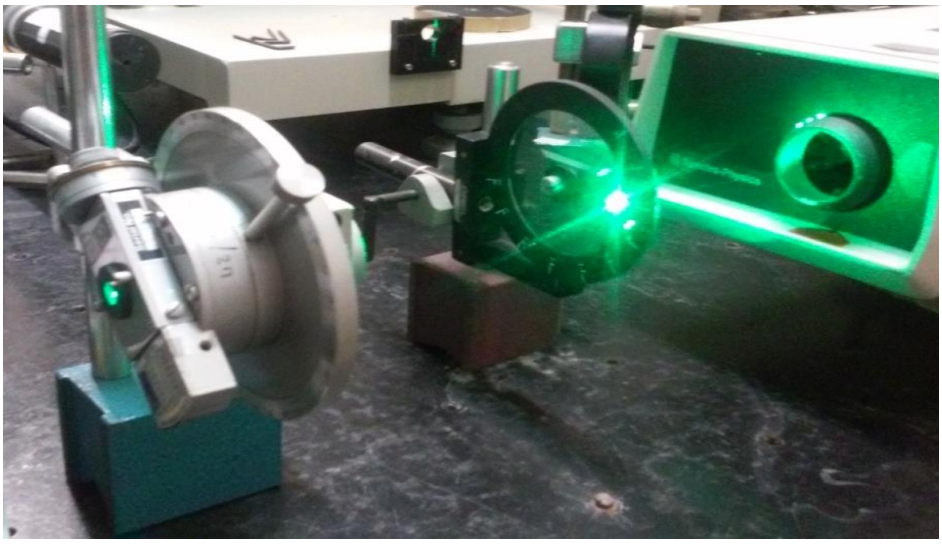


(*Εικόνα 3.2.3*)

Στη συνέχεια, αν θελήσουμε να μειώσουμε την ισχύ του λέιζερ χωρίς να πειράξουμε την έντασή του χρησιμοποιούμε ένα φίλτρο-**απορροφητή δέσμης.**

Στη διάταξή μας υπάρχει κι ένας **πολωτής** μέσω του οποίου ρυθμίζουμε την κατεύθυνση της πόλωσης του προσπίπτοντος φωτός στο δείγμα, τον οποίο όμως εμείς δεν χρησιμοποιήσαμε στο πειραμά μας.

Το φίλτο φαίνεται στα δεξιά και ο πολωτής στα αριστερά της φωτογραφίας:

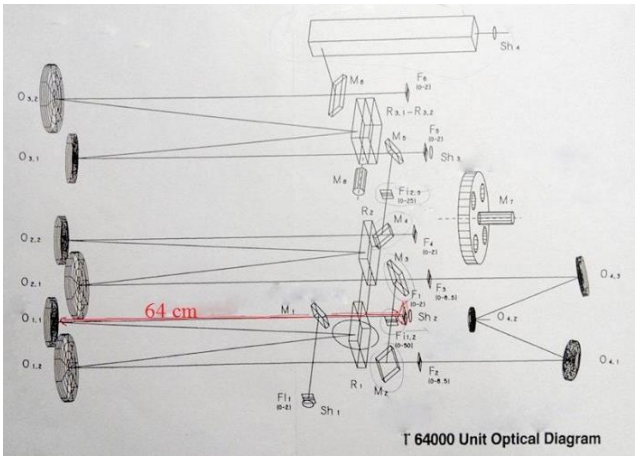


(*Είκονα 3.2.4*)

Η δέσμη είναι τώρα έτοιμη να οδηγηθεί στο μικροσκόπιο και να γίνει το πειραμά μας. Το σκεδαζόμενο φως από το δείγμα μας, συλλέγεται από έναν αντικειμενικό φακό και οδηγείται στην είσοδο του φασματομέτρου.

Το **φασματόμετρο** αποτελείται από τρεις μονοχρωμάτορες. Ο πρώτος αναλύει τη πολυχρωματική δέσμη στις επιμέρους συνιστώσες της με τη βοήθεια των ολογραφικών φραγμάτων περίθλασης. Η απόσταση από την είσοδο της δέσμης μέχρι τον πρώτο καθρέφτη στον οποίο και πέφτει εστιασμένη είναι 64 εκατοστά, εξού και το όνομα T64000. Στη συνέχεια περνάει στον δεύτερο μονοχρωμάτορα όπου και συντείθεται ξανά έχοντας όμως πλέον απορρίψει τις συχνότητες που αντιστοιχούν στη σκέδαση Rayleigh.

Παρακάτω παρατίθεται το σχέδιο του φασματόμετρου αναλυτικά με όλα τα οπτικά στοιχεία που χρειάζονται για subtractive (όταν κόβεται το εύρος των συχνοτήτων της Rayleigh) και για additive mode.



(*Είκονα 3.2.5*)

Η πολυχρωματική αυτή δέσμη περνάει στον τρίτο και τελευταίο μονοχρωμάτορα όπου και γίνεται η ουσιαστική ανάλυση της δέσμης, στην έξοδο της οποίας παίρνουμε τις συχνότητες που έχουμε επιλέξει να ερευνήσουμε (ανάλογα με τις διεγέρσεις φωνονίων που μας ενδιαφέρουν).

Το φασματόμετρο είναι εφοδιασμένο με κάμερα CCD για την καταγραφή του σήματος. Ο **ανιχνευτής CCD** είναι μια τεχνική πολυκαναλικής καταγραφής δεδομένων. Η δέσμη καταγράφεται πάνω σε μια επιφάνεια χωρισμένη σε πολύ μικρές φωτοευαίσθητες κυψελίδες (pixels). Σε κάθε κυψελίδα αναπτύσσεται ένα ηλεκτρικό φορτίο ανάλογο του αριθμού των προσπιπτόντων σε αυτό φωτονίων. Η CCD ψύχεται με υγρό άζωτο ώστε να διατηρείται κάτω από τους -133^{o} κελσίου, για να μειώνεται όσο γίνεται ο θερμικός θόρυβος. Το πλήθος και οι διαστάσεις των κυψελίδων του ανιχνευτή (άξονας ενέργειας ή διασποράς 1152 κυψελίδες x 22.5 μm/κυψελίδα=25.9 mm, άξονας εγκάρσιος στη διασπορά 298 κυψελίδες x 22.5 μm/κυψελίδα=6.7 mm) θέτουν όρια στη φασματική περιοχή που μπορεί ταυτόχρονα να σαρωθεί από το φασματόμετρο και καθορίζουν τη διακριτική ικανότητά του (15).

(*Εικόνα 3.2.6*)

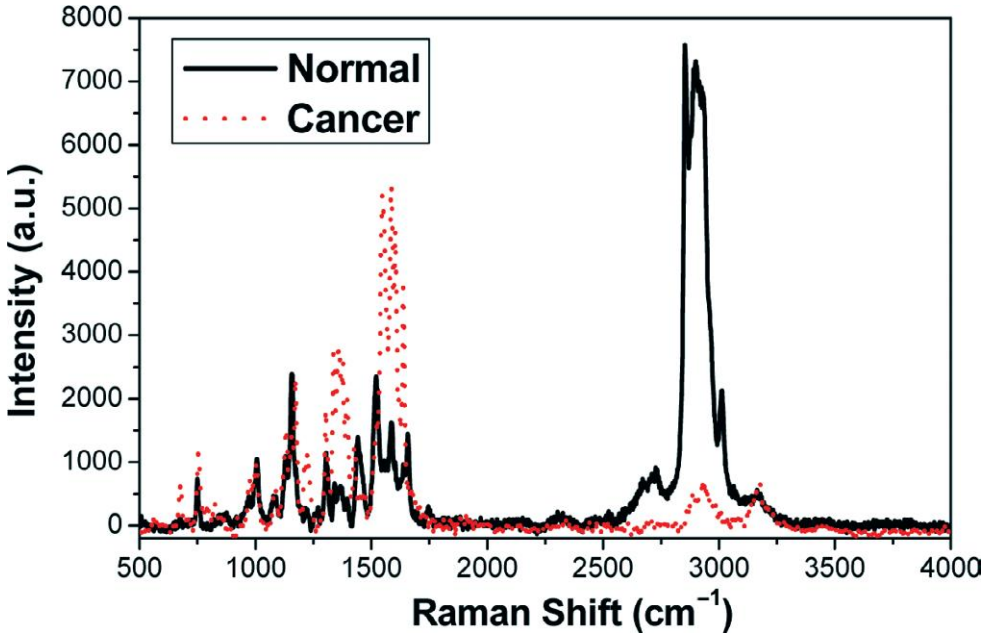
Στη συνέχεια μια μονάδα επικοινωνίας συνδέει τον ανιχνευτή CCD με τον υπολογιστή. Η δουλειά της μονάδας είναι να ψηφιοποιεί το σήμα που καταγράφεται στον ανιχνευτή, δηλαδή το ηλεκτρικό φορτίο ανά κυψελίδα. Το σήμα καταγράφεται με τη μορφή φάσματος: ο άξονας y είναι η ένταση του φωτός (ή αριθμός φωτονίων/κυψελίδα) και ο άξονας x είναι ο κυματαριθμός cm^{-1} (ή ισοδύναμα η συχνότητα ή η ενέργεια των φωτονών).

3.3 Αρχικά πειράματα

3.3.1 In-vivo πείραμα

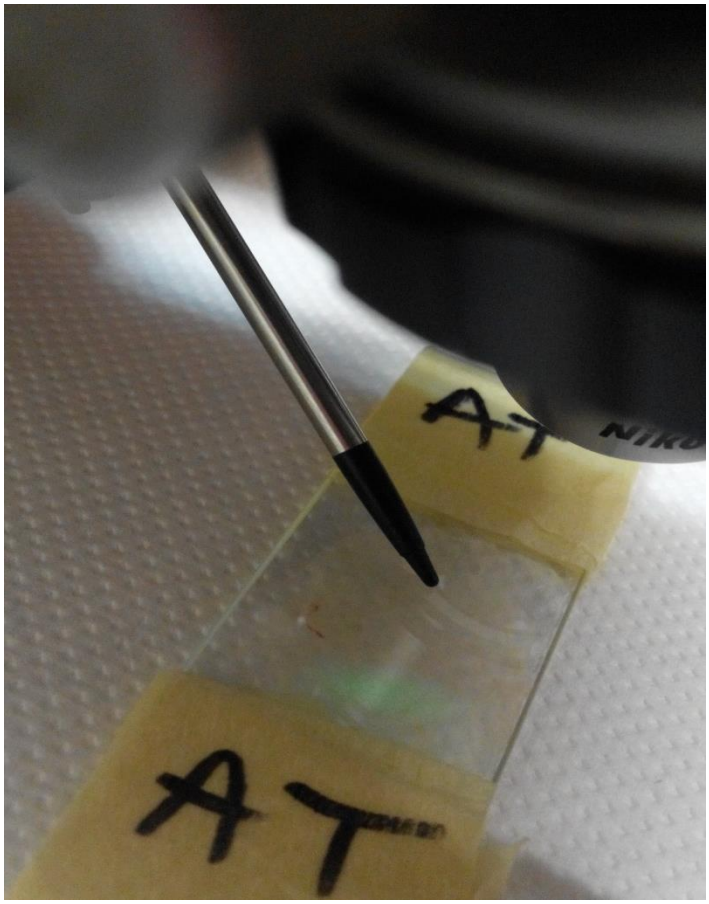
Το πρώτο σετ πειραμάτων μας περιελάμβανε μετρήσεις σε διάφορα στάδια καρκίνου από ιστό μαστού καφέ εργαστηριακών ποντικών καθώς και μετρήσεις σε υγιή ιστό. Ο μαστός χωρίζονταν με κάθετες τομές σε μικρότερα κομμάτια και τοποθετούνταν σε slides ώστε να μετρηθούν κάτω από το μικροσκόπιο. Οι μετρήσεις γίνονταν σε 3 διαφορετικές φασματικές περιοχές με εντάσεις από 0.2 mW μέχρι 5 mW, ώστε να βρούμε την κατάλληλη ένταση. Οι φασματικές περιοχές είχαν ως κέντρο τα 1000, 1500 και 2800 cm^{-1}.

Τα φάσματα σύμφωνα με τη βιβλιογραφία σε καρκινικό και υγιή ιστό μοιάζουν ως εξής (16):



(*Είκονα 3.3.1.1*)

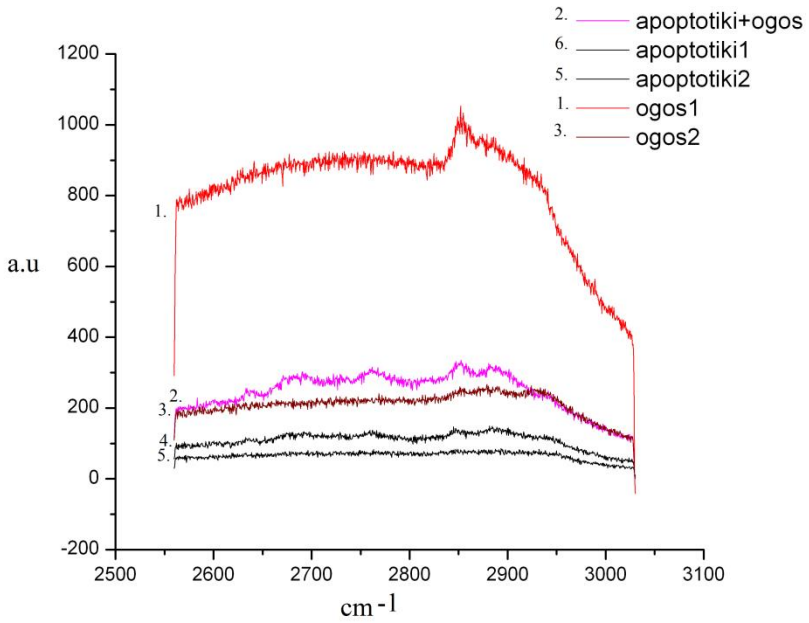
Τα δείγματα παρεσχέθησαν από το Ίδρυμα Ιατροβιολογικών Ερευνών της Ακαδημίας Αθηνών (ΙΙΒΕΑΑ).



(*Είκονα 3.3.1.2 Advanced tumor sample*)

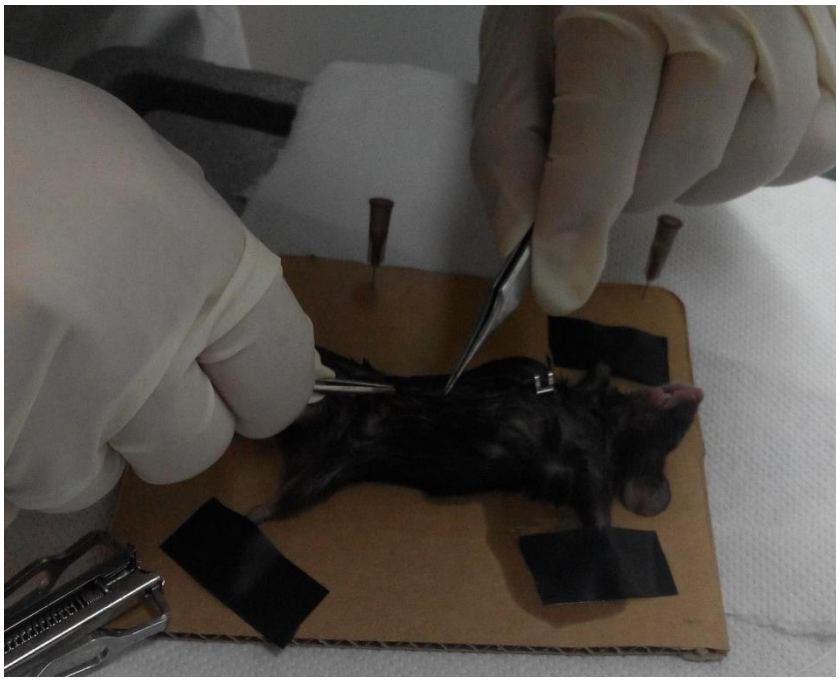
Αφού δοκιμάσαμε με διαφορετικές μεταβλητές καταλήξαμε να γίνουν οι μετρήσεις στα 3.5 mW και σε 2 κύκλους των 60 sec. Οι ιστοί ήταν τοποθετημένοι ανάμεσα σε 2 slides, γεγονός που μας δυσκόλευε αρκετά στην εστίαση. Τα εμπορικά slides γυαλιού προκαλούν πολύ φθορισμό και καλό θα ήταν να αποφεύγονται. Ενδεικτικά παρατίθεται ένα διάγραμμα που περιέχει φάσματα από όλα τα είδη των περιοχών που μελετήσαμε.

Το παρακάτω φάσμα προέρχεται από περιοχές με όγκο σε αρχικά στάδια, περιοχές αποπτωτικές, δηλαδή περιοχές στις οποίες έχει επέλθει κυτταρικός θάνατος. Λόγω της δυσκολίας που είχαμε στην εστίαση δεν μπορούσαμε να εστιάσουμε στις σωστές περιοχές και τα φάσματα των όγκων δεν αντιστοιχούν όλα στα βιβλιογραφικά (πχ το κόκκινο φασμα είναι φάσμα υγιούς κυτταρου). Η αποπτωτική περιοχή δίνει δικά της μοναδικά φάσματα τα οποία όμως στη συνέχεια δεν μελετήσαμε παραπάνω.



(*Εικονα 3.3.1.3*)

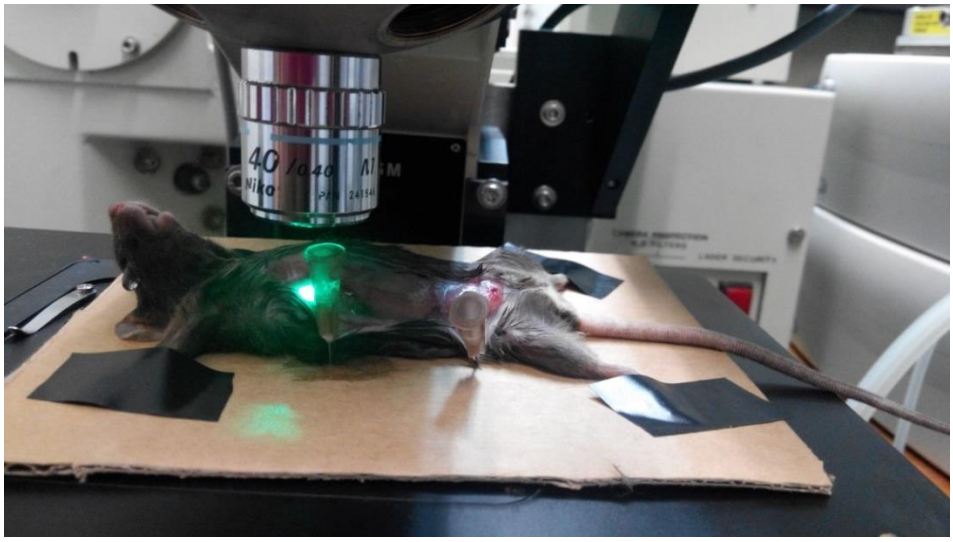
Αφού πήραμε κάποια αρχικά φάσματα In-vitro, δοκιμασαμε να κάνουμε μετρήσεις In-vivo σε ποντίκια. Τα ποντίκια ναρκώνονταν με προβλεπόμενες διαδικασίες και τοποθετούνταν πάνω σε μια επίπεδη επιφάνεια.



(Εικονα 3.3.1.4)

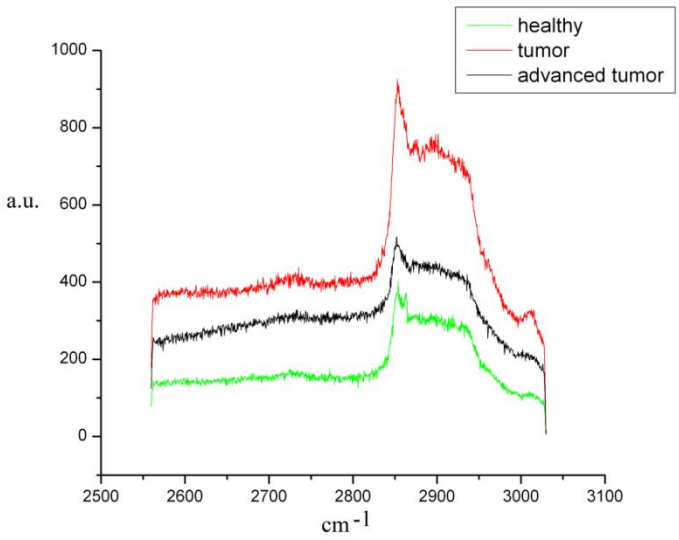
Στη συνέχεια τα ποντίκια τοποθετούνταν κάτω από το μικροσκόπιο με τη δέσμη του λέιζερ να εστιάζει πάνω στο όγκο σε κάποιον από τους μαστούς που νοσούσαν για τα

φάσμα του καρκίνου και πάνω σε κάποιον μαστό που δεν έχει όγκο για να φάσματα που αντιστοιχούν σε υγιή ιστό.

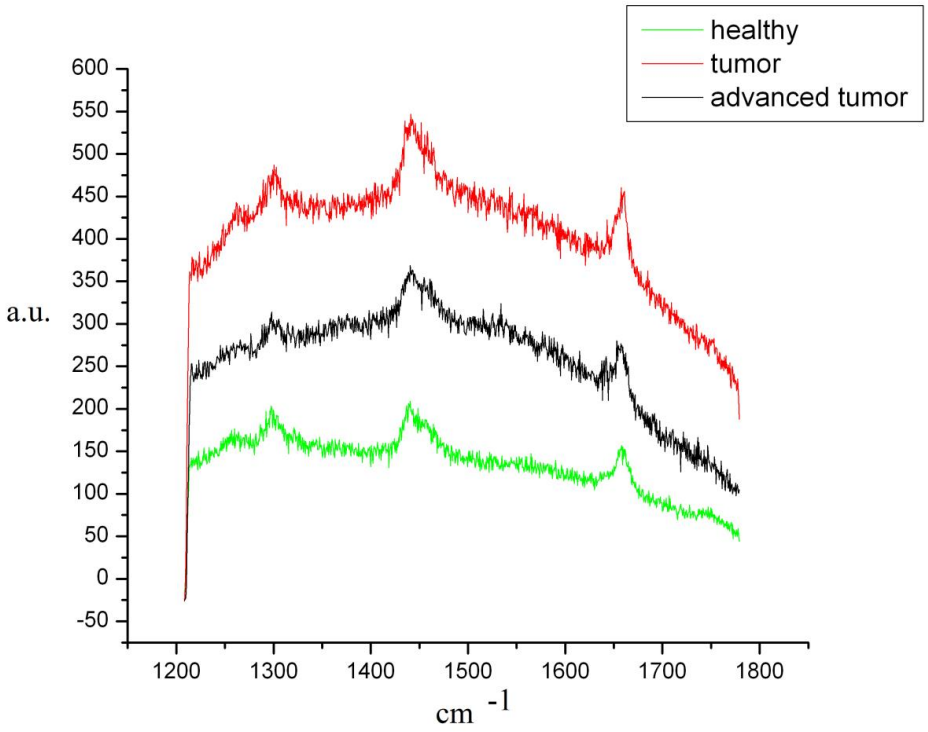


(*Εικονα 3.3.1.5*)

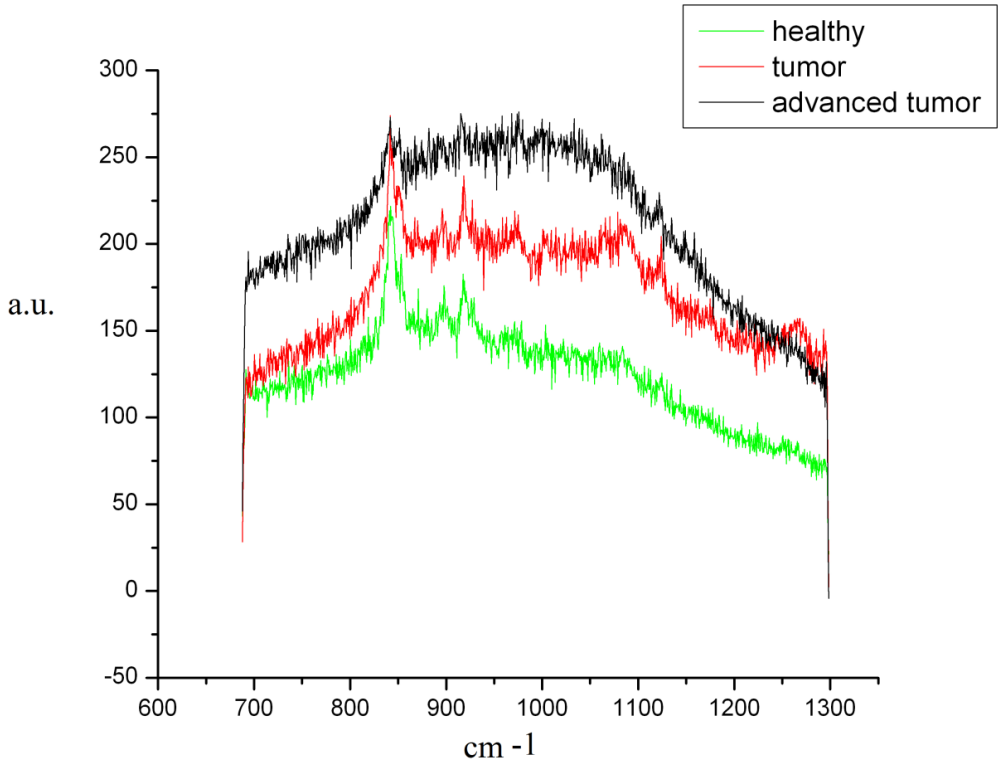
Οι μετρήσεις ήταν αρκετά προβληματικές. Λόγω της αναπνοής του ποντικού αλλά και του χτύπου της καρδιάς του ήταν πολύ δύσκολη η σωστή εστίαση και δεν μπορέσαμε να χρησιμοποιήσουμε το μικροσκόπιο για να εστιάσουμε. Γι' αυτό το λόγο πολλά από τα σημεία που μετρούσαμε, τα οποία αντιστοιχούν σε 1-2 κύτταρα περίπου, δεν έδειχναν ενδεικτικά φάσματα, είτε εστιάζαμε σε υγιή τμήματα ιστού είτε σε σημεία που δεν έδιναν σήμα όπως το αίμα. Ένα επιπλέον πρόβλημα ήταν ότι οι μετρήσεις δεν μπορούσαν να κρατήσουν πάνω από δύο ώρες γιατί το ποντίκι ξύπναγε από την νάρκωση και μια δεύτερη νάρκωση θα οδηγούσε πιθανόν σε θάνατο. Ενδεικτικά παραθέτουμε κάποια από τα φάσματα που πήραμε:



(*Εικονα 3.3.1.6*)



(*Εικονα 3.3.1.7*)



(*Εικονα 3.3.1.8)*

Όπως φαίνεται από αυτά τα φάσματα, αν τα συγκρίνουμε με τη βιβλιογραφία, δυσκολευόμαστε να βρούμε σημείο στο ιστό που να δίνει φάσμα που αντιστοιχεί σε φάσμα καρκίνου. Και τα 3 φάσματα (που αντστοιχούν σε υγιή μαστό, και σε μαστό με αρχικό και προχωρημένο καρκίνο) μοιάζουν ίδια ενώ θα έρπεπε να υπάρχουν διαφορές στις εντάσεις αλλά και στην ύπαρξη ή μη ορισμένων κορυφών.

Για τους λόγους αυτούς η ανάλυση In – vivo αποδείχτηκε προβληματική και εγκαταλήφθηκε.

3.3.2 SERS σε ιστό και κυτταροσειρές

Στη συνέχεια των πειραμάτων μας δοκιμάσαμε να ενισχύσουμε το σήμα μας με μια τεχνική που αποκαλείται Surface Enhanced Raman Spectroscopy (SERS-φασματοσκοπία Raman με επιφανειακή ενίσχυση)

Η SERS είναι μια πιο πολύπλοκη διαδικασία ενίσχυσης σήματος η οποία βελτιώνει το φαινόμενο Raman μέσω μορίων του δείγματος που απορροφώνται από μεταλλικές επιφάνειες ή από νανοδομές γενικότερα (17). Η παρουσία ενός τέτοιου παράγοντα μπορεί να δώσει πολύ μεγάλη ενίσχυση, μέχρι και 10^{15} (18) και έχει χρησιμοποιηθεί και σε βιολογικά δείγματα. Η αιτία αυτής της μεταβολής του σήματος δεν έχει διαλευκανθεί πλήρως ακόμα. Υπάρχουν δύο πιθανές εκδοχές για την εμφάνιση της συγκεκριμένης ενίσχυσης. Είτε ενίσχυση της πολωσιμότητας α , είτε η ενίσχυση της έντασης του ηλεκτρικού πεδίου Ε:

Η ενίσχυση της πολωσιμότητας μπορεί να επέλθει λόγω δημιουργίας χημικών δεσμών μεταξύ της μεταλλικής επιφάνειας και των υπό μελέτη φορτίων. Αυτού του τύπου η ενίσχυση αναφέρεται και ως χημική ενίσχυση.

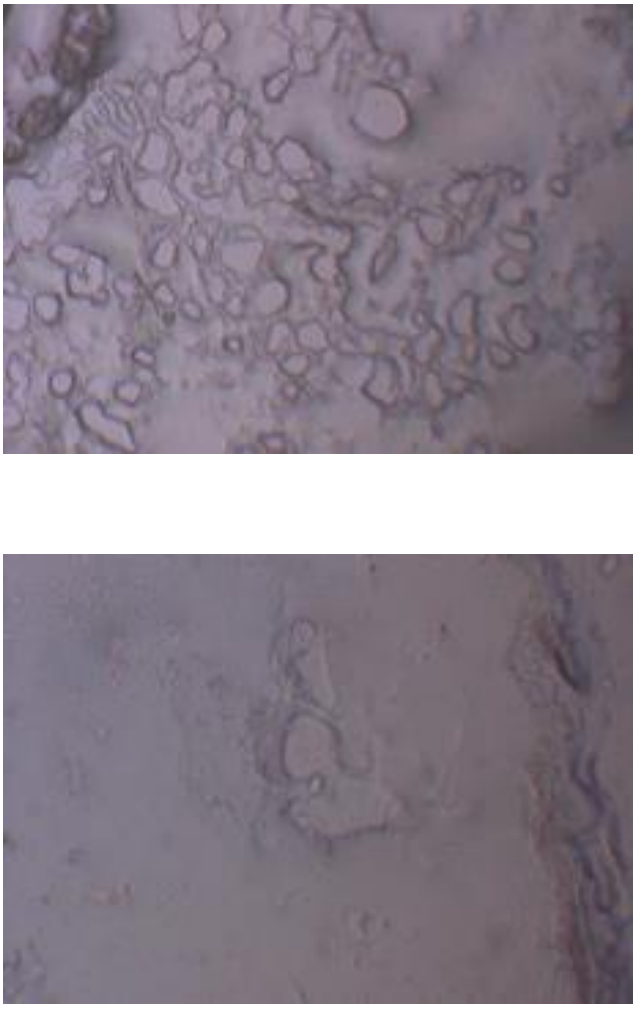
Ο δεύτερος τύπος ενίσχυσης γίνεται μέσω της έντασης του ηλεκτρικού πεδίου Ε και μπορεί να ερμηνευθεί αν ληφθεί υπόψη η αλληλεπίδραση της δέσμης laser με τα ελεύθερα ηλεκτρόνια του μετάλλου. Η δέσμη laser διεγείρει τα επιφανειακά τοπικά πλασμόνια. Αυτού του είδους η ενίσχυση του σήματος Raman αναφέρεται και ως ηλεκτρομαγνητική ενίσχυση και είναι σημαντικότερη, από άποψη ενίσχυσης του σήματος, ως προς τη χημική ενίσχυση.

Η SERS έχει μεγάλη δυσκολία στην προετοιμασία των δειγμάτων. Τα νανοσωματίδια που παραγγείλαμε ήταν μεγάλα (~100nm) και ήταν σωματίδια Αργύρου σε υγρή μορφή. Τα σωματίδια αυτά τα στάξαμε πάνω από τον τμήματα του ιστού και περιμέναμε από μερικά λεπτά εώς και μια μέρα για να απορροφηθούν. Παρόλ'αυτά δεν έδωσαν καμία ενίσχυση στο σήμα μας και είναι πιθανόν να μην απορροφήθηκαν σωστά ή να μην ήταν αρκετά κοντά στις υπό μελέτη ενώσεις μας ή το μέγεθος των νανοσωματιδίων να ήταν λάθος (18). Το ίδιο συνέβη και με τις κυτταρικές σειρές, από τις οποίες δεν μπορέσαμε να πάρουμε κανένα σήμα, πιθανόν γιατί χειριστήκαμε τα δείγματα με λάθος τρόπο (τα κύταρρα είναι εξαιρετικά ευαίσθητα) (19) (20).

3.4 Δείγμα

Μετά τα αρχικά πειράματα αλλάξαμε λίγο τα δείγματά μας ώστε να έχουμε όσο το δυνατό καλύτερο σήμα. Όλοι οι ιστοί πέρασαν από μια διαδικασία που αποκαλείται στερέωση (fixation). Η στερέωση είναι ένα κρίσιμο βήμα στην παρασκευή των τομών βιολογικών ιστών, με την οποία οι ιστοί διατηρούνται από τη φθορά, αποτρέποντας έτσι την αυτόλυση ή σήψη. Τα δικά μας δείγματα στερεώθηκαν με φορμαλδεΰδη την οποία ξεπλύναμε με PBS. Προηγουμένως οι ιστοί καταψύχθηκαν και υποβλήθηκαν σε κρυοτομή με πάχος από 10 μέχρο 50 μm.

Ενδεικτικά ένας καρκινικός ιστός στο μικροσκόπιο μοιάζει ως εξής:



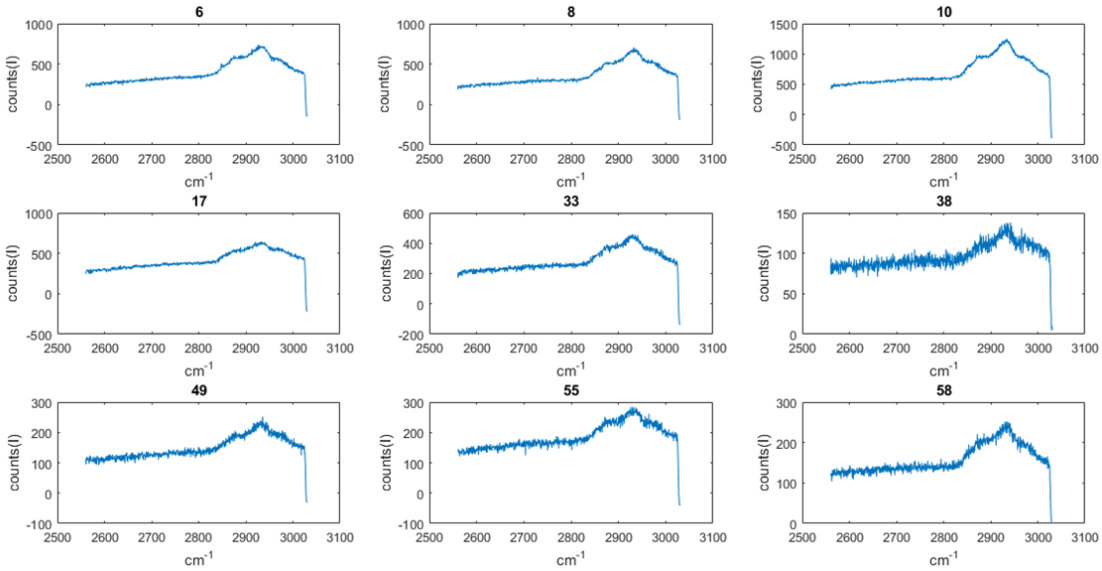
(*Εικόνα 3.4.1*)

Τα φάσματα που συλλέξαμε ήταν 146 στο σύνολο. Τα 49 προήλθαν από υγιή μαστικό ιστό, τα 39 από πρώιμο καρκινικό ιστό και τα 58 από προχωρημένο καρκινικό ιστό.

3.5 Φάσματα

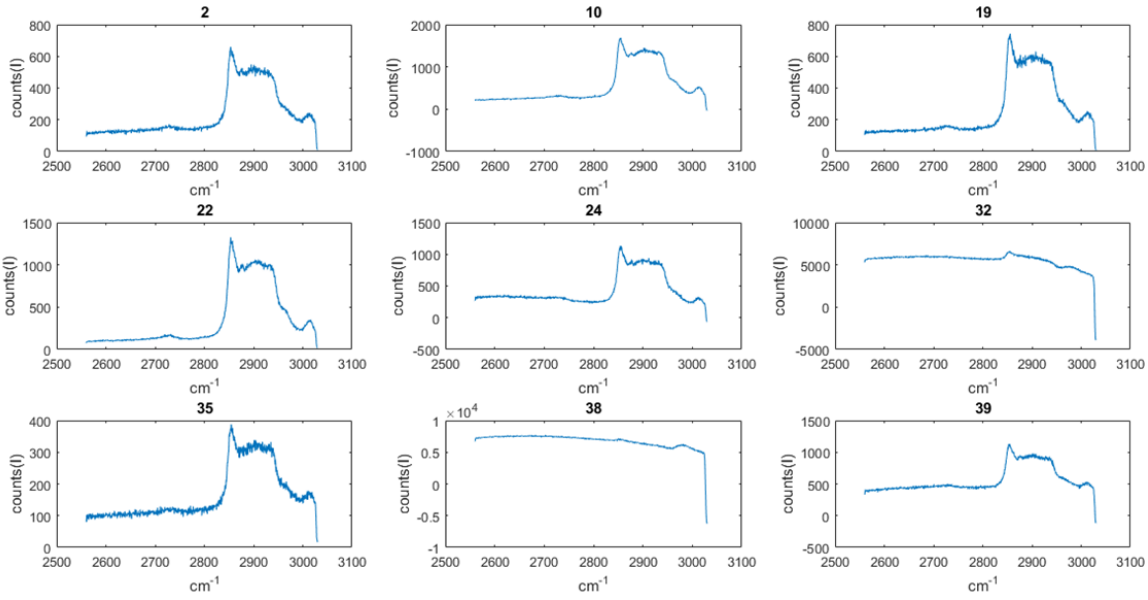
Για αρχή, παραθέτουμε τυχαία ανεπεξέργαστα φάσματα από ιστούς καρκινικούς αλλά και υγιείς:

Καρκινικοί ιστοί



(*Εικόνα 3.5.1*)

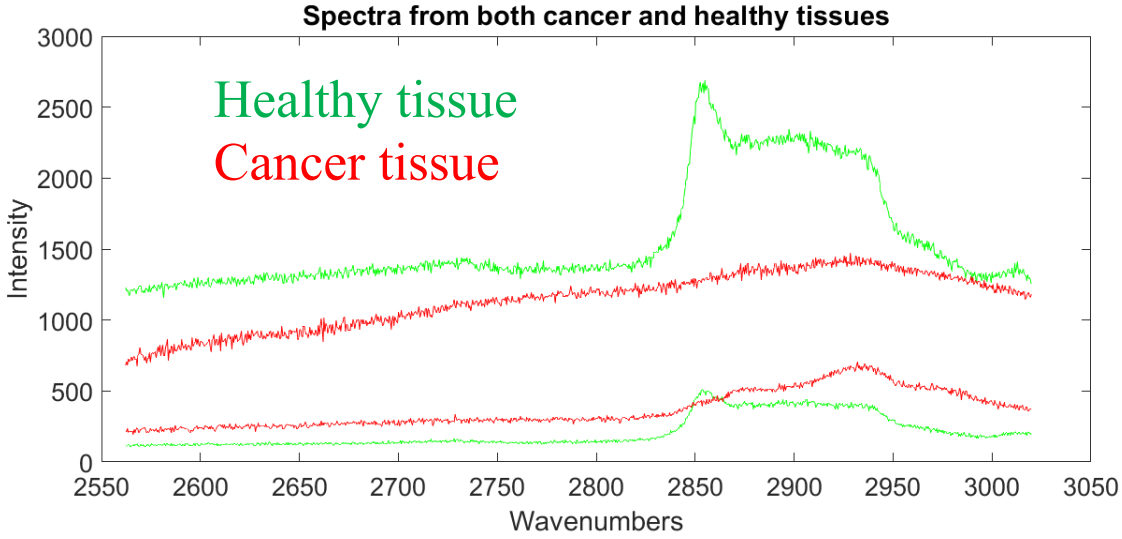
Υγιείς ιστοί



(*Εικόνα 3.5.2*)

Στο παρακάτω διάγραμμα παραθέτουμε 2 τυχαία φάσματα από καρκινικό μαστικό ιστό και άλλα 2 από υγιείς ιστούς. Η περιοχή συχνοτήτων που μελετήσαμε έχει ως

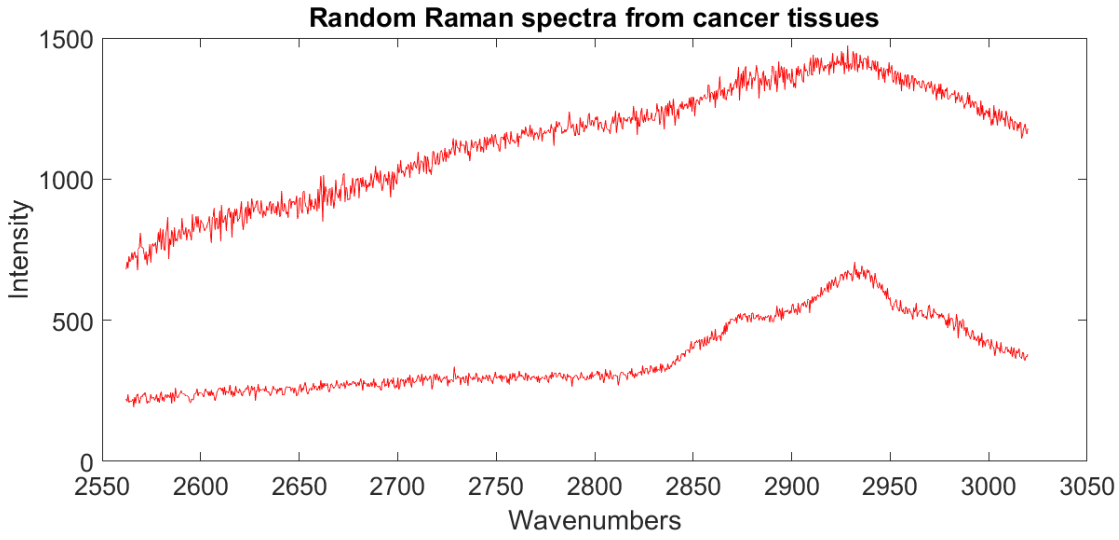
κέντρο τα 2800 cm^{-1}. Από,τι βλέπουμε στη βιβλιογραφία, αλλά φάνηκε και στα δικά μας φάσματα, οι διαφορές σε αυτές τις συχνότητες είναι πολύ μεγαλύτερες γι' αυτό και δεν κάναμε μετρήσεις (εκτός από μερικές δοκιμαστικές) στα 1500 και 1000 cm^{-1}.



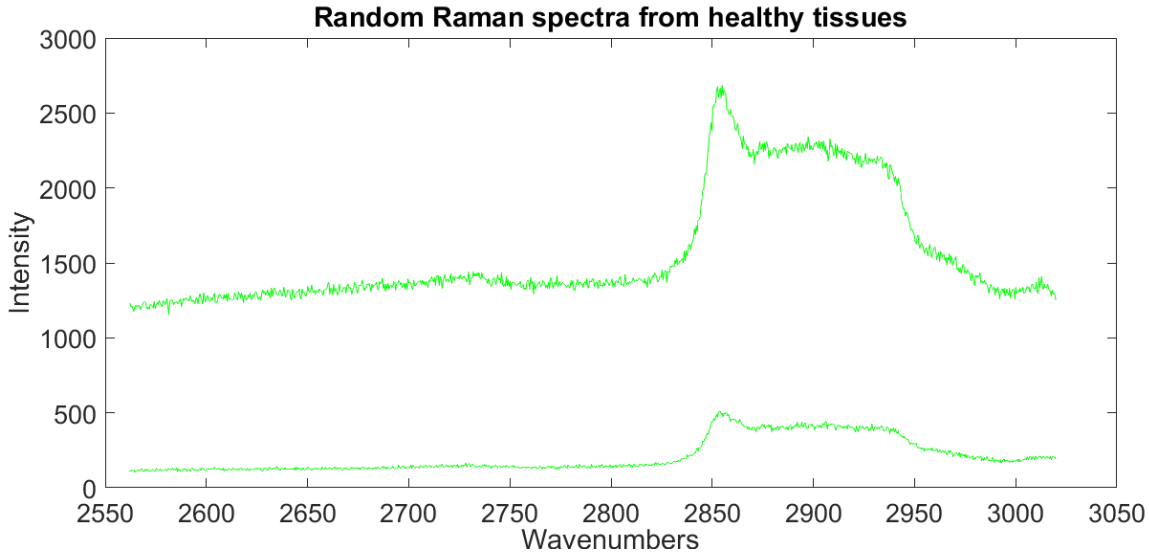
(*Εικόνα 3.5.3*)

Από μια πρώτη ματιά βλέπουμε πως τα φάσματα μοιάζουν πολύ στα βιβλιογραφικά αν και το σήμα είναι χειρότερο. Οι διαφορές ανάμεσα στα 2 είδη φασμάτων είναι αρκετά εμφανείς.

Πολλές κορυφές αλληλοεπικαλύπτονται και είναι δύσκολο να τις διακρίνουμε. Οι περιοχές αυτές αντιστοιχούν ως επί τον πλείστον σε λιπαρά οξέα. Θα αναφερθούμε στη συνέχεια λίγο πιο αναλυτικά στη χημική σύνθεση, αν και δεν είναι μέρος του σκοπού αυτής της εργασίας.



(*Εικόνα 3.5.4*)



(*Εικόνα 3.5.5*)

Κάποιες επιπλέον παρατηρήσεις που μπορούμε να κάνουμε έχουν να κάνουν με το σήμα των φασμάτων, το υπόβαθρο και τις εντάσεις των φασμάτων.

Ακόμη και ανάμεσα στου ίδου είδους φάσματα το σήμα μπορεί να είναι χειρότερο, το υπόβαθρο (background) διαφέρει αλλά και οι απόλυτες εντάσεις (παράδειγμα στα υγιή φάσματα που παραθέσαμε οι εντασεις διαφέρουν 2 και 3 φορές).

Για να απαλλαγούμε από αυτού του είδους τα προβλήματα **προ-επεξεργαζόμαστε** (**pre-processing**) τα δεδομένα μας για να μπορούμε μετά να περάσουμε στην τελική ανάλυση (21) (22) (23).

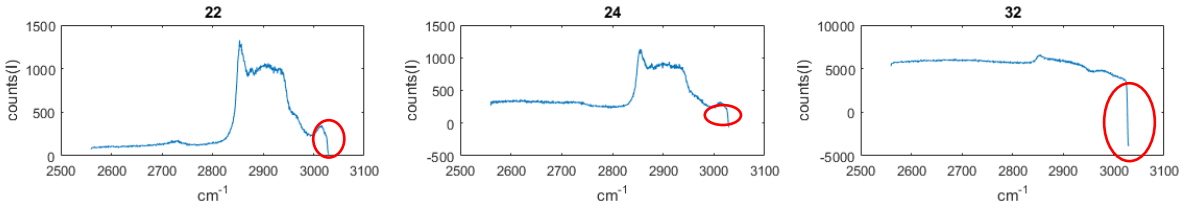
4. Στατιστική Ανάλυση

4.1 Pre-processing

Όλη η διαδικασία pre-processing έγινε στο Matlab.

4.1.1 Αφαίρεση των κακής ποιότητας φασμάτων

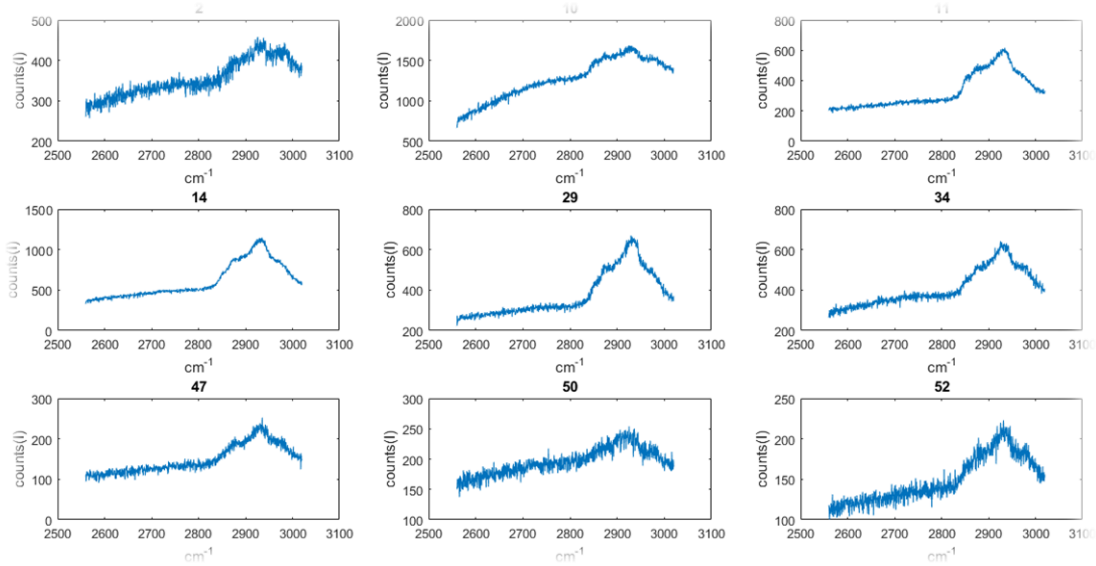
Ελέγξαμε τα φάσματα ένα ένα και αφαιρέσαμε κάποια λίγα τα οποία ήταν προβληματικά εξαιτίας της διαδικασίας της μέτρησης. Ο καρκινικός ιστός πιθανόν να περιέχει ίχνη από γάλα ή αίμα κτλ με αποτέλεσμα κάποιες μετρήσεις που κάναμε να μην ήταν σωστές. Έπειτα αφαιρεσαμε 4-5 τιμές από τις τελευταίες, ώστε να αποκόψουμε την απότομη κάθετη γραμμή που δημιουργείται λόγω κάποιου προσωρινού προβλήματος στη CCD.



(*Εικόνα 4.1.1.1*)

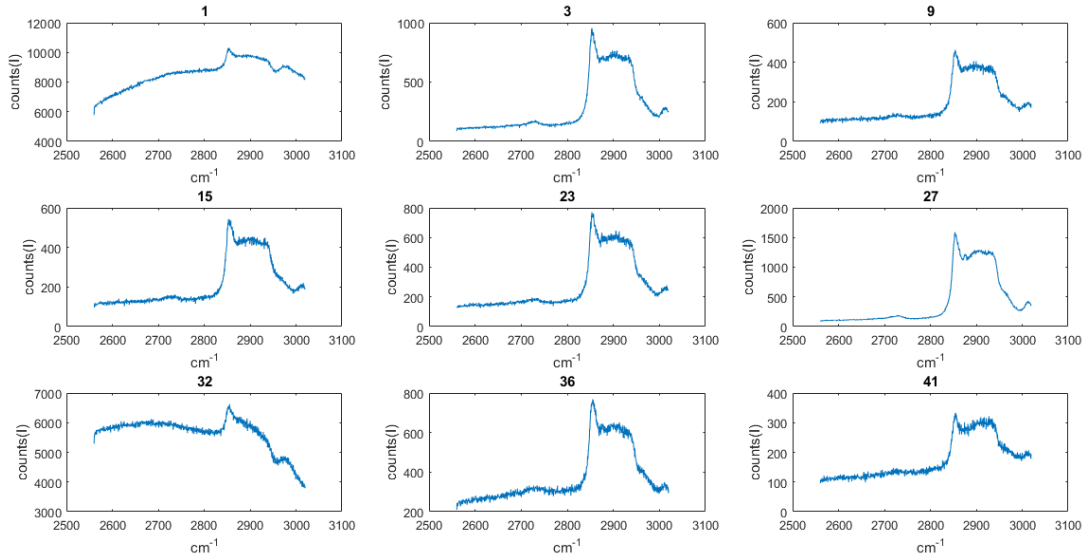
Τα αποτελέσματα μετά από αυτό το στάδιο επεξεργασίας μοιάζουν ως εξής:

Καρκινικοί ιστοί



(*Εικόνα 4.1.1.2*)

Υγιείς ιστοί



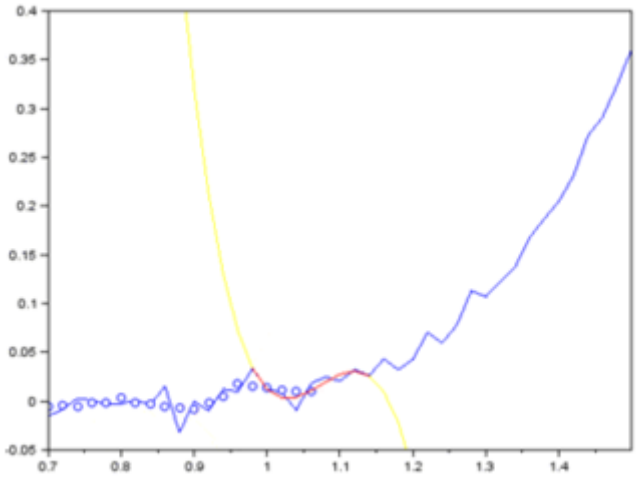
(*Εικόνα 4.1.1.3*)

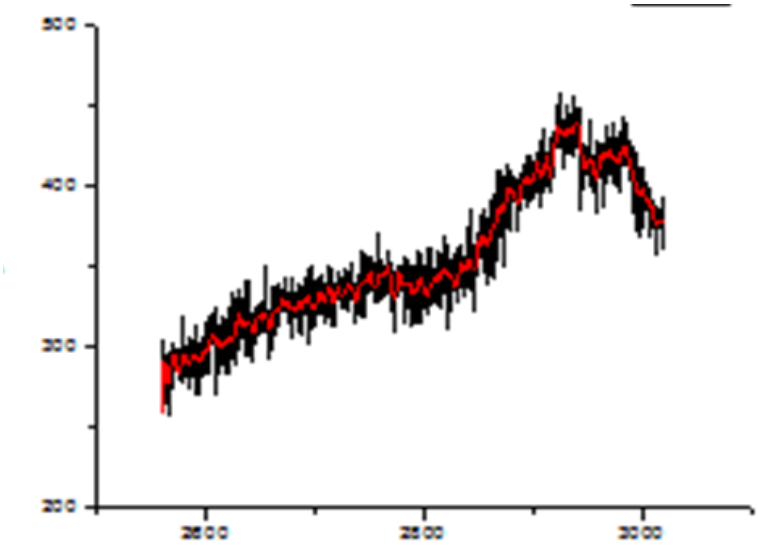
4.1.2 Μείωση Θορύβου (Noise reduction)

Τα φάσματα της Raman ειναι επιρρεπή στο θόρυβο και πολλές φορές απαιτείται μείωση του θορύβου για να ενισχυθεί η ποιότητα των φασμάτων. Εμείς μειώσαμε το θόρυβό μας με 2 τεχνικές. Η πρώτη μέθοδος που χρησιμοποιήσαμε είναι η Savitzky-Golay smoothing.

Αυτή η διαδικασία μπορεί υποβαθμίσει τα φασματικά χαρακτηριστικά, και γι αυτό συνίσταται προσεκτική χρήση του smoothing. Οι παράμετροι που χρησιμοποιήθηκαν επιλέχτηκαν με τέτοιο τρόπο ώστε να μην αλλοιωθούν όσο γίνεται τα χαρακτηριστικά του φάσματος.

Η μέθοδος αυτή προσαρμόζει ένα χαμηλού βαθμού πολυώνυμο στα δεδομένα μας, χρησιμοποιώντας γραμμικά ελάχιστα τετράγωνα. Όπως φαίνεται στο παρακάτω φάσμα η κόκκινη γραμμή είναι τα λεία (smoothed) δεδομένα και η μαύρη γραμμή αντιστοιχεί στα ανεπεξέργαστα δεδομένα. Η πρώτη προσέγγιση για τη βελτίωση της ποιότητας του σήματος Raman θα ήταν να αλλάξει κανείς τις ρυθμίσεις κατά τη διάρκεια της καταμέτρησης, χρησιμοποιώντας μεγάλους χρόνους (acquisition time) και υψηλότερη ισχύ λέιζερ. Εάν αυτές οι προσεγγίσεις δεν επαρκούν, τότε τα φάσματα πρέπει να επεξεργαστούν υπολογιστικά μετά την απόκτησή τους για τη βελτίωση του SNR. Η εξομάλυνση μιας συνάρτησης αφήνει την επιφάνεια κάτω από την συνάρτηση αμετάβλητη. Είναι αναπόφευκτο ότι το σήμα θα παραμορφωθεί κατά τη διαδικασία συνέλιξη. Προκύπτει λοιπόν όταν τα δεδομένα τα οποία έχουν μια κορυφή εξομαλύνονται, το ύψος της κορυφής θα μειωθεί και το half-width πλάτος θα αυξηθεί. Η πρώτη εικόνα παρακάτω δείχνει πώς λειτουργεί η Savitzky-Goley. Η μέθοδος sgolayfilt (Matlab) (24) προσαρμόζει ένα χαμηλού βαθμού πολυώνυμο στα δεδομένα μας, χρησιμοποιώντας ένα παράθυρο δεδομένου μεγεθους (εμείς διαλέξαμε 49) χρησιμοποιώντας γραμμικά ελάχιστα τετράγωνα .





(*Εικόνα 4.1.2.1 Κόκκινη γράμμη είναι η ομαλή γραμμή μέτα τη μέθοδο SG και η μαύρη τα raw data*)

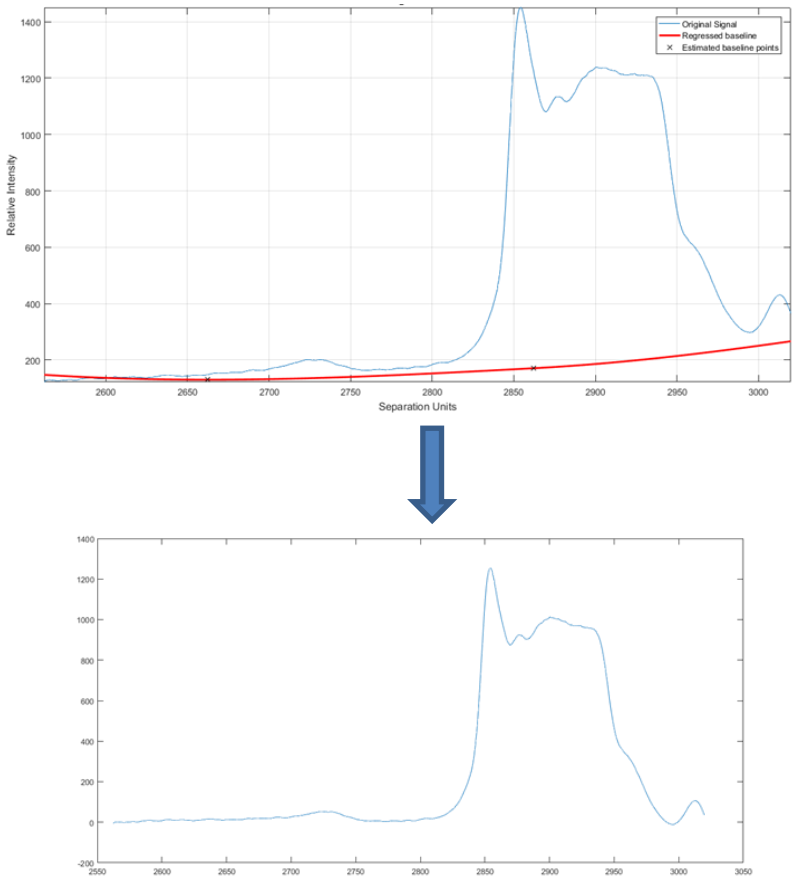
Οι παράμετροι που χρησιμοποιήσαμε είναι: πολυώνυμο βαθμού 5 και framelength 49.

Η δεύτερη και πολύ σημαντική τεχνική, η PCA, χρησιμοποιήθηκε για την ανακατασκευή φάσματων χρησιμοποιώντας την πιο ουσιώδη πληροφορία, διατηρώντας έτσι τα σημαντικά δεδομένα του φάσματος και αφαίροντας εμμέσως και το θορύβο του υποβάθρου. Θα αναφερθούμε και παρακάτω εκτενώς στην PCA.

4.1.3. Διόρθωση της γραμμής της βάσης (baseline)

Ο φθορισμός του δείγματος, καθώς και θερμικές διακυμάνσεις της CCD, μπορεί να επηρεάσουν σημαντικά τη φασματική γραμμή βάσης (baseline) και ως εκ τούτου η διόρθωση της γραμμής αυτής είναι απαραίτητη. Η πολυωνυμική προσαρμογή της βάσης (polynomial baseline fitting) μπορεί να εκτιμήσει το άγνωστο υπόβαθρο (background).

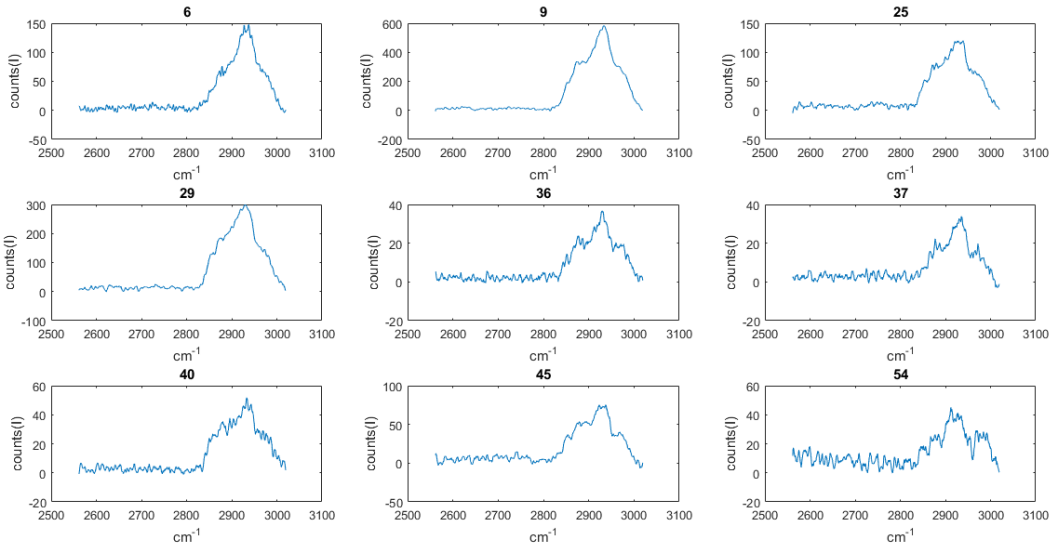
Για την αφαίρεση λοιπόν του υποβάθρου χρησιμοποιήσαμε μια συνάρτηση στο Matlab η οποία υπολογίζει κάποιες τιμές της baseline (baseline points) μέ κάποιες μεθόδους, εφαρμόζει παλινδρόμηση πάνω στις τιμές αυτές χρησιμοποιώντας μια προσέγγιση spline (spline approximation~μια συνάρτηση η οποία ορίζεται από πολλαπλές υπο-συναρτήσεις) και τέλος προσαρμόζει τη γραμμή βάσης (25).



(*Εικόνα 4.1.3.1*)

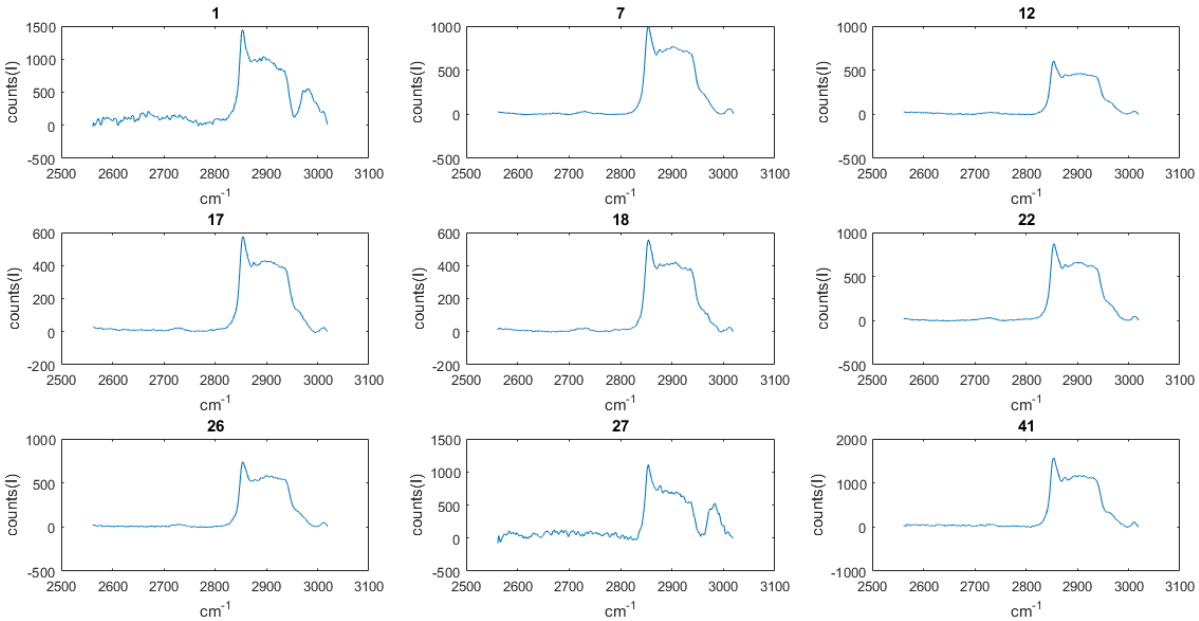
Μετά από την μείωση του θορύβου και την αφαίρεση της baseline παίρνουμε τα παρακάτω φάσματα.

<u>Καρκινικοί ιστοί</u>



(*Εικόνα 4.1.3.2*)

Υγιείς ιστοί



(*Εικόνα 4.1.3.3*)

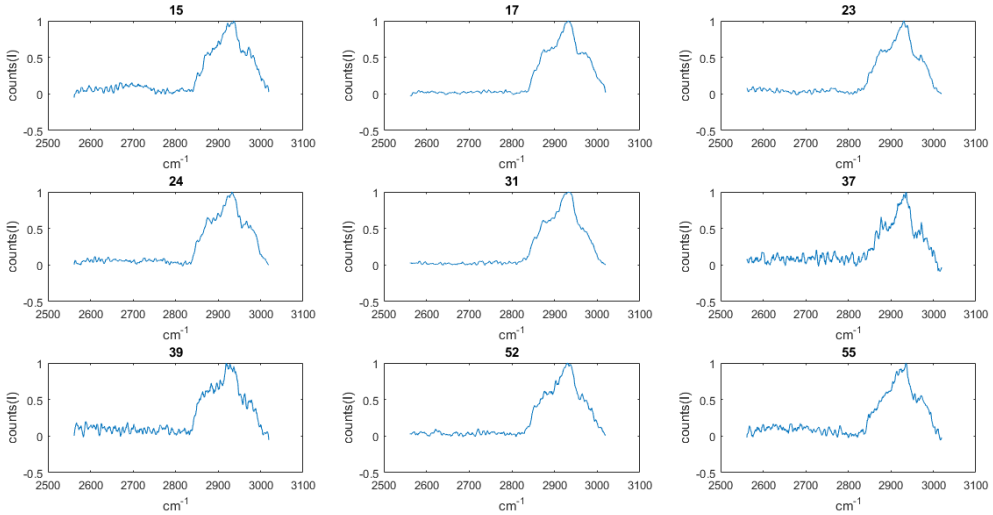
4.1.4 Κανονικοποίηση

Μετά τη διόρθωση της γραμμής βάσης τα φάσματα χρειάζονται κανονικοποίηση για να διορθωθεί η επίδραση από παράγοντες σχετικούς με το δείγμα ή το πείραμα όπως για παράδειγμα το πάχος, ή η πυκνότητα. Για το λόγο αυτό κανονικοποιήσαμε όλα τα φάσματά μας.

Υπάρχουν διάφοροι τρόποι κανονικοποίησης με συνηθέστερους το vector normalization, min-max normalization, Standard Normal Variate (SNV) Normalization, και Peak Normalization. Η κανονικοποίηση που κάναμε είναι μια πιο απλή εκδοχή του vector normalization. Σε κάθε φάσμα (που αντιστοιχεί σε ένα διάνυσμα) βρίκουμε τη μέγιστη τιμή (ενταση) και διαιρούμε κάθε τιμή που αντιστοιχεί σε μια φασματική συχνότητα με την μεγιστη τιμή. Το αποτέλεσμα είναι η μεγαλύτερη ένταση να έχει τιμή 1 και όλες οι υπόλοιπες να έχουν κανονικοποιηθεί με βάση αυτή, στο διάστημα 0-1. Με αυτόν τον τρόπο οι απόλυτες εντάσεις δεν υπάρχουν πια, ενώ οι σχετικές εντάσεις είναι αυτές που παίζουν ρόλο.

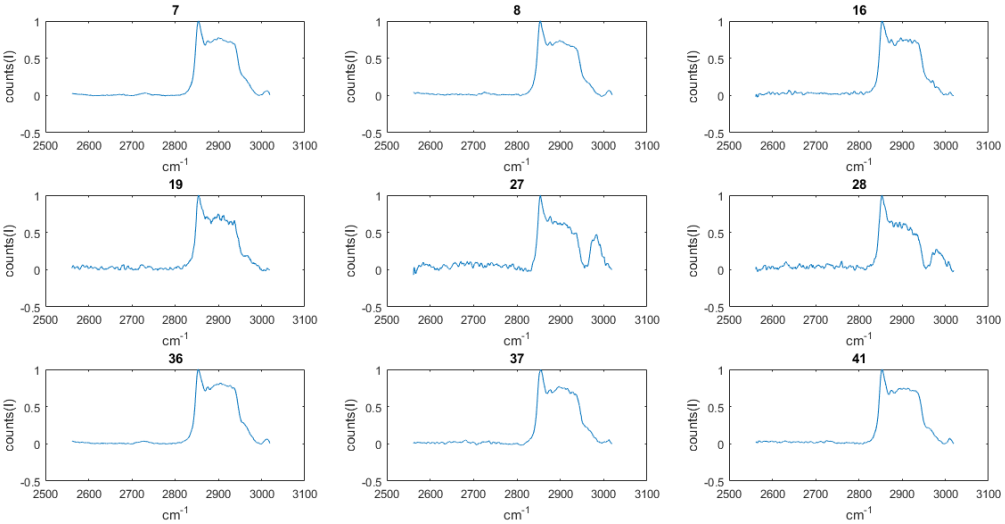
Μετά την κανονικοποίηση τα φάσματά μας μετασχηματίστηκαν ως εξής:

Καρκινικοί ιστοί:



(*Εικόνα 4.1.4.1*)

Υγιείς Ιστοί



(*Εικόνα 4.1.4.2*)

4.1.5 Μείωση δεδομένων- εξαγωγή χαρακτηριστικών.

Το στάδιο της Feature Extraction (FE - εξαγωγή χαρακτηριστικών) είναι υπεύθυνο για την παραγωγή ενός ***μικρότερου, νέου*** αριθμού μεταβλητών («χαρακτηριστικό» = «μεταβλητή εισόδου») που είναι πιο πληροφοριακές (Informative) από το αρχικό σύνολο των κυματάριθμών-μεταβλητών. Η FE γίνεται για να προετοιμάσει το σύνολο

των δεδομένων για ταξινόμηση (classification). Η αποτελεσματική FE μπορεί να ανακουφίσει το φορτίο του μετέπειτα μοντέλου του ταξινομητή. Στην αναγνώριση προτύπων, η «Κατάρα της διαστασιμότητας» (“curse of dimensionality”) αναφέρεται συχνά. Αυτή η «κατάρα» αναφέρεται στη δυσκολία να εκπαιδεύσεις έναν ταξινομητή σε χώρους μεγάλων διαστάσεων, όπου τα μοντέλα μπορούν εύκολα να γίνουν overfit ή να παραμείνουν undertrained. Αυτό μπορεί να είναι περισσότερο ή λιγότερο έντονο, ανάλογα με τον ταξινομητή. Η FE μπορεί να το αντιμετωπίσει αυτό αποτελεσματικά.

4.1.6 Στατιστική

Σε αυτό το σημείο κρίνεται απαραίτητη η γνώση ενός στοιχειώδους μαθηματικού υποβάθρου το οποίο θα μας βοηθήσει στην κατανόηση πολλών απ’τους αλγορίθμους που θα εξετάσουμε αναλυτικά στη συνέχεια.

Τυπική Απόκλιση:

Θεωρούμε ένα σύνολο δεδομένων *X* με στοιχεία *X1 ,X2 …Xn, n* ο αριθμός των στοιχείων του συνόλου.

Η τυπική απόκλιση ενός συνόλου δεδομένων είναι η μέτρηση του τρόπου με τον οποίο είναι κατανεμημένα τα δεδομένα στο σύνολο αυτό. Πιο συγκεκριμένα:

“ Τυπική απόκλιση ενός συνόλου δεδομένων είναι η μέση απόσταση του μέσου του συνόλου προς κάθε σημείο του συνόλου αυτού ”.

Ένας τρόπος υπολογισμού της τυπικής απόκλισης (για ένα δείγμα του πληθυσμού) είναι να υπολογίσουμε το άθροισμα των τετραγώνων των αποστάσεων κάθε σημείου του συνόλου με το μέσο αυτού, διαιρεμένο με το πλήθος n-1 και να βρούμε την τετραγωνική ρίζα του λόγου αυτού:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)}}$$

Διακύμανση (Variance):

Η διακύμανση είναι απλώς το τετράγωνο της τυπικής απόκλισης και αυτός μάλιστα είναι και ο συμβολισμός της .Και οι δύο είναι μονάδες μέτρησης του τρόπου κατανομής των δεδομένων με πιο συνηθισμένη την τυπική απόκλιση. Ο ορισμός όμως της διακύμανσης θα μας βοηθήσει να ορίσουμε και να κατανοήσουμε τη συμμεταβλητότητα .

$$var(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})}{(n-1)}$$

Συμμεταβλητότητα (Covariance):

Οι δύο προηγούμενες μονάδες μέτρησης ενεργούν σε μία μόνο διάσταση. Έτσι αν είχαμε ενα πολυδιάστατο σύνολο δεδομένων θα μπορούσαμε να υπολογίσουμε την τυπική απόκλιση μόνο για μία διάσταση ξεχωριστά. Είναι χρήσιμο παρ'όλα αυτά να μπορούμε να δούμε πως διαφοροποιούνται οι διαστάσεις απ'το μέσο, σε σχέση η μια απ'την άλλη. Ένα τέτοιο μέτρο είναι η συμμεταβλητότητα. Η συμμεταβλητότητα μετριέται πάντα μεταξύ δύο διαστάσεων. Αν υπολογίσουμε τη συμμεταβλητότητα μεταξύ μιας διάστασης και του εαυτού της θα πάρουμε τη διακύμανσή της. Ο τύπος της συμμεταβλητότητας είναι παρόμοιος με αυτόν της διακύμανσης:

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$

Αν η τιμή της συμμεταβλητότητας ανάμεσα σε δύο διαστάσεις είναι αρνητική τότε όταν η μία αυξάνεται η άλλη θα μειώνεται . Αν η συμμεταβλητότητα είναι ίση με 0 τότε αυτές οι δύο διαστάσεις είναι τελείως ανεξάρτητες μεταξύ τους.
Αν έχουμε πάνω από 2 διαστάσεις, έστω n, τότε υπολογίζουμε $\frac{n!}{(n-2)!*2}$ διαφορετικές τιμές συμμεταβλητότητας. Ένας πίνακας συνδιασποράς είναι ένας πίνακας του οποίου το στοιχείο στην i, j θέση είναι η συνδιακύμανση μεταξύ του i και του j στοιχείου ενός τυχαίου διάνυσματος.
Για παράδειγμα, η μεταβολή σε μια συλλογή από τυχαία σημεία στο διδιάστατο χώρο δεν μπορεί να χαρακτηριστεί πλήρως από έναν και μόνο αριθμό, ούτε οι διακυμάνσεις στις x και y κατευθύνσεις θα μπορούσαν να περιέχουν όλες τις απαραίτητες πληροφορίες. Στην περίπτωση αυτή μια θα ήταν αναγκαίος ένας 2x2 πίνακας για τον πλήρη χαρακτηρισμό της διδιάστατης περίπτωσης.

Ιδιοδιανύσματα και ιδιοτιμές:

Ένα ιδιοδιάνυσμα ενός γραμμικού μετασχηματισμού είναι ένα μη μηδενικό διάνυσμα που δεν αλλάζει την κατεύθυνσή του όταν αυτός ο γραμμικός μετασχηματισμός εφαρμόζεται σε αυτό. Πιο συγκεκριμένα αν T είναι ένας γραμμικός μετασχηματισμός από έναν διανυσματικό χώρο V πάνω σε ένα πεδίο F στον εαυτό του και ν είναι ένα διάνυσμα στο V που δεν είναι το μηδενικό διάνυσμα τότε το ν είναι ένας ιδιοδιάνυσμα του T αν T (ν) είναι ένα βαθμωτό πολλαπλάσιο του ν. Η συνθήκη είναι :

$$\boldsymbol{T(v) = \lambda v}$$

Όπου λ είναι ένα βαθμωτό μέγεθος που ονομάζεται ιδιοτιμή ή eigenvalue.

Αν ο διανυσματικός χώρος είναι πεπερασμένος, τότε ο γραμμικός μετασχηματισμός T απεικονίζεται ως ένα τετραγωνικός πίνακας A και το v είναι ένα διάνυσμα και η εξίσωση για τα ιδιοδιανύσματα γίνεται :

$$\mathbf{Av = \lambda v}$$

Ένα σημαντικό εργαλείο για τον υπολογισμό των ιδιοτιμών ενός τετραγωνικού πίνακα είναι το ***χαρακτηριστικό πολυώνυμο*** .
Αν θεωρήσουμε λ μία ιδιοτιμή ενός πίνακα A τότε το σύστημα των γραμμικών εξισώσεων $(A - \lambda I)v = 0$ (όπου I ο μοναδιαίος πίνακας) έχει μία μή μηδενική λύση v και άρα μπορούμε να πούμε ισοδύναμα ότι $\det(A - \lambda I) = 0$.
Έτσι οι ρίζες του πολυωνύμου $p(\lambda) = \det(A - \lambda I)$ είναι οι ιδιοτιμές του πίνακα A και το πολυώνυμο αυτό είναι το χαρακτηριστικό πολυώνυμο. Τα ιδιοδιανύσματα είναι πάντα ορθογώνια μεταξύ τους.

Ανάλυση πίνακα σε ιδιάζουσες τιμές. SVD:

Η ανάλυση σε ιδιάζουσες τιμές είναι μία παραγοντοποίηση ενός πίνακα με πραγματικά ή μιγαδικά στοιχεία. Η παραγοντοποίηση ενός πίνακα M, $m \times n$ είναι της μορφής

$$USV^*$$

Όπου ο U είναι ένας m x m πραγματικός ή μιγαδικός ορθομοναδιαίος πίνακας, S ένας $m \times n$ ορθογώνιος διαγώνιος πίνακας με μη αρνητικές τιμές στην διαγώνιο και V^* (ο συζυγής ανάστροφος του V, ή απλά ο ανάστροφος του V αν ο V είναι πραγματικός) ένας $n \times n$ πραγματικός ή μιγαδικός ορθομοναδιαίος πίνακας. Τα διαγώνια $S_{i,j}$ του S είναι γνωστά ως ιδιάζουσες τιμές του M.

Καθώς οι πίνακες U και V^* είναι ορθομοναδιαίοι, οι στήλες του καθενός σχηματιζουν ένα σύνολο ορθοκανονικών διανυσμάτων, που μπορούν να θεωρηθούν ως διανύσματα βάσης.

4.1.7 PCA

Η τεχνική που χρησιμοποιήσαμε γι' αυτό το σκοπό είναι η **PCA**. Η PCA αποκαλύπτει τα πιο σημαντικά μοτίβα διακύμανσης στα δεδομένα, χωρίς να έχει σημασία σε ποια κλάση ανήκουν (στην περίπτωσή μας καρκινικό-υγιές). Η PCA είναι μια στατιστική διαδικασία που χρησιμοποιεί ένα ορθογώνιο μετασχηματισμό για να μετατρέψει μια σειρά από παρατηρήσεις *πιθανώς συσχετιζόμενων μεταβλητών* σε ένα σύνολο τιμών από *γραμμικά ασυσχέτιστες* μεταβλητές που ονομάζονται κύριες συνιστώσες

(principal components –PC). Αυτός ο μετασχηματισμός ορίζεται κατά τέτοιο τρόπο ώστε η πρώτη κύρια συνιστώσα (PC1) να έχει τη μεγαλύτερη δυνατή διακύμανση (δηλαδή, να αντιπροσωπεύει όσο το δυνατόν περισσότερο τον διασκορπισμό-διασπορά (dispersity) των δεδομένων και κάθε επόμενη συνιστώσα με τη σειρά της έχει την μέγιστη δυνατή διακύμανση υπό τον περιορισμό να είναι ορθογώνια προς τις προηγούμενες συνιστώσες. Τα διανύσματα που προκύπτουν αποτελούν ένα ασυσχέτιστο ορθογώνιο σύνολο βάσης.

Με άλλα λόγια η PCA είναι ένας ορθογώνιος γραμμικός μετασχηματισμός που μετασχηματίζει τα δεδομένα σε ένα νέο σύστημα συντεταγμένων, τέτοιο ώστε η μεγαλύτερη διακύμανση από κάποια προβολή των δεδομένων να βρίσκεται πάνω στην πρώτη συντεταγμένη (που ονομάζεται πρώτη κύρια συνιστώσα), η δεύτερη μεγαλύτερη διακύμανση πάνω στη δεύτερη συντεταγμένη κ.ο.κ.

<u>Σύντομη μαθηματική περιγραφή της PCA</u>:

Παρακάτω παραθέτουμε πολύ συνοπτικά τον πιο συνηθισμένο τρόπο για να κάνει κανείς PCA ανάλυση. Για παραπάνω πληροφορίες μπορεί κανείς να ανατρέξει στη βιβλιογραφία (26) Μπορούμε με 2 τρόπους να κάνουμε PCA:

<u>Μέσω covariance matrix</u>: Έστω ότι έχουμε τον πίνακα με τα δεδομένα μας *X* μεγέθους *n x p*, όπου n είναι ο αριθμός των δειγμάτων (στην περίπτωσή μας 146) και *p* είναι ο αριθμός των μεταβλητων (στην περίπτωσή μας 1000 στήλες, που αντιστοιχούν στις μετρούμενες συχνότητες). Ας υποθέσουμε επίσης ότι ο πίνακας είναι κεντραρισμένος, δηλαδή ο μέσος της κάθε στήλη έχει αφαιρεθεί και αντιστοιχεί τώρα στο 0.

Ο $\boldsymbol{p\ x\ p}$ πίνακας συμμεταβλητότητας (covariance matrix) **C** δίνεται από $\boldsymbol{C = X^T X/(n-1)}$. Είναι συμμετρικός πίνακας και γι' αυτό μπορεί να διαγωνοποιηθεί ως εξής:

$$\boldsymbol{C = VLV^T,}$$

Όπου V είναι ένας πίνακας ιδιοδιανυσμάτων (κάθε στήλη είναι ένα ιδιοδιάνυσμα) και L είναι ένα διαγώνιος πίνακας με ιδιοτιμές λ_i σε φθίνουσα σειρά στη διαγώνιο. Τα ιδιοδιανύσματα λέγονται κύριοι άξονες ή κύριες διευθύνσεις των δεδομένων. Οι προβολές των δεδομένων πάνω στους κύριους άξονες ονομάζονται κύριες συνιστώσες (**principal components, PC's**). Αυτές τις συνιστώσες μπορούμε να τις δούμε σαν νέες, μετασχηματισμένες μεταβλητές. Η j-οστή κύρια συνιστώσα δίνεται από την j-οστή στήλη του XV. Οι συντεταγμένες του i-οστού σημείου δεδομένων στον νέο PC χώρο δίνονται από την i-οστή γραμμή του **XV**.

<u>Μέσω ανάλυσης σε Ιδιάζουσες Τιμές (SVD) στον **X**:</u> Όπου παίρνουμε:

$$\boldsymbol{X = USV^T}$$

Όπου S είναι ο διαγώνιος πίνακας των ιδιαζουσών τιμών S_i. Από εδώ μπορεί κανείς να δει ότι:

$$\boldsymbol{C = \frac{VSU^T USV^T}{n-1} = V\frac{S^2}{n-1}V^T}$$

Που σημαίνει ότι τα δεξιά ιδιάζοντα διανύσματα V είναι οι κύριες διευθύνσεις και ότι οι ιδιάζουσες τιμές σχετίζονται με τις ιδιοτιμές του πινακα συμμεταβλητότητας μέσω της σχέσης $\boldsymbol{\lambda_i = \frac{S_i^2}{n-1}}$. Οι κύριες συνιστώσες δίνονται από

$$\boldsymbol{XV = USV^T = US}$$

Συνοψίζοντας (27):

1. Αν $X = USV^T$, τότε οι στήλες του V είναι οι κύριοι άξονες/ διευθύνσεις (principal directions/axes).
2. Οι στήλες του US είναι οι κύριες συνιστώσες (principal components).
3. Οι ιδιάζουσες τιμές σχετίζονται με τις ιδιοτιμές του πίνακα συμμεταβλητότητας μέσω της σχέσης $\boldsymbol{\lambda_i = \frac{S_i^2}{n-1}}$. Οι ιδιοτιμές $\boldsymbol{\lambda_i}$ δείχνουν τις διακυμάνσεις των αντίστοιχων PC's.
4. Για να μειώσουμε τις διαστασιμότητα των δεδομένων από p σε k, με k < p, διαλέγουμε k πρώτες στήλες του U και το *k x k* πάνω αριστερά μέρος του S. Το γινόμενό τους $U_k S_k$ είναι ο απαιτούμενος *n x k* πίνακας που περιέχει τις k πρώτες PC's.
5. Αν πολλαπλασιάσουμε τις k πρώτες PC's με τους αντίστοιχους κύριους άξονες V_k^T παίρνουμε τον πίνακα $X_k = U_k S_k V_k^T$ που έχει το αρχικό μέγεθος *n x p* αλλά είναι μικρότερης τάξης (τάξης k). Αυτός ο πίνακας X_k παρέχει μια ανακατασκευή των αρχικών δεδομένων από τις k πρώτες PC's.

4.2 Μηχανική μάθηση. Ταξινόμηση

4.2.1 Μηχανική μάθηση

Η **μηχανική μάθηση** διερευνά τη μελέτη και την κατασκευή αλγοριθμων που μπορούν να μάθαίνουν από τα δεδομένα και να κάνουν προβλέψεις με καινούργια δεδομένα (2). Τέτοιοι αλγόριθμοι λειτουργούν κατασκευάζοντας μοντέλα από πειραματικά δεδομένα που στη συνέχεια παίρνουν αποφάσεις για νέα δεδομένα. Ο Arthur Samuel, πρωτοπόρος στην τεχνητή νοημοσύνη, ορίζει τη μηχανική μάθηση ως «Ο κλάδος της επιστήμης υπολογιστών που δίνει στους υπολογιστές την ικανότητα

να μάθαίνουν, χωρίς να έχουν ρητά προγραμματιστεί γι' αυτό». Σύμφωνα με τον Mitchell, «ένα πρόγραμμα υπολογιστή θεωρείται ότι μαθαίνει από εμπειρία , σε σχέση με κάποια κατηγορία εργασιών και μετρική αποτίμησης , εάν η απόδοση στις εργασίες του βελτιώνεται με την εμπειρία (28)»

Η μηχανική μάθηση μπορεί να διακριθεί στην επιβλεπόμενη μάθηση (**supervised learning**) και στη μάθηση χωρίς επίβλεψη (**unsupervised learning**). Η διαδικασία της μάθησης μοιάζει πολύ με τη διαδικασία της μάθησης στον άνθρωπο.

Ένα σύστημα επιβλεπόμενης μάθησης, εκπαιδεύεται αρχικά σε ένα σύνολο δεδομένων εκπαίδευσης (**train set**) όπου κάθε παράδειγμα χαρακτηρίζεται από μια κατηγορία (**class**). Στη συνέχεια το μοντέλο αυτό προσπαθεί να κάνει πρόβλεψη σε ένα νέο σύνολο δεδομένων (**test set**) σύμφωνα με αυτά που έχει μάθει. Τυπικό παράδειγμα επιβλεπόμενης μάθησης αποτελούν τα προβλήματα ταξινόμησης (**classification**).

Σε ένα πρόβλημα ταξινόμησης, κάθε στοιχείο του συνόλου εκπαίδευσης αντιστοιχεί σε ένα διάνυσμα. Ένα τέτοιο διάνυσμα είναι ένα σύνολο τιμών, το οποίο επιπλέον περιέχει και μια τιμή κατηγορίας ή κλάσης (class), η οποία περιγράφει σε ποιά κατηγορία ανήκει. Πληθώρα αλγορίθμων μηχανικής μάθησης είναι σχεδιασμένοι για προβλήματα ταξινόμησης, όπως είναι οι αλγόριθμοι SVD, k-NN, Naïve Bayes, decision trees κ.α. Το εκπαιδευμένο μοντέλο που προκύπτει από την εφαρμογή ενός αλγορίθμου ταξινόμησης σε ένα σύνολο δεδομένων καλείται ταξινομητής (**classifier**).

Στη μάθηση χωρίς επίβλεψη, δεν υπάρχει προκαθορισμένο σύνολο τιμών. Τα παραδείγματα εκπαίδευσης χωρίζονται σε, άγνωστες εκ των προτέρων, ομάδες με βάση τα χαρακτηριστικά τους, μια διαδικασία που συχνά αναφέρεται σαν κατηγοροι οποίηση (clustering). Στην εργασία μας ασχοληθήκαμε με classification και όχι clustering.

4.2.2 Ταξινόμιση

Ένας ταξινομητής (classifier) είναι ένα μαθηματικό μοντέλο που υπολογίζει τις κλάσεις (classes) ενός unclassed συνόλου δεδομένων, με βάση τη γνώση που έχει αποκτηθεί προηγουμένως από ένα σύνολο δεδομένων για εκπαίδευση (training data set). Ένας ταξινομητής έχει, ως εκ τούτου, μια φάση εκπαίδευσης (training phase) και μια φάση χρήσης (use phase). Ο όρος «Ταξινόμηση» (classification) αναφέρεται στη φάση χρήσης. Στην ανάλυση που παραθέτω χώρισα τα δεδομένα σε train test και test set (T-T). Αυτό σημαίνει ότι ο αλγόριθμος εκπαιδεύει ένα train set και δοκιμάζει να προβλέψει ένα Test set. Παρόμοια αποτελέσματα έδωσε και ο διαχωρισμός T-V-T με την αλλαγή κάποιων παραμέτρων. Σε αυτό το διαχωρισμό του data set εκπαιδεύουμε με το train set, και υπάρχει και ένας επιπλέον διαχωρισμός, το Validation set. Το validation test το χρησιμοποιούμε για να αλλάξουμε παραμέτρους στους αλγορίθμους

μας ώστε μετά όταν εφαρμόσουμε τον αλγόριθμο στο test set να ελέγξουμε ποιες αλλαγές ή αλγόριθμοι δώσαν τα καλύτερα αποτελέσματα.

Υπάρχει μια ποικιλία από ταξινομητές που διαφέρουν ως προς την προβλεπτική ικανότητα, την υπολογιστική τους δύναμη, την ευκολία της δημιουργίας του ταξινομητή κ.α. Το πρόγραμμα που χρησιμοποιήσαμε είναι το **WEKA** του πανεπιστημίου του Waikato (29). Στο πρόγραμμα αυτό επιλέγουμε από μια λίστα όποιον ταξινομητή θέλουμε και τον εκπαιδεύουμε με τα δεδομένα μας. Κάποιοι από τους ταξινομητές που χρησιμοποιήσαμε είναι ο SVM (Support Vector Machine), k-NN (k-Nearest Neighbors), και ο Naive Bayes.

Τέλος θα πρέπει να αναφέρουμε ότι ο τρόπος που γίνεται η εκπαίδευση είναι να χωρίζεται κατά κάποιον τρόπο το Train set σε 2 ποσοστά, ένα μέρος πάνω στο οποίο εκπαιδευεται ο αλγόριθμός και ένα στο οποίο τεστάρεται. Ύστερα ένα επόμενο μέρος χρησιμοποιείται ως test set κ.ο.κ μέχρι όλα τα στοιχεία να εχουν χρησιμοποιηθεί και για train και για test. Ο πιο αποτελεσματικός τρόπος είναι η λεγόμενη leave one out μέθοδος, δηλαδή εκπαιδεύουμε τον αλγόριθμο σε όλα τα δεδομένα μας εκτός από ένα το οποίο το χρησιμοποιούμε για test. Επείτα κάνουμε το ίδιο σε ένα ένα όλα τα data points. Ο διαχωρισμός αυτός μας δινεί τα καλύτερα αποτελέσματα και αυτόν χρησιμοποιήσαμε κι εμεις.

Overfitting :

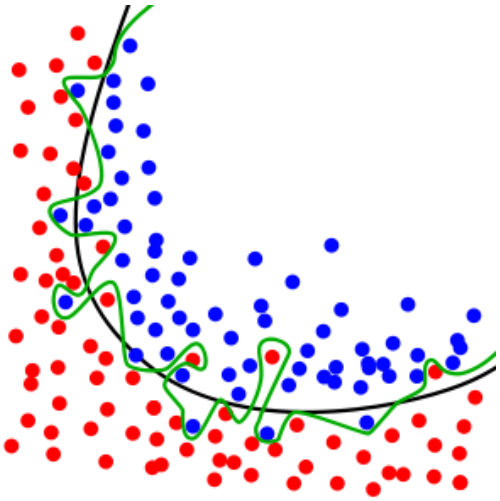
Ένας όρος που θα πρέπει να εισάγουμε σε αυτό το σημείο είναι το Overfitting (υπερπροσαρμογή). Στο overfitting, ένα στατιστικό μοντέλο περιγράφει έναν τυχαίο θόρυβο αντί για την υποκρυπτόμενη σχέση. Αυτό συμβαίνει όταν ένα μοντέλο είναι πολύ περίπλοκο, όπως για παράδειγμα να έχει πολλές παραμέτρους σε σχέση με τον αριθμό των παρατηρήσεων. Ένα τέτοιο μοντέλο έχει προφανώς κακή προβλεπτική ικανότητα μιας και αντιδρά υπερβολικά σε μικρές διακυμάνσεις των δεδομένων. Ενώ δηλαδή το σφάλμα στο Training set μας είναι μικρότερο, γίνεται μεγάλο στο test set γιατί δεν έχει μάθει να διαχωρίζει τις κλάσεις σωστά. Με λίγα λόγια, το overfitting συμβαίνει όταν ένα μοντέλο αρχίζει να «απομνημονεύει» αντί να «μαθαίνει» να γενικεύει.

Μια αναλογία για να το καταλάβει αυτό κανείς, είναι να σκεφτεί ένα μωρό που θέλουμε να του μάθουμε να ξεχωρίζει τι είναι κούπα. Θα του δείξουμε πολλές κούπες και θα ανακαλύψει ότι είναι όλες γυάλινες, με χερούλι και μπορείς να πιεις από εκεί. Αν του δείχνουμε τις ίδιες κούπες συνέχεια μπορεί λανθασμένα να υποθέσει ότι οι κούπες είναι μόνο πράσινες ή ότι όλα τα πράσινα είναι κούπες. Αυτό θα ήταν overfitting.

Έστω ότι θέλουμε να έχουμε τον πιο καθαρό ήχο και πιστό κι έτσι αγοράζουμε ένα υπερευαίσθητο μικρόφωνο ώστε να τα καταγράψουμε όλα. Εκτός όμως από τη μουσική, υπάρχουν και άλλοι ήχοι. Υπάρχουν οι θεατές που κινούνται στις θέσεις τους, κάποιος που βήχει, οι μουσικοί που γυρνανε τις σελίδες τους κτλ. Το να προσαρμόσεις λοιπόν ένα μοντέλο (fitting) είναι το να ξεχωρίσεις μόνο τη μουσική

στο κονσέρτο. Το overfitting είναι το να ακούς τα πάντα, ακόμα και το θόρυβο και αφήνεις αυτό να πνίγει τη μουσική και να πιστεύεις ότι αυτό είναι ένα κονσέρτο.

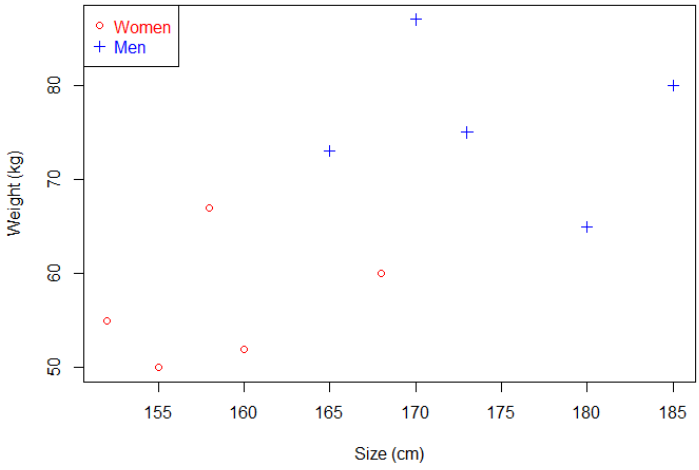
Για να το δείξουμε και σχηματικά, παρακάτω βλέπουμε ένα σύνολο δεδομένων τα οποία θέλουμε να διαχωρίσουμε. Η μαύρη γραμμή είναι μια καλή προσαρμογή, ένα καλό μοντέλο, παρότι κάποια λίγα σημεία κατατάσσονται λανθασμένα. Η πράσινη γραμμή από την άλλη, αν και δεν έχει σφάλμα και έχει χωρίσει όλα τα σημεία σωστά, εξαρτάται πάρα πολύ από αυτά και γι' αυτό αν δώσουμε ένα καινούργιο σημείο δε θα μπορέσει εύκολα να το κατατάξει σωστά.



(*Εικόνα 4.2.2.1*)

4.2.2.1 SVM (Support Vector Machine)

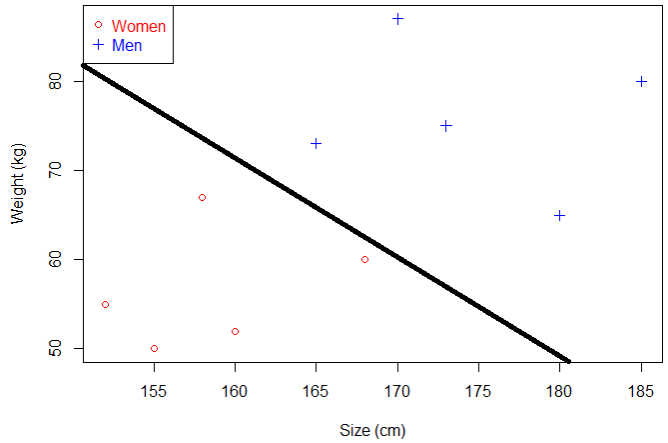
Έστω ότι έχουμε τα παρακάτω Training data:



(*Εικόνα 4.2.2.1.1*)

Έχουμε πλοτάρει το ύψος και το βάρος διαφόρων ανθρώπων και θέλουμε να δούμε αν μπορούμε να διαχωρίσουμε τα δεδομένα αυτά σε δυο κλάσεις (άντρες και γυναίκες). Ρωτάμε λοιπόν, για ένα συγκεκριμένο data point είναι αυτός ο άνθρωπος άντρας ή γυναίκα;

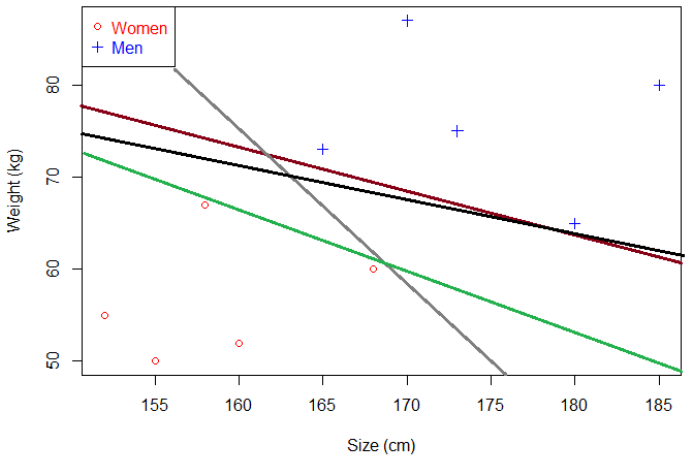
Κοιτάζοντας το διάγραμμα μπορούμε να δούμε αν είναι δυνατόν να διαχωρίσουμε τα δεδομένα μας. Για παράδειγμα θα μπορούσαμε να βρούμε μια γραμμή και όλα τα σημεία που ανιπροσωπεύουν τους άντρες να είναι πάνω από τη γραμμή και όσα αντιστοιχούν σε γυναίκες να είναι κάτω από τη γραμμή. Μια τέτοια γραμμή ονομάζεται υπερεπίπεδο διαχωρισμού (separating hyperplane):



(*Εικόνα 4.2.2.1.2*)

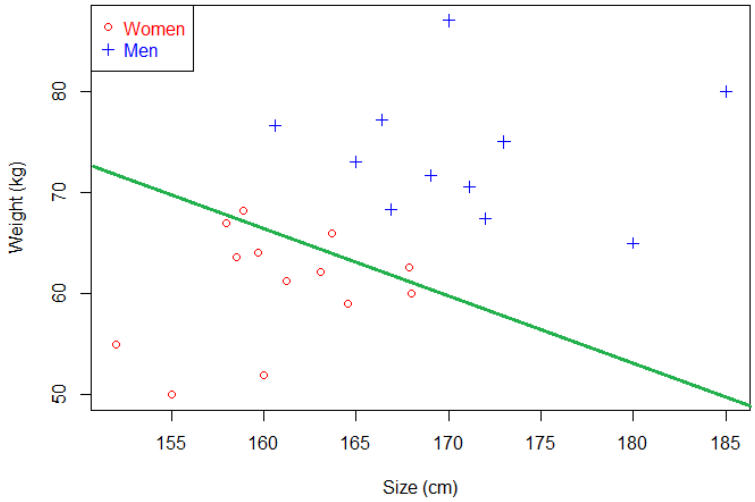
Όταν τα δεδομένα μας απεικονίζονται σε 2 διαστάσεις το υπερεπίπεδο είναι μια ευθεία.

Το γεγονός ότι μπορούμε να βρούμε ένα επίπεδο που να διαχωρίζει τα δεδομένα μας δεν σημαίνει ότι είναι και το βέλτιστο! Στο παρακάτω παράδειγμα φαίνονται διάφορα υπερεπίπεδα που διαχωρίζουν και κάθε ένα από αυτά είναι έγκυρο, καθώς διαχωρίζει με επιτυχία τους άντρες από τις γυναίκες.



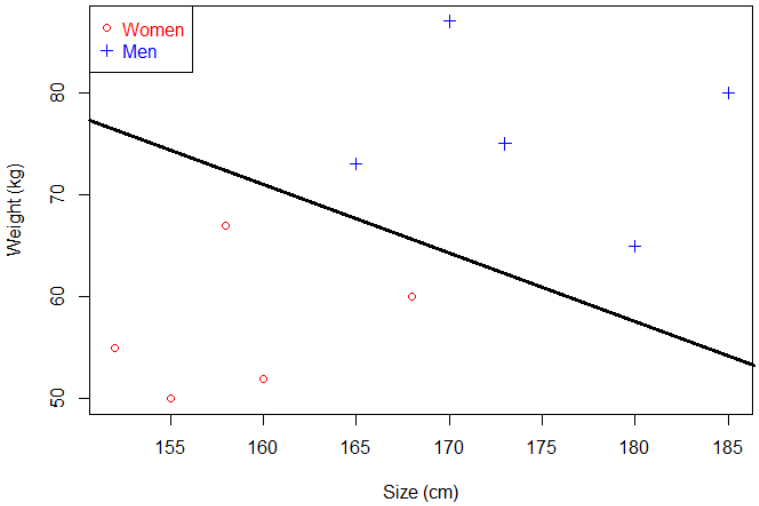
(*Εικόνα 4.2.2.1.3*)

Έστω ότι επιλέγουμε το πράσινο υπερεπίπεδο για να κατηγοριοποιήσουμε τα πραγματικά δεδομένα μας.



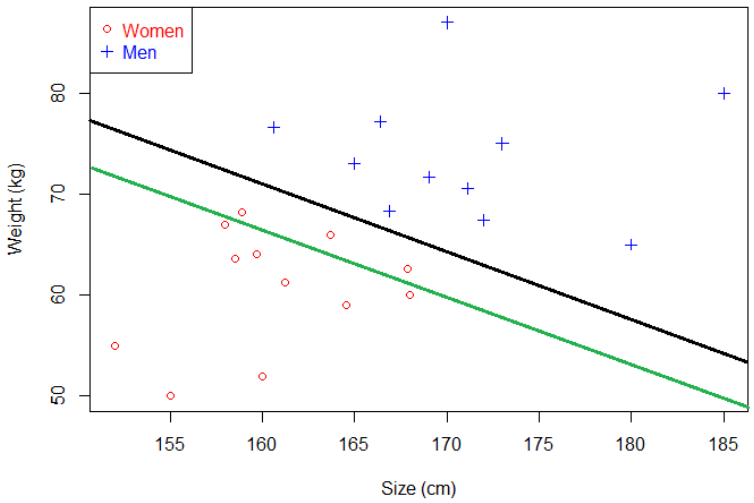
(*Εικόνα 4.2.2.1.4*)

Βλέπουμε ότι ο ταξινομητής (το υπερεπίπεδο δηλαδή) κατηγοριοποεί λανθασμένα 3 γυναίκες. Διαισθητικά, λοιπόν, καταλαβαίνουμε ότι αν διαλέξουμε ένα υπερεπίπεδο που είναι κοντά στα data points μιας από τις δύο κλάσεις, ίσως να μην γενικεύει σωστά. Γι΄αυτό το λόγο η μέθοδος SVD διαλέγει ένα υπερεπίπεδο που να είναι όσο πιο μακρυά γινεται από τα data points και των δύο κατηγοριών.



(*Εικόνα 4.2.2.1.5*)

Αν συγκρίνουμε τα δύο υπερεπίπεδα βλέπουμε ότι αυτό το υπερεπίπεδο είναι πολύ καλύτερος ταξινομητής από το πράσινο:

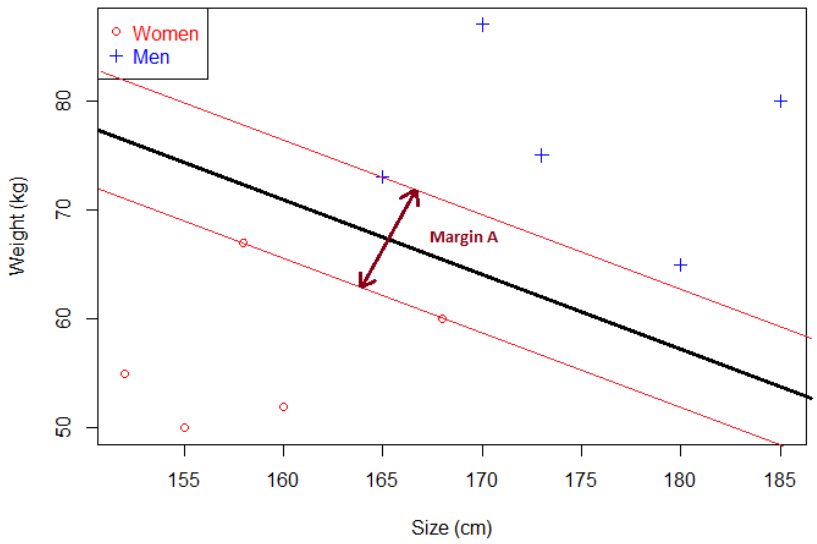


(*Εικόνα 4.2.2.1.6*)

Γι' αυτό το λόγο ο στόχος του SVM είναι να βρει το βέλτιστο (optimal) υπερεπίπεδο που διαχωρίζει τα δεδομένα μας:

- Επειδή ταξινομεί σωστά τα training data
- Και επειδή ταξινομεί καλύτερα τα δεδομένα που δεν έχει ξαναδεί (test data)

Με δεδομένο ένα συγκεκριμένο υπερεπίπεδο, υπολογίζουμε την απόσταση μεταξύ αυτού και των κοντινότερων data points. Αν αυτή την τιμή τη διπλασιάσουμε τότε παίρνουμε αυτό που αποκαλούμε περιθώριο (margin).



(*Εικόνα 4.2.2.1.7*)

Παρατηρούμε ότι όταν ένα υπερεπίπεδο είναι πολύ κοντά σε ένα data point, το περιθώριο θα είναι μικρό. Όσο πιο μακρυά είναι ένα υπερεπίπεδο από ένα data point, τόσο πιο μεγάλο είναι το περιθώριο. Μέσα στο περιθώριο δε θα υπάρχει ποτέ data point.

Όλα αυτά σημαίνουν ότι το βέλτιστο υπερεπίπεδο θα είναι αυτό με το μεγαλύτερο περιθώριο. Γι' αυτό ο SVM προσπαθεί να βρει ένα βέλτιστο υπερεπίπεδο που να μεγιστοποιεί το περιθώριο (margin) των training data.

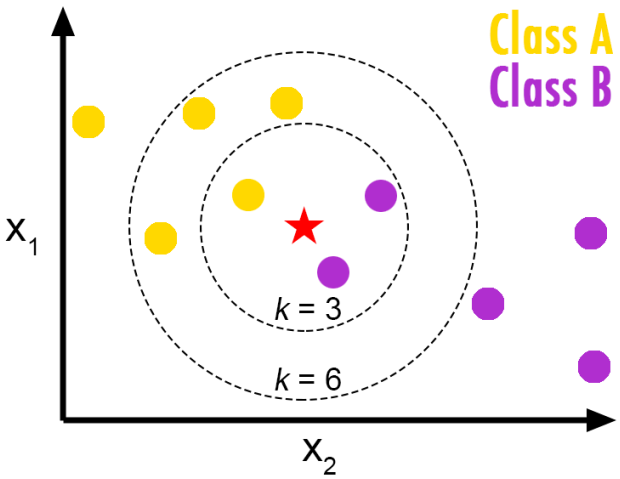
Για περισσότερες λεπτομέρειες και μαθηματικό φορμαλισμό παρατίθεται βιβλιογραφία (30).

4.2.2.2 k-NN (k-Nearest Neighbors)

Ο k-NN αλγόριθμος είναι πολύ απλός στο τρόπο που λειτουργεί και συγκεκριμένα ο 1-N-N ταξινομητής είναι από τις πιο παλιές μεθόδους. Η ιδέα είναι η εξής: για να ταξινομηθεί ο X βρίσκουμε τον κοντινότερο γείτονα (έστω X') ανάμεσα στα σημεία –δεδομένα μας(που είναι τα Training points) και ταξινομούμε το X στην ίδια κλαση με το X'.

Αν τώρα βάλουμε στον k τιμή μεγαλύτερη του 1, τότε χρησιμοποιώντας πάλι μια μετρική (πχ ευκλείδια) βρίσκουμε τους k κοντινότερους γείτονες του X και ταξονομούμε το X ανάλογα με την κλάση της πλειοψηφίας των γειτόνων.

Παράδειγμα:



(*Εικόνα 4.2.2.2.1*)

Το κόκκινο αστέρι θα ανήκε στην κλάση B αν το k μας ήταν 3 και στην κλάση A αν ήταν 6. Για να αποφύγουμε τον κίνδυνο ισοψηφίας χρησιμοποιούμε περιττό αριθμό k. Δεν υπάρχει k που να είναι απαραίτητα καλύτερο από άλλο, δοκιμάζουμε και

βλέπουμε ποιο μας δίνει καλύτερα αποτελέσματα στα δεδομένα μας. Για επιπλέον μαθηματικές αποδείξεις για το γιατί λειτουργεί ο αλγόριθμος παραπέμπουμε στη σχετική βιβλιογραφία (31) (32).

4.2.2.3 Naive Bayes

Αρχικά, πρέπει να ορίσουμε κάποιες έννοιες που σχετίζονται με την Naive Bayes.

Δεσμευμένη πιθανότητα (Conditional Probability): είναι η πιθανότητανα συμβεί κάτι, δεδομένου ότι κάτι άλλο έχει ήδη συμβεί.

Πάράδειγμα, έστω ότι έχουμε ένα σύνολο από γερουσιασές των ΗΠΑ. Μπορεί να είναι δημοκρατικοί ή Ρεπουμπλικάνοι. Επίσης είναι άντρες ή γυναίκες. Διαλέγουμε έναν γερουσιαστή στην τύχη. Ποια είναι η πιθανότητα να είναι γυναίκα δημοκρατική;

Η δεσμευμένη πιθανότητα μας βοηθάει σε αυτόν τον υπολογισμό.

P(Δημοκράτης & Γυναίκα) = P(Δημοκράτης) * P(Γυναίκα | Δημοκράτης) ==
P(Γυναίκα) * P(Δημοκράτης | Γυναίκα)

Νόμος του Bayes:

Είναι ένας τρόπος να πάμε από την πιθανότητα P(Στοιχεία| Γνωστό αποτέλεσμα) στην P(Αποτέλεσμα | Γνωστά στοιχεία).

Ή αλλιώς το πιο κλασσικό παράδειγμα για να κατανοήσουμε το νόμο του Bayes είναι:

Η πιθανότητα να έχει ασθένεια D δεδομένου ότι το τεστ είναι θετικό =

P (Test να είναι θετικό | D) * P (D)

(παράγοντας scaling) P (Test θετικό,με ή χωρίς ασθένεια)

Αν τώρα έχουμε πολλαπλά στοιχεία και όχι μόνο ένα τότε διαχειριζόμαστε το κάθε στοιχείο ως ανεξάρτητο. Αυτή είναι η προσέγγιση Naive Bayes. Δηλαδή:

P (Αποτέλεσμα|πολλαπλά στοιχεία) =
P(Στοιχείο1|Αποτέλεσμα) * P(Στοιχείο2|Αποτέλεσμα) * ... P(ΣτοιχείοΝ|Αποτέλεσμα) * P(Αποτέλεσμα)
(παράγοντας scaling) P(Πολλαπλά στοιχεία)
Ή αλλιώς:

$$P(\text{αποτέλεσμα} \mid \text{στοιχείων}) = \frac{P\ (\text{Πιθανότητα στοιχείων}) * P(\text{προγενέστερη πιθανότητα αποτελέσματος})}{P\ (\text{Στοιχείων})}$$

Μιας και διαιρούμε τα πάντα με το $P\ (Στοιχείων)$ δε χρειάζεται να το υπολογίσουμε. Αν τώρα προσπαθήσουμε να ταξινομήσουμε τα αποτελέσματά μας, κάθε αποτέλεσμα το αντιστοιχούμε σε κλάση και έχει ένα label (όπως είπαμε και νωρίτερα) 0 ή 1. Και πάλι η προσέγγισή μας είναι πολύ απλή: Η κλάση που έχει τη μεγαλύτερη πιθανότητα είναι η «νικητήρια» και εκχωρείται το 0 ή το 1 αντίστοιχα στο συνδυασμό των στοιχείων.
Παράδειγμα:

Έστω ότι έχουμε δεδομένα για 1000 κομμάια φρούτων (Μπανάνα, πορτοκάλι ή άλλο φρούτο). Γνωρίζουμε 3 χαρακτηριστικά για κάθε φρούτο:

1. Αν είναι μακρύ
2. Αν είναι γλυκό
3. Αν είναι κίτρινο

Το παρακάτω είναι το training set μας.

Type	Long	Not Long	Sweet	Not Sweet	Yellow	Not Yellow	Total
Banana	400	100	350	150	450	50	500
Orange	0	300	150	150	300	0	300
Other Fruit	100	100	150	50	50	150	200
Total	500	500	650	350	800	200	1000

Με βάση τις prior probabilities έχουμε:

P (μπανάνας) = 0.5 (500/1000)
P (πορτακαλιού) = 0.3
P (άλλου φρούτου) = 0.2

Πιθανότητα «Στοιχείων»:

P (μακρύ) = 0.5
P (γλυκό) = 0.65
P (κίτρινο) = 0.8

Πιθανότητες δεσμευμένες:

P (μακρύ|μπανάνα) = 0.8
P (μακρύ|πορτοκάλι) = 0
....
P (κίτρινο|Άλλο φρούτο) = 50/200 = 0.25
P (Όχι κίτρινο|Άλλο φρούτο) = 0.75

Ταξινόμηση με Naive Bayes:

Έστω ότι έχουμε τις ιδιότητες ενός αγνώστου φρούτου και θέλουμε να το ταξινομήσουμε. Το φρούτο είναι μακρύ, γλυκό και κίτρινο. Το φρούτο είναι;
Μπορούμε να τρέξουμε τα νούμερα ένα ένα. Τότε διαλέγουμε τη μεγαλύτερη πιθανότητα και ταξινομούμε το άγνωστο φρούτο σαν να ανήκει στην κλάση με τη μεγαλύτερη πθιανότητα, βασισμένο σην prior probability που βγάλαμε από το Train set.

- P (Μπανάνα | μακρύ, γλυκό & κίτρινο) =

$$\frac{P(\text{μακρύ} \mid \text{Μπανάνα}) * P(\text{Γλυκό} \mid \text{Μπανάνα}) * P(\text{Κίτρινο} \mid \text{Μπανάνα}) * P(\text{Μπανάνα})}{P(\text{μακρύ}) * P(\text{γλυκό}) * P(\text{κίτρινο})}$$

$$= 0.8 * 0.7 * 0.9 * 0.5 / P(\text{στοιχείων})$$

$$= \mathbf{0.252} / P(\text{στοιχείων})$$

- P (πορτοκαλί | μακρύ, γλυκό & κίτρινο) = **0**
- P (Άλλο φρούτο | μακρύ, γλυκό & κίτρινο) =

$$\frac{P(\text{μακρύ}|\text{Άλλο φρούτο}) * P(\text{γλυκό}|\text{Άλλο φρούτο}) * P(\text{Κίτρινο}|\text{Άλλο φρούτο}) * P(\text{άλλο φρούτο})}{P(\text{στοιχείων})}$$

$$= (100/200 * 150/200 * 50/200 * 200/1000) / P(\text{στοιχείων})$$

$$= \mathbf{0.01875} / P(\text{στοιχείων})$$

Βλέπουμε ότι 0.252 >> 0.01875 και άρα ταξινομούμε το φρούτο αυτό ως μπανάνα.

Ο Naive Bayes είναι λοιπόν διαδεδομένος, γιατί βασίζεται μόνο σε κάποιους απλούς πολλαπλασιασμούς και η ταξινόμηση γίνεται εύκολα, γρήγορα και αποτελεσματική.

4.2.3 PCA στο Matlab

Όλη η ανάλυση έγινε στο Matlab , το οποίο έχει μέσα την συνάρτηση έτοιμη και εμείς έπρεπε να ετοιμάσουμε κατάλληλα τα δεδομένα μας και να την τρέξουμε σωστά.

Η διαδικασία που ακολουθούμε είναι η εξής: (παρατίθεται η διαδικασία που περιλαμβάνει τα δείγματα που αντιστοιχούν σε προχωρημένα στάδια καρκίνου. Στη

συνέχει προστίθενται και τα φάσματα που αντιστοιχούν σε πρώιμο καρκίνο και η διαδικασία είναι πανομοιότυπη)

- Συγκεντρώσαμε τα δεδομένα μας σε έναν πίνακα 96x995 (1000 είναι οι συχνότητες στον xx' μείον 5 συχνότητες που κόψαμε στο τέλος λόγω θορύβου ηλεκτρονικών). 55x995 είναι ο πίνακας με γραμμές τα φάσματα του καρκινικού ιστού και 41x995 του υγιούς ακριβώς από κάτω. Δημιουργήσαμε με βάση τον πίνακα αυτόν μια στήλη με τις **κλάσεις** (κατηγορίες) μας στην οποία να αντιστοιχεί το «0» στα φάσματα από καρκινικό ιστό και το «1» στα φάσματα από υγιή ιστό.

- Στη συνέχεια ανακατέψαμε τις γραμμές των δεδομένων με τυχαία σειρά και με την ίδια σειρά ανακατέψαμε ξεχωριστά και τις γραμμές της στήλης με τις κλάσεις.

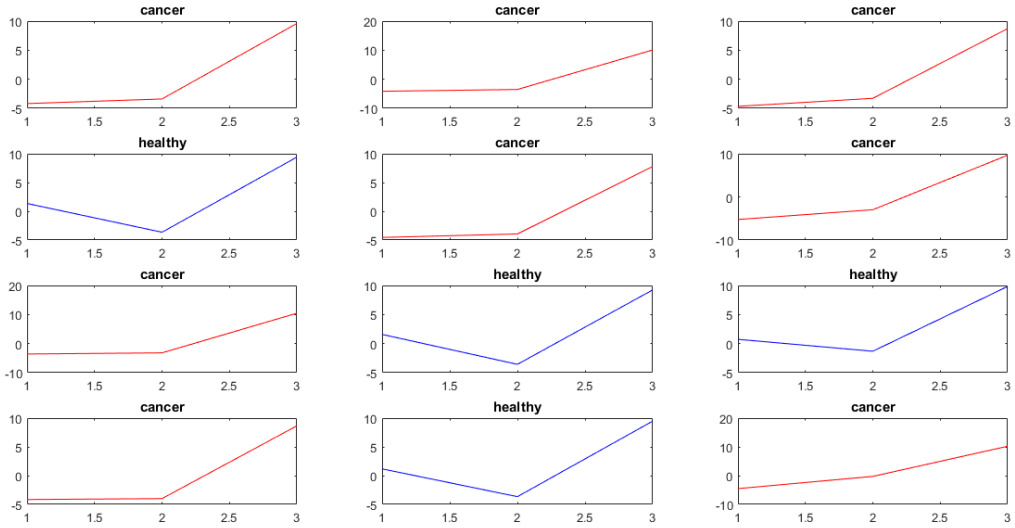
- Για να ταξινομήσουμε τα δεδομένα μας χρειάζεται να έχουμε ένα σύνολο δεδομένων που θα εκπαιδευτεί (**train set**) και ένα σύνολο δεδομένων στο οποίο θα δοκιμάσουμε το μοντέλο του ταξινομητή (**test set**). Σε μικρά δείγματα, σαν το δικό μας, συστήνεται να χωρίζουμε το data set σε train test, test set και **validation test**. Δοκιμάσαμε λοιπόν και αυτήν την προσέγγιση με διαχωρισμό 60% - 20% - 20%.Διαχωρίσαμε τα δεδομένα μας σε 80% train set και 20% test set. Με όμοιο τρόπο λειτουργήσαμε και στην περίπτωση T-V-T.

- Εφαρμόσαμε PCA στο train set και μετά από κάποιες δοκιμές αποφασίσαμε να κρατήσουμε 3 διαστάσεις, δηλαδή 3 κύριες συνιστώσες. Για να βρούμε τον μειωμένο σε διαστάσεις πίνακα, χρησιμοποιούμε τον πίνακα συντελεστών (coefficient matrix) που μας δίνει το Matlab αφού γίνει η PCA και πολλαπλασιάζουμε τα δεδομένα μας με αυτό τον πίνακα ώστε να μειώσουμε κατά πολύ την διάσταση των στηλών. Ο πίνακας που προκύπτει τελικά έχει διαστάσεις 77x3 και αφού προσθέσουμε και το διάνυσμα των κλάσεων γίνεται 77x4.

  Με βάση τον πίνακα συντελεστών που προέκυψε μετά την PCA στον train set κάναμε PCA στο test set (και στην περίπτωση που έχουμε validation test το ίδιο και σε αυτό). Προέκυψαν τελικά οι μειωμένοι πίνακες: 77x4 και 19x4 για T-T και 58x4,19x4 και 19x4 για T-V-T.

  Για να έχουμε μια διαίσθηση του τι μπορεί να σημαίνει οτι όλες οι μεταβλητες περιγράφονται από 3 σημαντικές, παραθέτω τα διαγραμματα των μειωμένων (reduced) φασμάτων, αφού δηλαδή έχουν υποστεί PCA. Τα διαγράμματα δείχνουν μαζεμένη πολλή πληροφορία. Όσο πιο πολλές PC είχαμε επιλέξει

τόσο πιο κοντά στα αρχικά μας φάσματα θα ήταν τα καινούργια διαγράμματα:



(*Εικόνα 4.3.2.1*)

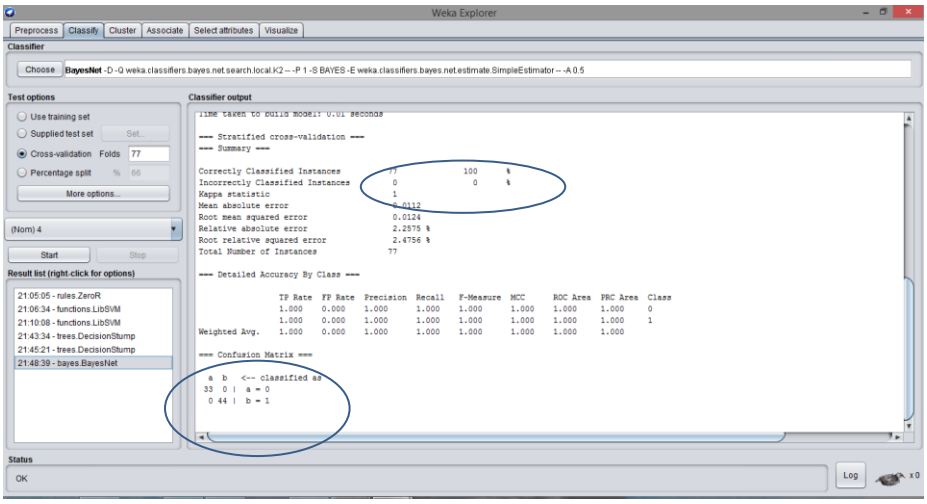
Αρχικά αναλύσαμε μόνο τα φάσματα που προήλθαν από προχωρημένο καρκινικό ιστό. Η τελική ανάλυσή μας περιλάμβανε και τα 146 φάσματα, δηλαδή και τον πρώιμο και τον προχωρημένο καρκίνο. Πιο αναλυτικά θα αναφερθούμε και παρακάτω.

5. Αποτελέσματα

5.1 Πρώτη φάση πειραμάτων

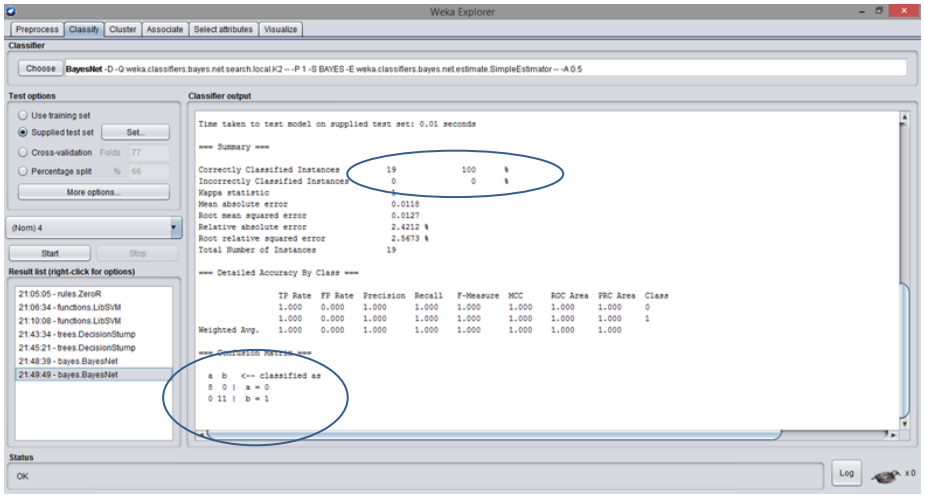
Στην πρώτη φάση των πειραμάτων μας τα δείγματά μας αποτελούνταν μόνο από υγιή και καρκινικό ιστό σε προχωρημένο στάδιο. Στα δείγματα αυτά οι διαφορές ήταν πολύ εμφανείς και το δείγμα πολύ μικρό (107 φάσματα). Αυτό οδήγησε στο να έχουμε 100% επιτυχία στο Training και στο test set μας. Ενδεικτικά παραθέτουμε τα αποτελέσματα του WEKA για Naive Bayes:

Train set:



(*Εικόνα 5.1.1*)

Test set:



(*Εικόνα 5.1.2*)

Πιο αναλυτικά στον παρακάτω πίνακα βλέπουμε τα σημαντικά στοιχεία των δυο παραπάνω αποτελεσμάτων.

Train set:

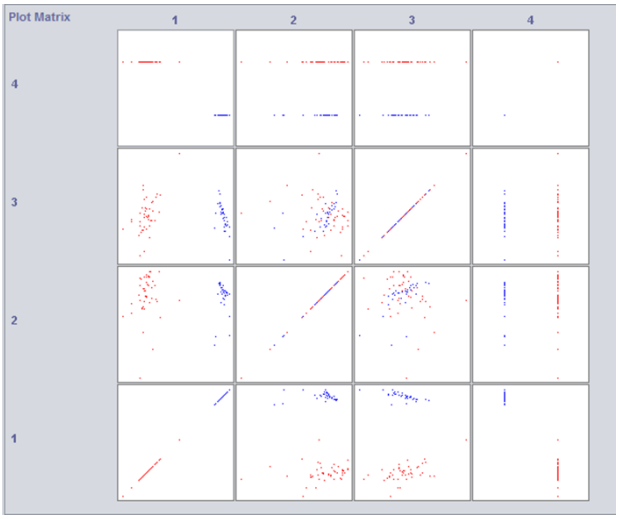
Σωστά ταξινομημένα δεδομένα	77	Επί τοις 100	100 %
Λανθασμένα ταξινομένα δεδομένα	0	Επί τοις 100	0%
Ταξινομημένα ως a (ένω ήταν b)	0		
Ταξινομημένα ως b (ένω ήταν a)	0		
Ταξινομημένα ως a (ένω ήταν a)	13		
Ταξινομημένα ως b (ένω ήταν b)	44		

Test set:

Σωστά ταξινομημένα δεδομένα	19	Επί τοις 100	100 %
Λανθασμένα ταξινομένα δεδομένα	0	Επί τοις 100	0%
Ταξινομημένα ως a (ένω ήταν b)	0		
Ταξινομημένα ως b (ένω ήταν a)	0		
Ταξινομημένα ως a (ένω ήταν a)	8		
Ταξινομημένα ως b (ένω ήταν b)	11		

Οι 4 τελευταίες γραμμές αντιστοιχούν στον πίνακα των πραγματικών κλάσεων σε σύγκριση με τις κλάσεις που κατατάχτηκαν από τον ταξινομητή (confusion matrix). Υπήρχαν, δηλαδή, 8 δεδομένα κλασης a=0 (υγιείς ιστοί) από τα οποία μηδέν προβλέφθηκαν ως b=1 (καρκινικοί ιστοί) και το ίδιο συνέβη και με τα 11 δεδομένα κλάσης b.

Από το διάγραμμα των 3 PC's και της 4^{ης} στήλης (των κλάσεων) μπορούμε επίσης να δούμε ότι τα δεδομένα μας είναι τέλεια διαχωρίσιμα.



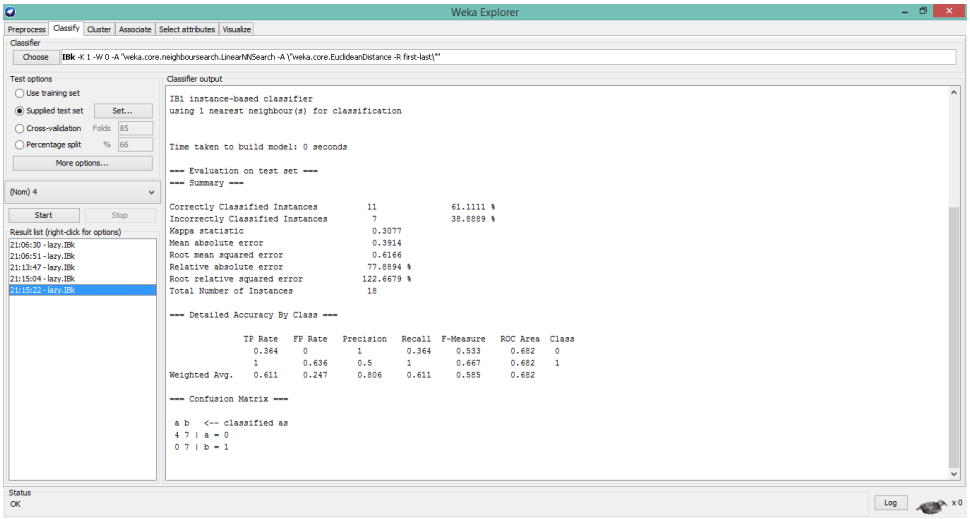
(*Εικόνα 5.1.3*)

Ίδια αποτελέσματα έδωσαν και ο k-NN και ο SVM. Γι' αυτό το λόγο αποφασίσαμε να κάνουμε και έναν δεύτερο γύρο πειραμάτων τα οποία να περιλαμβάνουν καρκινικό ιστό, αλλά σε πρώιμο στάδιο.

5.2 Δεύτερη φάση πειραμάτων

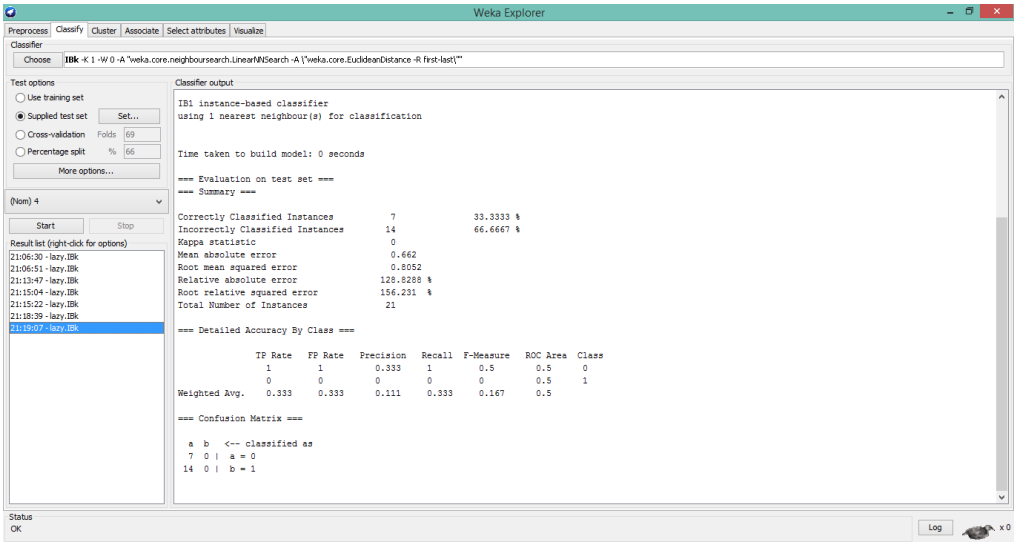
Αυτή είναι και η ολοκληρωμένη φάση των πειραμάτων μας. Έχουμε 146 διαφορετικά φάσματα, τα οποία αφού επεξεργαστήκαμε, κάναμε ταξινόμηση με τους 3 αλγορίθμους – ταξινομητές. Ο σωστός τρόπος για να γίνει κάτι τέτοιο είναι να ανακατεψουμε όλα τα δεδομένα και να κάνουμε όλη την επεξεργασία και την εκπαίδευση με αυτά.

Δοκιμάσαμε και τι θα γίνει αν εκπαιδεύσουμε τα δεδομένα μας μόνο με τα πρώτα φάσματα (δηλαδή του προχωρημένου καρκίνου) και προσπαθήσουμε τα ταξινομήσουμε τον πρώιμο καρκίνο με βάση αυτό το training. Δηλαδή το train set μας να είναι τα αρχικά δεδομένα και το test set μας τα πρώιμα καρκινικά. Τα αποτελέσματα όπως ήταν αναμενόμενο ήταν μέτρια, αφού μόνο ένα 61.1 % ταξινομήθηκε σωστά από τα δεδομένα μας.



(*Εικόνα 5.1.4*)

Δοκιμάσαμε επίσης να εκπαιδεύσουμε τον αλγόριθμό μας με δεδομένα τα φάσματα που πρώιμου καρκίνου και να χρησιμοποιήσουμε ως test σετ τα δεδομένα από την αρχική φάση του πειράματός μας. Τα αποτελέσματα ήταν πολύ κάτω του μετρίου σε αυτήν την περίπτωση (μόλις 33.3%!) . Τα φάσματα του πρώιμου καρκίνου κρατανε κάποια χαρακτηριστικά και από τα υγιή, ενώ του προχωρημένου καρκίνου εμφανίζουν νέες ιδιότητες που ο αλγόριθμος δεν είδε ποτέ κι έτσι δεν εκπαιδεύτηκε καθόλου σωστά.

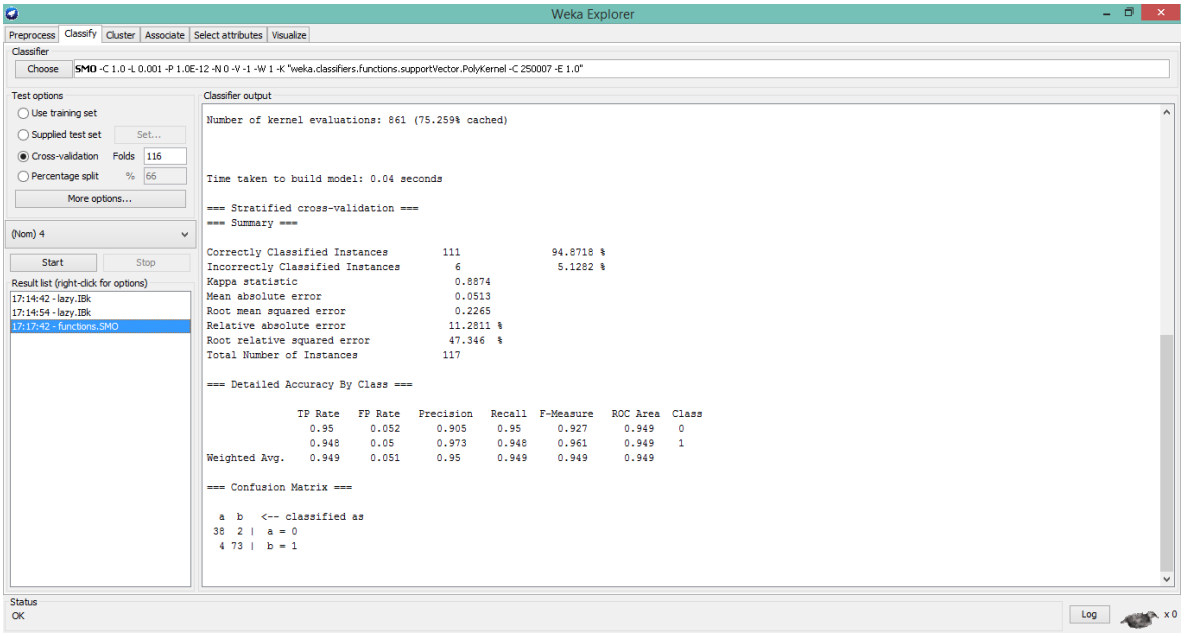


(*Εικόνα 5.1.5*)

Ας δούμε τώρα και την σωστή περίπτωση, δηλαδή αυτή στην οποία όλα τα φάσματα ανακατεύτηκαν τυχαία και εκπαιδεύτηκαν.

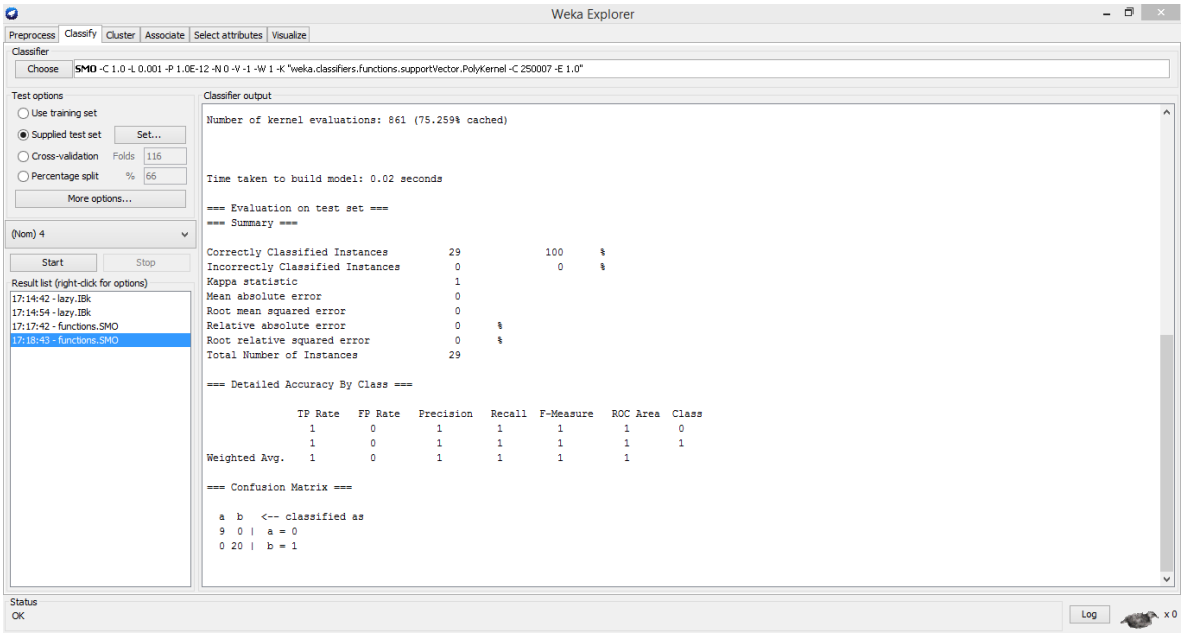
Αρχικά παραθέτουμε τα αποτελέσματα του SVM:

Train set:



(*Εικόνα 5.1.6*)

Test set:

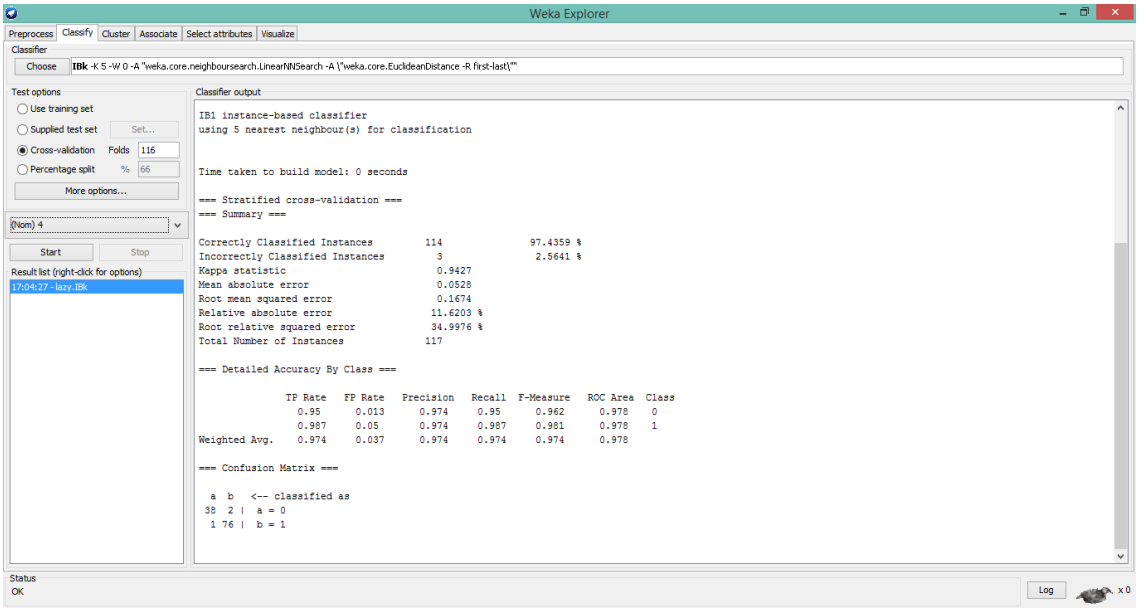


(*Εικόνα 5.1.7*)

Η ταξινόμηση λοιπόν έγινε με ποσοστό 100% στο test set και 94.8718% στο Train set.

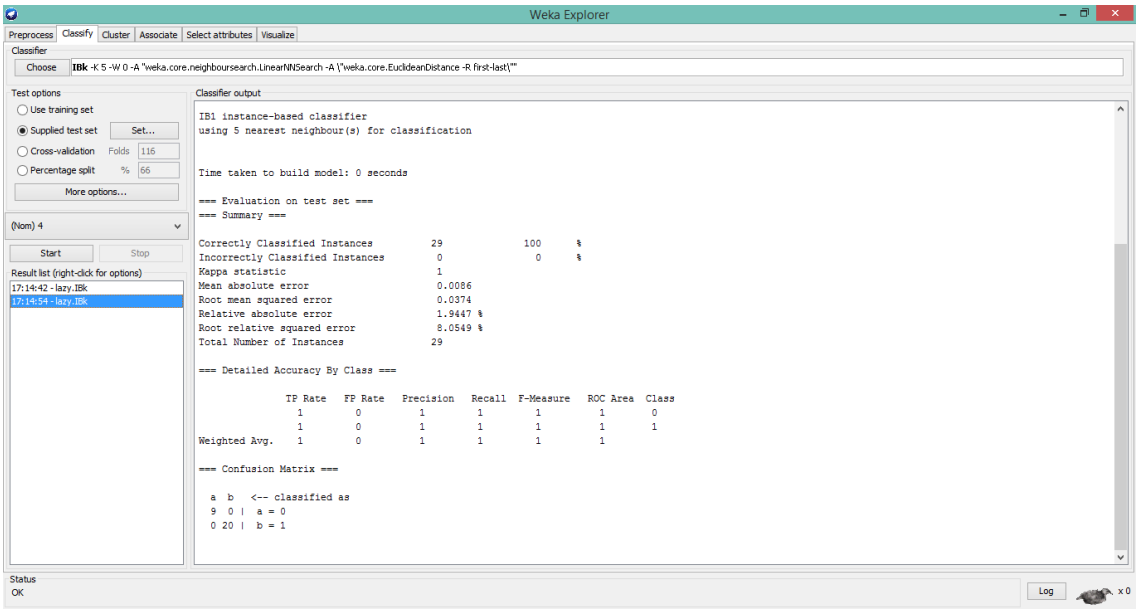
Αντίστοιχα αποτελέσματα έδωσε και ο k-NN με 97.4359% επιτυχία στο Train set και 100% επιτυχία στο Test set:

Train set:



(*Εικόνα 5.1.8*)

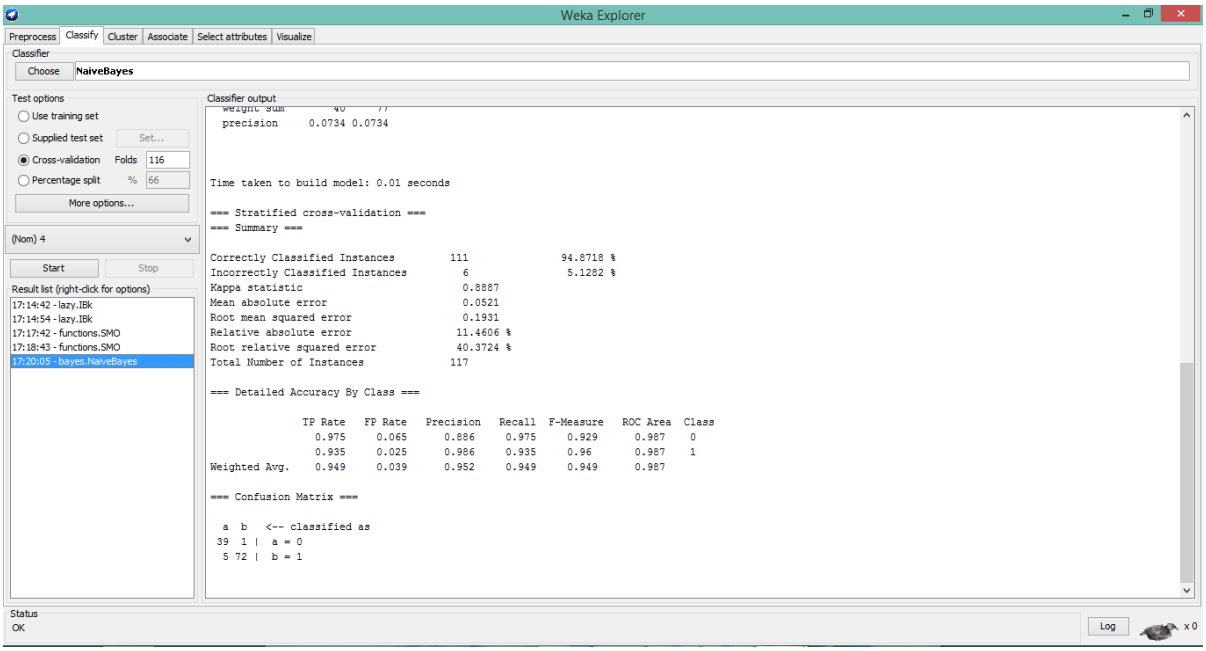
Test set:



(*Εικόνα 5.1.9*)

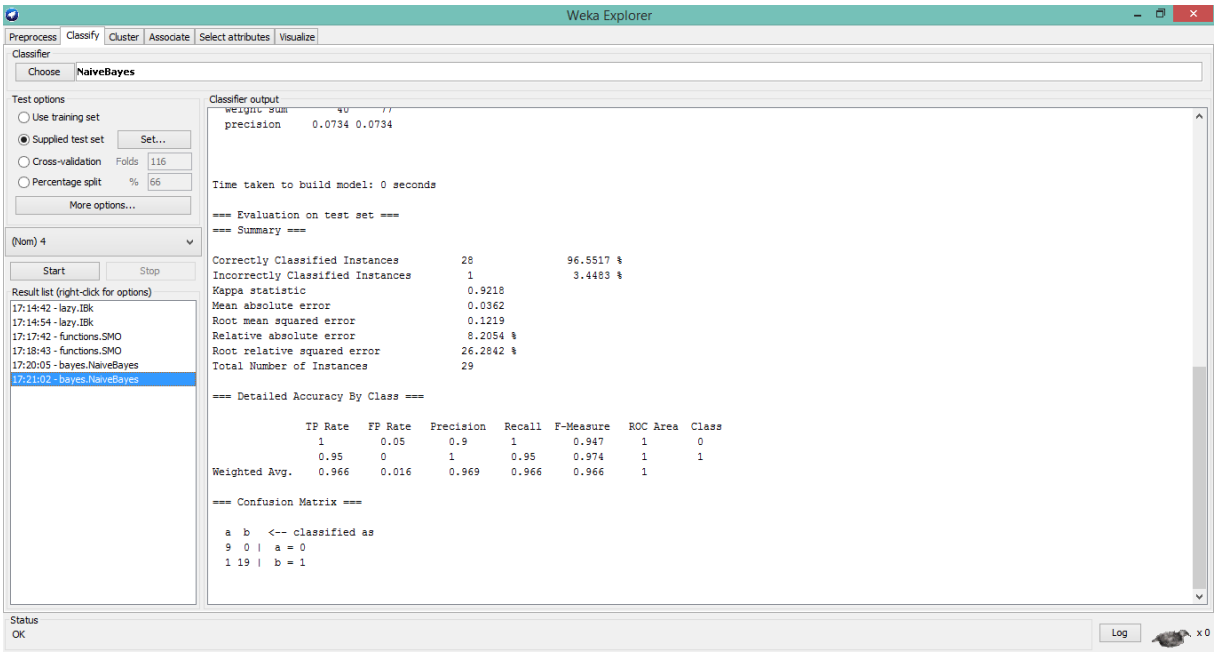
Ο Naive Bayes μας έδωσε λίγο χειρότερα αποτελέσματα. Συγκεριμένα στο train set ταξινόμησε 94.87% σωστά και στο test set 95.55%.

Train set:



(*Εικόνα 5.1.10*)

Test set:

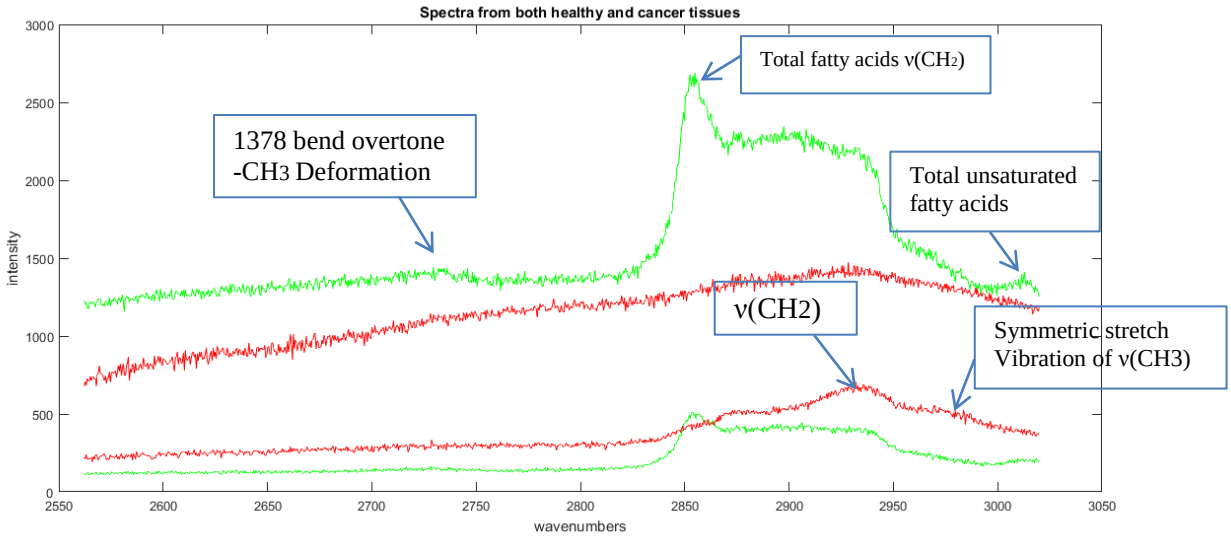


(*Εικόνα 5.1.11*)

Τέλος, αν και δεν είναι αυτό το θέμα αυτής της μελέτης, έχει ενδιαφέρον να δούμε ποιες κορυφές αλλάζουν από το υγιές στο καρκινικό φάσμα και σε τι χημικές ενώσεις αντιστοιχουν. Η πλειοψηφία αυτών των κορυφών αντιστοιχεί σε λιπίδια (16) (33).

Πιο αναλυτικά, οι δύο πολύ σημαντικές κορυφές στα υγιή φάσματα είναι αυτή που αντιστοιχεί στα 2845 cm^{-1} σε λιπαρά οξέα (Total fatty acids) TFA και αυτή στα 3009 cm^{-1} - 3015 cm^{-1} αντιστοιχεί σε ακόρεστα λιπαρά οξέα (Total unsaturated fatty acids) TUFA. Αυτές δεν εμφανίζονται έντονα στα καρκινικά φάσματα και γι' αυτό ένας μικρός λόγος TUFA/TFA έχει συσχετιστεί με καρκίνους (34) (35).
Άλλες κορυφές αντιστοιχούν από την βιβλιογραφία σε (16) (33) (36):

- 2727 cm^{-1}: είναι οι αρμονικές (overtone) μιας κάμψης που αντιστοιχεί στα 1378 cm^{-1} καθώς και σε παραμόρφωση (deformation) ενός CH_3 δεσμού.
- 2854 cm^{-1} : είναι πολύ διακριτή στα υγιή φάσματα και αντιστοιχεί στα TFU και πιο συγκεκριμμένα σε αλυσίδα από CH_2.
- 2933-2935 cm^{-1} : η κορυφή αυτή είναι πολύ διακριτή στο καρκινικό φάσμα και αντιστοιχεί σε αντισυμμετρικό τέντωμα (antisymmetric stretch) του CH_2 (λιπίδια και πάλι δηλαδή). Επιπλέον αυτές οι κορυφές αντιστοιχούν σε συμμερικό τέντωμα (symmetric stretch) του CH_3.
- 2985 cm^{-1}: Η κορυφή αυτή εμφανίζεται πιο πολύ στα καρκινικά φάσματα και αντιστοιχεί σε ομάδες CH_3 και σε πρωτεϊνες.
- 3009-3015 cm^{-1}: Οι κορυφές αυτές αντιστοιχούν σε TUFA και πιο συγκεκριμμένα σε olefinic =CH stretching.



(*Εικόνα 5.1.12*)

Ο υγιής ιστός αποτελείται στο μεγαλύτερο ποσοστό από λιπιδικό ιστό και οι συνεισφορές τους είναι πολύ εμφανείς σε αυτή την περιοχή. Η περιοχή αυτή περιλαμβάνει και αλλαγές σε πρωτείνες όπως ας πούμε αύξηση του της πρωτείνης HER2, που έχει παρατηρηθεί από πολλές μελέτες μέχρι τώρα στον καρκίνο του μαστού (36).

Τα λιπίδια εμπλέκονται στη ρύθμιση του πολλαπλασιασμού, της διαφοροποίησης (differentiation), της απόπτωσης, της φλεγμονής, και της ομοιόστασης της μεμβράνης και στον καρκίνο φαινεται όλα αυτά να απορρυθμίζονται (37) (38).

6. Συζήτηση

Για να χρησιμοποιήσουμε τη φασματοσκοπία Raman ως σύνηθες διαγνωστικό εργαλείο πρέπει να αναπτύξουμε ένα ιατρικό σύστημα που να χρησιμοποιηθεί με ασφάλεια σε ένα κλινικώς αποδεκτό χρονικό πλαίσιο για ένα φάσμα ασθενειών. Θα πρέπει να είναι ακριβές ώστε να μας δώσει αξιόπιστες πληροφορίες μέσω των οποίων οι ειδικοί θα μπορούν να κάνουν μια σωστή διάγνωση.

Συνοπτικά λοιπον, συλλέξαμε τα φάσματά μας χρησιμοποιώντας διάταξη micro-Raman. Στη συνέχεια τα επεξεργαστήκαμε στο Matlab για καλύτερη απόδοση και εκπαιδεύσαμε με αυτά αλγορίθμους. Τα άγνωστα φάσματά μας (test set) ταξινομήθηκαν με πολύ καλά αποτελέσματα. Τα αποτελέσματά μας έδειξαν πολύ μεγάλα ποσοστά σωστής διάγνωσης και με τον κατάλληλο αλγόριθμο έφτασαν στο 100%. Εδώ όμως θα πρέπει όμως να είμαστε λίγο προσεκτικοί. Τα φάσματά μας προήλθαν από πολύ λίγα ποντίκια και στη βιβλιογραφία (39) συνήθως χρησιμοποιούνται περισσότερα δεδομένα από τα δικά μας και από περισσότερα διαφορετικά δείγματα- ασθενείς. Αυτό θα μπορούσε να οδηγήσει σε overfitting αν και χρησιμοποιώντας τη μέθοδο leave-one-out και συμπληρωματικά test set αλλά και Validation set έχουμε μειώσει κατά πολύ τον κίνδυνο να έχει γίνει αυτό χωρίς να το καταλάβουμε.

Παρότι είναι λοιπόν πολλά υποσχόμενη η μέθοδος αυτή για διάγνωση καρκίνου, υπάρχουν κάποια πρακτικά προβλήματα που πρέπει να ξεπεραστούν, αλλά και βελτιώσεις που μπορούν να γίνουν στο μέλλον:

1. Το πρώτο βήμα για τη συνέχεια της έρευνάς μας είναι να περάσουμε από δείγματα ποντικών σε ανθρώπινους ιστούς.

2. Οι χρόνοι καταμέτρησης είναι καλό να μειωθούν πολύ, κατά προτίμηση σε λίγα sec για κάθε σημείο. Κάτι τέτοιο θα μπορούσε να γίνει με την τεχνική SERS (Surface Enhanced Raman Spectroscopy), με νανοσωματίδια δηλαδή, τα οποία μέσω ενός φαινομένου ενισχύουν το σήμα κατά πολλές τάξεις μεγέθους. Υπάρχουν ήδη αρκετές μελέτες πάνω σε αυτό, σε διάφορα ήδη καρκίνου και από ιστούς μέχρι βιολογικά υγρά (ούρα, αίμα, σάλιο, πλάσμα (40) (41) (42) (43) (44). Άλλες τέτοιες μέθοδοι είναι οι CARS και SRS.

3. Ένα σημαντικό ζήτημα είναι πώς θα δημιουργήσουμε διαγνωστικές μεθόδους μη επεμβατικές. Η απλή μέθοδος Raman δεν μπαινει σε βάθος μέσα στον ιστό και γι΄αυτό γίνεται μια προσπάθεια να αναπτυχθούν νέες τεχνικές όπως η deep Raman φασματοσκοπία η οποία μας επιτρέπει να γίνει η μέτρηση σε βάθος μέχρι και μερικά εκατοστά (45). Ένας άλλος τρόπος για να γίνει μέτρηση σε βάθος είναι ενδοσκοπική Raman φασματοσκοπία, κυρίως για καρκίνο του

στόματος αλλά και του γαστρεντερικού συστήματος (46). Επιπλέον υπάρχουν μελέτες για διάγνωση διαφόρων ειδών καρκίνου, συμπεριλαμβανομένου και του καρκίνου του μαστού (9), μέσω βιολογικών δεικτών σε βιολογικά υγρά (8).

4. Το τελευταίο βήμα θα ήταν φυσικά η ανάπτυξη ενός μηχανήματος που να δίνει τη δυνατότητα εύκολα και γρήγορα να γίνεται η διάγνωση ή ακόμη καλύτερα, να μπορει να γίνει προληπτική διάγνωση με μια εξέταση ούρων ή αίματος. Υπάρχει μια εταιρία που έχει αναπτύξει τέτοια τεχνολογία για καρκίνο του δέρματος και ονομάζεται Verisante (47).

7. Παράρτημα

Παρακάτω παραθέτουμε τους κώδικες του Matlab.

Στο Matlab γράψαμε όλους τις συναρτήσεις την main και τρέχουμε την main. Η main αποτελείται από τα εξής κομμάτια κώδικα:

Αρχικά διαβάζει τα δεδομένα από τον φάκελο στον οποίο βρίσκονται. Τα δεδομένα φορτώνονται σε έναν πίνακα με γραμμές όσες και τα φάσματά μας και γραμμές όσες οι συχνότητες που σκανάραμε.

```
disp('Reading data...');

% read data

% data_cancer = load_data('baseline correction/cancer/');

% [data_healthy, x] = load_data('baseline correction/healthy/');

data_cancer = load_data('cutted/cancer/');

[data_healthy, x] = load_data('cutted/healthy/');
```

Κανονικά η CCD καταγράφει 1024 διαφορετικές συχνότητες. Την περίοδο του πρώτου πειράματος υπήρχε ένα πρόβλημα με τα ηλεκτρονικά και δεν κατέγραφαν τόσες, αλλά λίγο λιγότερες συχνότητες. Γι' αυτό το λόγο μέσα στη συνάρτηση load_data αλλάζουμε λίγο το μέγεθος των δεδομένων αν χρειαστεί (παράδειγμα στην pca3) ώστε να μπορούν όλα τα δεδομένα να μπουν σε έναν πίνακα. Για τον ίδιο λόγο κόβουμε 5 από το τέλος παρακάτω στον κώδικα.

```
% cut 5 from front and 1 from the end

% data_cancer = data_cancer(:,5:end-1);

% data_healthy = data_healthy(:,5:end-1);

% x = x(:,5:end-1);

data_cancer = data_cancer(:,5:end);

data_healthy = data_healthy(:,5:end);

x = x(:,5:end);
```

Τα επόμενα βήματα αποτελούνται από το smoothing και την baseline correction. Οι παράμετροι επιλέχτηκαν μετά από δοκιμές.

```
% Smooth the data

data_cancer_smoothed = data_smooth(data_cancer, 5, 49);

data_healthy_smoothed = data_smooth(data_healthy, 5, 49);



% Baseline correction

data_cancer_corrected = msbackadj(x', data_cancer_smoothed', 'WINDOWSIZE', 50,
'STEPSIZE', 200)';

data_healthy_corrected = msbackadj(x', data_healthy_smoothed', 'WINDOWSIZE',
50, 'STEPSIZE', 200)';

% plot(x,data_healthy_smoothed(7,:));

% pause;
```

Έπειτα κανονικοποιούμε. Στα comments υπάρχει ένας κώδικας για κανονικοποίηση με τη μέθοδο AUC. Δεν χρησιμοποιήθηκε τελικά στην εργασία αυτή αλλά αν χρειαστεί να χρησιμοποιηθεί την κάνουμε uncomment και κάνουμε comment την απλή κανονικοποίηση:

```
disp('Normalization');


% Normalization

% % AUC normalization (AUC = 1)

% data_cancer_norm = auc_norm(x,data_cancer)';

% data_healthy_norm= auc_norm(x,data_healthy)';

%

% %  Max normalization (per column)

data_cancer_norm = max_norm(data_cancer_corrected);

data_healthy_norm = max_norm(data_healthy_corrected);

% plot(x,data_healthy_norm(7,:));

% pause;
```

Στο επόμενο βήμα βάζουμε 1 σε όλα τα φάσματα από καρκίνο και 0 σε όλα τα υγιή. Αυτό που κάνουμε είναι να βάζουμε την αντίστοιχη τιμή στο τέλος της κάθε γραμμής, που όπως είπαμε αντιστοιχεί στο αντίστοιχο φάσμα.

```
% Concentrate data

data = [data_cancer_norm;data_healthy_norm];

data_classes=[ones(size(data_cancer_norm,1),1);zeros(size(data_healthy_norm,1),1)];
```

Στη συνέχεια ανακατέβουμε τις γραμμές:

```
 % Permutate data

data_perm = randperm(size(data,1));

data = data(data_perm,:);

data_classes = data_classes(data_perm,:);
```

Ανά διαστήματα έχουμε βάλει να απεικονίζονται τα βήματα

```
disp('PCA & writing datasets...');
```

Πριν την PCA χωρίζουμε τα δεδομένα μας σε train και test set

```
% Create 2 datasets (a train-validate-test and train-test)

% % Train-Validate-Test

train = data(1:round(size(data,1)*0.6),:); % 60% train set

validation =
data((round(size(data,1)*0.6)+1):(round(size(data,1)*0.6)+round(size(data,1)*0.2)),:);
% 20% validation

test = data((round(size(data,1)*0.6)+round(size(data,1)*0.2))+1:end,:); % 20% test
```

και στη συνέχεια αποφασίζουμε πόσες κύριες συνιστώσες (principal components) θα κρατήσουμε.

```
% number of reduced dimensions

num_reducedDimensions = 3;
```

Κάνουμε PCA στο train-validate-test και το ίδιο κάνουε στη συνέχεια και για το Train-test:

```
% PCA

[train, coeff] = pca_reduction(train, num_reducedDimensions);

validation = pca_reduction(validation,num_reducedDimensions,coeff);

test = pca_reduction(test,num_reducedDimensions,coeff);
```

Προσθέτουμε τις κλάσεις στο τέλος, μετα την PCA.

```
% add classes at the end of the matrix (last col)

train = [train data_classes(1:round(size(data,1)*0.6))];

validation = [validation
data_classes((round(size(data,1)*0.6)+1):(round(size(data,1)*0.6)+round(size(data,1)
*0.2)))];

test = [test data_classes((round(size(data,1)*0.6)+round(size(data,1)*0.2))+1:end)];
```

Στο τελευταίο μέρος καταγράφουμε αυτά τα δεδομένα σε φακέλους.

```
% write files

dlmwrite('final_data/Train-Validation-Test/train.csv', train, 'delimiter',',');

dlmwrite('final_data/Train-Validation-Test/validation.csv', validation, 'delimiter',',');

dlmwrite('final_data/Train-Validation-Test/test.csv', test, 'delimiter',',');


arffwrite('final_data/Train-Validation-Test/train',train);

arffwrite('final_data/Train-Validation-Test/validation',validation);

arffwrite('final_data/Train-Validation-Test/test',test);
```

Ομοίως για το train-test:

```
% % Train-Test only

train = data(1:round(size(data,1)*0.8),:); % 80% train set

test = data(round(size(data,1)*0.8)+1:end,:); % 20% test set


% PCA
```

```
[train, coeff] = pca_reduction(train, num_reducedDimensions);
test = pca_reduction(test,num_reducedDimensions,coeff);

train = [train data_classes(1:round(size(data,1)*0.8))];
test = [test data_classes(round(size(data,1)*0.8)+1:end)];

% write files
dlmwrite('final_data/Train-Test/train.csv', train, 'delimiter',',');
dlmwrite('final_data/Train-Test/test.csv', test, 'delimiter',',');

arffwrite('final_data/Train-Test/train',train);
arffwrite('final_data/Train-Test/test',test);
```

Τα δεδομένα αυτά χρησιμοποιούνται στο WEKA.

Ανοίγοντας το WEKA πατάμε το κουμπί Explorer.

Στο Preprocess πατάμε Open file, ανοίγουμε το φάκελο final_data που περιέχει τα τελικά δεδομένα μας. Ανοίγουμε αρχικά το train. Στη συνέχεια πηγαίνουμε στο Classify. Κάτω από το Classifier επιλέγουμε από το choose τον classifier που επιθυμούμε. Επιλέγουμε στο cross validation μείον ένα από τα δεδομένα μας ώστε να γίνει το Training με τη μέθοδο leave-one-out και πατάμε Start. Μετά το training επιλέγουμε στο supplied test set το test σετ και επαναλαμβάνουμε.

Bibliography

1. **CDCP.** Centers for Disease Control and Prevention. [Online] https://www.cdc.gov/mmwr/preview/mmwrhtml/mm5846a5.htm.

2. **wikipedia.** wikipedia. *wikipedia.* [Online] https://www.wikipedia.org/.

3. **IARC.** International Agency for Research on Cancer. [Online] http://www.iarc.fr/.

4. *Raman spectroscopy for medical diagnostics - From in-vitro biofluid assays to in-vivo cancer detection.* **Kenny Kong, Catherine Kendall, Nicholas Stone, Ioan Notingher.** 2015, Advanced Drug Delivery Reviews.

5. *Modern Raman Spectroscopy – A Practical Approach.* **Ewen Smith, Geoffrey Dent.** 2005, John Wiley & Sons, Ltd.

6. *Diagnostic applications of Raman spectroscopy.* **Qiang Tu, Chang Chang.** 2015, Nanomedicine: Nanotechnology, Biology, and Medicine.

7. *Biomedical applications of Raman and infrared spectroscopy to diagnose tissues.* **Sergo, C. Krafft and V.** 2006, Biomedical applications of Raman and infrared spectroscopy.

8. *Raman spectroscopy for cancer diagnosis: how far have we come?* **Culha, Mustafa.** 2015, Bioanalysis.

9. *Preliminary Raman spectroscopic study of Urine: Diagnosis of breast cancer in animal models.* **T. Bhattacharjee, A. Khan, G. Marub, A. Ingle and C. Murali Krishna.** 2014, The Royal Society of Chemistry.

10. rsc. [Online] http://www.rsc.org/learn-chemistry/collections/spectroscopy/introduction#Introduction.

11. **Instruments, Princeton.** *Raman Spectroscopy Basics.* [Online] http://web.pdx.edu/~larosaa/Applied_Optics_464-564/Projects_Optics/Raman_Spectrocopy/Raman_Spectroscopy_Basics_PRINCETON-INSTRUMENTS.pdf.

12. **Hess, Christian.** *Raman spectroscopy: Basic principles and applications [presentation].* [Online] 2006. http://www.fhi-berlin.mpg.de/acnew/department/pages/teaching/pages/teaching__wintersemester__2006_2007/hess_raman_spectroscopy_101106.pdf.

13. **Γ. Βογιατζής, Σ. Γιαννόπουλος και Γ. Παπαθεοδώρου.** *ΣΤΟΙΧΕΙΑ ΦΑΣΜΑΤΟΣΚΟΠΙΑΣ RAMAN [Διάλεξεις].* [Online] 1999. http://tccc.iesl.forth.gr/AMS_EPEAEK/courses/VGP/ICEHT-Lab_Course.htm.

14. **Ράπτης, Ι.** *Συμπληρωματικές Σημειώσεις για το Εργαστήριο "Φασματοσκοπία Ραμαν" του μαθήματος "Μέθοδοι Χαρακτηρισμού Υλικών".* Αθήνα : ΕΜΠ, 2010.

15. **Παρασκευάς, Τσιακλάγκανος.** *Φασματοσκοπική μελέτη Raman νημάτων πολυεθυλένιου, Nylon και νανοσωλήνων άνθρακα.* Αθήνα : ΕΜΠ, 2014.

16. *Resonance Raman and Raman Spectroscopy for Breast Cancer Detection.* **C.-H. Liu, B.S., Y. Zhou, M.D., Ph.D., Y. Sun, Ph.D., J.Y.Li, M.D., Ph.D., L. X. Zhou, M.D., S. Boydston-White, Ph.D., V. Masilamani, Ph.D., K. Zhu, B.S., Yang Pu, Ph.D., R. R. Alfano, Ph.D.** 2013, Technology in Cancer Research and Treatment.

17. **Δήμητρα, Ξενιώτη.** *Ιδιότητες νανοδομών: Μελέτη polycarbonatek με φασματοσκοπία Raman.* Θεσσαλονίκη : ΑΠΘ Τμήμα Φυσικής, 2010.

18. *SERS Nanoparticles in Medicine: From Label-Free Detection to Spectroscopic Tagging.* **Lucas A. Lane, Ximei Qian, and Shuming Nie.** 2015, Chemical Reviews.

19. *Preparation of Tissues and Cells for Infrared and Raman Spectroscopy and Imaging.* **Fiona Lyng, Ehsan Gazi, Peter Gardner.** 2011, Medicine and health sciences commons.

20. *Infrared and Raman Microscopy in Cell Biology.* **Christian Matthäus, Benjamin Bird, Miloš Miljković, Tatyana Chernenko, Melissa Romeo, and Max Diem.** 2008, Methods Cell Biology.

21. *Review of multidimensional data processing approaches for Raman and infrared spectroscopy.* **Rekha Gautam, Sandeep Vanga, Freek Ariese, and Siva Umapathy.** 2015, EPJ Techniques and Instrumentation.

22. *Using Raman spectroscopy to characterize biological materials.* **Holly J Butler, Lorna Ashton, Benjamin Bird, Gianfelice Cinque, Kelly Curtis, Jennifer Dorney, Karen Esmonde-White, Nigel J Fullwood, Benjamin Gardner, Pierre L Martin-Hirsch, Michael J Walsh, Martin R McAinsh, Nicholas Stone and Francis L Martin.** 2016, Nature Protocols.

23. *Analyst.* **Julio Trevisan, Plamen P. Angelov, Paul L. Carmichael, Andrew D. Scott and Francis L. Martin.** 2012, Extracting biological information with computational analysis of Fouriertransform infrared (FTIR) biospectroscopy datasets: current practices to future perspectives.

24. sgolayfilt. [Online] https://www.mathworks.com/help/signal/ref/sgolayfilt.html.

25. msbackadj. *Mathworks.* [Online] https://www.mathworks.com/help/bioinfo/ref/msbackadj.html.

26. **I.T.Jolliffe.** *Principal Component Analysis, second edition.* New York : Springer-Verlag New York, 2002.

27. StackOverFlow. [Online] http://stackoverflow.com.

28. **Mitchell, Tom M.** *Machine learning.* 1997.

29. WEKA. [Online] http://www.cs.waikato.ac.nz/ml/weka/.

30. **Alexandre Kowalczyk.** *SVM tutorial.* [Online] http://www.svm-tutorial.com/.

31. **Castelli, Vittorio.** *Columbia Engineering.* [Online] 2005. http://www.ee.columbia.edu/~vittorio/lecture8.pdf.

32. **Jerome Friedman, Trevor Hastie, Robert Tibshirani.** *Book cover The Elements of Statistical Learning, Data Mining Inference and Prediction.* 2009.

33. *Raman and coherent anti-Stokes Raman scattering microscopy studies of changes in lipid content and composition in hormone-treated breast and prostate cancer cells.* **Potcoava, Mariana C.** s.l. : Journal of Biomedical Optics, 2014.

34. *The Lipid Phenotype of Breast Cancer Cells Characterized by Raman Microspectroscopy: Towards a Stratification of Malignancy.* **Claudia Nieva, Monica Marro, Naiara Santana-Codina, Satish Rao, Dmitri Petrov, Angels Sierra.** 2012, PLOS ONE.

35. *HOPE AND INNOVATIVE CANCER DIAGNOSTICS BY RAMAN SPECTROSCOPY AND RAMAN IMAGING.* **Abramczyk, Halina.** 2014, Fulbright Poland 55th Anniversary Conference.

36. *Raman imaging at biological interfaces: applications in breast cancer diagnosis.* **Jakub Surmacki, Jacek Musial, Radzislaw Kordek and Halina Abramczyk.** 2013, Molecular Cancer.

37. *Lipid biology of breast cancer.* **Jan Baumann, Christopher Sevinsky, Douglas S. Conklin.** 2013, Biochimica et Biophysica Acta.

38. **Αναστασιάδη, Γεωργία.** *Φασματοσκοπία Ραμαν ως διαγνωστικό μέσο καρκίνου του μαστού.* Αθήνα : ΕΜΠ, 2015.

39. *Advances in the clinical application of Raman spectroscopy for cancer diagnostics.* **Charlotte Kallaway, L. Max Almond, Hugh Barr, James Wood, Joanne Hutchings, Catherine Kendall, Nick Stone.** 2013, Photodiagnosis and Photodynamic Therapy.

40. *Surface-enhanced Raman spectroscopy (SERS): progress and trends.* **Dana Cialla, Anne März, René Böhme, Frank Theil, Karina Weber, Michael Schmitt, Jürgen Popp.** 2011, Analytical and Bioanalytical Chemistry.

41. *Surface Enhanced Raman Spectroscopic (SERS) study of saliva in the early detection of oral cancer.* **Kiang Wei Kho, Olivo Malini, Ze Xiang Shen, Khee Chee Soo.** 2005, Review of Scientific Instruments.

42. *Surface enhanced Raman spectroscopy (SERS): Potential applications for disease detection and treatment.* **Sarah McAughtrie, Karen Faulds, Duncan Graham.** s.l. : Journal of Photochemistry and Photobiology C: Photochemistry Reviews, 2014, Vol. 21.

43. *Characterization and noninvasive diagnosis of bladder cancer with serum surface enhanced Raman spectroscopy and genetic algorithms.* **Shaoxin Li, Linfang Li, Qiuyao Zeng, Yanjiao Zhang, Zhouyi Guo, Zhiming Liu, Mei Jin, Chengkang Su, Lin Lin, Junfa Xu and Songhao Liu.** 2015, Scientific Reports - Nature.

44. *Raman Technologies in Cancer Diagnostics.* **Lauren A. Austin, Sam Osseiran, and Conor L. Evans.** 2015, Analyst.

45. *Development of deep subsurface Raman spectroscopy for medical diagnosis and disease monitoring.* **Pavel Matousek, Nicholas Stone.** s.l. : Chemical Society Reviews, 2015, Vol. 45. 1794-1802.

46. *Endoscopic Raman Spectroscopy for Molecular Fingerprinting of Gastric Cancer: Principle to Implementation.* **Kim, Hyung Hun.** s.l. : BioMed Research International, 2015. 670121.

47. **Verisante.** Verisante. [Online] http://www.verisante.com/aura/medical_professional/.

48. *Introduction to Multivariate Analysis.* **Bloomfield, Peter.** 2008, http://www.stat.ncsu.edu/people/bloomfield/courses/st784/.

49. *Label-Free Imaging of Metal–Carbonyl Complexes in Live Cells by Raman Microspectroscopy.* **Konrad Meister, Johanna Niesel, Ulrich Schatzschneider, Nils Metzler-Nolte, Diedrich A. Schmidt, and Martina Havenith.** 2010, Angewandte Chemie International Edition.

50. *Automatic and objective oral cancer diagnosis by Raman spectroscopic detection of keratin with multivariate curve resolution analysis.* **Po-Hsiung Chen, Rintaro Shimada, Sohshi Yabumoto, Hajime Okajima, Masahiro Ando, Chiou-Tzu Chang, Li-Tzu Lee, Yong-Kie Wong, Arthur Chiou, & Hiro-o Hamaguchi.** 2016, Scientific Reports.

51. **Corporation, OriginLab.** *Origin 9.1 User Guide.* 2013.

52. *Model-based pre-processing in Raman spectroscopy of biological samples.* **Kristian Hovde Liland, Achim Kohler and Nils Kristian Afseth.** 2016, Journal of Raman Spectroscopy.

53. **Corporation, OriginLab.** *Origin 8.1 Getting Started Booklet.* 2009.

54. *Gold Nanoparticle Based Surface-Enhanced Raman Scattering Spectroscopy of Cancerous and Normal Nasopharyngeal Tissues Under Near-Infrared Laser Excitation.* **SHANGYUAN FENG, JUQIANG LIN, MIN CHENG, YONG-ZENG LI, GUANNAN CHEN, ZUFANG HUANG, YUN YU, RONG CHEN, and HAISHAN ZENG.** 2009, Applied Spectroscopy.

55. *Rapid detection of oral cancer using Ag–TiO2 nanostructured surface-enhanced Raman spectroscopic substrates.* **Chundayil Madathil Girish, Subramania Iyer, Krishnakumar Thankappan, V. V. Divya Rani, G. Siddaramana Gowd, Deepthy Menon, Shantikumar Naira and Manzoor Koyakutty.** 2013, Journal of Materials Chemistry B.

56. *From breast tissue diagnosis by Raman spectroscopy to femtosecond dynamics at the phospholipid membrane-water interface.* **Piotr Ciacka, Jakub Surmacki, Beata Brozek-Pluska, Joanna Jablonska, Radzislaw Kordek, Halina Abramczyk.** 2009, European Quantum Electronics Conference.

57. *Raman Imaging in Biochemical and Biomedical Applications. Diagnosis and Treatment of Breast Cancer.* **Brozek-Pluska, Halina Abramczyk and Beata.** 2012, Chemical Reviews.

58. *Raman Spectroscopy of Normal and Diseased Human Breast Tissues.* **McCreery, Christopher J. Frank and Richard L.** 1995, Analytical Chemistry.

59. *In vivo Raman spectroscopy for breast cancer: diagnosis in animal model.* **R. Bitar, M.A. Martins, D. Ribeiro, C. Carvalho, E.A.P. Santos, L.N.Z. Ramalho, F. Ramalho, H. Martinho, A.A. Martin.** 2008, Biomedical Optical Spectroscopy.

60. *Human breast tissue cancer diagnosis by Raman spectroscopy.* **H. Abramczyk, I. Placek, B. Brozek-Pluska, K. Kurczewski, Z. Morawiecc and M. Tazbir.** 2008, Spectroscopy.

61. *Diagnosing breast cancer by using Raman spectroscopy.* **Abigail S. Haka, Karen E. Shafer-Peltier, Maryann Fitzmaurice, Joseph Crowe, Ramachandra R. Dasari, and Michael S. Feld.** 2005, Proceedings of the National Academy of Sciences.

62. *Biochemical Correlation of Raman Spectra of Normal, Benign and Malignant Breast Tissues: A Spectral Deconvolution Study.* **M.V.P. Chowdary, K. Kalyan Kumar, Stanley Mathew, Lakshmi Rao, C. Murali Krishna, Jacob Kurien.** 2009, Biopolymers.

63. *Raman 'optical biopsy' of human breast cancer.* **Halina Abramczyk, Beata Brozek-Pluska, Jakub Surmacki, Joanna Jablonska-Gajewicz, Radzis1aw Kordek.** 2011, Progress in Biophysics and Molecular Biology.

64. *Raman spectroscopic analysis differentiates between breast cancer cell lines.* **A.C.S. Talari, C.A.Evans, I. Holen, R. E. Coleman and Ihtesham Ur Rehman.** 2001, Journal Raman Spectroscopy.

65. *Littrow Configuration Tunable External Cavity Diode Laser with Fixed Direction Output Beam.* **Scholten, C. J. Hawthorn and K. P. Weber and R. E.** 2001, Review of Scientific Instruments.

66. **Αθανάσιος, Αδάμος.** *Αλγόριθμοι ταξινόμησης δεδομένων υπερφασματικής απεικόνισης για την ανίχνευση , τμηματοποίηση και ταυτοποίηση χαρακτηριστικών διαγνωστικής σημασίας.* Χάνια : ΤΜΗΜΑ ΗΛΕΚΤΡΟΝΙΚΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ, 2006.

67. **Μαριάνθη, Παναγοπούλου.** *Ανάπτυξη Υποστρωμάτων SERS μέσω Αυτο-οργάνωσης Νανοσωματιδίων Αργύρου σε Περιοδικές Δομές.* Αθήνα : ΕΜΠ, 2011.

68. **Νικόλαος, Παντίσκος.** *Υπόστρωμα SERS νανοσωματιδίων αργύρου – Φαινόμενα γήρανσης, χρόνου εναπόθεσης, θερμικής ανόπτησης.* Αθήνα : ΕΜΠ, 2011.

69. **Τσαμποδήμου, Μαρία.** *Επιφανειακή ενίσχυση σκέδασης Ράμαν (SERS) μορίων ροδαμίνης από νανοσωματίδια αργύρου.* Αθήνα : ΕΜΠ, 2011.

70. *Raman microspectroscopy as a biomarking tool for in vitro diagnosis of cancer: a feasibility study.* **Aleksandra Pavićević, Sofija Glumac, Jelena Sopta, Ana Popović- Bijelić, Miloš Mojović, Goran Bačić.** 2012, Croatian Medical Journal.

71. *Raman spectroscopic analysis of human skin tissue sections Exvivo: Evaluation of the effects of tissue processing and dewaxing.* **S. M. Ali, F. Bonnier, A. Tfayli, H. Lambkin, K. Flynn, V. McDonagh, C. Healy, T.C. Lee, F.M. Lyng, Hugh J. Byrne.** 2013, Journal of Biomedical Optics.

72. *Biomedical Raman Spectroscopy.* **Motz, Jason T.** 2006, Presentation for Harvard Medical School and The Wellman Center for Photomedicine Massachusetts General Hospital.

73. *Diagnosing breast cancer by using Raman spectroscopy.* **Abigail S. Haka, Karen E. Shafer-Peltier, Maryann Fitzmaurice, Joseph Crowe, Ramachandra R. Dasari, and Michael S. Feld.** 2005. National Academy of Sciences.

74. *Diagnosing breast cancer using Raman spectroscopy: prospective analysis.* **Abigail S. Haka, Zoya Volynskaya, Joseph A. Gardecki, Jon Nazemi, Robert Shenk, Nancy Wang, Ramachandra R. Dasari, Maryann Fitzmaurice, Michael S. Feld.** 2009, Journal of Biomedical Optics 14(5).

75. *The Use of Raman Spectroscopy in Cancer Diagnostics.* **Katherine A. Bakeev, Robert Thomas, Robert Chimenti, Michael Claybourn.** 2013, Spectroscopy online, Volume 28, Issue 9.

76. *Raman spectroscopy for physiological investigations of tissues and cell.* **Thomas Huser, James Chanc.** 2015, Advanced Drug Delivery Reviews.

77. *Introductory Raman Spectroscopy.* **John R. Ferraro, Kazuo Nakamoto and Chris W. Brown.** 2003, Elsevier.

78. **Αλέξανδρος, Κατσαούνης.** Εισαγωγή στις φασματομετρικές τεχνικές. [Online] 2016. http://www.chemeng.upatras.gr/sites/default/files/users/alex.katsaounis/Chapter%204.pdf.

79. **Maniu, D.** Raman Spectroscopy. [Online] http://www.phys.ubbcluj.ro/~dana.maniu/OS/BS_5.pdf.

80. **Hirata, So.** General issues of spectroscopies [presentation]. *Department of Chemistry, University of Illinois at Urbana-Champaign.* [Online]

81. **Καραγιάννη, Μ.** *Φυσική μελέτη ακτινοβλίας LASER.* [Online] http://e-physics.teipir.gr/HN/physics1_files/laser.pdf.

82. **Γκιώνης, Βασίλης Χρυσικός, Γεώργιος Δ.** *Συνεστιακή (confocal) μικροσκοπία Raman Βασικές αρχές και εφαρμογές [Τεχνική Αναφορά ΤΔ006.01].* Αθήνα : http://www.eie.gr/nhrf/institutes/tpci/researchteams/mspc/epidiktikesefarmoges/td006.pdf, 2008.

83. **Θεόδωρος, Δρ. Γκανέτσος.** *Φασματοσκοπία Raman και εφαρμογές.* [Online] 2013. http://users.teilam.gr/~etzoutzis/Raman%20spectroscopy%20and%20applications.pdf.

84. *Practical Group Theory and Raman Spectroscopy, Part I: Normal Vibrational Modes.* **Tuschel, David.** 2014, Spectroscopy online.

85. *Surface-Enhanced Raman Scattering, Physics and Applications.* **Katrin Kneipp, Martin Moskovits, Harald Kneipp.** 2006, Topics in Applied Physics, Volume 103.

86. *Actively Targeted In Vivo Multiplex Detection of Intrinsic Cancer Biomarkers Using Biocompatible SERS Nanotags.* **U. S. Dinish, Ghayathri Balasundaram, Young-Tae Chang and Malini Olivo.** 2014, Scientific Reports - Nature.

87. **Κ.Α. Χαριτίδης, Ι.Α. Καρτσωνάκης, Ε. Μηλιώνη.** *Νανοδομές μηδενικών διαστάσεων: Νανοσωματίδια [παρουσίαση].* Αθήνα : ΕΜΠ, Σχολή Χημικών μηχανικών, Τομέας επιστήμης και τεχνικής των υλικών, 2014.

88. *Raman Spectroscopy in Nanomedicine: Current Status and Future Perspectives.* **Hugh J. Byrne, Mark E. Keating.** 2013, NanoMedicine, 8.

89. *Surface-enhanced Raman scattering in cancer detection and imaging.* **Marc Vendrell, Kaustabh Kumar Maiti, Kevin Dhaliwal and Young-Tae Chang.** 2013, Trends in Biotechnology, Volume 31, Number 4.

90. *Biocompatible surface-enhanced Raman scattering nanotags for in vivo cancer detection.* **Animesh Samanta, Santanu Jana, Raj Kumar Das and Young Tae Chang.** 2014, Nanomedicine.

91. *Applications and limits of Raman spectroscopy in the study of colonic and pulmonary malformations.* **Codrich, Daniela.** 2006, UNIVERSITA' DEGLI STUDI DI TRIESTE.

92. *Raman spectroscopy of lipids: a review.* **K. Czamara, K. Majzner, M. Z. Pacia, K. Kochan, A. Kaczor and M. Baranska.** 2014, Journal of Raman Spectroscopy.

93. *Raman and coherent anti-Stokes Raman scattering microscopy studiesof changes in lipid content and composition in hormone-treated breast and prostate cancer cells.* **et.al, Mariana C. Potcoava.** 11, s.l. : Journal of Biomedical Optics, 2014, Vol. 19. 111605.

94. *Collagen VI in cancer and its biological mechanisms.* **Peiwen Chen, Matilde Cescon , Paolo Bonaldo.** 7, Padova : Trends in Molecular Medicine, 2013, Vol. 19.