ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Φορητός FT-NIR φασματικός αισθητήρας για την ανίχνευση χημικών πρόδρομων εκρηκτικών ουσιών με την χρήση ανεπτυγμένων αλγοριθμικών μοντέλων

Αδαμαντία Μαρία Γραμματικάκη

Επιβλέπων:

Ηρακλής Αβραμόπουλος

Καθηγητής Ε.Μ.Π

Αθήνα, Μάρτιος 2022

Περίληψη

Στην εργασία αυτή παρουσιάζεται η ανάπτυξη και αξιολόγηση διάφορων μοντέλων κατηγοριοποίησης, με στόχο να ανιχνεύουν και να ταξινομούν συγκεκριμένους χημικούς προδρόμους εκρηκτικών. Οι πρόδρομοι που επιλέξαμε είναι το νιτρικό αμμώνιο (*ammonium nitrate*), η ουρία (*urea*), το νιτρικό κάλιο (*potassium nitrate*) και το νιτρικό νάτριο (*sodium nitrate*), οι οποίοι χρησιμοποιούνται συχνά σαν λιπάσματα ή συντηρητικά, ενώ σε πολλές χώρες η παραγωγή, χρήση και πώλησή τους υπόκεινται σε περιορισμούς. Η αξιολόγηση των ουσιών πραγματοποιείται με βάση το φάσμα τους στο εγγύς υπέρυθρο, το οποίο παράγεται με την βοήθεια ενός φορητού φασματογραφικού αισθητήρα FT-NIR. Η τεχνική *Fourier-transform* (FT) καθώς και η φασματοσκοπία εγγύς υπερύθρου (NIR) έχουν πολλά προτερήματα, τα οποία φαίνονται ιδιαίτερα χρήσιμα στην συγκεκριμένη εφαρμογή. Συνοπτικά, επιτρέπουν επί τόπου αναλύσεις μεγάλων ταχυτήτων, δεν χρειάζεται προηγούμενη προετοιμασία του δείγματος για την ανάλυσή του, μπορούν να αναλυθούν ουσίες σε πολλές διαφορετικές φυσικές καταστάσεις και τέλος, η σύνδεση της πηγής του αισθητήρα με μία οπτική ίνα καθιστά το σύστημα ασφαλέστερο, καθώς ο χρήστης δεν χρειάζεται να έρθει σε άμεση επαφή με την ουσία. Τα μοντέλα που αναπτύχθηκαν αφορούν τους αλγορίθμους μηχανικής μάθησης *Random Forest* και *Support Vector Machine*, μόνους αλλά και σε συνδυασμό με τον αλγόριθμο PCA για μείωση της διαστατικότητας, καθώς και με έναν αλγόριθμο βασισμένο στον συντελεστή συσχέτισης HQI. Τα NIR δεδομένα προεπεξεργάστηκαν με τις μεθόδους *Standard Normal Variate* και *Savitzky-Golay (*1st *derivative*). Για την εκπαίδευση και αξιολόγηση των μοντέλων παρήχθησαν 2911 δείγματα προδρόμων αλλά και "αρνητικών ουσιών", ενώ μετρικές αξιολόγησης αποτέλεσαν οι καμπύλες Open-Set *Classification Rate* και *Binary Open-Set Classification Rate*, καθώς και η μέγιστη ακρίβεια. Τελικά, το μοντέλο *Random Forest* σε συνδυασμό με την τεχνική προεπεξεργασίας Standard *Normal Variate* έδειξε τα καλύτερα αποτελέσματα, επιτυγχάνοντας περισσότερο από 83% *Correct Classification Rate* για 1% *False Positive Rate* και μέγιστη ακρίβεια πάνω από 96%.

Λέξεις κλειδιά εγγύς υπέρυθρο, πρόδρομοι εκρηκτικών, φασματοσκοπικός αισθητήρας, HQI, *Random Forest*, *Support Vector Machine*, FT-NIR

Abstract

This project presents the development and evaluation of different classification models, which aim to detect and classify specific chemical precursors of explosives. The precursors we chose are Ammonium Nitrate, Urea, Potassium Nitrate and Sodium Nitrate, which are often used as fertilizers or preservatives, while in many countries their porduction,sale and use face legal limitations. The evaluation of the substances is based on their near-infrared (NIR) spectrm, which is produced with the help of a portable spectroscopic FT-NIR sensor. The Fourier-transform (FT) technique, as well as NIR spectroscopy consider many advantages, which seem particularly useful in this appication. In short, they allow on-site high-speed analyzes, no prior sample manipulation is required, substances can be analyzed in many different physical forms, and finally, attaching an optical fiber to the sensor source makes the system safer as the user does not need to come in direct contact with the substance. The models were developed using machine learning algorithms, including Random Forest and Support Vector Machine, alone but also in combination with the PCA algorithm for dimensionality reduction, as well as an algorithm based on the HQI correlation coefficient. NIR data were preprocessed using the Standard Normal Variate and Savitzky-Golay (1st derivative) methods. For the training and evaluation of the models, 2911 samples of precursors and "negative substances" were produced, while the evaluation metrics regarded the Open-Set Classification Rate and binary Open-Set Classification Rate curves, as well as the maximum accuracy. Finally, the Random Forest model in combination with the Standard Normal Variate pretreatment technique showed the best results, yielding more than 83% Correct Classification Rate for 1% False Positive Rate and maximum accuracy over 96%.

Keywords near-infrared, precursors of explosives, spectroscopic sensor, HQI, Random Forest, Support Vector Machine, FT-NIR,

Ευχαριστίες

Η παρούσα διπλωματική εκπονήθηκε στο Εργαστήριο Φωτονικών Επικοικωνιών του Εθνικού Μετσόβιου Πολυτεχνείου την περίοδο 2021-2022. Θα ήθελα να ευχαριστήσω τον καθηγητή κ. Ηρακλή Αβραμόπουλο για την ευκαιρία που μου έδωσε να διαβάσω και να εργαστώ στο εργαστήριο και να έρθω σε επαφή με τον τομέα της φωτονικής, καθώς και να συνεργαστώ με τα άτομα που εργάζονται σε αυτό. Θα ήθελα να ευχαριστήσω επίσης τον ερευνητή Δρ. Λευτέρη Γουναρίδη και τον ερευνητή Αδάμ Ραπτάκη που παρά τον ιδιαίτερα υψηλό φόρτο εργασίας την περίοδο εκείνη, μου έδειξαν εμπιστοσύνη και δέχτηκαν να αναλάβουν την καθοδήγησή μου, γεγονός πραγματικά πολύ σημαντικό για εμένα. Τους ευχαριστώ ακόμα που με καθοδήγησαν και μου προσέφεραν γνώσεις βασικές για την εκπόνηση της διπλωματικής και που κατασκεύασαν τον φωτονικό αισθητήρα χωρίς τον οποίο δεν θα υπήρχε η εργασία αυτή.

Ευχαριστώ επίσης τον ερευνητή και φίλο Στάθη Ανδριανόπουλο για την καλή επικοινωνία και την έμμεση, αλλά πολύ σημαντική βοήθειά του, στο πλαίσιο της οποίας με έφερε και σε επαφή με τον πλέον φίλο μου Αντρέα Αθανασόπουλο, ο οποίος έπαιξε πολύ σημαντικό ρόλο στην εξέλιξη της εργασίας μου. Ένα σημαντικό εργαλείο που εντάσσεται στην παρούσα εργασία αφορά την Μηχανική Μάθηση, και η διαδικασία να το κατανοήσω και να το χρησιμοποιήσω σωστά θα ήταν φοβερά πιο δύσκολη, χωρίς την συμβολή του Αντρέα. Η βοήθεια του Αντρέα έχει σίγουρα ένα πολύ σημαντικό ποιοτικό αποτύπωμα στην διπλωματική μου εργασία.

Τέλος, θα ήθελα να ευχαριστήσω όλους τους φίλους και την μητέρα μου, που ήταν πάντα παρόντες και υποστηρικτικοί και πάντα αποτελούν το θεμέλιο κάθε μου προσπάθειας.

Στην μητέρα μου,

Βασιλική

Περιεχόμενα

1. Εισαγωγή

Η φασματοσκόπια εγγύς υπερύθρου (NIR *spectroscopy*) βασίζεται στην απορρόφηση των ουσιών μεταξύ 800 και 2500 nm. Τις τελευταίες δεκαετίες χρησιμοποιείται ολοένα περισσότερο σε ποικίλες εφαρμογές καθότι προσφέρει δυνατότητες οι οποίες την καθιστούν ιδιαίτερα πολύτιμο και ευέλικτο εργαλείο. Χαρακτηριστικά, η NIR φασματογραφία αποδεσμεύει τους αναλυτές από την μεταφορά της υπό μελέτης ουσίας σε κάποιο εργαστήριο, καθώς προσφέρεται για επί τόπου αναλύσεις, όπως επίσης έχει την δυνατότητα να αναλύει ουσίες σε διάφορες φυσικές καταστάσεις, χωρίς οποιαδήποτε προηγούμενη μεταχείριση τους. Έπειτα, συχνά υπέρυθροι φασματογράφοι συνδέονται με μια οπτική ίνα, το οποίο προσφέρει στους χρήστες πιο ασφαλείς μετρήσεις. Αξίζει επίσης να σημειωθεί ότι συγκεκριμένα οι NIR φασματογράφοι είναι συμβατοί με οικονομικότερες οπτικές ίνες και με υψηλότερο σηματοθορυβικό λόγο σε σχέση με αυτές του μέσου υπέρυθρου. Ο λόγος που παλαιότερα η NIR φασματοσκοπία δεν αξιοποιούνταν όσο σήμερα είναι η ποιότητα των φασμάτων που παράγει. Συγκεκριμένα οι ζώνες που παρατηρούνται σε ένα NIR φάσμα είναι ευρείες, σχετικά ασθενείς και συχνά αλληλοκαλυπτόμενες. Αυτό συμβαίνει γιατί οι ζώνες αποδίδονται σε χημικά φαινόμενα που αποτελούν υπερτονικές και ζώνες συνδυασμού μεταβάσεων δονητικών καταστάσεων, των οποίων οι θεμελιώδεις μεταβάσεις βρίσκονται στο μέσο υπέρυθρο. Η δυσκολία αυτή αντιμετωπίστηκε με την ανάπτυξη των υπολογιστών και της χημειομετρίας, επιστημονικό πεδίο που αναφέρεται ειδικά σε υπολογιστικές τεχνικές για την ερμηνεία και μεταχείριση φασματικών δεδομένων. Για την εξαγωγή χρήσιμης πληροφορίας από NIR φάσματα είναι συχνά απαραίτητη η χρήση τεχνικών χημειομετρίας, μηχανικής μάθησης και νευρωνικών δικτύων.

Στην παρούσα εργασία, κληθήκαμε να αντιμετωπίσουμε το πρόβλημα της ταυτοποίησης κάποιων χημικών πρόδρομων εκρηκτικών, δηλαδή χημικών ουσιών από τις οποίες μπορούν να παραχθούν εκρηκτικά. Ανά τον καιρό καταγράφονται διάφορες χημικές ουσίες που παράγουν χειροποίητα εκρηκτικά και πολλοί φορείς ανανεώνουν σχετικές λίστες. Οι χημικοί πρόδρομοι που επιλέξαμε να ανιχνεύουμε είναι το νιτρικό αμμώνιο, το νιτρικό νάτριο, η ουρία και το νιτρικό κάλιο, καθότι περιέχουν πληροφορία στο εγγύς υπέρυθρο και βρίσκονται σε στερεά μορφή. Για την ταυτοποίηση τους κατασκευάστηκε αντίστοιχα ένας FT-NIR αισθητήρας και το αντίστοιχο λογισμικό που θα αξιολογεί την πληροφορία που παράγει ο αισθητήρας και που περιγράφεται αναλυτικά στην παρούσα εργασία. Η τεχνολογία Fourier Transform (FT) αποτελεί σημαντική ανάπτυξη στην οργανολογία της φασματοσκοπίας υπερύθρου. Με την χρήση ενός συμβολομέτρου, ο φασματογράφος μπορεί σε μία λήψη να παράγει και να ακτινοβολεί την ουσία με ένα φάσμα μηκών κύματος. Αυτή η τεχνική έχει βασικό πλεονέκτημα την υψηλή ταχύτητα μέτρησης του φάσματος της ουσίας, με αποτέλεσμα να δίνει την δυνατότητα σε αλλεπάλληλες μετρήσεις.

Για την ανάπτυξη του λογισμικού κατασκευάσαμε και συγκρίναμε μεταξύ τους διαφορετικά μοντέλα τα οποία αφενός διακρίνουν αν η ουσία υπό μελέτη είναι κάποιος πρόδρομος, αφετέρου μπορούν να αξιολογούν την ταυτότητα του προδρόμου. Τα μοντέλα αυτά βασίστηκαν σε δυο αλγόριθμους μηχανικής μάθησης, τον *Random Forest* και τον *Support Vector Machine*, οι οποίοι συνδυάστηκαν και με την μέθοδο *Principal Component Analysis*, όπως επίσης και στην χρήση ενός συντελεστή συσχέτισης και την μετρική *Hit Quality Index*. Το ενδιαφέρον σε αυτό το πρόβλημα, όπως και σε αρκετά προβλήματα που καλείται να αντιμετωπίσει η μηχανική μάθηση, ήταν η προσπάθεια να ελαχιστοποιούμε την επίδραση διαφόρων παραγόντων οι οποίοι καθιστούν απαιτητική την ανάπτυξη αποτελεσματικών μοντέλων. Αρχικά, η φύση του προβλήματος, το οποίο αποτελεί ένα *open-set* πρόβλημα, ήδη συνδέεται με αρκετές δυσκολίες στην ικανότητα γενίκευσης του μοντέλου. Με αλλά λόγια, το γεγονός πως ο αισθητήρας θα έρχεται σε επαφή με πληθώρα ουσιών τις οποίες είναι αδύνατο να φέρουμε σε επαφή με το μοντέλο εκ των προτέρων, δυσκολεύει την δυνατότητα του μοντέλου, τα αποτελέσματα που δίνει στο πλαίσιο του εργαστηρίου να τα ανάγει στο πλαίσιο του πραγματικού κόσμου. Μια άλλη δυσκολία έρχεται από το ίδιο το NIR φάσμα. Διάφοροι παράγοντες καθιστούν την αναπαραγωγή ενός NIR φάσματος δύσκολη. Αυτό σημαίνει πως με την μεταβολή διαφόρων συνθηκών κατά την μέτρηση, η ίδια ουσία μπορεί να παράγει φάσματα με αρκετές διάφορες μεταξύ τους. Το πρόβλημα κατά το οποίο τα δεδομένα στα οποία θα εκτεθεί το μοντέλο έρχονται

από διαφορετικό σύνολο από αυτό με το οποίο έχει εκπαιδευτεί, συχνά αναφέρεται ως *distribution shift*. Μία ακόμα πρόκληση που προκύπτει σε αυτή την εφαρμογή είναι «η κατάρα της διαστατικότητας» (*curse of dimensionality*). Πρόκειται για διάφορες επιπλοκές που ενδέχεται να δημιουργούνται στα μοντέλα μηχανικής μάθησης όταν μεταχειρίζονται δεδομένα μεγάλων διαστάσεων. Εδώ καθότι διάσταση των φασματικών δεδομένων αποτελεί το πλήθος των μηκών κυμάτων στα οποία ο φασματογράφος καταγράφει την απορρόφηση, καταλαβαίνουμε ότι το παραπάνω φαινόμενο θα μπορούσε να επηρεάσει την λειτουργία των μοντέλων. Για την αντιμετώπιση όλων αυτών των ζητημάτων, αναζητήσαμε και επινοήσαμε τεχνικές, οι οποίες περιγράφονται αναλυτικά στα επόμενα κεφάλαια.

2. Θεωρία Φασματοσκοπίας υπέρυθρου

Η φασματοσκοπία υπέρυθρου είναι αναλυτική τεχνική που χρησιμοποιείται περισσότερο από έναν αιώνα. Το υπέρυθρο φάσμα εκτείνεται από 0.800 έως 1000 μm και συνηθίζεται να διαιρείται σε τρείς φασματικές υποπεριοχές. Αυτές είναι το κοντινό ή εγγύς υπέρυθρο (*near infrared* ή NIR), το οποίο βρίσκεται μεταξύ 800 – 2500 nm, το μέσο υπέρυθρο (*mid-infrared* ή *mid*-IR), που αντιστοιχεί σε μήκη κύματος 2500 – 25,000 nm και το μακρινό ή άπω υπέρυθρο (*far infrared*) μεταξύ 25 μm - 1 mm, εκ των οποίων οι πιο συχνά χρησιμοποιούμενες περιοχές είναι αυτές του κοντινού και μέσου υπέρυθρου. Στην μελέτη υπέρυθρου είναι σύνηθες η συχνότητα να εκφράζεται μέσω του κυματαριθμού (cm^{-1}), μέγεθος το οποίο είναι αντίστροφο του μήκους κύματος [11].

Η υπέρυθρη φασματοσκοπία είναι μία τεχνική δονητικής φασματοσκοπίας και προκύπτει από την αλληλεπίδραση της εκπεμπόμενης υπέρυθρης ηλεκτρομαγνητικής ακτινοβολίας με τους δεσμούς της ουσίας. Οι δονήσεις που λαμβάνουν χώρα σε ένα μόριο διακρίνονται σε δύο κατηγορίες: τις δονήσεις τάσης (*stretching modes*), όταν τα άτομα πλησιάζουν και απομακρύνονται μεταξύ τους κατά μήκος του δεσμού και τις δονήσεις κάμψης (*bending modes*), όταν τα άτομα των δεσμών κινούνται με τέτοιο τρόπο, ώστε να αλλάζει η γωνία των δεσμών. Κάθε ουσία απορροφά συγκεκριμένα μήκη κύματος, γεγονός που εξαρτάται από την χημική της σύνθεση. Η παρακάτω σχέση δίνει την εξάρτηση της συχνότητας απορρόφησης ενός δεσμού:

Equation 1

$$u = \frac{h}{2\pi}\sqrt{\frac{k}{\mu}},$$

όπου u είναι η συχνότητα, h η σταθερά του *Planck*, k η σταθερά δύναμης, μ η ανηγμένη μάζα.

Παρατηρούμε πως η συχνότητα απορρόφησης εξαρτάται από την δύναμη και τις μάζες των ατόμων που σχηματίζουν τον δεσμό.

Η φασματοσκοπία εγγύς υπερύθρου (NIR *Spectroscopy*) είναι η φασματοσκοπία στην περιοχή των 12,500–4,000 cm^{-1} και επιτρέπει την μελέτη των υπερτονικών (*overtones*) και των αρμονικών δονήσεων ή δονήσεων συνδυασμού (*harmonic or combination bands*). Στο μέσο υπέρυθρο φάσμα (MIR), βρίσκονται οι θεμελιώδεις ζώνες απορρόφησης των μορίων, οι βασικές μεταβολές στη δόνησή τους, ενώ στο άπω υπέρυθρο εξετάζονται οι δονήσεις βαρέων ατόμων. Στην NIR περιοχή τα φασματικά χαρακτηριστικά των μορίων είναι λιγότερα και συνεπώς το φάσμα είναι αρκετά δύσκολο να μελετηθεί, ωστόσο τις τελευταίες δεκαετίες με την ανάπτυξη των υπολογιστών, η NIR φασματοσκοπία χρησιμοποιείται ολοένα και περισσότερο, ενώ έχει αναπτυχθεί νέος επιστημονικός κλάδος, η χημειομετρία, που ασχολείται με την εξαγωγή πληροφορίας από την επεξεργασία τέτοιων περίπλοκων χημικών δεδομένων. Συνολικά για την εξαγωγή της χρήσιμης φασματικής πληροφορίας από NIR δεδομένα, χρησιμοποιούνται τόσο χημειομετρικές τεχνικές όσο και τεχνικές μηχανικής μάθησης και νευρωνικά δίκτυα [2].

2.1 Ζώνες απορρόφησης NIRS

Οι απορροφήσεις που παρατηρούνται στην NIR περιοχή προέρχονται από υπερτονικές ζώνες απορρόφησης (*overtones*) και ζώνες συνδυασμού (*combination bands*). Υπερτονικές, στην δονητική φασματοσκοπία, είναι οι φασματικές ζώνες που προκύπτουν από την μετάβαση ενός μορίου από την θεμελιώδη κατάσταση ν=0 σε μία κατάσταση μεγαλύτερη της πρώτης, Δν > 1. Τα ενεργειακά άλματα ν=0 → ν=2 και ν=0 → ν=3 αποτελούν την πρώτη και τη δεύτερη υπερτονική αντίστοιχα. Οι ζώνες συνδυασμού συμβαίνουν όταν περισσότερες από μία δονητικές θεμελιώδεις καταστάσεις διεγείρονται ταυτόχρονα. Σε αυτές τις περιπτώσεις η συχνότητα απορρόφησης θα είναι ίση με κάποιον γραμμικό συνδυασμό των συχνοτήτων των επιμέρους ενεργειακών μεταβάσεων, παραδειγματικά $\nu_1+\nu_2$, $\nu_1-\nu_2$, $\nu_1+2\nu_2$. Όπως οι υπερτονικές, έτσι και οι ζώνες συνδυασμού παρουσιάζουν χαρακτηριστικά χαμηλές εντάσεις, ενώ όσο οι απορροφήσεις είναι μεγαλύτερων τάξεων το φασματικό σήμα τείνει να είναι περισσότερο ασθενές. Κατά συνέπεια το NIR φάσμα δομείται από ευρείες, υπερκαλυπτόμενες και ασθενείς ζώνες απορρόφησης (Figure 1) και έτσι η απόδοση των απορροφήσεων σε συγκεκριμένα χημικά χαρακτηριστικά είναι δύσκολη [2].

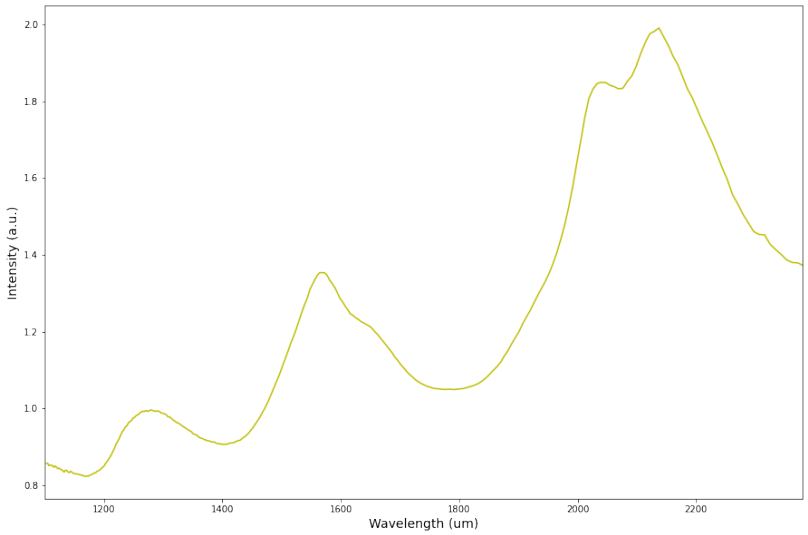

Figure 1 Φάσμα NIR Νιτρικής Αμμωνίας

Στον παρακάτω πίνακα (Table 1) φαίνονται τα μήκη κύματος, ο κυματαριθμός και οι εντάσεις των ζωνών που αφορούν την θεμελιώδη, την πρώτη, δεύτερη και τρίτη υπερτονική για τον δεσμό άνθρακα-υδρογόνου που περιέχεται στο χλωροφόρμιο. Παρατηρείται πως όσο μεγαλώνει η τάξη της υπερτονικής, τόσο ασθενέστερη και η απορρόφηση.

Table 1 Τα μήκη κύματος, οι κυματαριθμοί και οι εντάσεις των ζωνών που προκύπτουν από το θεμελιώδες, το πρώτο, δεύτερο και τρίτο υπέρτονο για τον δεσμό άνθρακα-υδρογόνου που περιέχεται στο χλωροφόρμιο.

	Band position/nm	*Band position/cm^{-1}*	*Intensity/cm^{2}mol^{-1}*
υ	3290	3040	25000
2υ	1693	5907	1620
3υ	1154	8666	48
4υ	882	11338	1.7
5υ	724	13831	0.15

Οι περισσότερες ζώνες απορρόφησης στο NIR, πηγάζουν από λειτουργικές ομάδες που περιέχουν άτομα υδρογόνου (π.χ. OH, CH, NH). Επιπλέον, μπορούν να παρατηρηθούν δεύτερης τάξης υπερτονικά από δονήσεις τάσης των δεσμών C = O και C ≡ N. Στην παρακάτω εικόνα (Figure 2) παρουσιάζονται οι σημαντικότερες ζώνες απορρόφησης στο NIR και τα μήκη κύματος όπου αυτές συναντώνται. Οι μαύρες γραμμές υποδεικνύουν τις περιοχές απορρόφησης.

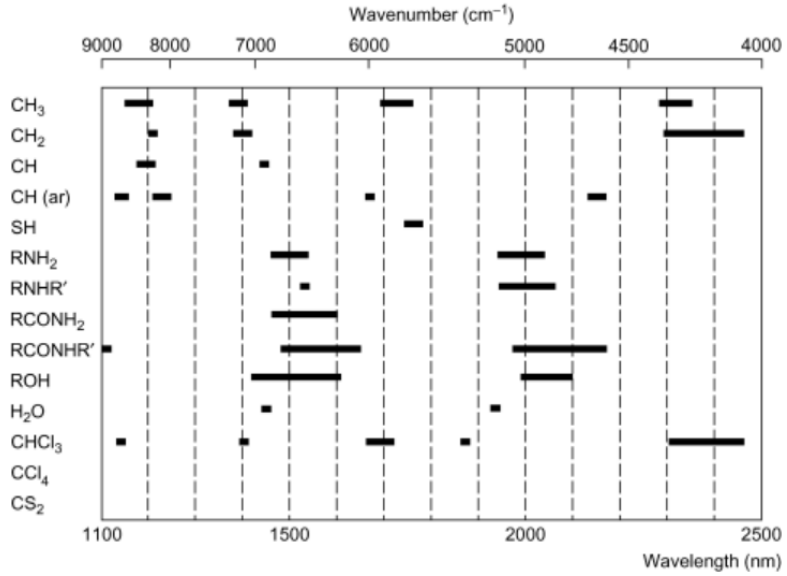

Figure 2 Μήκη κύματος των σημαντικότερων απορροφήσεων στο εγγύς υπέρυθρο

2.2 Πλεονεκτήματα NIR Φασματοσκοπίας

Στο πλαίσιο των εφαρμογών, η NIR φασματογραφία έχει πολλά χρήσιμα χαρακτηριστικά. Αρχικά αποτελεί ένα μη καταστρεπτικό και in situ εργαλείο ανάλυσης. Μπορεί να είναι φορητή συσκευή για αναλύσεις πεδίου και ο χρόνος που απαιτεί για τον έλεγχο είναι πολύ μικρός. Επίσης μπορεί να εξετάζει εσωτερικά το δείγμα, χωρίς να επεμβαίνει σε αυτό. Η δυνατότητα των φασματογράφων εγγύς υπερύθρου να συνδέεται με μία οπτική ίνα για την λήψη των φασματικών δεδομένων, επίσης αποτελεί ένα βασικό πλεονέκτημα, καθώς προσφέρει την δυνατότητα να αναλυθούν επικίνδυνες ή απομακρυσμένες ουσίες. Επιπλέον, είναι δυνατή η εφαρμογή NIR φασματοσκοπίας σε δείγματα διαφόρων καταστάσεων, σχημάτων και παχών. Υπάρχουν και πιο ειδικά πλεονεκτήματα της NIR φασματοσκοπίας έναντι της IR φασματοσκοπίας που θα μπορούσαν να αφορούν την παρούσα εργασία καθώς και μετέπειτα εξέλιξή της.

Οι ίνες για την IR φασματοσκοπία είναι λιγότερο ανθεκτικές και πολύ ακριβότερες. Αυτό δίνει προβάδισμα στην NIRS, όπου πιο οικονομικός και μακρύς καθετήρας μπορεί να χρησιμοποιηθεί, καθιστώντας δυνατή την ανάλυση σε πιο επικίνδυνα περιβάλλοντα. Έπειτα, οι δεσμοί του νερού δίνουν αρκετά ισχυρότερο φάσμα στην IRS, σε αντίθεση με την NIRS, η οποία και κρίνεται καταλληλότερη για ανάλυση υδατικών διαλυμάτων. Τέλος, όλα τα οπτικά υλικά που χρησιμοποιούνται για την NIRS είναι φθηνότερα από αυτά για την IRS [2].

2.3 Τρόπος λειτουργίας NIR φασματογράφων

Τα βασικά τμήματα ενός NIR φασματογράφου είναι η πηγή ακτινοβολίας, ο επιλέκτης μήκους κύματος (*wavelength selector*) ή το συμβολόμετρο (*interferometer*) και ο ανιχνευτής, τα οποία συνδέονται μεταξύ τους με οπτικά υλικά. Αντίστοιχα, οι δύο βασικές κατηγορίες φασματογράφων είναι αυτή της διασποράς μήκους κύματος και αυτή του μετασχηματισμού Fourier (FT). Η πρώτη, πλέον λιγότερο χρησιμοποιούμενη, βασίζεται στον επιλογέα μήκους κύματος (*wavelength selector*), ο οποίος επιτρέπει ένα πολύ στενό φάσμα ακτινοβολίας να φτάσει στον ανιχνευτή σε συγκεκριμένο χρόνο. Αυτή η πρακτική συνδέεται με διάφορα μειονεκτήματα, όπως η μικρή ταχύτητα σάρωσης, καθώς και η χαμηλή απόδοσή τους. Η δεύτερη κατηγορία είναι η πλέον χρησιμοποιούμενη και σε αντίθεση με τα φίλτρα μηκών κύματος, ένα συμβολόμετρο επιτρέπει την ταυτόχρονη εμφάνιση ενός φάσματος μηκών κύματος στον ανιχνευτή και στη συνέχεια το φάσμα αποκτάται μέσω μετασχηματισμού Fourier. Τα FT-IR φασματόμετρα αποτελούν τα κύρια όργανα μέτρησης του υπέρυθρου φάσματος μέχρι σήμερα. Τις δύο τελευταίες δεκαετίες, μελετώνται τρόποι, ώστε να μειωθούν οι διαστάσεις των FT-IR και να αναπτυχθούν ακόμη και φορητές διατάξεις.

Στην πλειοψηφία των FT-IR φασματογράφων χρησιμοποιείται ως πυρήνας το συμβολόμετρο *Michelson*, για τον λόγο πως συνδυάζει δύο βασικά πλεονεκτήματα. Το πρώτο πλεονέκτημα είναι το *Fellgett* ή πολλαπλό πλεονέκτημα. Αφορά την δυνατότητα του συμβολομέτρου να εξετάζει ολόκληρο το φάσμα με μόνο μία σάρωση. Αυτό βελτιώνει σημαντικά τον χρόνο μέτρησης του φάσματος όπως ενδεικτικά γίνεται φανερό από τον Griffiths στο βιβλίο *Fourier Transform Infrared Spectrometry* [4]. Εκεί αναφέρεται πως απαιτούνται τουλάχιστον 30 λεπτά για την λήψη ενός φάσματος καλής ανάλυσης στο μέσο υπέρυθρο με την χρήση πρίσματος ή φράγματος. Αυτός ο χρόνος είναι ιδιαίτερα μεγάλος συγκριτικά με τα λίγα δευτερόλεπτα που χρειάζεται ένας FT-IR φασματογράφος και οφείλεται στις πολλαπλές σαρώσεις που απαιτούνται, καθώς με κάθε σάρωση μετράται μόνο ένα μικρό κομμάτι φάσματος. Το δεύτερο πλεονέκτημα είναι το πλεονέκτημα *Jacquinot* ή συμβολομετρικό πλεονέκτημα και αφορά την υψηλή συγκέντρωση φωτός στον ανιχνευτή σε σύγκριση με τον συμβατικό φασματογράφο. Στους συμβατικούς φασματογράφους χρησιμοποιούνται πρίσματα και σχισμές εισόδου, τα οποία περιορίζουν σημαντικά το φως που εκπέμπεται από την πηγή. Η απουσία αυτών στον FT-IR επιτρέπει μεγαλύτερο σηματοθορυβικό λόγο και συνεπώς μεγαλύτερη ευαισθησία του οργάνου σε μικρές απορροφήσεις, πλεονέκτημα που ενδιαφέρει σημαντικά την φασματοσκοπία εγγύς υπερύθρου [3].

2.3.1 Συμβολόμετρο Michelson

Το συμβολόμετρο *Michelson* σχεδιάστηκε από τον φυσικό Albert Abraham Michelson (1852-1931) το 1891 και είναι ταυτόχρονα το πιο συχνά χρησιμοποιούμενο εργαστηριακά φασματόμετρο, αλλά

και το παλαιότερο. Ο πυρήνας του σχηματίζεται από ένα σταθερό (fix*ed mirror*) και ένα κινούμενο (*moving mirror*) κάτοπτρο, πάνω στα οποία κατευθύνεται ταυτόχρονα η πολυχρωμική δέσμη αφού πρώτα έχει περάσει από έναν διαχωριστή δέσμης (*beamsplitter*). Οι δύο δέσμες είναι ίσης έντασης και διαδρομές είναι κάθετες μεταξύ τους. Η μία καταλήγει στον σταθερό και η άλλη στον κινούμενο καθρέπτη, ο οποίος λόγω της κίνησης, προσαρμόζει το μήκος διαδρομής της δεύτερης δέσμης. Οι δέσμες αυτές συμβάλλουν ξανά, ενώ έχει εισαχθεί μία διαφορά μήκους διαδρομής. Η διαφορά διαδρομής μεταξύ των δεσμών που έχει εισαχθεί κατά τη διάρκεια της σάρωσης οδηγεί σε περιοδικά εναλλασσόμενες παρεμβολές (διαφορές φάσεων), από τις οποίες μπορεί να ανακατασκευαστεί το φάσμα. Μία απλή σχεδίαση του συμβολόμετρου Michelson φαίνεται στην παρακάτω εικόνα (Figure 3).

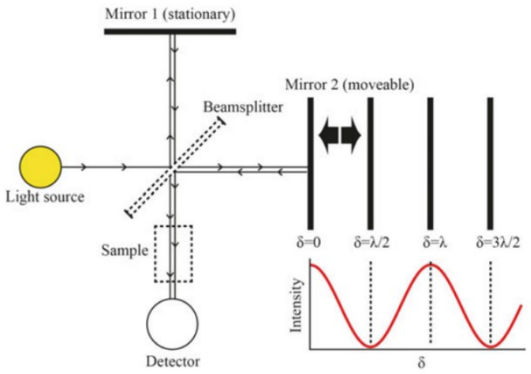

Figure 3 Σχηματική αναπαράσταση ενός συμβολόμετρου michelson [2]

Οι συμβατικές φασματομετρικές συσκευές απαιτούν συνεχώς βαθμονόμηση προκειμένου να επιτευχθεί ακριβής έλεγχος του μήκους κύματος/κυματαριθμού. Αντίθετα, σε συσκευές που αξιοποιούν συμβολόμετρα, διατηρείται εύκολα και συνεχώς έλεγχος στον άξονα του μήκους κύματος, με την βοήθεια ενός λέιζερ αναφοράς (συνήθως λέιζερ He–Ne). Μια υψηλής ακρίβειας βαθμονόμηση κυματαριθμών επιτυγχάνεται μέσω συσχέτισης του μήκους κύματος του λέιζερ με τις τομές μηδενισμού του συμβολομέτρου. Η επιλογή του ανιχνευτή εξαρτάται από την περιοχή μήκους κύματος που ερευνάται. Υπάρχουν δύο τύποι ανιχνευτών, ανιχνευτές φωτονίων και θερμικοί ανιχνευτές. Λόγω της ικανότητας λειτουργίας σε μια ευρεία περιοχή NIR, η πρώτη κατηγορία κυριαρχεί σχεδόν αποκλειστικά στα φασματόμετρα που αξιοποιούνται επιστημονικά.

2.3.2 Συμβολόγραμμα

Όπως περιεγράφηκε παραπάνω, ο κινούμενος καθρέπτης καθορίζει το μήκος διαδρομής της δεύτερης δέσμης ακτινοβολίας, ενώ για την πρώτη δέσμη που προσπίπτει στον σταθερό καθρέπτη, το μήκος διαδρομής παραμένει σταθερό. Η συνολική διαφορά του μήκους διαδρομής μεταξύ των δύο δεσμών λέγεται διαφορά οπτικής διαδρομής (*Optical Path Difference* – OPD). Με την παραδοχή ότι ο διαχωριστής δέσμης είναι ιδανικός, δηλαδή δεν απορροφά ακτινοβολία και χωρίζει την προσπίπτουσα δέσμη σε δύο δέσμες ίσου πλάτους, μπορούμε να θεωρήσουμε πως για την θέση του κινούμενου καθρέπτη που ορίζει την διαφορά οπτικής διαδρομής ίση με μηδέν (*Zero Path Difference*), οι δύο δέσμες συμβάλλουν με

συμφωνία φάσης και συνεπώς δημιουργούν ξανά το αρχικό σήμα. Αντίστοιχα όταν η διαφορά οπτικής διαδρομής είναι ίση με $\lambda_0/2$, όπου λ_0 το μήκος κύματος μίας συνιστώσας της ακτινοβολίας, για την συνιστώσα αυτή οι δύο δέσμες θα είναι εκτός φάσης, η συμβολή θα λειτουργήσει αναιρετικά και θα έχει ως αποτέλεσμα μηδενικό σήμα. Αν θεωρήσουμε μονοχρωματική ακτινοβολία και για σταθερή ταχύτητα του κινούμενου καθρέπτη, το συμβολόμετρο παράγει ένα ημιτονοειδές σήμα, τα μέγιστα του οποίου βρίσκονται στις θέσεις όπου η διαφορά οπτικής διαδρομής είναι ίση με ακέραιο πολλαπλάσιο του μήκους κύματος της πηγής. Επαγωγικά, για πηγή συνεχούς φάσματος το συμβολόγραμμα θα έχει την μορφή που φαίνεται στην *Figure 4*. Αρχικά όλα τα μήκη κύματος συμβάλλουν προσθετικά και όσο αυξάνεται η διαφορά οπτικής διαδρομής και συνεπώς η διαφορά φάσης μεταξύ τους, συμβάλλουν αρνητικά και έτσι προκύπτει σήμα με εκθετικά μειούμενη ένταση, η οποία καταλήγει να μηδενίζεται.

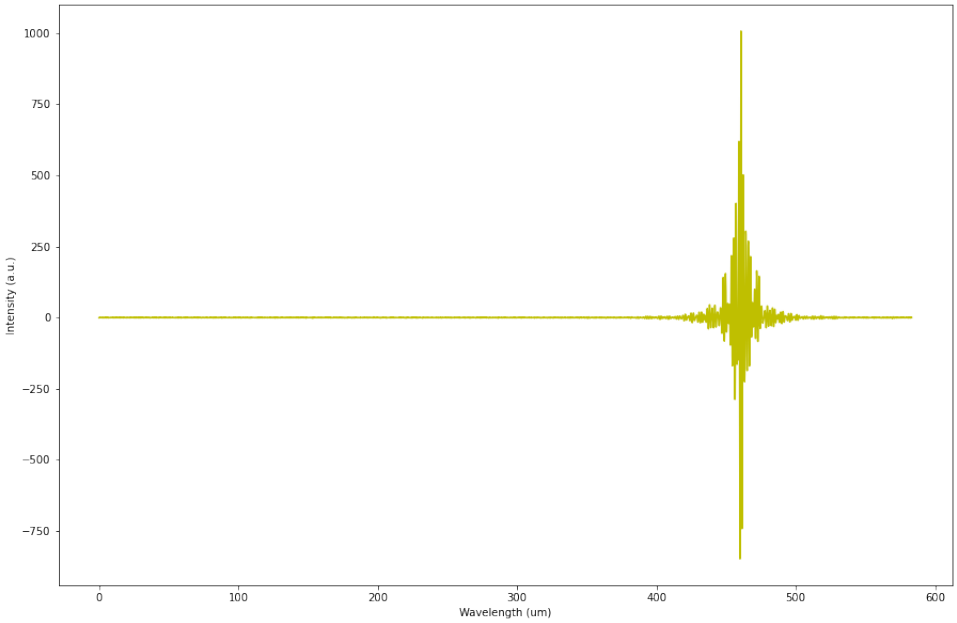

Figure 4 Μορφή συμβολογράμματος

2.3.3 Παραγωγή φάσματος απορρόφησης και διαπερατότητας

Πριν μετρηθεί το φάσμα του δείγματος, με την χρήση μίας πρότυπης λευκής επιφάνειας αναφοράς, μετράται το συμβολόγραμμα της πηγής. Στην συνέχεια γίνεται λήψη του συμβολογράμματος του δείγματος μαζί με την πηγή. Και τα δύο συμβολογράμματα, μέσω μετασχηματισμού *Fourier*, παράγουν αντίστοιχα το φάσμα πηγής και το συνολικό φάσμα. Διαιρώντας το συνολικό φάσμα με αυτό της πηγής, απομονώνεται η χημική πληροφορία του δείγματος, και προκύπτει το φάσμα διαπερατότητας του δείγματος. Η διαπερατότητα του δείγματος εκφράζει τον λόγο της ακτινοβολίας που εξέρχεται από το δείγμα και της προσπίπτουσας ακτινοβολίας.

Τον δέκατο όγδοο αιώνα, ο Pierre Bouguer συνείσφερε σημαντικά στην κατανόηση του τρόπου με τον οποίο το φως αλληλοεπιδρά με την ύλη. Παρατηρώντας πως το φως περνώντας από την ατμόσφαιρα γινόταν ασθενέστερο, συσχέτισε την εναπομένουσα ένταση του φωτός με το πάχος της ατμόσφαιρας μέσα από την οποία οδηγήθηκε περιγράφοντάς την με μια λογαριθμική σχέση πρώτης τάξης.

Equation 2

$$\frac{I}{I_o} = exp(-\in t),$$

όπου I_0 η ένταση του φωτός που προσπίπτει στην ατμόσφαιρα, I η ένταση του φωτός που φτάνει τον ανιχνευτή και t το πάχος της ατμόσφαιρας.

Μία ακόμα σημαντική συνεισφορά για την κατανόηση που έχουμε σήμερα για την αλληλεπίδραση της ακτινοβολίας με την ύλη, αποτέλεσε τον 19ο αιώνα αυτή του Beer, σύμφωνα με την οποία τα χημικά στοιχεία έχουν την ικανότητα να απορροφούν φως σε συγκεκριμένες συχνότητες και η απορρόφηση του φωτός είναι ανάλογη της συγκέντρωσης της ουσίας. Η μαθηματική έκφραση για την απορρόφηση είναι:

Equation 3

$$Absorbance = -\log_{10}\left(\frac{I}{I_0}\right),$$

όπου I_0 η ένταση του φωτός που προσπίπτει στο δείγμα, I η ένταση του φωτός που φτάνει τον ανιχνευτή.

Τελικά ο " Νόμος του Beer" μπορεί να εκφραστεί ως:

Equation 4

$$T(\nu) = \frac{I(\nu)}{I_0(\nu)} = exp\left[-a(\nu)\, b\right],$$

όπου α(ν) ο συντελεστής γραμμικής απορρόφησης και b το πάχος του δείγματος.

Τέλος, το φάσμα απορρόφησης προκύπτει μέσα από το φάσμα διαπερατότητας ως εξής:

Equation 5

$$A(\nu) = \log_{10}\frac{1}{T(\nu)} = \frac{1}{\ln 10}a(\nu)\, b.$$

Το μεγαλύτερο μέρος του κεφαλαίου 2 περιγράφεται αναλυτικά στο βιβλίο των Ozaki Y. et al., "Near-Infrared Spectroscopy", Part 1 & Part 3.

3. Τεχνικές επεξεργασίας και ανάλυσης NIR φασμάτων

3.1 Χημειομετρία

Όπως έχει αναφερθεί στο κεφάλαιο 2, στην NIR φασματοσκοπία συναντώνται κάποιες δυσκολίες, κυρίως όσον αφορά την μελέτη και κατανόηση των φασμάτων. Αυτές κυρίως οφείλονται στην πολυπλοκότητα του εξαγόμενου φάσματος λόγω των υπερτονικών και των δονήσεων συνδυασμού. Ευρείες και υπερκαλυπτόμενες ζώνες συχνά προκύπτουν στα NIR φάσματα και καθιστούν δύσκολη την εξαγωγή σχετικής πληροφορίας σε συγκεκριμένα μήκη κύματος. Επιπλέον, η κατάσταση του δείγματος καθώς και οι συνθήκες περιβάλλοντος κατά την μέτρηση επηρεάζουν το φάσμα, δυσκολεύοντας ακόμα περισσότερο την κατανόησή του.

Όλες αυτές οι ιδιαιτερότητες των NIR φασμάτων καθιστούν αναγκαία την αξιοποίηση μεθόδων ανάλυσης δεδομένων και της χημειομετρίας (*chemometrics*) για την εξαγωγή χρήσιμης πληροφορίας από τα φάσματα [5]. Χημειομετρία είναι ο κλάδος της χημείας που ασχολείται με την εφαρμογή μεθοδολογιών της στατιστικής, των μαθηματικών και της τυπικής λογικής στην χημεία με στόχο τον σχεδιασμό ή την επιλογή των βέλτιστων πειραματικών διαδικασιών και μετρήσεων, την εξαγωγή της μέγιστης σχετική χημικής πληροφορίας από την ανάλυση χημικών δεδομένων, όπως δίνεται στο *Chemometrics and Intelligent Laboratory Systems* [6]. *Η* χημειομετρία βρίσκει εφαρμογή σε πολλούς κλάδους πέραν της χημείας, κι έτσι τείνει να γίνει αυτόνομος επιστημονικός κλάδος. Συνολικά η χημειομετρία και η μηχανική μάθηση (*machine learning* - ML) έχουν κάποια επικαλυπτόμενα σημεία, όπου μέθοδοι μηχανικής μάθησης αξιοποιούνται από την χημειομετρία. Από μία σκοπιά θα μπορούσαμε να δούμε την χημειομετρία να βρίσκεται μέσα στο πλαίσιο της μηχανικής μάθησης.

3.2 Μηχανική Μάθηση

Η μηχανική μάθηση είναι επιστημονικό πεδίο που βρίσκει εφαρμογές σε πάρα πολλά πεδία όπως την πρόγνωση καιρού, την βιοϊατρική, το χρηματιστήριο και άλλα. Αποτελεί κομμάτι της τεχνητής νοημοσύνης, είναι πρακτικά ένα σύνολο μεθόδων και χρησιμοποιεί αλγορίθμους που μπορούν να επεξεργαστούν και να αναλύσουν δεδομένα, βασιζόμενοι στην συσχέτιση της πληροφορίας τους. Το 1959, ο Arthur Samuel περιέγραψε την μηχανική μάθηση ως “το επιστημονικό πεδίο που δίνει στους υπολογιστές την δυνατότητα να μάθουν, χωρίς να είναι ρητά προγραμματισμένοι”. Κατέληξε στο ότι η δυνατότητα των υπολογιστών να μαθαίνουν εμπειρικά, τελικά μπορεί να βοηθήσει ώστε να μην υπάρχει η ανάγκη για τόσο λεπτομερή προγραμματισμό. Ένας άλλος ορισμός που αφορά τη μηχανική μάθηση, από τον Tom M. Mitchell το 1997, αναφέρει ότι “ένα πρόγραμμα υπολογιστών λέγεται ότι μαθαίνει από την εμπειρία E, σε σχέση με κάποια τάξη εργασιών T και μέτρηση απόδοσης P (*Performance Measure*) εάν η απόδοσή του σε εργασίες στο T, όπως μετράτε από το P, βελτιώνεται με την εμπειρία E.” [7].

3.2.1 Βασικές κατηγορίες προβλημάτων και αλγορίθμων μηχανικής μάθησης

Υπάρχουν διάφορα κριτήρια με τα οποία μπορούν να κατηγοριοποιηθούν οι διάφοροι αλγόριθμοι στην μηχανική μάθηση. Ένας βασικός διαχωρισμός αφορά τους εξής τρεις τύπους μάθησης: την επιβλεπόμενη μάθηση (*Supervised Learning*), την μη επιβλεπόμενη μάθηση (*Unsupervised Lear*ning) και την ενισχυτική μάθηση (*Reinforcement Learning*).

3.2.1.1 Επιβλεπόμενη μάθηση

Βασικό χαρακτηριστικό αυτής της κατηγορίας είναι πως τα δεδομένα εισόδου που θα εκπαιδεύσουν τον αλγόριθμο είναι συνδεδεμένα με τα επιθυμητά αποτελέσματα και ο στόχος είναι το μοντέλο να μάθει έναν γενικό κανόνα προκειμένου να αντιστοιχίσει τις εισόδους με τα αποτελέσματα. Τα προβλήματα στα οποία εφαρμόζεται επιβλεπόμενη μάθηση μπορούν να χωριστούν σε προβλήματα κατηγοριοποίησης/ταξινόμησης (*Classification problems*) και προβλήματα παλινδρόμησης (*Regression problems*). Μερικοί γνωστοί αλγόριθμοι οι οποίοι εφαρμόζουν επιβλεπόμενη μάθηση είναι η γραμμική παλινδρόμηση (*Linear Regression*), τα νευρωνικά δίκτυα (*Neural Networks*), οι μηχανές διανυσμάτων στήριξης (*Support Vector Machines* – SVMs), η μάθηση κατά Bayes (*Bayesian Learning*), τα δένδρα απόφασης (*Decision Trees*), ο k πλησιέστεροι γείτονες (*k Nearest Neighbors* – kNN), η λογιστική παλινδρόμηση (*Logistic Regression*) και τα τυχαία δάση *(Random Forests*).

- Προβλήματα κατηγοριοποίησης: Το μοντέλο δέχεται σαν είσοδο δεδομένα τα οποία χωρίζονται σε δύο ή περισσότερες (*multi-label classification*) κλάσεις. Σκοπός είναι να μπορεί να κατηγοριοποιεί τα εισαγόμενα δεδομένα στις δεδομένες κλάσεις – κατηγορίες.
- Προβλήματα παλινδρόμησης: Το μοντέλο προσπαθεί να προβλέψει αποτέλεσμα που είναι μία συνεχή τιμή. Εκπαιδεύεται από ένα σύνολο δεδομένων, των οποίων τα χαρακτηριστικά επηρεάζουν την ζητούμενη τιμή. Σκοπός του μοντέλου τελικά είναι να μπορεί για νέα δείγματα, με βάση τις νέες τιμές των χαρακτηριστικών τους, να μπορεί να προβλέπει την ζητούμενη τιμή.

3.2.1.2 Μη επιβλεπόμενη μάθηση

Σε αντίθεση με την επιβλεπόμενη μάθηση, όπου εξαρχής τα δεδομένα εισόδου είναι αντιστοιχισμένα με την σωστή απάντηση, στη μη επιβλεπόμενη μάθηση χρησιμοποιούνται μη κατηγοριοποιημένα δεδομένα ή με δεδομένα με άγνωστη δομή. Οι τεχνικές μη επιβλεπόμενης μάθησης είναι κατάλληλες για διερευνητική ανάλυση δεδομένων (*Exploratory Data Analysis*), δηλαδή για να εξάγει πληροφορία που αφορά την εσωτερική δομή των δεδομένων. Τα προβλήματα που μπορεί να αντιμετωπίσει η μη επιβλεπόμενη μάθηση μπορούν να είναι προβλήματα μείωσης διαστατικότητας (*Dimensionality Reduction problems*) και προβλήματα συσταδοποίησης (*Clustering problems*). Αλγόριθμοι που εφαρμόζουν μη επιβλεπόμενη μάθηση είναι η ανάλυση κυρίων συνιστωσών (*Principal Components Analysis* - PCA), η γραμμική διαχωριστική ανάλυση (*Linear Discriminant Analysis* - LDA), η τυποποίηση (*Standardization*), η κανονικοποίηση (*Normalization*), ο *K-Means*, ο *Fuzzy C-Means*, ο PAM (*Partitioning Around Medoids*), ο BIRCH και ο DBSCAN.

- Προβλήματα μείωσης διαστατικότητας: Σκοπός του μοντέλου είναι να δημιουργήσει νέες διαστάσεις για να περιγράψει τα δεδομένα εισόδου. Ο αριθμός των νέων διαστάσεων είναι μικρότερος και κατασκευάζονται με τέτοιον τρόπο ώστε να υπάρχει η μεγαλύτερη δυνατή διασπορά των δεδομένων σε αυτές.
- Προβλήματα συσταδοποίησης: Το μοντέλο προσπαθεί να χωρίσει τα δεδομένα εισόδου σε ομάδες (*clusters*). Η διαφορά με τα προβλήματα κατηγοριοποίησης είναι ότι εδώ δεν υπάρχουν δεδομένες

κλάσεις, αλλά αντίθετα ο αλγόριθμος εξερευνάει την κατανομή των δεδομένων στον χώρο και τα χωρίζει σε ομάδες που είναι άγνωστες εκ των προτέρων. Γι' αυτό άλλωστε πρόκειται για τυπική κατηγορία προβλημάτων που συναντάται στην μη επιβλεπόμενη μάθηση.

3.2.1.3 Ενισχυτική μάθηση

Είναι μάλλον ο λιγότερο συχνά χρησιμοποιούμενος τύπος μηχανικής μάθησης. Σκοπός του μοντέλου είναι να μάθει να αλληλοεπιδρά με ένα δυναμικό περιβάλλον, να παίρνει δηλαδή σωστές αποφάσεις με δεδομένο ένα περιβάλλον που αλλάζει συνεχώς. Για παράδειγμα τέτοιο σύστημα θα μπορούσε να δίνει την δυνατότητα σε ένα πρόγραμμα να παίζει σκάκι με έναν υπαρκτό αντίπαλο που συνεχώς παίρνει νέες αποφάσεις.

3.2.2 Ανίχνευση ανωμαλιών (*Outlier Detection*)

Ανωμαλίες θεωρούνται τα δείγματα που έχουν διαφορετικό μοτίβο σε σύγκριση με τα άλλα δείγματα ίδιας ομάδας στο σύνολο δεδομένων. Στο πλαίσιο της NIR φασματοσκοπίας, ανωμαλία θεωρείται κάποιο φάσμα ουσίας που αποκλίνει σημαντικά σε σύγκριση με τα υπόλοιπα φάσματα που έχουν μετρηθεί από την ίδια ουσία. Οι ανωμαλίες είναι πάντα σημαντικό να εντοπίζονται είτε γιατί μπορεί να περιγράφουν κάτι νέο, ένα ιδιαίτερο νέο στοιχείο με νέα χαρακτηριστικά και λειτουργίες, είτε γιατί μπορεί να καταστρέψουν το μοντέλο το οποίο θα εκπαιδεύσουν. Στην δεύτερη περίπτωση οι αιτίες που προκαλούν την εμφάνιση των ανωμαλιών είναι ο έντονος θόρυβος, λάθος μέτρηση, λάθος αντιστοίχιση φάσματος με ουσία, και θα πρέπει να απομακρύνονται. Μετά την απομάκρυνση των ανωμαλιών πρέπει να ξαναγίνει έλεγχος του συνόλου των δεδομένων.

Equation 6

$$T^2 = \sum_{\alpha=1}^{\alpha=A} \left(\frac{t_{i,a}}{S_a} \right)^2$$

Συχνά, σε ανάλυση φασματικών δεδομένων είναι δύσκολο να ανιχνευτούν τα φάσματα αυτά με απλή παρατήρηση τους και σε αυτή την περίπτωση χρησιμοποιούνται διάφορες τεχνικές χημειομετρίας. Η PCA μέθοδος κι εδώ μπορεί να φανεί χρήσιμη καθώς με την οπτικοποίηση των δειγμάτων μπορούν εύκολα να παρατηρηθούν οι ανωμαλίες. Υπάρχουν δύο μετρικές απόστασης που περιγράφουν το κατά πόσο απέχει ένα δείγμα από τα υπόλοιπα της ομάδας του, το μέγεθος του καταλοίπου (*residual*) και το *Hotelling's* $\mathbf{T^2}$. Το μέγεθος το καταλοίπου μπορεί απευθείας να υπολογιστεί από τον πίνακα **E**. Υπολογίζεται το τετράγωνο του αθροίσματος της κάθε στήλης, και η μετρική αφορά την απόκλιση του κάθε φάσματος από το αποτέλεσμα του προηγούμενου υπολογισμού. Ένα φάσμα με υψηλότερη διασπορά καταλοίπου θα έχει ένα μοτίβο στα χρήσιμα δεδομένα, που δεν θα είναι όμοιο με τα υπόλοιπα φάσματα. Το *Hotelling's* $\mathbf{T^2}$ βασίζεται στα βάρη των φασμάτων **T** και πρακτικά υπολογίζεται η απόστασή τους από το κέντρο του συνόλου (Equation 6). Συνδυάζοντας αυτά τα δύο μεγέθη παίρνουμε τις σημαντικότερες γραφικές παραστάσεις της PCA.

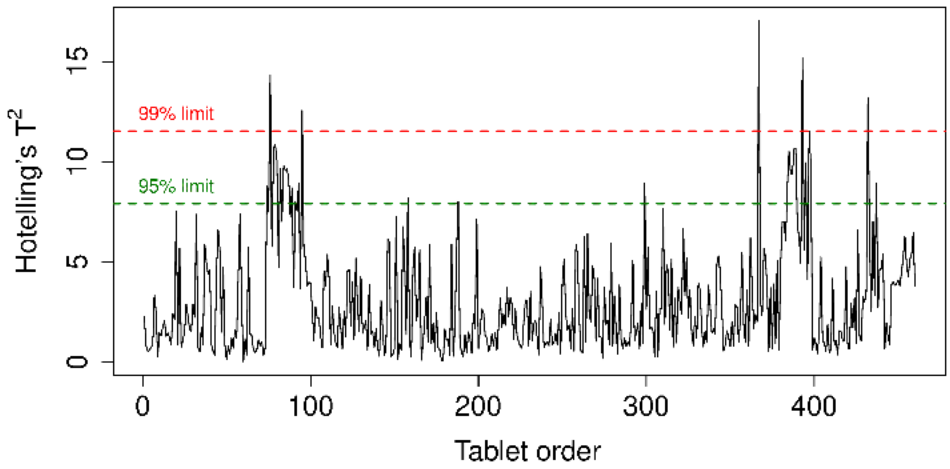

Figure 5 Χρησιμοποίηση μετρικής Hotelling's T^2 για τον προσδιορισμό των ανωμαλιών [8]

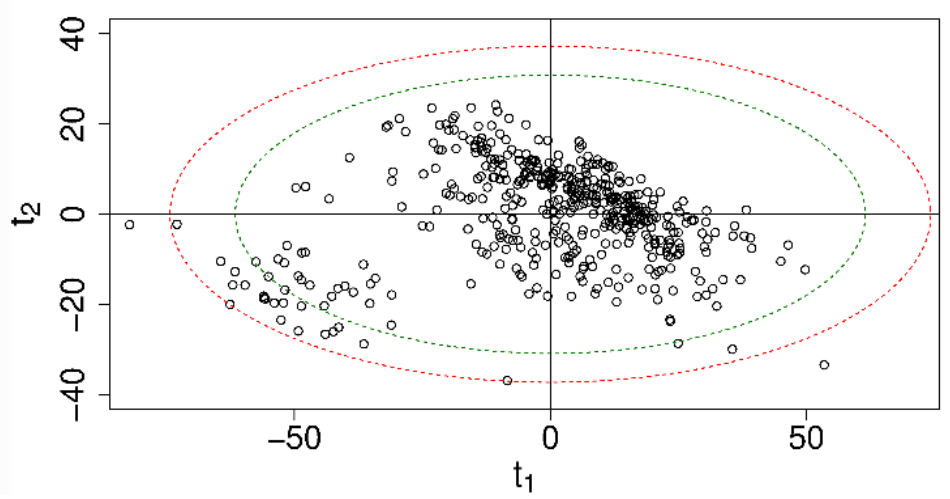

Figure 6 Οπτικοποίηση της κατανομής των δεδομένων στις συνιστώσες t_1 και t_2, για τον προσδιορισμό των ανωμαλιών [8]

3.2.3 Αξιολόγηση κι επιλογή μοντέλου

Η σωστή χρήση των μεθόδων αξιολόγησης κι επιλογής του μοντέλου και του αλγορίθμου μηχανικής μάθησης είναι πολύ σημαντική τόσο σε επίπεδο έρευνας στην μηχανική μάθηση, όσο και στις εφαρμογές της. Συνολικά, όταν χτίζουμε έναν κώδικα μηχανικής μάθησης θέλουμε να ξέρουμε την ικανότητα γενίκευσης (*generalization ability*) του μοντέλου, να επιλέξουμε το πιο αποδοτικό μοντέλο μέσα από ένα σύνολο υπερπαραμέτρων οι οποίες ρυθμίζουν τον αλγόριθμο που έχουμε επιλέξει και να καταλήξουμε στον καταλληλότερο αλγόριθμο, αυτόν που μας δίνει το καλύτερο μοντέλο. Η ικανότητα γενίκευσης είναι η ικανότητα του μοντέλου να παράγει αξιόπιστα αποτελέσματα όταν πάρει σαν είσοδο ένα σύνολο δεδομένων, με τα οποία δεν έχει έρθει προηγουμένως σε επαφή. Έπειτα, επιλέγοντας διαφορετικές υπερπαραμέτρους για έναν αλγόριθμο, παράγονται αντίστοιχα διαφορετικά μοντέλα με αντίστοιχα διαφορετικές αποδόσεις.

3.2.3.1 Ικανότητα γενίκευσης

Στην επιβλεπόμενη μάθηση, αρχικά το σύνολο δεδομένων χωρίζεται σε δύο υποσύνολα δεδομένων, το σύνολο εκπαίδευσης (*training set*) και το σύνολο ελέγχου (*test set*). Το πρώτο αποσκοπεί στην εκπαίδευση του αλγορίθμου και το δεύτερο στην αξιολόγησή του. Για ένα πρόβλημα κατηγοριοποίησης για παράδειγμα, υποθέτουμε ότι το αρχικό σύνολο δεδομένων έχει μία συγκεκριμένη αναλογία ως προς τον αριθμό των δεδομένων που αντιστοιχούν σε κάθε κλάση. Όταν γίνει ο χωρισμός στα δύο υποσύνολα, είναι πολύ πιθανό, αυτή η αναλογία να αλλάξει. Εάν το μοντέλο δεν είναι ανθεκτικό σε τέτοιου είδους διαταραχές, αυτό μπορεί να δημιουργήσει σημαντικό πρόβλημα. Το πρόβλημα είναι ακόμα μεγαλύτερο όταν το αρχικό σύνολο δεδομένων έχει ήδη άνιση κατανομή στις κλάσεις. Αυτό το πρόβλημα αντιμετωπίζεται με μία μέθοδο που λέγεται στρωματοποίηση (*stratification*). Αυτή η μέθοδος χρησιμοποιείται προκειμένου να διατηρεί τις αρχικές αναλογίες στα δύο υποσύνολα. Συχνά θεωρείται πως για μεγάλα σύνολα δεδομένων ο μη στρωματοποιημένος διαχωρισμός δεν αποτελεί σημαντικό ζήτημα, ωστόσο είναι πολύ εύκολος στην υλοποίηση και ακόμα και για πολύ μεγάλα σύνολα θα μπορούσε να φανεί χρήσιμος.

3.2.3.2 *Holdout* μέθοδος

Αρχικά χωρίζουμε τα δεδομένα μας στο σύνολο εκπαίδευσης και στο σύνολο ελέγχου. Έπειτα διαλέγουμε έναν αλγόριθμο που θέλουμε να μελετήσουμε, επιλέγουμε τις παραμέτρους του και τον εφαρμόζουμε στα δεδομένα εκπαίδευσης. Στη συνέχεια αξιολογούμε την ικανότητα γενίκευσης του μοντέλου αξιοποιώντας τα δεδομένα ελέγχου, τα οποία το μοντέλο δεν έχει δει ακόμα. Εάν δεν έχει εξαντλήσει την χωρητικότητά του σε δεδομένα και το εκπαιδεύσουμε στην συνέχεια με ολόκληρο το σύνολο των δεδομένων η ικανότητα γενίκευσής του θα είναι αυξημένη. Η χωρητικότητα αναφέρεται σε μία τιμή αριθμού δεδομένων εκπαίδευσης πάνω από την οποία η απόδοση του αλγορίθμου μένει σταθερή. Ειδικά για περιπτώσεις που το σύνολο δεδομένων είναι περιορισμένο, η επανεκπαίδευση του αλγορίθμου με όλα τα διαθέσιμα δεδομένα είναι ιδιαίτερα χρήσιμη. Εκεί, βέβαια, λαμβάνουμε υπ' όψει πως η νέα ικανότητα γενίκευσης θα είναι μεγαλύτερη. Ο τρόπος που γίνεται η αξιολόγηση είναι μετρώντας το ποσοστό των λάθος εκτιμήσεων για το σφάλμα (*error*), και αντίθετα το ποσοστό των σωστών εκτιμήσεων για την ακρίβεια (*accuracy*) (Equation 7).

Equation 7

$$ERR_S = \frac{1}{n}\sum_{i=1}^{n} L\left(\hat{y}_i, y_i\right) = 1 - ACC_S$$

Equation 8

$$L\left(\hat{y}_i, y_i\right) = \begin{cases} 0 & \hat{y}_i = y_i \\ 1 & if\ \ \hat{y}_i \neq y_i \end{cases}$$

όπου ERR_S και ACC_S το σφάλμα και η ακρίβεια σε ένα σύνολο δεδομένων S, και το $L(\cdot)$ δίνει αποτέλεσμα 1 και 0 σύμφωνα με τον τύπο (Equation 8), που υπολογίζεται από την προβλεπόμενη τάξη ($\hat{y}_i$) και την ετικέτα από την γνωστή κλάση (y_i) για i = 1, ..., n στο σύνολο δεδομένων S.

Η ίδια τακτική μπορεί να εφαρμοστεί για τα προβλήματα επιβλεπόμενης μάθησης, όπως και τα προβλήματα παλινδρόμησης. Σε αυτή τη περίπτωση για τον υπολογισμό του σφάλματος, υπολογίζουμε το μέσο τετραγωνικό σφάλμα (*Mean Squared Error* – MSE) (Equation 9).

Equation 9

$$MSE_S = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2$$

3.2.3.3 Επαναλαμβανόμενο *holdout* και *bootstrapping*

Figure 7 Απεικόνιση διαχωρισμού training και test σετ στην μέθοδο Leave-One-Out Bootstrap

Προκύπτει όμως το εξής ζήτημα: Στην περίπτωση που η εκπαίδευση και η αξιολόγηση γίνουν με το ίδιο σύνολο δεδομένων, τότε έχουμε πολύ αισιόδοξη πρόβλεψη ως προς την απόδοση του μοντέλου (*optimistic bias*). Αντίθετα, όταν ένα αρκετά μεγάλο τμήμα των δεδομένων το δεσμεύουμε για την αξιολόγηση και έπειτα εκπαιδεύουμε ξανά το μοντέλο με όλα τα δεδομένα (*holdout method*), τότε η αρχική εκτίμησή μας για την ακρίβεια είναι απαισιόδοξη (*pessimistic bias*). Από την άλλη, εάν χρησιμοποιήσουμε ένα πολύ μικρό τμήμα δεδομένων για αξιολόγηση, ενώ το bias μειώνεται, έχουμε μεγαλύτερη διακύμανση στην εκτίμηση της απόδοσης του μοντέλου. Ένας τρόπος για να αντισταθμίσουμε κάπως αυτό το πρόβλημα είναι η επαναλαμβανόμενη μέθοδος holdout, δηλαδή να επαναλάβουμε τον διαχωρισμό, την εκπαίδευση και την αξιολόγηση και να πάρουμε τον μέσο όρο της εκτιμώμενης ακρίβειας του μοντέλου. Συχνά αυτή η μέθοδος αναφέρεται ως *Monte Carlo Cross-Validation*. Μία άλλη μέθοδος που στοχεύει σε μία πιο αξιόπιστη εκτίμηση της ακρίβειας του μοντέλου είναι η μέθοδος *bootstrap*. Αυτή η μέθοδος δημιουργεί b διαφορετικά σύνολα δεδομένων αριθμού n, όσο δηλαδή και ο αριθμός των δεδομένων του αρχικού συνόλου. Συνεπώς κάθε ένα από τα b σύνολα έχει δημιουργηθεί, παίρνοντας σειριακά, τυχαία δείγματα από το αρχικό σύνολο με αποτέλεσμα κάποια δείγματα να μην υπάρχουν καθόλου, ενώ κάποια άλλα να επαναλαμβάνονται. Τελικά, κάθε ένα από τα σύνολα αξιολογούνται και ο μέσος όρος αποτελεί την τελική εκτίμηση της ακρίβειας του μοντέλου (Equation 10).

Equation 10

$$ACC_{boot} = \frac{1}{b}\sum_{j=1}^{b}\frac{1}{n}\sum_{i=1}^{n}\left(1 - L\left(\hat{y}_i, y_i\right)\right)$$

Επειδή όμως, όπως έχουμε αναφέρει, η επαναχρησιμοποίηση των ίδιων δεδομένων για την αξιολόγηση του μοντέλου είναι *optimistically biased*, εναλλακτικά χρησιμοποιείται η μέθοδος *Leave-One-Out Bootstrap*. Σε αυτή την μέθοδο, κάθε από b σύνολα αξιολογείται από το σύνολο δεδομένων που δεν

χρησιμοποιήθηκε για την παραγωγή του (Figure 7). Εκτιμάται πως 50 μέχρι 200 *bootstrap* σύνολα είναι επαρκή για μία ικανοποιητική αξιολόγηση του μοντέλου.

3.2.3.4 Επιλογή βέλτιστου μοντέλου και *Cross-Validation*

Αλλάζοντας τις υπερπαραμέτρους σε έναν αλγόριθμο, προκύπτουν διαφορετικά μοντέλα και ένα από τα βασικά ζητήματα είναι να βρεθεί το βέλτιστο μοντέλο μέσα από την ρύθμιση των υπερπαραμέτρων (*hyperparameters tuning*). Όπως έχουμε αναφέρει, η αξιολόγηση ενός μοντέλου με βάση τα δεδομένα με τα οποία έχει εκπαιδευτεί δίνει πολύ αισιόδοξα αποτελέσματα γι' αυτό και αποφεύγεται. Αν υποθέσουμε πως για την εύρεση του βέλτιστου μοντέλου χρησιμοποιηθεί η *holdout* μέθοδος κι έτσι χωριστούν τα δεδομένα σε σύνολο εκπαίδευσης και σύνολο ελέγχου, είναι πολύ πιθανό πάλι να έρθουμε αντιμέτωποι με το πρόβλημα του *optimistic bias*. Αυτό μπορεί να συμβεί, καθώς αν αξιολογηθεί το μοντέλο κι έπειτα γίνει κάποια τροποποίηση στις υπερπαραμέτρους, πληροφορία από το σύνολο ελέγχου μπορεί να έχει έρθει σε επαφή με το μοντέλο. Συνεπώς, καθίσταται σημαντική η δημιουργία ενός τρίτου συνόλου, του συνόλου της επικύρωσης (*validation set*). Το μοντέλο θα εκπαιδεύεται στο σύνολο εκπαίδευσης, θα αξιολογείται στο σύνολο επικύρωσης, και όταν βρεθεί το βέλτιστο μοντέλο θα εκπαιδευτεί στο σύνολο εκπαίδευσης και επικύρωσης και θα αξιολογηθεί στο σύνολο ελέγχου. Αυτό ο διαχωρισμός των δεδομένων έχει σαν αποτέλεσμα το η εξερεύνηση του βέλτιστου μοντέλου να γίνεται με πολύ μειωμένο αριθμό δεδομένων εκπαίδευσης συγκριτικά με τα διαθέσιμα δεδομένα. Σε αυτό το πρόβλημα έρχονται να δώσουν λύση οι *cross-validation* (CV) μέθοδοι. Υπάρχουν διάφορες CV μέθοδοι, ωστόσο η βασική ιδέα είναι κοινή για όλες τις μεθόδους. Η ίσως δημοφιλέστερη μέθοδος είναι η *k-fold* CV. Και σε αυτή τη μέθοδο ένα σύνολο ελέγχου χωρίζεται για την τελική αξιολόγηση. Το σύνολο εκπαίδευσης χωρίζεται σε k υποσύνολα. Για k επαναλήψεις, τα k-1 υποσύνολα λειτουργούν σαν σύνολο εκπαίδευσης, και το 1 υποσύνολο που απομένει αξιολογεί το μοντέλο. Η τελική ακρίβεια του μοντέλου προκύπτει από τον μέσο όρο των τιμών που έχουν προκύψει από τις k επαναλήψεις. Η μέθοδος αυτή μπορεί να είναι υπολογιστικά ακριβή, αλλά αξιοποιεί με πολύ συμφέρων τρόπο τα δεδομένα. Ιδιαίτερα όταν τα δεδομένα είναι περιορισμένα, το πλεονέκτημα αυτό είναι ιδιαίτερα σημαντικό.

3.2.3.5 Επιλογή αλγορίθμου

Ο κάθε αλγόριθμος μηχανικής μάθησης διαθέτει κάποια χαρακτηριστικά και κάποιες προϋποθέσεις κάτω από τις οποίες έχει την καλύτερη απόδοση. Με βάση αυτές τις πληροφορίες μπορεί να γίνει αρχικά η επιλογή ενός αλγορίθμου. Το μέγεθος του συνόλου δεδομένων για την εκπαίδευση του αλγορίθμου, η ταχύτητα εκπαίδευσης που απαιτείται, η ακρίβεια που χρειάζεται η πρόβλεψη, η κατανομή των δεδομένων στον χώρο, ο αριθμός των χαρακτηριστικών (*features*) κάθε δείγματος, είναι παράγοντες που παίζουν ρόλο στην επιλογή του κατάλληλου αλγορίθμου. Για παράδειγμα, τα νευρωνικά δίκτυα (*Artificial Neural Networks*) και η βαθιά μάθηση (*Deep Learning*) απαιτούν πολύ μεγάλο αριθμό δεδομένων για εκπαίδευση για να έχουν βέλτιστη απόδοση. Στην περίπτωση, όμως, που χρειάζεται να επιλέξουμε ανάμεσα σε κάποιους αλγορίθμους, ώστε να βρούμε τον πλέον αποδοτικό, υπάρχουν ορισμένες μέθοδοι που μπορούν να βοηθήσουν σε αυτή την κατεύθυνση. Επιγραμματικά αναφέρονται οι *Nested Cross Validation* και *Combined 2x5 cv F test*.

Παρακάτω θα περιγράψουμε κάποια στοιχεία της χημειομετρίας και της μηχανικής μάθησης που έχουν φανεί χρήσιμα είτε για την εκπόνηση της συγκεκριμένης εργασίας είτε ως προς την καλύτερη κατανόηση της ανάλυσης των NIR φασμάτων.

3.2.4 Ανάλυση Κύριων Συνιστωσών - PCA

Μία από τις βασικότερες χημειομετρικές τεχνικές αποτελεί η *Principal Component Analysis* ή Ανάλυση Κύριων Συνιστωσών (PCA) [2, κεφάλαιο 7]. Η PCA χρησιμοποιείται κυρίως για οπτικοποίηση πολυπαραμετρικών δεδομένων και αποτελεί αλγόριθμο μείωσης διαστατικότητας. Είναι χρήσιμο να κάνουμε την εξής θεώρηση: Για την επεξεργασία ενός συνόλου φασμάτων θεωρούμε δείγμα (*sample*) το κάθε φάσμα, και παραμέτρους /χαρακτηριστικά (*features*) την ένταση του φάσματος για τα μήκη κύματος που έχει ληφθεί. Τα φάσματα, δηλαδή, αποτελούν δείγματα, με παραμέτρους τις εντάσεις του φάσματος στα συγκεκριμένα μήκη κύματος που δειγματολειπτούνται, και συνεπώς είναι πολυπαραμετρικά.

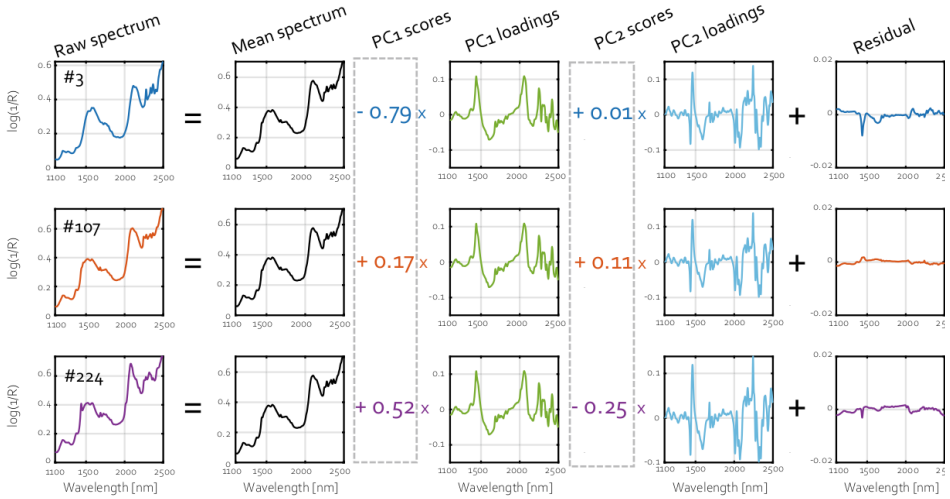

Figure 8 Ανάλυση φασματικών δεδομένων με την μέθοδο PCA. Η πρώτη στήλη αποτελεί τα δεδομένα εισόδου, η δεύτερη στήλη τον μέσο όρο των φασμάτων και είναι ίδιος για όλα τα φάσματα εισόδου, η τρίτη και η τέταρτη το πρώτο και το δεύτερο Loading αντίστοιχα, όπου είναι ίδια για όλα τα φάσματα εισόδου, αλλά με διαφορετικά βάρη και η τελευταία στήλη τα residuals, δηλαδή το σήμα που απομένει όταν από τα φάσματα εισόδου αφαιρεθούν τα πρώτα δύο Loadings. [2]

Η PCA δημιουργεί νέες συντεταγμένες, όσες αποφασίσει ο χρήστης, κάθε μία από τις οποίες, με φθίνουσα σειρά εκφράζει καλύτερα τα δεδομένα. Αυτές οι συντεταγμένες είναι τα λεγόμενα *Principal Components* (PCs) ή *Loadings*, όπου το PC1 εκφράζει μεγαλύτερο μέρος πληροφορίας από το PC2 κ.ο.κ.. Με άλλα λόγια, η μέθοδος αυτή προσπαθεί να κατασκευάσει τους άξονες πάνω στους οποίους τα δεδομένα θα έχουν την μέγιστη διασπορά. Τα *Principal Components* είναι γραμμικοί συνδυασμοί των αρχικών συντεταγμένων, δηλαδή εδώ των φασμάτων απορρόφησης (*Figure 8*). Ο συνολικός αριθμός των components είναι όσος και ο αριθμός των αρχικών διαστάσεων, δηλαδή όσα και τα δείγματα. Ωστόσο τα πρώτα λίγα components, περιέχουν πάνω από το 90% της πληροφορίας. Τελικά τα δείγματα μπορούν να περιγραφούν ως εξής:

$$\boldsymbol{A}=\mathbf{T}\cdot\mathbf{F}+\mathbf{E},$$

όπου A είναι ο πίνακας των φασμάτων απορρόφησης, **T** ο πίνακας των βαρών (*score values)* των principal component, F ο πίνακας των principal component και E ο πίνακας που περιγράφει τα κατάλοιπα (*residuals*), τα οποία συνήθως έχουν ιδιαίτερα μικρή ένταση.

Γενικά, η PCA μέθοδος μπορεί να αξιοποιηθεί για διαφορετικούς σκοπούς. Αρχικά, σε κάθε περίπτωση, δίνει την δυνατότητα για οπτικοποίηση των δεδομένων, δίνοντας πληροφορία για την κατανομή των δειγμάτων. Με την απεικόνιση των δειγμάτων σε έναν χώρο πολύ λιγότερων διαστάσεων, για παράδειγμα σε χώρο δύο διαστάσεων και άξονες τα PC1 και PC2, μπορεί να μελετηθεί εάν τα δείγματα ομαδοποιούνται, πληροφορία πολύ χρήσιμη για παράδειγμα σε αυτή την εργασία, όπου η ικανότητα των φασμάτων να ομαδοποιούνται αναδεικνύει την δυνατότητα για κατηγοριοποίηση (*Classification*). Έπειτα, η μέθοδος είναι χρήσιμη για την ανίχνευση ανωμαλιών (*Outlier Detection*) στο σύνολο των δειγμάτων. Επίσης σημαντική είναι η συνεισφορά της μεθόδου σε προβλήματα συμπίεσης δεδομένων και ελαχιστοποίηση του θορύβου.

3.2.5 Decision Trees και Random Forest

Τα δέντρα απόφασης (*Decision Trees*) είναι μοντέλα μηχανικής μάθησης που μπορούν να χρησιμοποιηθούν τόσο για προβλήματα κατηγοριοποίησης όσο και για προβλήματα παλινδρόμησης. Πρακτικά το δέντρο απαντά συγκεκριμένες ερωτήσεις με βάση τις οποίες καταλήγει σε κάποια απόφαση. Το κάθε κλαδί αντιστοιχεί σε μία συνθήκη που αφορά τα χαρακτηριστικά του δείγματος και τα φύλλα αντιστοιχούν στην τελική απόφαση, σε μία κλάση. Όταν οι τιμές εξόδου στα φύλλα είναι διακριτές τιμές έχουμε πρόβλημα κατηγοριοποίησης, ενώ όταν είναι συνεχείς, έχουμε πρόβλημα παλινδρόμησης.

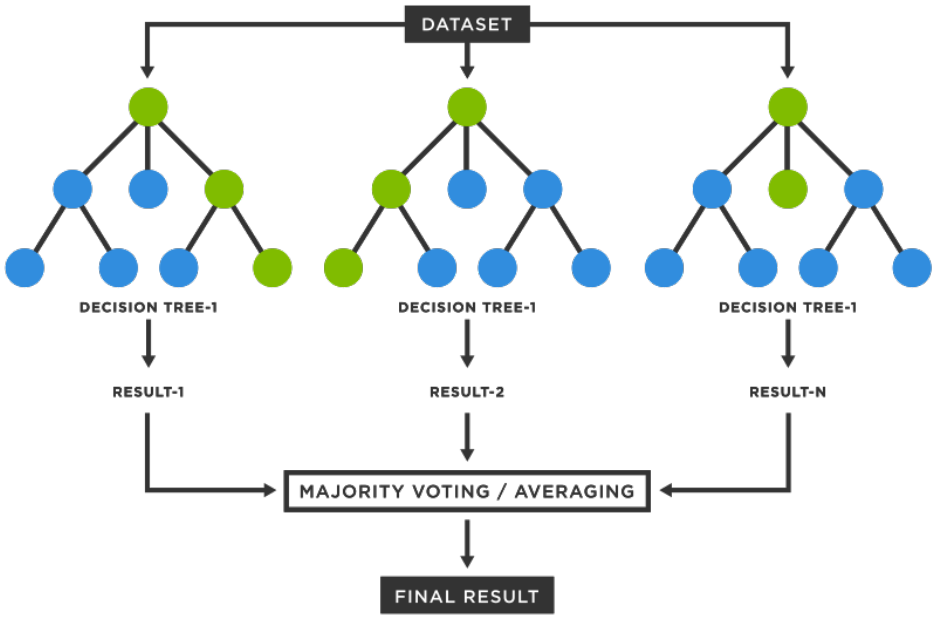

Figure 9 Η μέθοδος Random Forest. [9]

Πολλές φορές προκειμένου να ενισχύσουμε την εγκυρότητα ενός μοντέλου, μπορούμε να συνδυάσουμε περισσότερους αλγορίθμους. Αυτή η τακτική λέγεται συλλογική μάθηση (*Ensemble Learning*). Το μοντέλο που συνδυάζει πολλαπλά δέντρα απόφασης λέγεται τυχαία δάση (*Random Forests*) και χρησιμοποιείται αντίστοιχα για προβλήματα κατηγοριοποίησης και παλινδρόμησης. Πιο συγκεκριμένα, εκπαιδεύει πολλά διαφορετικά δέντρα απόφασης και αποφασίζει βάσει της πλειοψηφίας για προβλήματα κατηγοριοποίησης και βάσει του μέσου όρου για προβλήματα παλινδρόμησης (Εικόνα 7). Τα

διαφορετικά δέντρα κατασκευάζονται με την επιλογή διαφορετικών χαρακτηριστικών με τυχαίο τρόπο. Όσο αυξάνεται ο αριθμός των δέντρων το σφάλμα γενίκευσης συγκλίνει σε μία σταθερή τιμή η οποία εξαρτάται από την δύναμη των μεμονωμένων δέντρων, αλλά και της σχέσης μεταξύ τους [10].

Συνολικά ο αλγόριθμος Random Forests είναι μία καλή επιλογή για προβλήματα παλινδρόμησης ή κατηγοριοποίησης όταν ο αριθμός των δειγμάτων δεν είναι πολύ μεγάλος και αντίθετα ο αριθμός των χαρακτηριστικών των δειγμάτων είναι σχετικά μεγάλος. Σε γενικές γραμμές είναι ένας πολυχρησιμοποιούμενος αλγόριθμος, εύκολος στην αντίληψη και με πολύ καλή απόδοση.

3.2.6 Support Vector Machine

Ο αλγόριθμος *Support Vector Machine* (SVM) είναι αλγόριθμος επιβλεπόμενης μάθησης, αποτελεί μία από τις πλέον γνωστές και χρησιμοποιούμενες τεχνικές κατηγοριοποίησης [11]. Δημιουργήθηκε από τους Vladimir N. Vapnik και Alexey Ya. Chervonenkis το 1963 και αρχικά είχε εφαρμογή σε δυαδικά προβλήματα κατηγοριοποίησης [12]. Για την κατανόηση του αλγορίθμου είναι καλό να γνωρίζουμε τέσσερις βασικές έννοιες: την διαχωριστική επιφάνεια (*separation hyperplane*), το *maximum-margin hyperplane* ή αλλιώς *hard margin*, το *soft margin* και το *kernel trick*. Αν υποθέσουμε ότι έχουμε ένα πρόβλημα κατηγοριοποίησης με δύο κλάσεις και το κάθε παράδειγμα έχει δύο χαρακτηριστικά. Μπορούμε να απεικονίσουμε τα παραδείγματα σε έναν δισδιάστατο χώρο, όπου ο κάθε άξονας αντιστοιχεί σε ένα χαρακτηριστικό και οι συντεταγμένες των παραδειγμάτων είναι οι τιμές των χαρακτηριστικών.

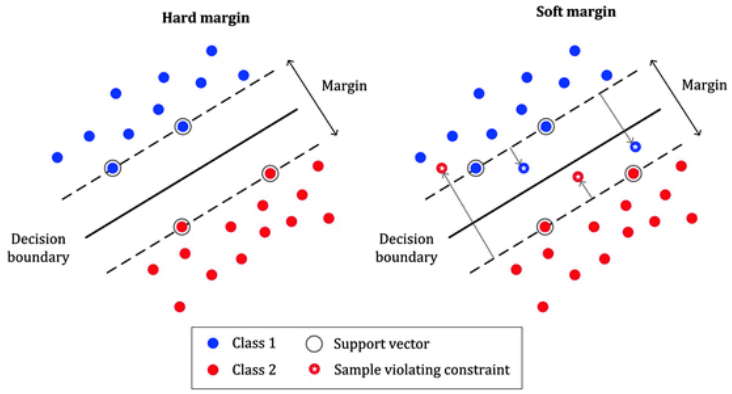

Figure 10 Η μέθοδος Support Vector Machine [13]

Ο αλγόριθμος SVM προσπαθεί να βρει την γραμμή αυτή που διαχωρίζει καλύτερα τις διαφορετικές κλάσεις σε αυτόν τον χώρο. Η υπόθεση αυτή μπορεί να αναχθεί σε χώρους μεγαλύτερων διαστάσεων, όπου πάλι ο αλγόριθμος προσπαθεί να βρει το επίπεδο που διαχωρίζει καλύτερα τις διαφορετικές κλάσεις. Αυτή αποτελεί και την διαχωριστική επιφάνεια. Έπειτα, θεωρούμε margin, την απόσταση μεταξύ του *hyperplane* και του κοντινότερου παραδείγματος στον χώρο. Για τον διαχωρισμό δύο ομάδων σε έναν χώρο υπάρχουν πολλά διαφορετικά *hyperplanes*. Ο SVM αλγόριθμος επιλέγει το *hyperplane* που συνεπάγεται την μεγαλύτερη τιμή *margin* (*maximum-margin hyperplane*). Διαλέγοντας αυτό το *hyperplane* ο αλγόριθμος μεγιστοποιεί την πιθανότητα να κάνει μία σωστή πρόβλεψη.

Η μέθοδος του *maximum-margin hyperplane* πολλές φορές αποτυγχάνει όταν τα δεδομένα των διαφορετικών κλάσεων δεν είναι γραμμικά διαχωρίσιμα. Για την υπέρβαση αυτού του προβλήματος προτάθηκαν δύο λύσεις: το soft-margin και η συνάρτηση *Kernel*. Το *soft-margin* πρακτικά λειτουργεί όπως

και το *hard-margin* με την διαφορά ότι επιτρέπει την ύπαρξη μερικών αστοχιών .Συχνά αυτό λειτουργεί προστατευτικά από *overfitting* του μοντέλου στα δεδομένα (Figure 11). *Overfitting* έχουμε όταν το μοντέλο προσαρμόζεται με απόλυτο τρόπο στα δεδομένα με τα οποία εκπαιδεύεται, χάνοντας την ικανότητα γενίκευσης, δηλαδή την ικανότητά του να προβλέπει σωστά δεδομένα που δεν έχει ξαναδεί και βρίσκονται πιθανώς εκτός κατανομής. Αυτό συμβαίνει συχνά όταν ο αριθμός των δεδομένων είναι περιορισμένος.

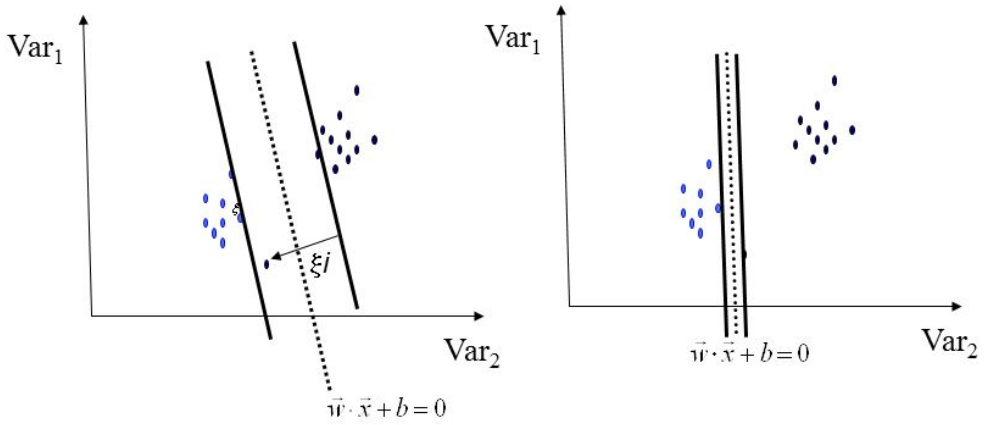

Figure 11 Δεξιά η κατασκευή επιτρέπει κάποιες αστοχίες, στα αριστερά το hard-margin αποτυγχάνει να παράγει ένα μοντέλο με καλή ικανότητα γενίκευσης [14]

Το *kernel trick* αποτέλεσε την πρώτη λύση στο πρόβλημα το *maximum-margin hyperplane* και εισήχθη από τους Bernhard E. Boser, Isabelle M. Guyon και Vladimir N. Vapnik το 1992. Η μέθοδος αυτή μέσω κάποιου μη γραμμικού μετασχηματισμού (*polynomial, gaussian, sigmoid,* κτλ.) μεταφέρει τα δεδομένα σε έναν διανυσματικό χώρο μεγαλύτερων διαστάσεων, όπου πλέον τα δεδομένα είναι γραμμικά διαχωρίσιμα (Figure 12). Στον νέο χώρο κατασκευάζει το hard-margin και το μεταφέρει στις αρχικές διαστάσεις. Και με αυτή τη μέθοδο, όμως, το μοντέλο είναι πιθανό να υποφέρει από αδυναμία γενίκευσης κάνοντας *overfitting* στα δεδομένα.

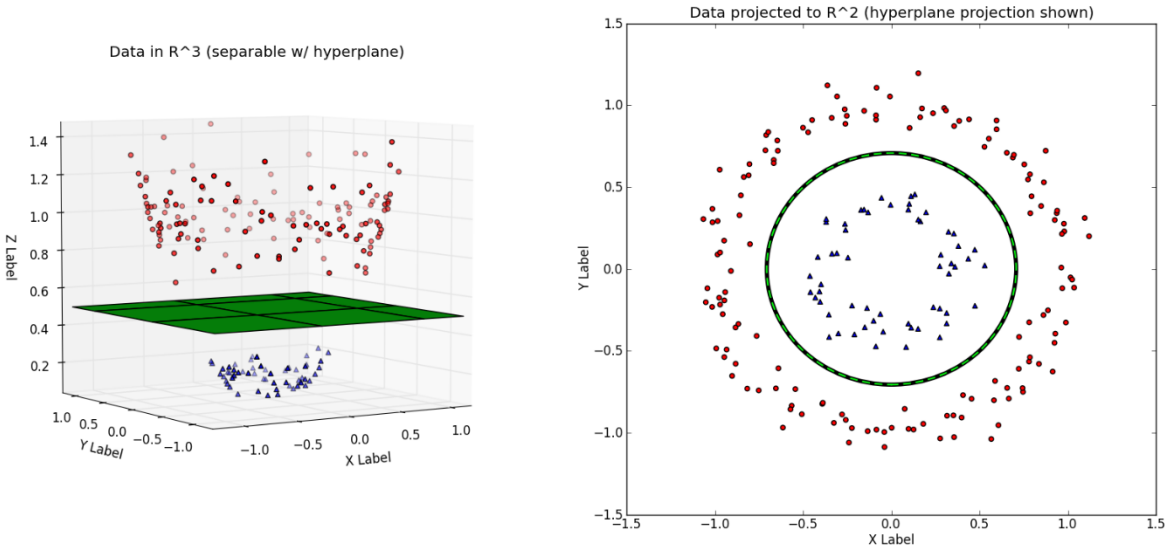

Figure 12 Kernel trick [15]

3.2.7 Μέθοδοι αξιολόγησης μοντέλων σε προβλήματα κατηγοριοποίησης

Ο πίνακας σύγχυσης (*Confusion Matrix*) [16] αποτελεί βασικό εργαλείο όσον αφορά την αξιολόγηση προβλημάτων κατηγοριοποίησης. Για προβλήματα κατηγοριοποίησης με περισσότερες από δύο κλάσεις (*multiclass*), ο πίνακας σύγχυσης είναι ένας τετραγωνικός πίνακας, όπου στην θέση (i,j) βρίσκεται το πλήθος των δειγμάτων κλάσης i τα οποία έχουν ταξινομηθεί στην κλάση j (Figure 13).

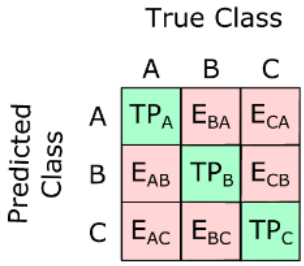

Figure 13 Πίνακας σύγχυσης για multiclass προβλήματα [16]

Αυτός ο πίνακας, συνεπώς, μας δείχνει τα λάθη του μοντέλου και το κατά πόσο δύο κλάσεις μπορεί να συγχέονται. Με βάση αυτόν τον πίνακα, η συνολική ακρίβεια είναι ίση με το άθροισμα της διαγώνιου του πίνακα προς τον συνολικό αριθμό των δειγμάτων (Equation 11). Έπειτα δύο ακόμα μεγέθη που προκύπτουν από τον πίνακα είναι η ακρίβεια (*precision*) και η ανάκληση (*recall*). Η ακρίβεια περιγράφει το ποσοστό των δειγμάτων που κατηγοριοποιούνται σε μία κλάση και πράγματι ανήκουν σε αυτή (Equation 12). Η ανάκληση αφορά το ποσοστό των δειγμάτων που ανήκουν σε μία κλάση και ταξινομήθηκαν σωστά στην κλάση αυτή (Equation 13). Για να γίνει πιο κατανοητή η διαφορά των δύο διαφορετικών μεγεθών μπορούμε να διατυπώσουμε τις εξής ερωτήσεις: η ερώτηση στην οποία απαντά η ακρίβεια είναι "Δοθέντος ενός δείγματος που ταξινομήθηκε στην κλάση x, ποια η πιθανότητα η ταξινόμηση να είναι σωστή;", αντίστοιχα για την ανάκληση η ερώτηση είναι "Δοθέντος ενός δείγματος, ποια η πιθανότητα να ταξινομηθεί σωστά στην κλάση του;". Συνολικά, για ένα δυαδικό πρόβλημα κατηγοριοποίησης μπορούμε

να εισάγουμε τις παρακάτω τέσσερις έννοιες. *true positive* είναι τα δείγματα που ανήκουν στην *positive* κλάση, και σωστά κατηγοριοποιήθηκαν εκεί. *false positive* είναι τα δείγματα που ταξινομήθηκαν στην *positive* κλάση, αλλά δεν ανήκουν εκεί. Αντίστοιχα, *true negative* είναι τα δείγματα που σωστά ταξινομήθηκαν σαν negative, ενώ *false negative* είναι τα δείγματα που είναι *positive* αλλά ταξινομήθηκαν στην *negative* κλάση.

Equation 11

$$OA = \frac{\sum_{i=1}^{M} A(i,i)}{N}$$

Equation 12

$$P_i = \frac{A(i,i)}{\sum_{j=1}^{N} A(j,i)},$$

Equation 13

$$R_i = \frac{A(i,i)}{\sum_{j=1}^{N} A(i,j)}$$

, όπου OA η συνολική ακρίβεια, P η ακρίβεια, R η ανάκληση, A ο πίνακας σύγχυσης, i και j οι συντεταγμένες του πίνακα, N ο συνολικός αριθμός των στοιχείων.

Τα μοντέλα που αντιμετωπίζουν προβλήματα κατηγοριοποίησης παράγουν μία τιμή, η οποία αντικατοπτρίζει την πιθανότητα το δείγμα υπό εξέταση να βρίσκεται στις δεδομένες κλάσεις. Συχνά μπορούμε να εφαρμόσουμε μία οριακή τιμή (*threshold*) στις πιθανότητες που παράγει το μοντέλο, βάσει της οποίας θα αποφασίζει αν το δείγμα ανήκει στην δεδομένη κλάση. Πιο συγκεκριμένα, αν η τιμή πιθανότητας για μία δεδομένη κλάση ξεπερνά την οριακή τιμή, τότε το δείγμα ανήκει στην δεδομένη κλάση. Αυτός είναι και ένας τρόπος για να αντιμετωπίζονται τα *open-class* προβλήματα, τα οποία θα παρουσιαστούν παρακάτω.

Με την εισαγωγή μίας οριακής τιμής στο μοντέλο, μπορούμε να καταλάβουμε πως για διαφορετικές τιμές οριακής τιμής, το μοντέλο επιτυγχάνει διαφορετικές τιμές *false positive, true positive, false negative* και *true negative*. Συχνά χρησιμοποιούνται δύο καμπύλες για να αποτυπώσουν την αποτελεσματικότητα ενός μοντέλου κατηγοριοποίησης. Η πρώτη λέγεται *Receiver Operating Characteristic curve* (Figure 15 A) και απεικονίζει πως αλλάζει το ποσοστό *true positive* όσο μεταβάλλεται το ποσοστό *false positive* για τις διαφορετικές οριακές τιμές. Η δεύτερη λέγεται *Precision-Recall curve* (Figure 15 B).

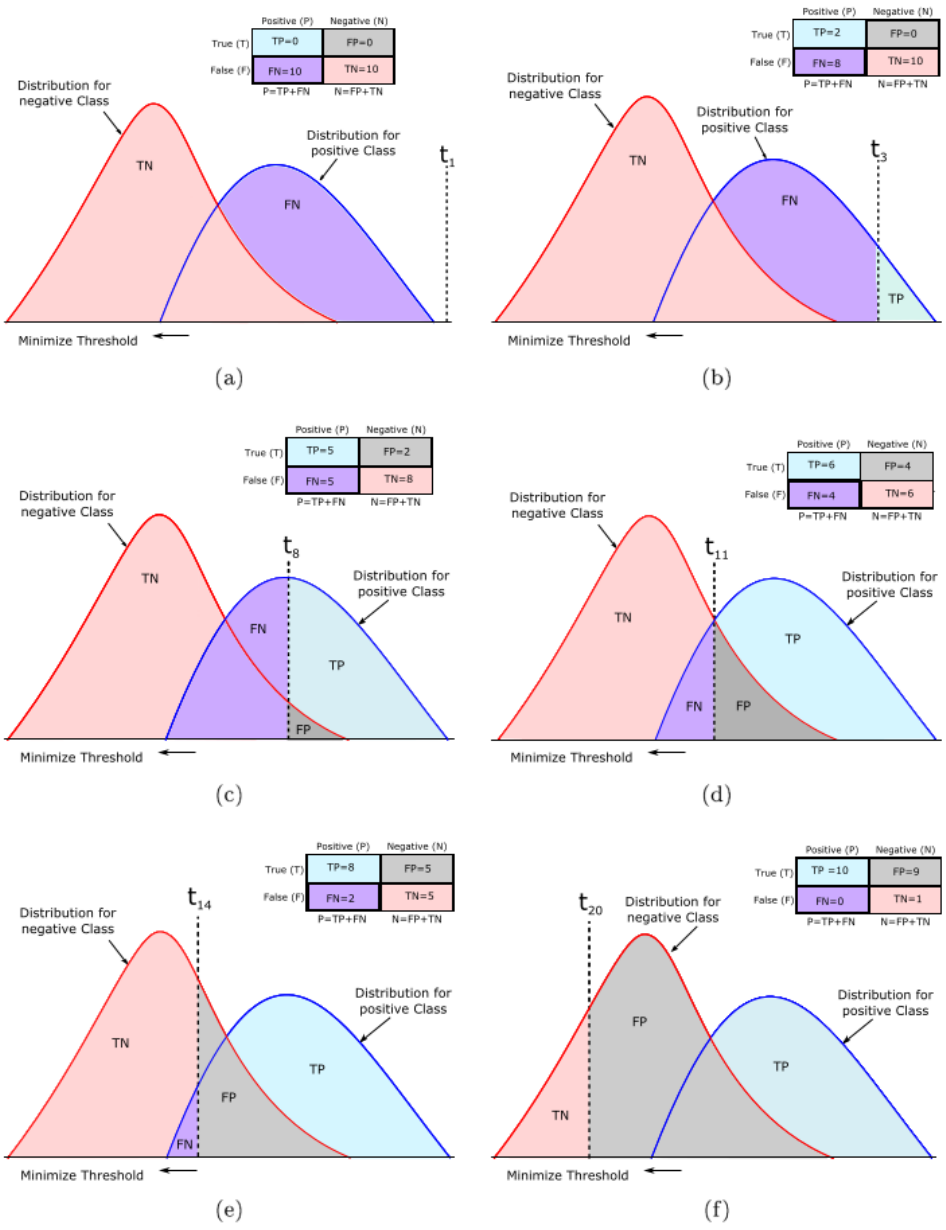

Figure 14 Ο τρόπος με τον οποίο μεταβάλλονται τα μεγέθη True Positive, False Positive, False Negative και True Negative, όσο αυξάνεται η οριακή τιμή (threshold) [16]

Πάλι, μεταβάλλοντας την οριακή τιμή, σχεδιάζουμε την τιμή *Precision* σε συνάρτηση με την τιμή *Recall*. Εδώ αξίζει να αναφέρουμε οι τιμές πιθανότητας που παράγουν διαφορετικά μοντέλα, δεν είναι συγκρίσιμες μεταξύ τους. Συνεπώς, για προβλήματα σύγκρισης μοντέλων μεταξύ τους, καμπύλες που αφορούν άμεσα την οριακή τιμή δεν παράγουν αξιόπιστα αποτελέσματα. Αντίθετα οι καμπύλες *Receiver Operating Characteristic* και *Precision-Recall*, που δεν περιέχουν αυτή την παράμετρο είναι κατάλληλες για τέτοιες συγκρίσεις. Συνήθως οι *Receiver Operating Characteristic* καμπύλες χρησιμοποιούνται όταν τα δεδομένα από κάθε κλάση είναι σχετικά ίσα σε αριθμό, ενώ οι *Precision-Recall* όταν έχουμε ανισορροπία στον αριθμό δεδομένων ανά κλάση.

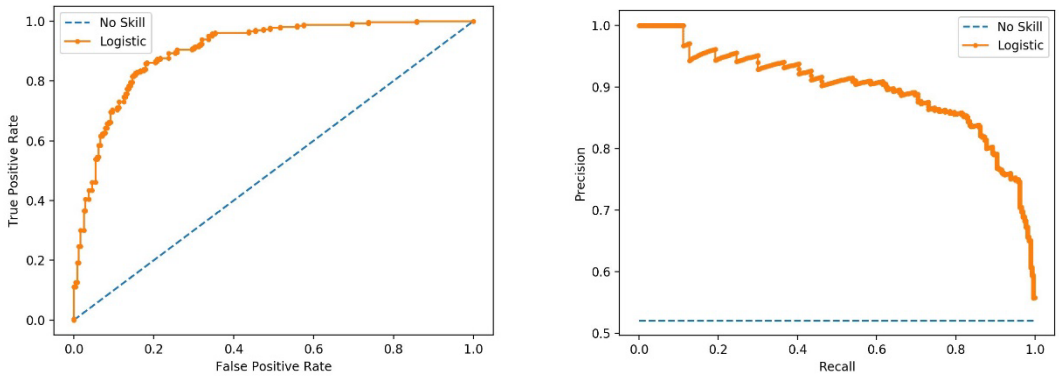

Figure 15 A. Receiver Operating Characteristic curve and B. Precision-Recall curve [17]

3.2.8 Open-Set Recognition problems

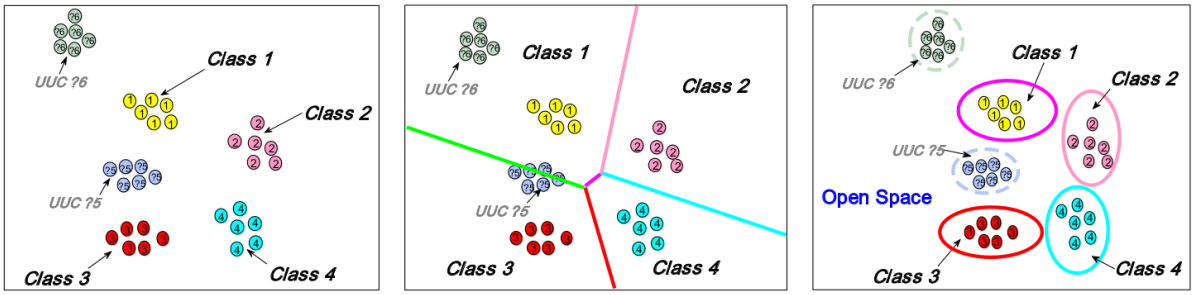

Figure 16 Σύγκριση μεταξύ κλασικού προβλήματος classification και open-set recognition. Στην πρώτη εικόνα φαίνεται η κατανομή του συνόλου δεδομένων, όπου UUC σημαίνει Unknown Unknown Class και αφορά δεδομένα με τα οποία ο αλγόριθμος δεν έχει έρθει σε επαφή. Η δεύτερη εικόνα δείχνει την κατηγοριοποίηση που θα έκανε ένας κλασικός αλγόριθμος και την αστοχία του να αντιμετωπίσει τις άγνωστες κλάσεις. Στην τρίτη εικόνα φαίνεται η προσέγγιση ενός αλγορίθμου για ένα ορισμένο open-set recognition πρόβλημα, όπου αφήνεται χώρος για να υπάρχουν δεδομένα εκτός γνωστών κλάσεων [18]

Συχνά στην μηχανική μάθηση, αντιμετωπίζουμε προβλήματα για τα οποία είναι εγγενώς αδύνατο ο αλγόριθμος να έρθει σε επαφή, να εκπαιδευτεί με όλα τα δυνατά δεδομένα και κλάσεις που υπάρχουν στον κόσμο. Περιορισμοί στην παραγωγή/συλλογή των δεδομένων καθιστούν ιδιαίτερα δύσκολη την εξάντληση όλων των πιθανών κλάσεων κατά την διάρκεια της εκπαίδευσης. Η ατελής γνώση που έχουμε την στιγμή της εκπαίδευσης και η επαφή με πρωτόγνωρα δεδομένα κατευθείαν κατά την διαδικασία της αξιολόγησης, ζητά από τους αλγορίθμους όχι μόνο να κατηγοριοποιούν σωστά τα γνωστά δεδομένα, αλλά και να αντιμετωπίζει αποδοτικά τα άγνωστα. Ορίζουμε *close-set* πρόβλημα, όταν η εκπαίδευση και η αξιολόγηση ενός αλγορίθμου συμβαίνει με δεδομένα από τον ίδιο χώρο. Πολλοί κλασικοί αλγόριθμοι μηχανικής μάθησης έχουν ήδη δείξει εξαιρετική λειτουργεία σε τέτοιου τύπου προβλήματα. Αντίθετα, στις περιπτώσεις, όπου ο αλγόριθμος μπορεί να έρθει σε επαφή με δεδομένα τα οποία δεν ανήκουν σε κάποια από τις γνωστές κλάσεις, αυτό αποτελεί open-set (Figure 16). Ο αλγόριθμος πρέπει αφενός να αναγνωρίσει αν το δείγμα υπό εξέταση ανήκει σε κάποια κλάση, και έπειτα να το αντιστοιχίσει σωστά σε μία. Κατά συνέπεια, η αξιολόγηση και σύγκριση τέτοιου είδους μοντέλων γίνεται πιο περίπλοκή. Αυτή η ανάγκη απαντήθηκε από τους Akshay Raj Dhamija et al. [19] και αφορά μία νέα καμπύλη, την Open-Set

Classification Rate curve (OSCR curve). Πρόκειται για την απεικόνιση του μεγέθους *Correct Classification rate* (CCR) σε συνάρτηση με το μέγεθος *False Positive rate* (FPR). Σύμφωνα με την μέθοδο, για δεδομένη οριακή τιμή θ, το CCR αφορά το ποσοστό των δειγμάτων όπου η μέγιστη πιθανότητά τους είναι μεγαλύτερη από θ και αντιστοιχεί στην σωστή κλάση ĉ (Equation 14). Αντίστοιχα το FPR περιγράφει το ποσοστό των "αρνητικών" δειγμάτων, των οποίων η μέγιστη πιθανότητα είναι μεγαλύτερη από θ και αντιστοιχεί σε κάποια γνωστή κλάση (Equation 15).

Equation 14

$$CCR(\theta) = \frac{\left|\left\{x \mid x \in D_c \wedge \arg\max_c P(c \mid x) = \hat{c} \wedge P\left(\hat{c} \mid x\right) \geq \theta\right\}\right|}{|D_c|}$$

Equation 15

$$FPR(\theta) = \frac{\left|\left\{x \mid x \in D_u \wedge \max_c P(c \mid x) \geq \theta\right\}\right|}{|D_u|}$$

, όπου D_u το σύνολο των "αρνητικών" δειγμάτων, D_c το σύνολο των δειγμάτων που αντιστοιχούν σε μία κλάση c, θ η οριακή τιμή και X ĉ η σωστή κλάση.

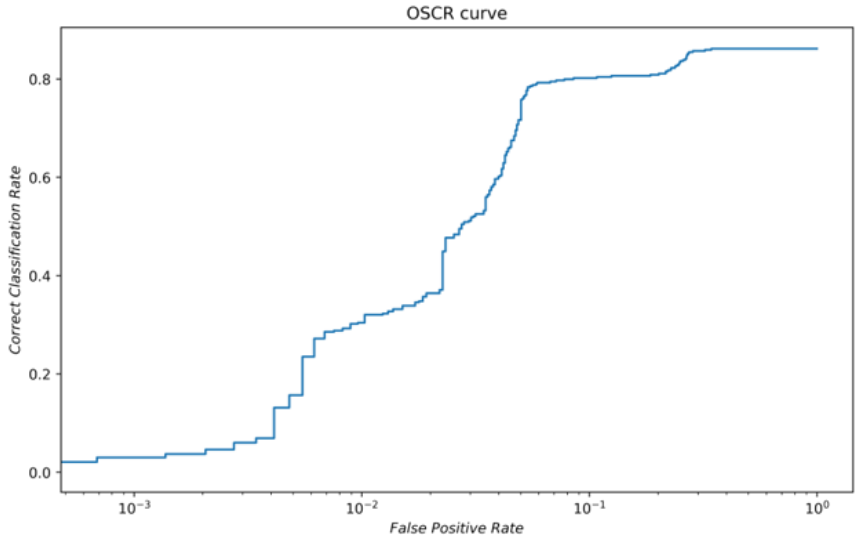

Figure 17 Open-Set Classification Rate curve

3.3 Μέθοδοι προεπεξεργασίας NIR σημάτων

Η NIR φασματοσκοπία αποτελεί φασματοσκοπική τεχνική με τις περισσότερες και ποικίλες μεθόδους προεπεξεργασίας των φασματικών σημάτων, κυρίαρχα γιατί τα NIR φάσματα συνδέονται με έναν μεγάλο αριθμό ανεπιθύμητων παρεμβολών. Υπάρχουν διάφοροι παράγοντες που οδηγούν τα NIR

φάσματα να υποφέρουν από ανεπιθύμητες φασματικές αλλοιώσεις και μεταβολές στην γραμμή βάσης (*baseline*). Αρχικά, σημαντικά αίτια σε αυτή την κατεύθυνση αποτελούν η σκέδαση του φωτός από στερεά ή θολά υγρά δείγματα, καθώς το μήκος κύματος του φωτός στο εγγύς υπέρυθρο είναι συχνά συγκρίσιμο με το μέγεθος των ατόμων που αποτελούν το δείγμα. Επίσης αλλαγές στην θερμοκρασία, υγρασία, αλλά και πυκνότητα του υλικού προς εξέταση, δημιουργούν αλλοιώσεις [20]. Όπως έχει αναφερθεί σε προηγούμενο κεφάλαιο, η NIR ηλεκτρομαγνητική ακτινοβολία έχει την ικανότητα να διεισδύει κάτω από την επιφάνεια του δείγματος. Η απόσταση που διανύει μέσα στο δείγμα αναφέρεται ως *path length*. Διαφορετικά *path lengths* προκαλούν ελαφριές μεταβολές στο φάσμα, καθιστώντας δύσκολη την ακριβή αναπαραγωγή του. Τέλος, παρεμβολές μπορούν επίσης να δημιουργηθούν από τα όργανα που χρησιμοποιεί η NIR φασματοσκοπία. Θόρυβος από τον ανιχνευτή και τον ενισχυτή μπορεί να προστεθεί, καθώς επίσης η χρησιμοποίηση οπτικής ίνας μπορεί να επιφέρει μεταβολές στην γραμμή βάσης [2, κεφάλαιο 4]. Όλοι αυτοί οι παράγοντες καθιστούν την προ-επεξεργασία των NIR φασμάτων να αποτελεί αναπόσπαστο κομμάτι της NIR ανάλυσης. Οι πιο συχνά χρησιμοποιούμενες μέθοδοι προεπεξεργασίας μπορούν να χωριστούν σε δύο κατηγορίες. Αυτή της διόρθωσης των φαινομένων σκέδασης (scatter - correction) και αυτή των φασματικών παραγώγων (*spectral derivatives - derivative spectroscopy*). Οι κυρίαρχες τεχνικές διόρθωσης σκεδάσεων είναι η *Standard Normal Variate*, η *Multiplicative Scatter Correction* και η μέθοδος κανονικοποίησης (*normalization*). Όσον αφορά τις φασματικές παραγώγους δύο τεχνικές που τις υλοποιούν είναι οι *Norris-Williams* και η *Savitzky-Golay* [21]. Παρακάτω θα αναλύσουμε συνοπτικά τις μεθόδους που χρησιμοποιήθηκαν στην συγκεκριμένη μελέτη.

3.3.1 Standard Normal Variate

Η Standard Normal Variate αποτελεί πολύ γνωστή μέθοδο προεπεξεργασίας των φασματικών NIR σημάτων ενάντια του θορύβου των σκεδάσεων. Ο μετασχηματισμός των φασμάτων προκύπτει από την παρακάτω εξίσωση (Equation 16).

Equation 16

$$x_{corr} = \frac{x_{org} - a_0}{a_1}$$

όπου x_{corr} το διορθωμένο φάσμα, x_{org} το αρχικό φάσμα, α_0 ο μέσος όρος του αρχικού φάσματος και α_1 η τυπική απόκλιση του αρχικού φάσματος.

Πρακτικά η SNV μέθοδος φτιάχνει τα φασματικά δεδομένα υπό κλίμακα και τα κεντράρει. Στην εικόνα (Figure 18) φαίνεται η επίδραση της μεθόδου σε ένα σύνολο φασμάτων ουρίας.

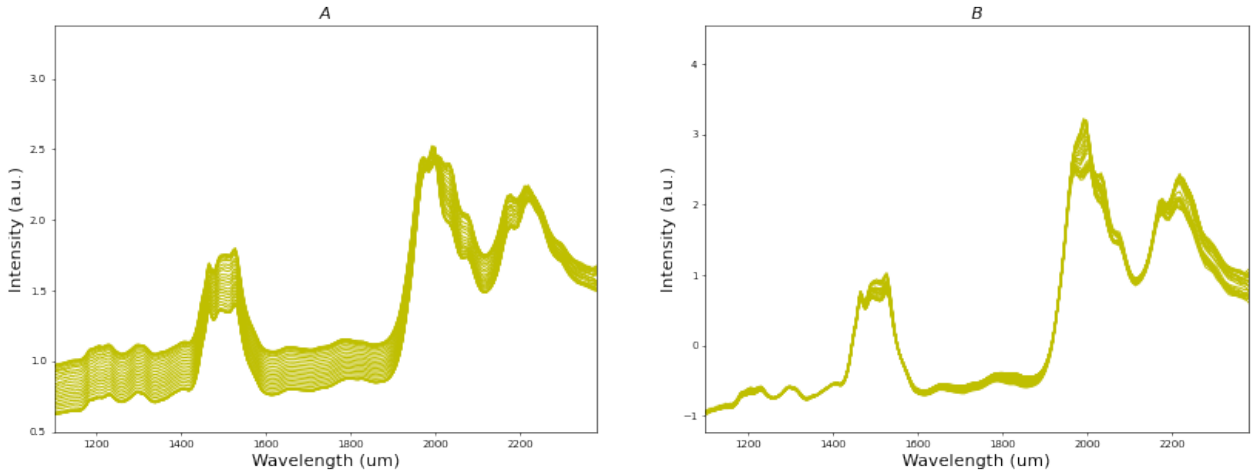

Figure 18 Η επίδραση της μεθόδου SNV σε ένα σύνολο φασμάτων ουρίας. A Τα αρχικά φάσματα, B Τα φάσματα μετά την επεξεργασία

3.3.2 Παράγωγοι φάσματος

Οι φασματικές παράγωγοι έχουν την ικανότητα να περιορίζουν τόσο αθροιστικές όσο και πολλαπλασιαστικές αλλοιώσεις στην γραμμή βάσης και χρησιμοποιούνται στην NIR φασματοσκοπία ήδη για δεκαετίες. Οι πιο συχνά χρησιμοποιούμενες παράγωγοι είναι η πρώτη και η δεύτερη, ενώ η χρήση μεγαλύτερων τάξεων παραγώγων συχνά αποφεύγεται λόγω αύξησης του σηματοθορυβικού λόγου. Ειδικά για φασματικά σήματα χαμηλών εντάσεων, η φασματικές παράγωγοι μπορεί να μην αποτελέσουν ικανές λύσεις.

$$Zero\ Order: \qquad A = f(\lambda)$$

$$First\ Order: \qquad \frac{dA}{d\lambda} = f'(\lambda)$$

$$Second\ Order: \qquad \frac{d^2A}{d\lambda^2} = f''(\lambda)$$

Η παράγωγος πρώτου βαθμού εκφράζει την αλλαγή του φάσματος απορρόφησης ως προς το μήκος κύματος. Ξεκινάει και τελειώνει στο μηδέν, ενώ μηδενίζει στα μέγιστα του φάσματος. Στα σημεία καμπής του φάσματος έχουμε τα τοπικά μέγιστα και ελάχιστα στην παράγωγο. Αυτό το μοτίβο περιγράφει όλες τις παραγώγους μονής τάξης. Το χαρακτηριστικό της δεύτερης παραγώγου είναι μία αρνητική ζώνη με ελάχιστο στο σημείο που το αρχικό φάσμα έχει το μέγιστο (Figure 19). Όσο υπολογίζονται μεγαλύτερες τάξεις παραγώγων του φάσματος, αυξάνεται ο αριθμός των ζωνών. Αυτή η αύξηση στην πολυπλοκότητα του φάσματος μπορεί να είναι πολύ χρήσιμη στην ποιοτική ανάλυση [22]. Φάσματα απορρόφησης που μοιάζουν πολύ, μπορεί να έχουν σημαντικές διαφορές στην παράγωγό τους.

Ένα συχνά ανεπιθύμητο φαινόμενο στην φασματοσκοπία είναι η μετατόπιση γραμμής βάσης (*baseline shift*). Είναι μία μορφή θορύβου που προκαλείται από αστάθειες των οργάνων, ή από τον χειρισμό

του δείγματος. Η φασματοσκοπία παραγώγων συχνά χρησιμοποιείται για διόρθωση γραμμής βάσης (*baseline correction*). Η παράγωγος των σταθερών είναι μηδέν, συνεπώς η παραγώγιση του φάσματος απορρόφησης διορθώνει σημαντικά την γραμμή βάσης (Figure 20).

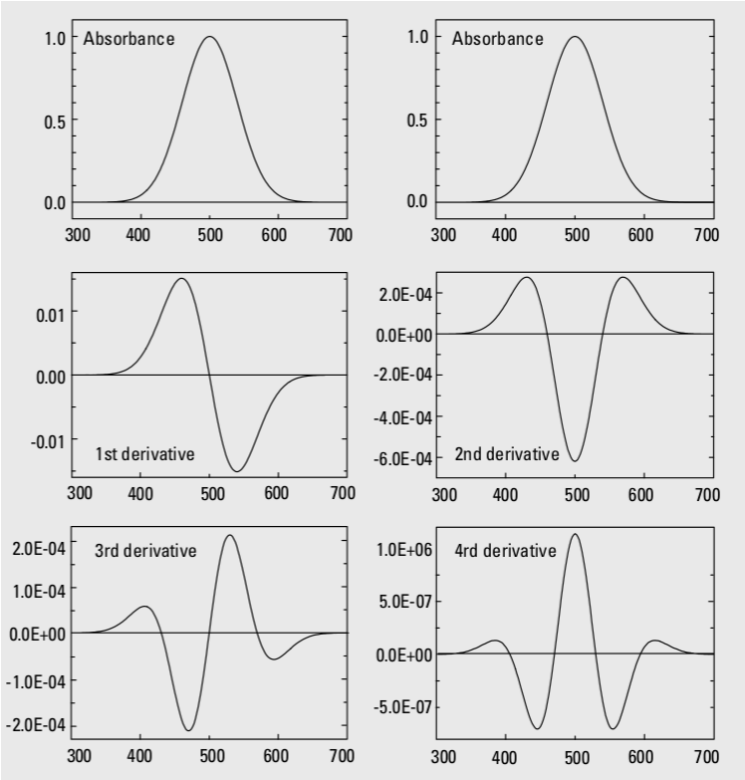

Figure 19 Οι παράγωγοι μίας gaussian φασματικής ζώνης [22]

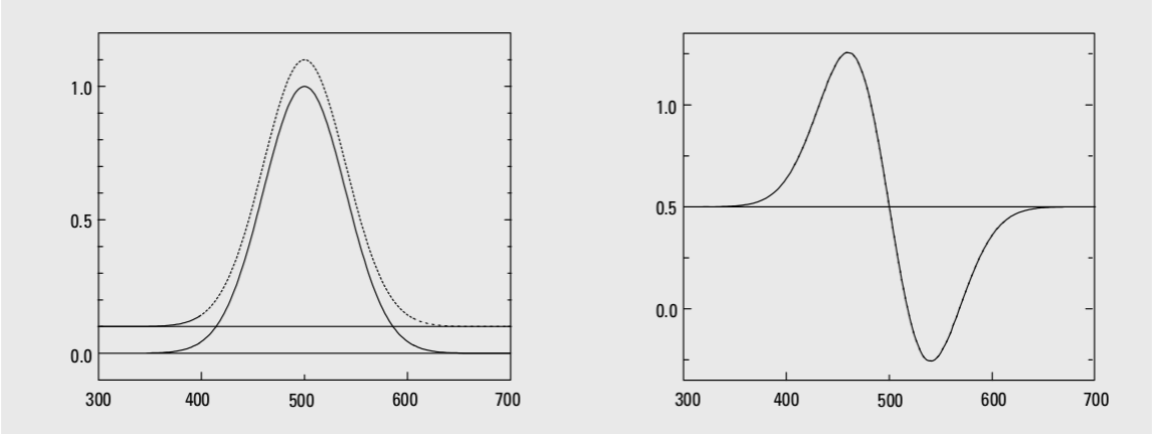

Figure 20 Ελαχιστοποίηση της μετατόπισης γραμμής βάσης με χρήση πρώτης παραγώγου [22]

Ένα ακόμα σημαντικό πλεονέκτημα των φασματικών παραγώγων είναι η εξασθένηση που δείχνουν ανάλογα με την ευρύτητα των ζωνών απορρόφησης. Πιο συγκεκριμένα οι παράγωγοι ευρειών ζωνών τείνουν να έχουν πολύ χαμηλότερη ένταση σε σύγκριση με παραγώγους οξειών ζωνών (Figure 21). Αυτό οφείλεται στο γεγονός πως το πλάτος D_n μίας Γκαουσιανής ζώνης στη n-ιοστή παράγωγο είναι αντιστρόφως ανάλογο με το αρχικό εύρος υψωμένο στην n (Equation 17).

Equation 17

$$D^n = \frac{1}{W^n}.$$

Αυτό το χαρακτηριστικό είναι ιδιαίτερα χρήσιμο στην NIR φασματοσκοπία, όπου κυριαρχούν ευρείες ζώνες. Εδώ, βέβαια, πρέπει να σημειωθεί ότι σε ασθενή φάσματα, η παραγώγιση μπορεί να μειώσει επικίνδυνα την ένταση του φάσματος, καθώς επίσης και η παραγώγιση μεγάλων βαθμών να δημιουργήσει πλευρικούς λοβούς επικίνδυνους για την ανάλυση του φάσματος.

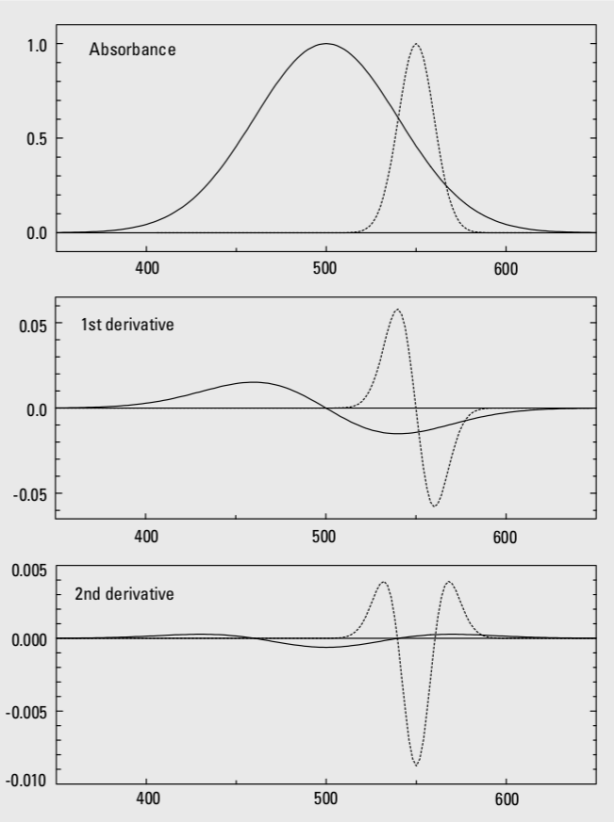

Figure 21 Εξασθένηση ευρειών ζωνών στις φασματικές παραγώγους [22]

3.4 Hit-Quality index και φασματική βιβλιοθήκη

Με την ανάπτυξη του υλικού των υπολογιστών, έγινε δυνατή η ταυτοποίηση υλικών, ειδικά από φορητά φασματοσκοπικά συστήματα. Η ταυτοποίηση προϋποθέτει την δημιουργία μιας βιβλιοθήκης φασμάτων και τον προσδιορισμό κάποιων κριτηρίων, τα οποία θα αποφασίζουν την κλάση στην οποία ανήκει ένα υλικό [2, κεφάλαιο 18.5]. Η βιβλιοθήκη θα πρέπει να αναπτυχθεί με τέτοιον τρόπο ώστε να

αντιπροσωπεύει την ποικιλία των φασμάτων που πρόκειται να ταυτοποιηθούν. Εάν τα μελλοντικά δείγματα ενδέχεται να μετρηθούν σε διαφορετικές καταστάσεις, αυτό πρέπει να συμπεριλαμβάνεται στην βιβλιοθήκη, διαφορετικά το πρόγραμμα θα αδυνατεί να κατηγοριοποιεί σωστά τις ουσίες. Καθότι είναι δύσκολο εκ των προτέρων να προβλεφθούν όλες οι αιτίες που μπορούν να μεταβάλλουν το φάσμα μίας ουσίας, είναι χρήσιμο να θεωρηθούν και να συμπεριληφθούν στην βιβλιοθήκη πολλά και διαφορετικά χημικά εύρη. Όταν τελικά δημιουργηθεί μία αντιπροσωπευτική βιβλιοθήκη φασμάτων, δημιουργείται ένας αλγόριθμος για την κατηγοριοποίηση των φασμάτων.

Όπως έχει αναλυθεί στην εισαγωγή, το NIR φάσμα είναι αρκετά περίπλοκο, συνεπώς και για την ταυτοποίηση ουσιών χρησιμοποιούνται μέθοδοι αναγνώρισης προτύπων [23]. Γενικά, μέθοδοι αναγνώρισης προτύπων αποτελούν και αλγόριθμοι μηχανικής μάθησης, ωστόσο σε αυτό το κεφάλαιο θα ασχοληθούμε με τους συντελεστές συσχέτισης (*Correlation Coefficient*). Οι συντελεστές χρησιμοποιούνται για να αποδώσουν αριθμητικά την ομοιότητα ενός άγνωστου φάσματος με κάποιο γνωστό από την βιβλιοθήκη, την τιμή *Hit Quality Index* (HQI). Η τιμή αυτή γίνεται 1 για ίδια φάσματα ενώ μειώνεται όσο η μεγαλώνει η διαφορά των φασμάτων. Υπάρχουν διάφοροι τρόποι υπολογισμού της τιμής συσχέτισης. Συχνά χρησιμοποιούμενος είναι ο *Correlation Derivative* που έχει τύπο [24]:

Equation 18

$$HQI = \frac{(library \cdot unknown)^2}{(library \cdot library)(unknown \cdot unknown)},$$

όπου *library* είναι το φάσμα αναφοράς από την βιβλιοθήκη και *unknown* το φάσμα υπό εξέταση.

4. Περιγραφή πειράματος

Σκοπός της παρούσας εργασίας ήταν η ανάπτυξη ενός αλγορίθμου ικανού να ανιχνεύει συγκεκριμένους χημικούς προδρόμους εκρηκτικών ουσιών (*precursors of explosives*), ο οποίος στην συνέχεια θα ήταν δυνατό να εγκατασταθεί στο λογισμικό ενός φορητού FT-NIR αισθητήρα. Πιο συγκεκριμένα, στόχος του αλγορίθμου ήταν να εντοπίζει και να κατηγοριοποιεί τους εξής προδρόμους: νιτρικό αμμώνιο (*ammonium nitrate*), ουρία (*urea*), νιτρικό κάλιο (*potassium nitrate*) και νιτρικό νάτριο (*sodium nitrate*). Όλες αυτές οι ουσίες υπάρχουν στις συνήθεις περιβαλλοντικές συνθήκες σε μορφή λευκής σκόνης, είναι εμπορικά διαθέσιμες και μπορούν να χρησιμοποιηθούν για την κατασκευή χειροποίητων, ισχυρών εκρηκτικών. Υπάρχουν και εναλλακτικές χρήσεις των ουσιών, όπως για παράδειγμα η χρησιμοποίησή τους σαν λιπάσματα και συντηρητικά κρεάτων. Ήδη σε πολλά διαφορετικά νομοθετικά πλαίσια ανά τον κόσμο, οι πρόδρομοι αυτοί συνδέονται με προβλέψεις που θέτουν περιορισμούς σχετικά με την χρήση, μετακίνηση και αγοροπωλησία τους [26][27]. Προσθετικά στους προαναφερθείς λόγους, επιλέξαμε αυτούς τους προδρόμους καθότι παρουσιάζουν πληροφορία στο φάσμα τους εγγύς υπέρυθρου [25]. Εκτός της βιβλιογραφίας που το επιβεβαιώνει, αυτό είναι αναμενόμενο καθότι οι ουσίες αυτές περιέχουν O-H, N-H και C-H δεσμούς, οι οποίοι δίνουν υπερτονικές και ζώνες συνδυασμού στο εγγύς υπέρυθρο φάσμα.

Όσον αφορά το κομμάτι του αλγορίθμου, η εργασία εξελίχθηκε σε διαφορετικά στάδια. Αρχικά δουλέψαμε με έναν κλασικό αλγόριθμο ο οποίος χρησιμοποιούσε διαφορετικούς συντελεστές συσχέτισης (*correlation coefficients*), συνέκρινε κάποια πρότυπα φάσματα με το άγνωστο, παρήγαγε μία τιμή HQI και κατέληγε σε κάποια πρόβλεψη σχετικά με την φύση της ουσίας. Σε αυτή την φάση, συγκρίναμε τους διαφορετικούς συντελεστές συσχέτισης μεταξύ τους για να καταλήξουμε στον πλέον αποδοτικό. Οι εν λόγω συντελεστές συσχέτισης ήταν οι *First Derivative Correlation search*, *Pearsons Correlation search*, *Euclidean Distance search*, *Absolute Value search* και *Least Squares search* [24]. Οι *search* αλγόριθμοι που απέδιδαν καλύτερα ήταν οι *First Derivative Correlation search* και *Pearsons Correlation search* σύμφωνα με μία επιδερμική ανάλυση.

Τα δεδομένα μας σε εκείνη την φάση ήταν λίγα σε αριθμό, περίπου 100 μετρήσεις-φάσματα που συμπεριλάμβαναν φάσματα τριών από τους προδρόμους (νιτρικό αμμώνιο, ουρία, νιτρικό νάτριο) καθώς και φάσματα από αλεύρι, ζάχαρη και αλάτι. Βασισμένοι στην δουλειά των Zapata F. et al. [25] αποφασίσαμε να προμηθευτούμε επιπλέον νιτρικό κάλιο, καθώς έδειχνε κάποια πληροφορία στο NIR και βρίσκεται σε μορφή λευκής σκόνης. Επιπλέον προμηθευτήκαμε νιτρικό αμμώνιο δημητρίου (*ammonium cerium nitrate*) και *diazolidinyl urea* τα οποία έχουν ιδιαίτερα όμοια χημική δομή με τους προδρόμους ammonium nitrate και urea, είναι και αυτά σε μορφή σκόνης. Οι ουσίες αυτές χρησιμοποιήθηκαν σαν "αρνητικά δείγματα" (*negative samples*), αποτελώντας σοβαρή πρόκληση για τον εκάστοτε αλγόριθμο να τα μπερδέψει με τους κοντινά χημικούς προδρόμους.

Ταυτόχρονα, με βάση μία βιβλιογραφική αναζήτηση, σε εφαρμογές ανάλυσης NIR φασμάτων μίας συγκεκριμένης ουσίας ή ομάδας ουσιών, πιο συνήθης ήταν η χρήση αλγορίθμων μηχανικής μάθησης [28][29][30][31][32]. Συνεπώς, συνέχεια της παρούσας εργασίας ήταν η αναζήτηση και ανάπτυξη καταλλήλων μοντέλων μηχανικής μάθησης και η μεταξύ τους σύγκριση, καθώς επίσης και η σύγκρισή τους με έναν συντελεστή συσχέτισης με σκοπό την εύρεση του πιο αποδοτικού μοντέλου, και την εξαγωγή

σχετικών συμπερασμάτων. Για την τελική αξιολόγηση των μοντέλων που αναπτύχθηκαν, προμηθευτήκαμε με ακόμα επτά αρνητικά δείγματα, λευκές σκόνες που οι άνθρωποι συχνά χρησιμοποιούν και έχουν πάνω τους. Αυτές αποτελούν η βρεφική πούδρα, η βανίλια, το μαχαλέπι, η μαστίχα, η μαγειρική σόδα, η μαγειρική αμμωνία και η παρακεταμόλη. Η τελική μορφή της έρευνας συμπεριλάμβανε συνολικά δώδεκα ουσίες αρνητικών δειγμάτων και τους τέσσερις προδρόμους εκρηκτικών, ενώ τέσσερα μοντέλα μηχανικής μάθησης και ένα μοντέλο με συντελεστή συσχέτισης απάρτισαν το τελικό σύνολο αλγορίθμων που αξιολογήθηκαν.

4.1 Παραγωγή δειγμάτων

Στην τελική του μορφή, το σύνολο δεδομένων αποτελείται από 2911 δείγματα από τους χημικούς προδρόμους εκρηκτικών, τα οποία μπορούν να χωριστούν σε δύο υποσύνολα. Το πρώτο σύνολο D_c περιέχει 1456 δείγματα από τους προδρόμους νιτρικό αμμώνιο , ουρία, νιτρικό κάλιο και νιτρικό νάτριο. Στο πλαίσιο του προβλήματος της ταυτοποίησης και κατηγοριοποίησης των προδρόμων, θεωρούμε πως κάθε πρόδρομος αποτελεί και μία κλάση, οι οποίες απαρτίζονται από 364 δείγματα έκαστην. Υπάρχουν αναφορές σχετικά με την αλλοίωση των φασματικών δεδομένων στο εγγύς υπέρυθρο, μεταβάλλοντας την θερμοκρασία και την σχετική υγρασία του περιβάλλοντος κατά την μέτρηση. Στην πράξη, ο φορητός φασματογραφικός αισθητήρας αναμένεται να χρησιμοποιηθεί σε περιβάλλοντα με ποικίλες συνθήκες θερμοκρασίας και υγρασίας, γι αυτόν τον λόγο για τον κάθε πρόδρομο δημιουργήσαμε 8 διαφορετικές συνθήκες περιβάλλοντος στις οποίες υποβάλαμε τα δείγματα, πριν και κατά την διάρκεια της μέτρησης. Πιο συγκεκριμένα, χρησιμοποιήσαμε ένα ψυγείο με ψύξη στους 7°C, έναν καταψύκτη με ψύξη περίπου στους -18°C. Για τις υψηλές θερμοκρασίες χρησιμοποιήθηκε μία κλειστή, πλαστική και διαφανής συσκευασία στην οποία τοποθετήθηκαν οι ουσίες. Η συσκευασία αυτή ακτινοβολούταν από μία θερμή λάμπα και όταν αυτό ήταν αναγκαίο, χρησιμοποιήθηκε ένα πιστόλι θερμού αέρα. Για την αύξηση της υγρασίας στις υψηλές θερμοκρασίες, χρησιμοποιήσαμε ένα πλατύ δοχείο με ζεστό νερό το οποίο τοποθετήσαμε μέσα στην συσκευασία με τα δείγματα. Το νερό από το δοχείο εξατμιζόταν σταδιακά, αυξάνοντας την σχετική υγρασία. Για τις χαμηλές θερμοκρασίες, χρησιμοποιήθηκε ένα θερμομονωτικό δοχείο για την μεταφορά των ουσιών από την ψύξη στο σημείο του εργαστηρίου που συγκεντρώθηκαν οι μετρήσεις. Και σε αυτή την περίπτωση χρησιμοποιήσαμε ένα πλατύ δοχείο με νερό για να αυξήσουμε την υγρασία περιβάλλοντος. Σε ορισμένες μετρήσεις ψεκάσαμε επίσης ελαφρά απιονισμένο νερό μέσα στις συσκευασίες, αφενός για την περεταίρω αύξηση της θερμοκρασίας, αλλά και γιατί γνωρίζουμε ότι το νερό δίνει πληροφορία στο εγγύς υπέρυθρο, το οποίο μπορεί να αποτελέσει πρόβλημα για την ταυτοποίηση μίας ουσίας η οποία φέρει στην επιφάνειά της υγρασία. Το δεύτερο σύνολο D_u απαρτίζεται από τις "αρνητικές" ουσίες. Πιο συγκεκριμένα, πραγματοποιήσαμε 1455 μετρήσεις που αφορούν χημικές ουσίες, οι οποίες δεν είναι πρόδρομοι εκρηκτικών. Τις επιλέξαμε με βάση δύο κριτήρια. Οι ουσίες νιτρικό αμμώνιο δημητρίου (*ammonium cerium nitrate*) και *diazolidinyl* urea επιλέχθηκαν λόγω την χημικής ομοιότητάς τους με τους προδρόμους νιτρικό αμμώνιο και ουρία αντίστοιχα. Παρήχθησαν 265 δείγματα από νιτρικό αμμώνιο δημητρίου σε έξι διαφορετικές συνθήκες περιβάλλοντος και 200 δείγματα από *diazolidinyl urea* σε 5 συνθήκες περιβάλλοντος. Αναφορικά με την τελευταία ουσία δεν ενδείκνυται να βρίσκεται σε υψηλές θερμοκρασίες, συνεπώς την αποκλείσαμε από τις διαδικασίες μέτρησης όπου προκαλούσαμε θερμό περιβάλλον. Έπειτα, στην παραγωγή δειγμάτων συμπεριλάβαμε δέκα ουσίες, που βρίσκονται σε μορφή λευκής σκόνης και χρησιμοποιούνται και μεταφέρονται καθημερινά από τους ανθρώπους, το οποίο αποτέλεσε και το δεύτερο κριτήριο. Αυτές οι ουσίες συμπεριλαμβάνουν το αλεύρι, την ζάχαρη, το αλάτι,

την μαγειρική αμμωνία, την μαστίχα, την πούδρα για μωρά, το μαχαλέπι, την παρακεταμόλη, την βανίλια και την μαγειρική σόδα. Παρήχθησαν 99 δείγματα από κάθε ουσία.

4.2 Ο φασματογραφικός αισθητήρας

Τα δείγματα παρήχθησαν με την βοήθεια του φασματογραφικού αισθητήρα που κατασκευάστηκε στο εργαστήριο. Ο αισθητήρας αυτός έχει την δυνατότητα να μεταφέρεται, και με την βοήθεια μίας μπαταρίας είναι αυτόνομος. Το σύστημα FT-NIR αισθητήρα αποτελείται από διαφορετικά μέρη τα οποία απεικονίζονται στην φωτογραφία (Figure 22).

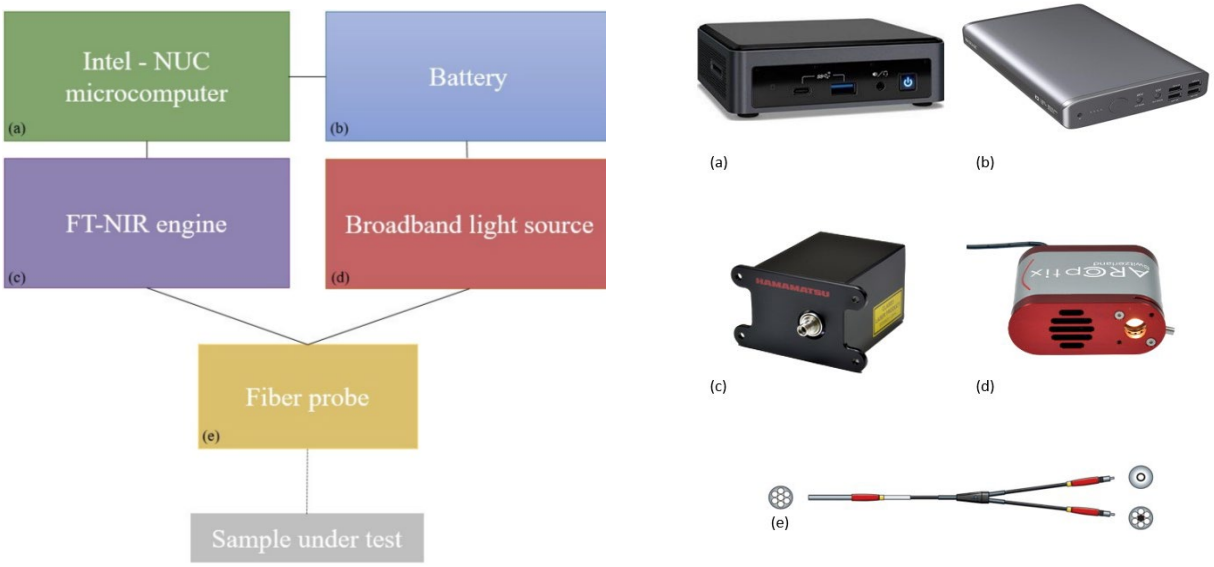

Figure 22 Τα μέρη που απαρτίζουν τον φασματογραφικό αισθητήρα FT-NIR

Figure 23 Ο φασματογραφικός αισθητήρας FT-NIR

Το φως στον αισθητήρα παράγεται από μία σταθεροποιημένη πηγή βολφραμίου-αλογόνου (*Tungsten-Halogen*) [33], η οποία συνδέεται με μία οπτική ίνα 6:1. Η πηγή εκπέμπει ακτινοβολία μελανού σώματος 20 mWatt στο φάσμα 400 με 4000 nm μηκών κύματος. Συνεπώς, καθότι η FT-NIR μηχανή λειτουργεί στα μήκη κύματος 1100 με 2500 nm, η πηγή αυτή καλύπτει ικανά τις ανάγκες του συστήματος. Η πηγή βολφραμίου-αλογόνου αποτελεί είδος λάμπας πυρακτώσεως. Το ρεύμα περνάει μέσα από το νήμα βολφραμίου-αλογόνου και έτσι το νήμα θερμαίνεται μέχρι περίπου 2850 Kelvin. Σε αυτή τη θερμοκρασία το βολφράμιο παράγει ορατό και υπέρυθρο φως, και κατά συνέπεια προκύπτει μία πηγή φωτός, η οποία θα μπορούσε να παρομοιαστεί με πηγή μελανού σώματος. Η οπτική ίνα που χρησιμοποιήθηκε έχει συνολικό μήκος 2 μέτρα και το μέγεθος του πυρήνα είναι 600 μm [34]. Τα μήκη κύματος ηλεκτρομαγνητικής ακτινοβολίας που μπορεί να μεταφέρει είναι από 400 μέχρι 2500 nm. Αποτελείται από 6 περιφερειακές ίνες μέσω των οποίων η δέσμη φωτός μεταφέρεται από την πηγή στο δείγμα και από μία κεντρική, η οποία λαμβάνει το φως που ανακλάται στο δείγμα. Έπειτα το φως μέσα από την κεντρική ίνα οδηγείται στον FT-NIR φασματογράφο [35]. Το διάχυτο ανακλώμενο φως που επιστρέφει από την ουσία και μετράται από το FT-NIR χρειάζεται πρώτα ένα σήμα αναφοράς. Πρόκειται για το φάσμα που παράγεται από τα όργανα μέτρησης, καθότι για παράδειγμα η πηγή μπορεί να μην ακτινοβολεί στην ίδια ένταση για όλα τα μήκη κύματος. Έτσι, προτού ακτινοβοληθεί η ουσία υπό μελέτη, ακτινοβολεί πρώτα μία λευκή λεία επιφάνεια, μία επιφάνεια αναφοράς που έχει τοποθετηθεί στο τέλος του καθετήρα της οπτικής ίνας. Ο πυρήνας του FT-NIR αισθητήρα είναι ο FT-NIR φασματογράφος, ο οποίος είναι υπεύθυνος για την λήψη και επεξεργασία της φασματικής απόκρισης από το ανακλώμενο φως στο δείγμα που αφορά τα μήκη κύματος στο εγγύς υπέρυθρο. Το μέγεθός του είναι αρκετά μικρό, τόσο που θα μπορούσε να μεταφέρεται με το ένα χέρι. Τα χαρακτηριστικά του βοηθούν ιδιαίτερα στην συγκεκριμένη εφαρμογή, καθότι μπορεί να χρησιμοποιηθεί σε οποιαδήποτε τοποθεσία χρειαστεί, προσφέροντας συνεχείς μετρήσεις σε άμεσο χρόνο. Αυτό αποδεσμεύει τους χειριστές του οργάνου από το να μεταφέρουν τα δείγματα σε μίας εργαστηριακή μονάδα για ανάλυση. Στο σύστημα αυτό ανήκει το συμβολόμετρο *Michelson* και ένα κύκλωμα ελέγχου. Το λογισμικό από τον FT-NIR φασματογράφο φορτώθηκε στον μικροϋπολογιστή, ο οποίος συνδέθηκε με τον φασματογράφο μέσω USB. Το λογισμικό πλέον μπορεί να εξάγει την πληροφορία από το συμβολόγραμμα που έχει παραχθεί στην FT-NIR μηχανή. Τελικά, ο FT-NIR φασματογράφος παράγει έξι καμπύλες, το συμβολόγραμμα (*interferometer*) πηγής και δείγματος μαζί με πηγή, το φάσμα που προκύπτει μετά τον μετασχηματισμό Fourier από την πηγή και από την πηγή μαζί με το δείγμα, το φάσμα απορρόφησης (*absorption spectrum*) του δείγματος και την δεύτερη παράγωγο του φάσματος απορρόφησης. Επιπλέον, το λογισμικό του φασματογράφου περιέχει κάποιες ακόμα δυνατότητες, όπως είναι ο ορισμός διαφόρων παραμέτρων, η αποθήκευση των δεδομένων και η γραφική αναπαράστασή τους. Ο μικροϋπολογιστής που χρησιμοποιήθηκε είναι ο Intel® NUC [36], ο οποίος επίσης εφοδιάζει τον φασματογράφο με ενέργεια μέσω USB. Επίσης στον μικροϋπολογιστή μπορεί να εγκατασταθεί το μοντέλο που κατασκευάστηκε στην παρούσα εργασία, με σκοπό να ανιχνεύει τους διαφορετικούς χημικούς προδρόμους εκρηκτικών.

4.3 Προεπεξεργασία και οπτικοποίηση δεδομένων

Όπως έχει αναλυθεί προηγουμένως στην θεωρία, τα φάσματα των ουσιών στο NIR είναι περίπλοκα, αλλά και δύσκολα να αναπαραχθούν ακόμα και όταν πρόκειται για την ίδια ουσία [2, κεφάλαιο 4.3]. Κατά συνέπεια, η ανάλυση των NIR φασμάτων έχει συνδεθεί με την εφαρμογή μεθόδων προεπεξεργασίας στα δεδομένα, πριν από την ανάλυση και γενικότερα την χρησιμοποίησή τους. Στην

παρούσα εργασία χρησιμοποιήσαμε δύο διαφορετικούς τρόπους προεπεξεργασίας. Ο πρώτος αφορά την εφαρμογή της μεθόδου *Standard Normal Variate* στα δεδομένα για την κανονικοποίησή τους. Ο δεύτερος αφορά την εφαρμογή της SNV και ακολούθως την παραγώγιση των φασμάτων με την μέθοδο *Savitzky-Golay* (*1st derivative, 19-datapoint smoothing window*). Με την βοήθεια του αλγορίθμου PCA οπτικοποιήσαμε την κατανομή των δεδομένων στον χώρο πριν αλλά και μετά την προεπεξεργασία.

4.4 Ανάπτυξη μοντέλων

Για την αναγνώριση και κατηγοριοποίηση των χημικών προδρόμων εκρηκτικών αναπτύξαμε 5 διαφορετικά μοντέλα και τα συνδυάσαμε και με τους δύο τρόπους προεπεξεργασίας, με σκοπό την εύρεση του πιο αξιόπιστου μοντέλου. Τα πρώτα τέσσερα μοντέλα αφορούν τους αλγορίθμους *Random Forest* και *Support Vector Machine*. Δύο μοντέλα αποτελούν οι αλγόριθμοι αυτοί, και τα άλλα δύο είναι ο συνδυασμός των αλγορίθμων με την μέθοδο *Principal Component Analysis* για μείωση της διαστασιμότητας, κρατώντας τις 200 πρώτες συνιστώσες οι οποίες εξέφραζαν περισσότερο από το 99% της πληροφορίας. Συνδυάσαμε τους αλγορίθμους με την PCA για την αποφυγή της κατάρας της διαστατικότητας (*curse of dimensionality*) [38]. Αυτό αφορά δεδομένα με μεγάλο αριθμό χαρακτηριστικών που ο αλγόριθμος καλείται να λάβει υπ' όψει. Στα φασματικά δεδομένα, η διάσταση αντιστοιχεί στον αριθμό μηκών κύματος που δειγματοληπτούνται κατά την σάρωση του φάσματος. Στην μηχανική μάθηση όσο αυξάνεται ο αριθμός των χαρακτηριστικών, τα μοντέλα μπορούν να κάνουν καλύτερες προβλέψεις. Μετά από κάποιο σημείο όμως, η απόδοσή τους μειώνεται εκθετικά με την αύξηση της διαστατικότητας. Συχνά τα μοντέλα υπερ-προσαρμόζονται στα δεδομένα εκπαίδευσης (*overfitting*) βρίσκοντας την "τέλεια λύση" για το συγκεκριμένο σετ δεδομένων, αλλά υστερούν στην ικανότητα γενίκευσης. Έτσι συχνά η PCA συνδυάζεται με τους αλγορίθμους, για να προλάβει τέτοια προβλήματα. Όσον αφορά την επιλογή των υπερπαραμέτρων, χρησιμοποιήσαμε την μέθοδο *GridSearch* ενώ η αξιολόγησή τους έγινε με την *3-Fold Cross-Validation*. Το πέμπτο μοντέλο αποτελεί η σταθερά συσχέτισης *Hit Quality Index*, σε συνδυασμό με μία φασματική βιβλιοθήκη που κατασκευάσαμε, η οποία περιέχει ένα φάσμα από κάθε διαφορετική συνθήκη μέτρησης για κάθε πρόδρομο.

Για να αντιμετωπίσουμε την επιπλέον "αρνητική" κλάση του open-set προβλήματος, σε όλα τα μοντέλα ορίσαμε την μεταβλητή *threshold*, η οποία ρυθμίζει την οριακή τιμή, όπου δείγματα με μέγιστη πιθανότητα μικρότερη από αυτή αξιολογούνται ως "αρνητικά".

4.5 Μέθοδοι αξιολόγησης

Οι περισσότεροι αλγόριθμοι μηχανικής μάθησης κατασκευάζονται με την υπόθεση πως τα δεδομένα στα οποία θα εκτεθούν, έχουν την ίδια κατανομή με τα δεδομένα με τα οποία εκπαιδεύτηκαν. Στην πραγματικότητα όμως, οι συνθήκες κάτω από τις οποίες έχουν παραχθεί τα δείγματα εκπαίδευσης μπορεί να μην είναι οι αναμενόμενες όταν αργότερα ο αλγόριθμος χρησιμοποιηθεί στον πραγματικό κόσμο. Αυτό μπορεί να δημιουργήσει πρόβλημα και ο αλγόριθμος να μην μπορεί να αναγνωρίσει τα νέα δείγματα. Αυτή η διαφορετικότητα στις κατανομές των δεδομένων εκπαίδευσης και των πραγματικών δεδομένων συχνά αναφέρεται ως *distribution shift* [37]. Για να αντιμετωπίσουμε αυτή τη πρόκληση, χωρίσαμε τα δεδομένα σε σετ εκπαίδευσης και αξιολόγησης, με τέτοιο τρόπο ώστε το μοντέλο να εκτίθεται σε νέα δεδομένα και πιθανώς εκτός κατανομής. Πιο συγκεκριμένα, από σετ εκπαίδευσης αποκλείσαμε δύο σετ δεδομένων μετρημένων σε δύο διαφορετικές συνθήκες περιβάλλοντος. Τελικά, το σετ εκπαίδευσης (*training set*) απαρτίστηκε από το 92% των δειγμάτων των υπόλοιπων έξι σετ συνθηκών περιβάλλοντος

από τους χημικούς πρόδρομους των εκρηκτικών, ενώ το σετ αξιολόγησης (*test set*) από όλα τα δείγματα των δύο σετ που αποκλείστηκαν από το σύνολο εκπαίδευσης, το 8% των υπόλοιπων έξι σετ για κάθε πρόδρομο, όπως επίσης ολόκληρο το σύνολο των αρνητικών δειγμάτων. Με αυτόν τον τρόπο μπορούμε να εξασφαλίσουμε πως κατά την αξιολόγηση το μοντέλο θα κληθεί να αντιμετωπίσει φάσματα των προδρόμων σε καταστάσεις που δεν έχει δει πιο πριν.

Έπειτα, σαν μετρικές αξιολόγησης, χρησιμοποιήσαμε την καμπύλη *Open-Set Classification*, την καμπύλη *Binary Open-Set Classification* και την μέγιστη ακρίβεια (*Maximum Accuracy*). Η *Binary Open-Set Classification* αποτελεί καμπύλη που επινοήσαμε εμείς, προσαρμόζοντας την *Open-Set Classification* στις απαιτήσεις του προβλήματός μας. Ένα ξεχωριστά σημαντικό χαρακτηριστικό στην απόδοση ενός μοντέλου, είναι η δυνατότητά του να ξεχωρίζει αν η ουσία που μελετά είναι κάποιος χημικός πρόδρομος εκρηκτικών ή ασφαλής, ανεξάρτητα από την ταυτότητα του προδρόμου, βασισμένοι σε αυτή την αντίληψη μετατρέψαμε το πρόβλημά μας σε δυαδικό πρόβλημα. Πιο συγκεκριμένα, θεωρήσαμε ένα νέο μέγεθος, το *Binary Correct Classification rate* (Equation 19), το οποίο μετράει το ποσοστό των χημικών προδρόμων που αληθώς κατηγοριοποιήθηκαν ως τέτοιοι, ανεξάρτητα αν η πρόβλεψη για την ταυτότητα του προδρόμου ήταν σωστή.

Equation 19

$$BCCR(\theta) = \frac{\left|\left\{x \mid x \in D_c \wedge \max_c P(c \mid x) \geq \theta\right\}\right|}{|D_c|}$$

Τελικά η καμπύλη *Binary Open-Set Classification* προκύπτει απεικονίζοντας το μέγεθος *Binary Correct Classification rate* σε συνάρτηση με το ποσοστό *False Positive* για διαφορετικές τιμές οριακής τιμής, σε συμφωνία με την καμπύλη *Open-Set Classification*. Τελικά η καμπύλη αυτή, μας βοηθά να αξιολογήσουμε πως ανταποκρίνεται το μοντέλο σε διαφορετικές τιμές *False Positive* στο να αναγνωρίζει αν το δείγμα υπό μελέτη είναι κάποιος χημικός πρόδρομος. Τέλος, η μετρική *Maximum Accuracy* αποτέλεσε μία μετρική που δίνει μία πιο επιδερμική, αλλά συνοπτική εκτίμηση της βέλτιστης λειτουργίας του μοντέλου. Συγκεκριμένα, δίνει το μέγιστο ποσοστό σωστών εκτιμήσεων από το μοντέλο, το οποίο επιτυγχάνεται σε μία συγκεκριμένη, βέλτιστη οριακή τιμή.

5. Αποτελέσματα και συζήτηση

5.1 Προεπεξεργασία και οπτικοποίηση δεδομένων

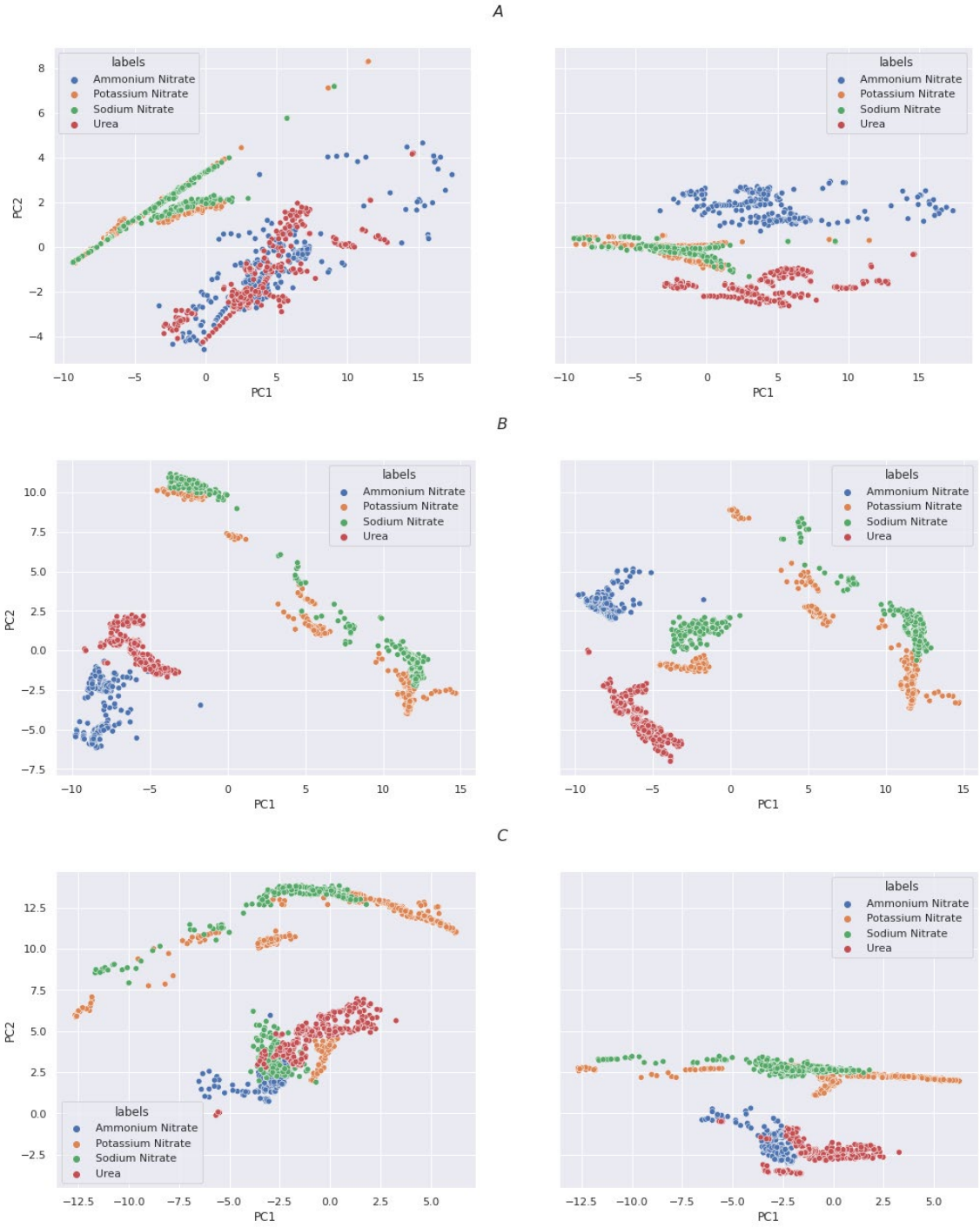

Figure 24 Οπτικοποίηση της κατανομής των δεδομένων των χημικών προδρόμων A. αρχικά, B. μετά από την εφαρμογή της SNV και C. μετά την εφαρμογή της SG 1st derivative

Στην παραπάνω εικόνα (Figure 24) βλέπουμε την κατανομή των δεδομένων πριν από την προεπεξεργασία, μετά την προεπεξεργασία με τον πρώτο τρόπο (*Standard Normal Variate*) και μετά την προεπεξεργασία με τον δεύτερο τρόπο (*Standard Normal Variate & Savitzky-Golay 1st derivative*). Πιο συγκεκριμένα στα διαγράμματα διασποράς στα αριστερά ο οριζόντιος άξονας αφορά την πρώτη συνιστώσα του PCA (την *Principal Com*ponent - PC1) και ο κάθετος την δεύτερη συνιστώσα (PC2). Αντίστοιχα, στα διαγράμματα στα δεξιά, ο οριζόντιος άξονας αφορά την πρώτη συνιστώσα και ο κάθετος την τρίτη συνιστώσα (PC3). Διαλέγουμε την πρώτη συνιστώσα σε όλα τα διαγράμματα καθότι περιέχει το μεγαλύτερο ποσοστό της πληροφορίας των δεδομένων.

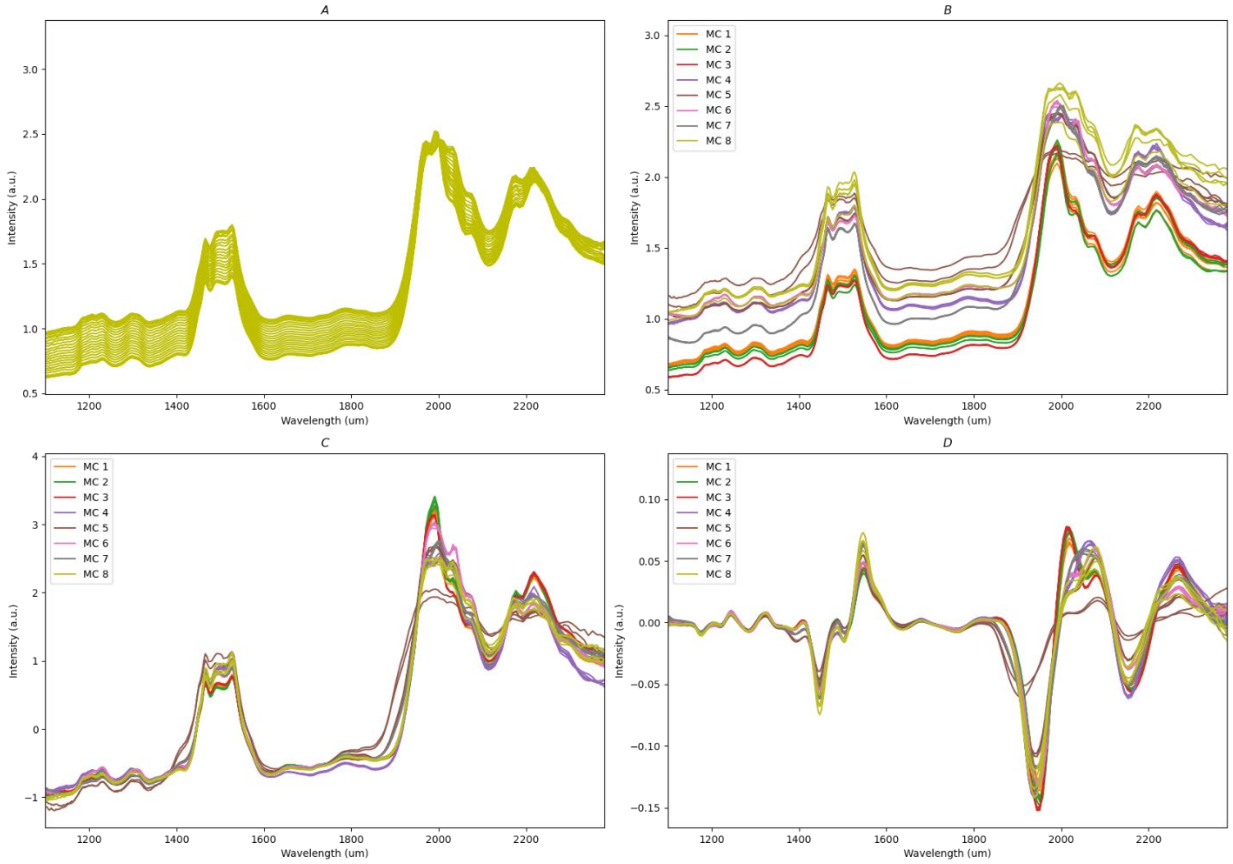

Figure 25 Στο A απεικονίζονται όλα τα φάσματα ουρίας που παράχθηκαν σε μία συνθήκη περιβάλλοντος. Στο B απεικονίζονται 5 δείγματα ουρίας ανά συνθήκη περιβάλλοντος. Η επίδραση των μεθόδων προεπεξεργασίας C. μετά από την SNV και D. μετά από την εφαρμογή της πρώτης παραγώγου

Από την απεικόνιση της κατανομής των δεδομένων, μπορούμε εύκολα να διακρίνουμε την επίδραση της μεθόδου SNV. Παρατηρούμε τα δεδομένα χωρίζονται καλύτερα σε ομάδες, συνεπώς και η κατηγοριοποίησή τους φαίνεται ευκολότερη. Αντίστοιχα, για την κατανομή των δεδομένων μετά την εφαρμογή των μεθόδων SNV και SG (*1st derivative*), η κατανομή φαίνεται πάλι να γίνεται πιο περίπλοκη. Αξίζει να σημειωθεί πως διαφορετικοί αλγόριθμοι μηχανικής μάθησης χρησιμοποιούν διαφορετικά μαθηματικά μοντέλα για την επίλυση προβλημάτων κατηγοριοποίησης, συνεπώς το γεγονός ότι γεωμετρικά έχουμε έναν καλύτερο διαχωρισμό των δεδομένων σε ομάδες, δεν εξασφαλίζει την καλύτερη αποτελεσματικότητα οποιουδήποτε αλγορίθμου μηχανικής μάθησης. Αυτή η λογική μπορεί ευκολότερα να

εφαρμοστεί για το μοντέλο *Support Vector Machines*, όπου ο τρόπος λειτουργίας του βασίζεται στην εύρεση κάποιου επιπέδου που να χωρίζει καλύτερα τις διαφορετικές κλάσεις. Εντούτοις, με μία πιο επιδερμική προσέγγιση, βλέπουμε την SNV να δημιουργεί μία πιο διακριτή κατανομή στα δεδομένα, ενώ με την εφαρμογή της πρώτης παραγώγου, οι διαφορετικές κλάσεις να έρχονται κοντύτερα και να επικαλύπτονται. Αυτό μας προδιαθέτει απαισιόδοξα για την επίδραση αυτής της μεθόδου στα μοντέλα παρακάτω, ως τρόπου προεπεξεργασίας. Στην εικόνα (Figure 25) βλέπουμε την επίδραση των διαφορετικών μεθόδων σε ένα σύνολο φασμάτων ουρίας.

5.2 Αξιολόγηση μοντέλων

Αρχικά, από τον πίνακα (Table 2) και παρατηρώντας τις τιμές *Maximum Accuracy* βλέπουμε την δυνατότητα των μοντέλων να παράγουν ποσοτικά αρκετά ικανοποιητικά αποτελέσματα. Αυτό αποτελεί επιπρόσθετα ένα ιδιαίτερα υποσχόμενο αποτέλεσμα αναφορικά με την επιτυχία του πειράματος, και την δυνατότητα της φασματοσκοπίας εγγύς υπερύθρου και του αισθητήρα που αναπτύχθηκε στην συνέχεια, να αναγνωρίζουν τους συγκεκριμένους χημικούς προδρόμους. Επίσης, μία βασική παρατήρηση αποτελεί η υπεροχή που δείχνει το μοντέλο *Random Forest* σε συνδυασμό με την μέθοδο προεπεξεργασίας των φασματικών δεδομένων *Standard Normal Variate,* πετυχαίνοντας 96.1% μέγιστη ακρίβεια και κατηγοριοποιώντας σωστά πάνω από το 80% των χημικών προδρόμων για μόλις 1% *False Positive* προβλέψεων.

Table 2 Αποτελέσματα CCR και BCCR των μοντέλων για τις διαφορετικές τιμές FPR, καθώς και η μέγιστη ακρίβεια τους

Model	*Preprocessing Method*	*CCR (%)*			*BCCR (%)*			*Max. Accuracy*
		@1	@2	@5	@1	@2	@5	
RF	SNV	**83.5 ± 1.7**	**84.7 ± 1.5**	**85.9 ± 0.4**	**83.5 ±1.8**	**84.8 ± 1.6**	**86.2 ± 0.7**	**96.1 ± 0.4**
	SNV + SG	76.6 ± 1.2	77.2 ± 1.4	80.0 ± 0.6	76.6 ± 1.2	77.2 ± 1.4	80.0 ± 0.7	94.5 ± 0.2
RF + PCA	SNV	74.1 ± 0.2	74.3 ± 0.2	75.4 ± 0.9	77.7 ± 2.9	78.9 ± 2.7	81.8 ± 2.7	93.8 ± 0.1
	SNV + SG	76.5 ± 2.1	76.8 ± 2.1	78.3 ± 2.6	76.5 ± 2.1	77.0 ± 2.2	78.9 ± 3.1	94.4 ± 0.5
SVM	SNV	30.6 ± 1.9	37.9 ± 5.9	74.7 ± 1.5	30.6 ± 1.9	37.9 ± 5.9	79.4 ± 5.0	90.7 ± 0.1
	SNV + SG	54.0 ± 3.8	61.4 ± 1.8	75.8 ± 0.3	54.0 ± 3.8	61.4 ± 1.8	75.8 ± 0.4	91.2 ± 0.1
SVM + PCA	SNV	31.7 ± 0.5	36.7 ± 2.8	76.0 ± 0.1	31.7 ± 0.4	36.7 ± 2.8	82.8 ± 1.6	90.8 ± 0.1
	SNV + SG	45.1 ± 1.6	55.5 ± 2.5	75.8 ± 0.3	45.1 ± 1.6	55.0 ± 2.5	75.8 ± 0.4	91.3 ± 0.1
HQI	SNV	71.2	72.6	73.9	71.2	72.6	73.9	92.6

Αξίζει εδώ να σημειώσουμε, πως αυτό σε καμία περίπτωση δεν αποτελεί κάποιο απόλυτο αποτέλεσμα, καθότι το σύνολο των δεδομένων και ο τρόπος που τα χωρίσαμε σε σετ εκπαίδευσης και σετ αξιολόγησης, παίζουν μεγάλο ρόλο στην απόδοση των μοντέλων. Ίσως, για παράδειγμα, συγκεκριμένα αρνητικά δείγματα στο σετ αξιολόγησής μας να επηρεάζουν αρνητικά κάποιο άλλο μοντέλο και στον πραγματικό κόσμο, όπου η κατανομή των δεδομένων θα είναι διαφορετική, κάποιο άλλο μοντέλο με παραπλήσια αποτελέσματα να δείξει σχετική υπεροχή. Το ίδιο θα μπορούσε να συμβεί, εάν αποφασίζαμε να αποκλείσουμε από το σετ εκπαίδευσης κάποια διαφορετικά δύο σετ μετρήσεων. Ωστόσο, σε κάθε

περίπτωση, είναι χρήσιμη η παρατήρηση της αποτελεσματικής λειτουργίας του αλγορίθμου *Random Forest.* Στην αντίθετη κατεύθυνση βρίσκεται η αποδοτικότητα του αλγορίθμου *Support Vector Machine,* ο οποίος για χαμηλές τιμές *False Positive* έδειξε αδυναμία στην αναγνώριση και σωστή κατηγοριοποίηση των προδρόμων, με μόλις 37.9% ικανότητα να αναγνωρίζει αν μία ουσία είναι πρόδρομος για 1% *False Positive* προβλέψεις. Η συμπεριφορά των SVM μοντέλων μπορεί να γίνει καλύτερα κατανοητή με την παρατήρηση των καμπυλών (Figure 26). Παρατηρούμε πως μέχρι κάποια τιμή *False Positive rate,* περίπου 4%, τα μοντέλα διατηρούν χαμηλές τιμές, ενώ στην συνέχεια παρατηρείται ραγδαία αύξηση. Με αφορμή αυτή την παρατήρηση, μπορούμε να υπογραμμίσουμε την χρησιμότητα των δύο καμπυλών *Open-Set Classification Binary* και *Open-Set Classification.* Βλέπουμε, πως για το SVM μοντέλο που έχει αρκετά χαμηλή απόδοση σε χαμηλές τιμές *False Positive*, η μέγιστη ακρίβεια είναι πάνω από 90%. Συνεπώς αν βασίζαμε την αξιολόγησή μας μόνο στην μέγιστη ακρίβεια ενός μοντέλου, μπορεί να καταλήγαμε σε λάθος συμπεράσματα, και λανθασμένη επιλογή μοντέλου.

Για μία μέση τιμή *False Positive 5%,* παρατηρούμε πως προηγείται το μοντέλο RF, ακολουθεί το μοντέλο SVM και τέλος έρχεται ο αλγόριθμος HQI. Ήδη βλέπουμε πως οι αλγόριθμοι μηχανικής μάθησης μπορούν να έχουν καλύτερα αποτελέσματα από την τεχνική HQI. Επίσης αναλογιζόμενοι την δυνατότητα των αλγορίθμων μηχανικής μάθησης να βελτιώνονται συνεχώς, είτε ρυθμίζοντας καλύτερα τις υπερπαραμέτρους τους, είτε επιλέγοντας διαφορετικούς αλγορίθμους, είτε αναβαθμίζοντας το σετ εκπαίδευσης, συμπεραίνουμε πως η μηχανική μάθηση μπορεί να έχει υψηλότερη αποδοτικότητα στο πρόβλημά μας. Σχετικά με την επίδραση της επίδρασης της πρώτης παραγώγου στην προεπεξεργασία των δεδομένων ως προς την αποδοτικότητα των μοντέλων συνολικά δεν μπορούμε να εξάγουμε κάποιο συμπέρασμα. Στο μοντέλο RF, για παράδειγμα, είχαμε πτώση της απόδοσής του όταν εκπαιδεύτηκε και αξιολογήθηκε σε δεδομένα που είχαν προεπεξεργαστεί με τον δεύτερο τρόπο προεπεξεργασίας.

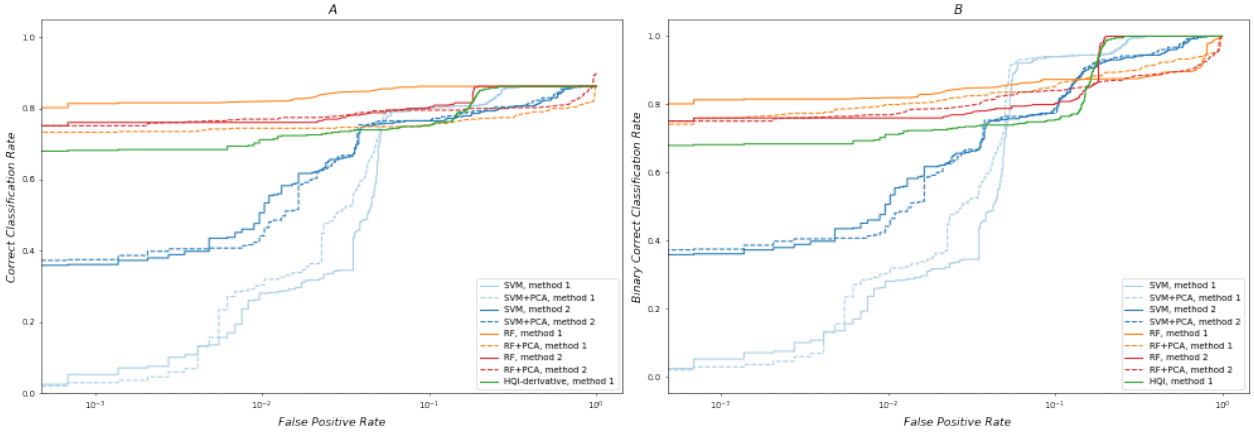

Figure 26 *A Open-Set Classification Rate curve, B Binary Open-Set Classification Rate curve για τα μοντέλα που αναπτύχθηκαν*

Αντίθετα, όταν το μοντέλο RF σε συνδυασμό με το PCA, είχε βελτίωση στην απόδοσή του όταν χρησιμοποίησε δεδομένα που είχαν παραγωγιστεί. Ωστόσο, όσον αφορά το κλειστό πρόβλημα κατηγοριοποίησης, στην καμπύλη *Open-Set Classification,* πάνω δεξιά φαίνεται το μοντέλο RF + PCA σε συνδυασμό με τον δεύτερο τρόπο προεπεξεργασίας να έχει την υψηλότερη αποτελεσματικότητα. Μπορούμε να υποθέσουμε ότι συνολικά στην ανάλυση φασματοσκοπικών δεδομένων στο εγγύς υπέρυθρο, η παραγώγιση είναι αρκετά βοηθητική, αλλά στην προκειμένη, όπου τα φασματικά σήματα του νιτρικού

νατρίου και καλίου ήταν ασθενή, η πρώτη παράγωγός τους είχε σαν αποτέλεσμα ακόμα πιο εξασθενημένα σήματα και κατά συνέπεια, η εξόρυξη της πληροφορίας υπήρξε δυσκολότερη.

Αξίζει να αναφέρουμε πως τα αποτελέσματα προέκυψαν από την επανάληψη του πειράματος δέκα φορές για κάθε μοντέλο. Πιο συγκεκριμένα, οι αλγόριθμοι μηχανικής μάθησης που χρησιμοποιήσαμε συνδέονται με κάποια στοχαστικότητα κατά την ανάπτυξή τους. Συνεπώς, εάν επαναληφθεί το πείραμα αναπτύσσοντας το μοντέλο και επιλέγοντας τις ίδιες υπερπαραμέτρους, μπορεί να υπάρχει κάποια απόκλιση στα αποτελέσματα. Σε αυτό το γεγονός, άλλωστε, ακουμπάει η τυπική απόκλιση που αναφέρεται σε κάθε μέγεθος στον πίνακα. Έτσι, λοιπόν, οι υποθέσεις και οι παρατηρήσεις που κάναμε μπορούν να θεωρηθούν αξιόπιστα για την συγκεκριμένη πειραματική διάταξη και διαδικασία.

6. Συμπεράσματα

Συνοψίζοντας, σύμφωνα με την παρούσα εργασία, η αναγνώριση και ταυτοποίηση του νιτρικού αμμωνίου, νιτρικού καλίου, ουρίας και νιτρικού νατρίου είναι δυνατή με την χρησιμοποίηση ενός σχετικά οικονομικού και φορητού φασματογραφικού αισθητήρα. Ο συγκεκριμένος αισθητήρας, ως αισθητήρα FT-NIR, συνδέεται με ιδιαίτερα χρήσιμα χαρακτηριστικά για τις απαιτήσεις της συγκεκριμένης εφαρμογής. Πιο συγκεκριμένα, προσφέρει γρήγορη σάρωση τους δείγματος και εξαγωγή αποτελέσματος, επιτρέπει την εξέταση δειγμάτων χωρίς προηγουμένως να έχουν επεξεργαστεί, όπως επίσης η σύνδεσή του με την οπτική ίνα προσφέρει την δυνατότητα στον χειριστή να κρατάει κάποια ασφαλή απόσταση από την ουσία. Η εργασία αυτή εστίασε στην ανάπτυξη του αλγορίθμου, με βάση τον οποίο ο αισθητήρας θα εξάγει συμπεράσματα. Πιο συγκεκριμένα, κάνοντας συστηματικές δοκιμές διαφορετικών αλγορίθμων με δύο τρόπους προεπεξεργασίας των φασματικών δεδομένων, καταλήξαμε σε ορισμένες χρήσιμες παρατηρήσεις. Τα μοντέλα που αναπτύχθηκαν χρησιμοποίησαν τους αλγορίθμους *Random Forest* και *Support Vector Machine*, οι οποίοι συνδυάστηκαν και με τον *Principal Component Analysis* για μείωση της διαστατικότητας των δεδομένων, καθώς και τον *Hit Quality Index*, βασισμένο σε έναν συντελεστή συσχέτισης των φασμάτων με μία δεδομένη, εξατομικευμένη βιβλιοθήκη. Οι δύο τρόποι επεξεργασίας αφορούν την μέθοδο *Standard Normal Variate*, και την SNV σε συνδυασμό με την μέθοδο *Savitzky-Golay* (*1st derivative, 19-point smoothing window*). Από τα αποτελέσματα, το μοντέλο RF σε συνδυασμό με την μέθοδο SNV φάνηκε το πλέον αποδοτικό, επιτυγχάνοντας περισσότερο από 83% *Correct Classification rate* για μόλις 1% *False Positive rate*. Αντίθετα, ο SVM δεν ανταποκρίθηκε καλά στην απαίτηση για χαμηλό *False Positive rate*. Επίσης, όπως ήταν αναμενόμενο, μπορέσαμε να διακρίνουμε μία υπεροχή των αλγορίθμων μηχανικής μάθησης έναντι του HQI. Συνολικά, η παρούσα έρευνα απέδειξε πως ο φασματογραφικός αισθητήρας θα μπορούσε να αποτελέσει ένα αποτελεσματικό εργαλείο στην ανίχνευση των συγκεκριμένων χημικών προδρόμων εκρηκτικών. Έπειτα, στην εργασία αυτή καταγράφεται και αναλύεται εκτενώς ο τρόπος ανάπτυξης μοντέλων που επεξεργάζονται φασματικά δεδομένα στο εγγύς υπέρυθρο.

Σίγουρα, η παρούσα έρευνα μπορεί να αποτελέσει βάση για μία μελλοντική έρευνα που αφορά την ανίχνευση χημικών προδρόμων εκρηκτικών ουσιών. Κάποιοι επιπρόσθετοι πρόδρομοι που θα μπορούσαν να αναλυθούν και ενδεχομένως να ανιχνευθούν είναι τα *hexamine* και sodium nitrite καθώς περιέχουν δεσμούς που προμηνύουν υπερτονικά και ζώνες συνδυασμού στο εγγύς υπέρυθρο. Επίσης, προτείνουμε την συλλογή περισσότερων δειγμάτων έτσι ώστε να χρησιμοποιηθούν νευρωνικά δίκτυα και βαθιά μάθηση, τα οποία συχνά έχουν μεγαλύτερη αποτελεσματικότητα. Έπειτα, αρκετό ενδιαφέρον προσφέρει η παρατήρηση του φορητού φασματογραφικού αισθητήρα σε πραγματικές εφαρμογές, καθότι τα *open-set* προβλήματα αποτελούν σοβαρή πρόκληση για τα μοντέλα αναγνώρισης και κατηγοριοποίησης.

Βιβλιογραφία

1. Gunasekaran, S., "Nondestructive Food Evaluation: Techniques to Analyze Properties and Quality," CRC Press (2018).
2. Ozaki, Y., Huck, C., Tsuchikawa, S., Engelsen, S. B., "Near-Infrared Spectroscopy," Springer Nature Singapore Pte Ltd. (2021).
3. Saptari, V., "Fourier Transform Spectroscopy Instrumentation Engineering," SPIE Press (2004).
4. Griffiths, P.R., de Haseth, J.A., "Fourier Transform Infrared Spectrometry," John Wiley & Sons, Inc. (2007).
5. Luypaert, J., Massart, D. L., Vander Heyden, Y., "Near-infrared spectroscopy applications in pharmaceutical analysis,"Talanta 72,865-883 (2007).
6. Vékey, K., Telekes, A., Vertes, A., "Chemoinformatics-multivariate mathematical-statistical methods for data evaluation," Medical Applications of Mass Spectrometry, 141-169 (2008).
7. Awad, M., Khanna, R., "Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers," Apress (2015).
8. https://learnche.org/pid/latent-variable-modelling/principal-component-analysis/hotellings-t2-statistic
9. www.tibco.com/reference-center/what-is-a-random-forest
10. Breiman, L., "Random Forests," Machine Learning 45, 5–32 (2001).
11. Noble, W. S., "What is a support vector machine?," Nature Biotechnology 24, 1565–1567(2006).
12. Cortes, C., Vapnik, V., "Support-Vector Networks," Machine Leaming 20, 273-297 (1995).
13. https://ankitnitjsr13.medium.com/math-behind-svm-support-vector-machine-864e58977fdb
14. https://datascience.stackexchange.com/questions/25977/using-svm-classifier-with-c-0-and-c-infinity-what-would-be-the-effect-on-classi
15. https://www.researchgate.net/figure/Non-linear-classifier-using-Kernel-trick-16_fig4_340610860
16. Tharwat, A., "Classification assessment methods," Applied Computing and Informatics 17, 168-192 (2021).
17. https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/
18. Geng, C., Huang, S. J., Chen, S., "Recent Advances in Open Set Recognition: A Survey," IEEE TPAMI (2020).
19. Dhamija, A. R., Günther, M., Boult, T.E., "Reducing Network Agnostophobia,"32nd Conference on Neural Information Processing Systems (2018).
20. Xu, X., Xie, L., Ying, Y., "Factors influencing near infrared spectroscopy analysis of agro-products: a review," Front. Agr. Sci. Eng.,105–115(2019).
21. Rinnan, Å., van den Berg, F., Engelsen, S. B., "Review of the most common pre-processing techniques for near-infrared spectra," Trends in Analytical Chemistry, 1201-1222(2009).
22. Owen, A. J., "Uses of Derivative Spectroscopy," Agilent Technologies Application Note (1995).
23. Blanco, M., Romero, M. A., "NIR libraries in the pharmaceutical industry: a solution for identity confirmation," Analyst 126, 2212-2217 (2001).
24. http://ftirsearch.com/Help/algo.asp

25. Zapata, F., Ferreiro-González, M., García-Ruiz, C., “Interpreting the near infrared region of explosives,” Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 204,81-87 (2018).
26. Zapata, F., García-Ruiz, C., “Chemical Classification of Explosives,” Critical Reviews in Analytical Chemistry, 656-673 (2021).
27. WCO Programme Global Shield (PGS) - E-book No .01, Updated as on 18.09.2015
28. Liu, C., Yang, S. X., Deng, L., “Determination of internal qualities of Newhall navel oranges based on NIR spectroscopy using machine learning,” Journal of Food Engineering 161, 16-23(2015).
29. Liu, J., Sun, S., Tan, Z., Liu, Y., “Nondestructive detection of sunset yellow in cream based on near-infrared spectroscopy and interval random forest,” Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 242 (2020).
30. Cruz-Tirado, J. P., da Silva Medeiros, M. L., Barbin, D. F., “On-line monitoring of egg freshness using a portable NIR spectrometer tandem with machine learning,” Journal of Food Engineering 306 (2021).
31. Teye, E., Amuah, C. L Y., McGrath, T., Elliott, C., “Innovative and rapid analysis for rice authenticity using hand-held NIR spectrometry and chemometrics,” Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 217, 147-154(2019).
32. Liu, C. M., Han, Y., Min, S. G., Jia, W., Meng, X., Liu, P. P., “Rapid qualitative and quantitative analysis of methamphetamine, ketamine, heroin, and cocaine by near-infrared spectroscopy”, Forensic Science International, 162-168(2018).
33. http://www.arcoptix.com/nir_ir_lamp.htm
34. https://www.oceaninsight.com/products/fibers-and-probes/probes/reflectionbackscatter-probes/nanoq-rprobe-600-vis-nir/?qty=1
35. https://www.hamamatsu.com/eu/en/product/type/C15511-01/index.html
36. https://ark.intel.com/content/www/us/en/ark/products/95061/intel-nuc-kit-nuc7i5bnk.html
37. Malinin, A., Band, N., Gal, Y., Gales, M. J. F., Ganshin, A., Chesnokov, G., Noskovet al., “Shifts: A Dataset of Real Distributional Shift Across Multiple Large-Scale Tasks,”35th Conference on Neural Information Processing Systems (2021).
38. Verleysen, M., François, D., “The Curse of Dimensionality in Data Mining and Time Series Prediction,” Springer-Verlag Berlin Heidelberg (2005).

Portable FT-NIR spectroscopic sensor for detection of chemical precursors of explosives using advanced prediction algorithms

A. M. Grammatikaki [b], A. Raptakis [a,b], L. Gounaridis [*a,b], A. Athanasopoulos [b], D. Gounaridis [a,b], P. Groumas [a,b], A. Dadoukis [a], E. Maltezos [a], L. Karagiannidis [a], E. Ouzounoglou [a], A. Amditis [a], H. Avramopoulos [a,b], C. Kouloumentas [a,b]

[a]Institute Communication and Computer System, 42 Patission, Athens, Greece

[b]National Technical University of Athens, 9 Iroon Polytechneiou, Zografou, Greece

ABSTRACT

Near-infrared (NIR) spectroscopy has acquired widespread adoption in various sectors as a result of its benefits over other analytical techniques, the most notable of which is the ability to record spectra for solid samples without any prior manipulation. Furthermore, advances in instrumentation have led to the creation of compact and high-speed spectrometers that can be used in a variety of scenarios, including hazardous materials identification. Fourier Transform NIR (FT-NIR) technology is one of the most useful tools for onsite analysis of chemical and biological substances. Herein, we propose a compact, portable FT-NIR spectroscopic sensor for field measurements, based on commercial broadband light source and spectrometer for detection of chemical precursors of explosives. We mainly focus on four compounds, ammonium nitrate, potassium nitrate, sodium nitrate and urea, some of the best-known chemical precursors of explosives with NIR content. A customized spectral library is constructed, including the forementioned substances under different environmental conditions. We emphasize on two basic factors that can affect the NIR spectra: the relative humidity and the ambient temperature. For the unknown spectrum identification, we evaluate prediction models which involve the use of Random Forest and Support Vector Machine, as well as the Hit Quality Index (HQI) value. The FT-NIR spectroscopic sensor additionally includes an integrated communication module that provides measurement spectra and results to a novel edge computing platform, called DECIoT. We demonstrate the operation of the FT-NIR spectroscopic sensor in real settings under humidity, straight sunlight, and temperature fluctuations, achieving maximum accuracy of 0.96.

Keywords: Precursors of explosives, portable FT-NIR sensor, near-infrared, Random Forest, Support Vector Machine, Hit Quality Index, DECIoT, Open-Set Classification Rate curve

1. INTRODUCTION

Over the years, NIR spectroscopy has been demonstrating remarkable progress in both spectral analysis treatment and instrumentation and has been versatile regarding its applications. Fourier Transform (FT) technology is one common tool in NIR spectroscopy and almost all benchtop spectrometers have adopted the principle. The NIR spectrum is extracted through a simultaneous incidence of all wavelengths on the detector, followed by Fourier transformation. FT NIR spectrometers are often preferred, as they accomplish short scan times, significantly low signal-to-noise ratio and high precision. The aforementioned features, as well as the NIR advantages, make FT NIR a very useful tool regarding the onsite analysis and identification of chemical and biological substances. Previously, these material analysis applications had to be done in a lab and necessitated the removal of a component of the object in issue. These measurements may now be made in situ without removing a sample, giving real, nondestructive analysis regardless of the form, position, or orientation of the object in question, thanks to the introduction of portable FTIR. Over the last 10 years a lot of products have been fabricated toward this direction like [1], [2], [3] and [4].

Near-infrared (NIR) spectroscopy is based on the absorption of the light in the region of 800 – 2500 nm (12,500 – 4000 cm^{-1}). The most prominent absorption bands come from overtones and combination bands of fundamental vibrational motions, originating in the mid-infrared. These bands principally involve O-H, N-H and C-H vibrational

transitions and typically are broad, overlapping and less intense than the corresponding fundamental absorption bands. Accordingly, multivariate analysis methods from chemometrics, machine learning (ML) and neural networks are often required for extracting useful information from the spectra. NIR spectroscopy comes with numerous advantages. First, it is a non-invasive and in situ analysis method. Second, NIR analysis can be applied to materials in various physical conditions and even examine the sample internally. Third, an optical fiber can be attached to the spectrometer, which enables remote examination of the sample. Especially, compared to infra-red, NIR light-fibers are cheaper and more robust [5], [6]. The advantages of NIR spectroscopy are of great importance, nevertheless, there is the difficulty in analyzing and evaluating the NIR spectra, thus major efforts are being devoted to expanding our knowledge on multivariate analysis methods and their contribution to different NIR spectroscopy applications. A big part of these methods involves machine learning methods combined with preprocessing techniques [7], [8], [9], [10], [11].

In this study, we attempted to explore the NIR spectra of four precursors of explosives, Ammonium Nitrate, Potassium Nitrate, Sodium Nitrate and Urea, and propose a compact, portable FT NIR sensor combined with an AI model, able to identify these compounds on-scene. The aforementioned precursors come in a form of white powder and have information in NIR spectrum [12]. Moreover, they consist some of the most famous precursors of explosives, monitored by Programme Global Shield by World Customs Organization, with Ammonium Nitrate already facing some restrictions regarding its manufacture, sale and use [13]. While the model could potentially encounter with unknown and unseen samples as well, this certain task exceeds the limits of a close-set classification problem, and the generalization performance of the model gets disputed. In other words, the set of samples, which will be seen by the model during training (training set) is derived from a different label space than the set of samples, which will finally evaluate the model (test set). The possibility of encountering with an unknown class defines an open-set recognition/classification problem [14], [15]. One common way to deal with these tasks is to consider a threshold value, which will define the rejection class. Thus, if the prediction of the model comes with a probability lower than the threshold value, the sample is classified as 'negative'. After developing a set of models, a mindful selection of the models' evaluation protocol is vital. In our case, we decided to focus on the need of the application to minimize the false positive alerts i.e., the predictions according to which a negative sample is classified as any precursor. Finally, since the precursors, as well as other unknown substances probable for examination can be hazardous when contacting the skin or inhaled, the ability of NIR spectrometers for non-contact analysis, offers a safe measurement environment for the user.

2. MATERIALS AND METHODS

2.1. Precursors and Negative Challenge samples

The dataset consists of NIR reflectance spectra acquired from 2911 precursors and negative challenge samples. The data can be divided into two sets of samples, one containing the precursors and a second one composed by the negative challenge samples. The first set (D_c) consists of 1456 samples from Ammonium Nitrate, Potassium Nitrate, Sodium Nitrate and Urea, 364 samples per class, measured in 8 different measurement conditions (MC). It has been suggested that ambient temperature and humidity can affect the NIR spectra [16], therefore these have been the main factors that constitute the different measurement conditions. In particular, each MC correspond to temperature and relative humidity as follows: MC1 ~ (20°C, 18%RH), MC2 ~ (7°C, 25%RH), MC3 ~ (40°C, 26%RH), MC4 ~ (38°C, 60%RH), MC5 ~ (14°C, 90%RH), MC6 ~ (-18°C, 60%RH), MC7 ~ (23°C, 78%RH), MC8 ~ (21°C, 33%RH). Moreover, moisture content shows information in NIR spectrum, which can definitely alter the intended measurement, under specific conditions [17], [18]. The second set (D_u) consists of 1455 negative challenge samples, derived from 12 different compounds, which form the "unknown class". This set was composed and used only for evaluating the final model, while the chemical substances were chosen with regard to their chemical similarity with the precursors, as well as the frequency with which people carry and use these compounds. Ammonium Cerium Nitrate and Diazolidinyl Urea were the most challenging negative compounds, as they show very similar chemical structure with Ammonium Nitrate and Urea respectively. We focused on constructing a relatively challenging negative challenge testing set, therefore 265 samples of Ammonium Cerium Nitrate in 7 different conditions and 200 samples of Diazolidinyl Urea in 5 different measurement conditions were obtained. Additionally, 990 samples from 10 different commonly found substances were collected. These substances include salt, sugar, flour, mastic, baking soda, ammonium carbonate (cooking ammonia), mahalepi, paracetamol, talc, and synthetic vanillin.

2.2. Instruments and settings

2.2.1 The spectroscopic sensor

The spectroscopic sensor is a portable, autonomous device that analyzes the NIR spectrum of chemicals utilizing a revolutionary Microelectromechanical system (MEMS) based spectrometer. The FT-NIR sensor consists of different components that are presented in *Figure 1*. Light is incident on the FT-NIR engine from a stabilized Tungsten-Halogen light source coupled to a 6:1 reflection fiber/probe (a bundle of seven optical fibers with one at the center for receiving light and six at the perimeter for irradiation). The diffuse reflection light is measured by the FT-NIR engine using a reference plate positioned at the tip of the reflection probe.

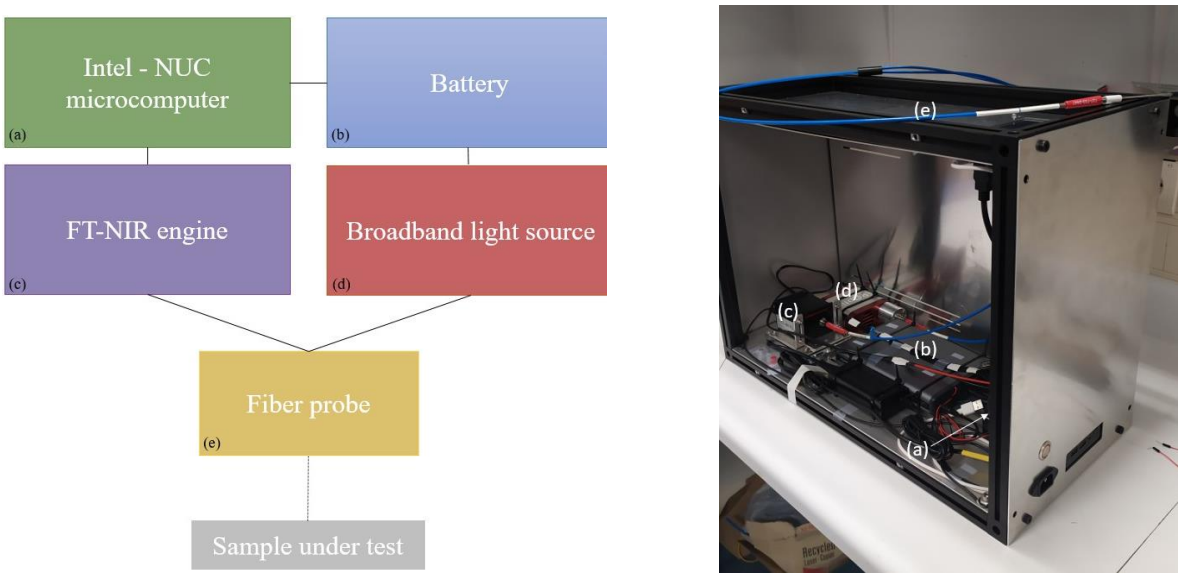

Figure 1. FT-NIR sensor's components. a) Intel NUC microcomputer, b) battery, c) FT-NIR engine, d) broadband light source, and e) fiber probe.

The core of this sensor is the FT-NIR engine spectroscopic module [19] (*Figure 1*c), which is responsible for receiving and processing the spectral response of the diffusely reflected light of a sample under NIR illumination. The FT-NIR engine is portable and thus could be carried in one hand. In a palm-sized box, a Michelson optical interferometer and control circuit are housed. By attaching a microcomputer via USB, the absorbance spectrum of the sample under test can be extracted providing vital information about its chemical signature. It can be used for on-site, real-time measurements and continuous monitoring without bringing the measurement sample into the analysis room. The optical interferometer in the FT-NIR engine has a movable and a fixed mirror that leverages MEMS technology. Spectrum acquisition with high wavelength accuracy is possible due to the built-in semiconductor VCSEL, that monitors the adjustable mirror position. The engine includes assessment software that allows the measurement parameters definition, along with other features such as data storage and graphical representation. The stabilized Tungsten-Halogen light source [20] (*Figure 1*d) emits a 20 mW blackbody radiation spectrum with a wavelength range of 400 to 4000 nm. So, it is suited for integration into optical measurement setup since the FT-NIR engine operates from 1100 to 2500 nm. A Tungsten-Halogen light source is a form of an incandescent lamp. The current flows through the Tungsten-Halogen filament during operation, which heats up to roughly a couple of 2850 Kelvin. At this temperature, tungsten generates visible and infrared light, generating a light source that can be approximated by a black body radiator. In *Figure 1*e the reflection/backscatter probe is presented [21]. It consists of 6 peripheral fibers through which the beam from the light source hits the sample and one fiber in the center of the probe, which receives the diffusely reflected light from the sample and propagates it to the FT-NIR engine for further processing. The wavelength range that is propagating in the fiber probe ranges from 400 to 2500 nm. The fiber core size is 600 μm, and the total length of the probe is 2 m. The probe ferrule material at the end of the fiber is stainless steel which extend the operation temperature of the probe from -65 to 300º C. As a microcomputer, the Intel® NUC is used [22] (*Figure 1*a). Intel NUC acts also as a power supplier for the FT-NIR engine via a USB port. Moreover, it is responsible for executing the software which controls the FT-NIR engine as well as the operation of the Python-based algorithm which is analyzed

in this work. All the different components are sharing a common housing. In the same housing there is a battery which acts as the power supplier of the FT-NIR sensor, which provides the capability of autonomous operation in the field.

2.2.2 The network

In this study, the FT-NIR system has been integrated with a novel edge computing platform called Distributed Edge Computing Platform (DECIoT) [23], [24]. In this context, the FT-NIR system is able to share in real-time chemical precursors detection alerts to other systems or platforms. Thus, the FT-NIR system has a great potential to be exploited from decision-makers under a common operational picture (COP) aspect. In general, the DECIoT platform is able to address among others the problem of gathering, filtering and aggregating data, interacts with the IoT devices, provides security and system management, provides alerts and notifications, executes commands, stores data temporarily for local persistence, transforms/process data and in the end exports the data in formats and structures that meet the needs of other platforms. This whole process is being done by using open source microservices that are state-of-the-art in the area of distributed edge IoT solutions. The DECIoT is based on the EdgeX foundry open-source framework [25]. The EdgeX foundry is considered in the bibliography as a highly flexible and scalable edge computing framework facilitating the interoperability between devices and applications at the IoT edge such industries, laboratories, and datacenters [26], [27]. The DECIoT platform consist of multiple layers and each layer contains multiple microservices. The communications between the micro-service within the same or different layers can be done either directly with the use of REST APIs or with the use of a message bus that follows a pub/sub mechanism. Both are being exploited in FT-NIR system. DECIoT consists of a collection of reference implementation services and software development kit tools. The different layers that have been adapted in the FT-NIR sensor are a) the Device Service, b) the Core Services, c) the Support Services and d) the Application Services layer. A detailed documentation and implementation of the DECIoT has been provided in [24].

The crucial information associated with the chemical precursor alert *Figure 2* are sent from the Application Service to the specified topic of Apache Kafka (this is considered here as the middleware of a smart city platform) where: i) “source” a description of the system/sensor that produces the data, ii) “event_id” the unique identification of the event, iii) “time” the date and timestamp for the creation of the event, iv) “localization” the coordinates (latitude, longitude, elevation) to which the FT-NIR system make measurements (pre-defined for cases that the FT-NIR goes to specific location according to stakeholder’s command or dynamically changed in case that a GNSS receiver is integrated with the FT-NIR system), v) “event_type” the type of the event, and vi) “metadata “ the data that further describe the event.


```
{
  "source": "FTNIR",
  "event_id": "70eee248-7d09-430-bcd8-b08a69bc2313",
  "time": "2022-02-14T14:41:51+02:00",
  "localization": {
    "lat": 37.979158,
    "lon": 23.780678,
    "elevation": 10
  },
  "event_type": "FTNIR.jdl_1.event.ammonium_nitrate_detected",
  "metadata": "ICCS, FTNIR, ammonium_nitrate_detected"
}
```


Figure 2. Crucial information associated with the chemical precursor detection alert sent from the Application Service of DECIoT to Apache Kafka

2.3. Data Analysis and Problem Formulation

2.3.1. Exploratory analysis and data preprocessing

Pretreatment of the NIR spectral data is an integral part of the analysis aiming to eliminate artifacts caused by undesired physical phenomena during sampling. In particular, the interactions between NIR electromagnetic radiation and sample particles create undesired scattering effects, which in combination with changes in the temperature, particle size and path length generate unwanted spectral variations and baseline shifts [5]. NIR pre-processing methods can be divided into scatter correction and derivative methods. We used Standard Normal Variate (SNV), a common scatter correction

technique for eliminating multiplicative interferences of scatter and particle size [28], [29], which centers and scales the data. It removes baseline variation, while the shape of the spectra remains the same. Derivative spectroscopy is widely used, as it efficiently removes the baseline and enhances the resolution of overlapping peaks. First and second order derivatives are mostly used, however in cases where the spectral signals are weak, the derivatives are not always a suitable tool, due to the fact that signal-to-noise ratio (SNR) decreases as higher order derivatives are used [30]. Savitzky-Golay (SG) is a derivation technique, which embodies the smoothing of the new signal with the aim of eliminating the SNR reduction. We computed Savitzky-Golay first order derivative with a 19-datapoint smoothing window.

2.3.2. Model development

Classification models were built based on two different approaches. The first approach regards the construction of a customized spectral library and the correlation between the unknown sample and the elements of the library [31], [32], [33]. The library includes one sample per precursor and per measurement condition from the training set, which are considered reference spectra. Furthermore, we used the Hit Quality Index (HQI) for characterizing the correlation between the unknown and the reference spectrum [34]. While HQI is often a part of commercial spectrometer's built-in algorithm, regarding customized applications ML models are usually preferred, as HQI might fail distinguish very similar spectra [35]. HQI along with an optimal threshold form the final model of the first approach. More specifically, for every sample input, the search algorithm comes with a numeric value, which reflects the similarity between the sample and a reference spectrum from the library. In a close-set classification problem, the reference spectrum yielding the highest correlation value would equate with the system's prediction. Considering an open-set recognition problem we use a threshold value, with the purpose of rejecting the negative samples. Threshold defines the lowest value required to assign a given spectrum to a specific class. Once the maximum correlation value between the given spectrum and spectra in the library exceeds the threshold, then it can be assigned to the corresponding class [31].

In the second approach we used machine learning algorithms for building, applying and evaluating different ML models. Support Vector Machine (SVM) [36] and Random Forest (RF) [37][38] models were developed. SVM is regarded as one of the most robust, thus commonly used ML classification algorithms. Considering the data in a high-dimensional space, SVM constructs a set of hyperplanes to separate the different classes, creating different subspaces corresponding to every class. RF algorithm is an ensemble method, i.e., a combination of decision trees. A multitude of decision trees is built during training, and the final prediction of the model arises from the set of the individual decision trees' predictions. We chose this algorithm for two reasons. First, it shows a stable error rate as the number of the features increases, which concerning our case, it is a promising attribute. Second, it intrinsically avoids overfit, which in small datasets can be a threat. Additionally, the models were combined with Principal Component Analysis (PCA), which reduced the dimensions of the collected data, using the first two hundred Principal Components (PCs), as they expressed more than 99% of the sample's variance. Lastly, since we had a relatively small number of samples, we chose the threshold method to deal with the rejection class [15]. Namely, a threshold value was applied to each of the models' higher score values to indicate whether the examined sample belongs to the predicted class D_c or to the unknown class D_u.

2.3.3. Evaluation protocol

A model dealing with an open-set recognition problem has a two-fold goal. Firstly, it needs to reject the samples which belong to an unknown class D_u, and secondly to classify the samples which belong to a correct class D_c. This makes evaluating the prediction model more complex. The metric we chose for estimating the performance of our models was Open-Set Classification Rate curve (OSCR) [15]. According to this evaluation method, the samples are divided into the samples from the known classes D_c and the samples from the unknown classes D_u. Given a threshold θ, the following two rates are computed: Correct Classification Rate (CCR) (Eq. 1) is the fraction of the samples where the maximum probability is greater than θ and corresponds to the correct class ĉ. False Positive Rate (FPR) (Eq. 2) is the fraction of samples from D_u, where the maximum probability is greater than θ and corresponds to any known class in D_c.

$$CCR(\theta) = \frac{\left| \left\{ x \mid x \in D_c \wedge \arg\max_c P(c \mid x) = \hat{c} \wedge P\left(\hat{c} \mid x \right) \geq \theta \right\} \right|}{\left| D_c \right|} \quad (1)$$

$$FPR(\theta) = \frac{\left|\left\{x \mid x \in D_u \wedge \max_c P(c \mid x) \geq \theta\right\}\right|}{\left|D_u\right|} \tag{2}$$

We compute these rates for several threshold values, specifically, for every threshold value that provokes different CCR and FPR values. Finally, we plot CCR values versus FPR. In view of the fact that we attempt to minimize the possibility of a wrong classification of a negative sample, we consider as our final metrics, the CCR rates which correspond to three low FPR values: 1%, 2% and 5%. Supplementally, two more metrics were used for the final evaluation. First, considering the task as a binary classification problem, where the positive class is any class D_c and the negative class is the unknown class D_u, we extract a binary Open-Set Classification Rate curve. Again, the horizontal axis regards the FPR values, while the vertical axis regards the Binary Correct Classification Rate (BCCR) (Eq. 3), i.e., the fraction of the samples belonging to any known class D_c and the maximum probability, greater than θ corresponds to a known class in D_c. In accordance with OSCR, we take the BCCRs for the FPR values: 1%, 2% and 5%. This metric focalizes on the ability of the model to correctly assign a negative sample to the unknown class and correctly alert for a precursor, regardless its class. Lastly, we consider the maximum accuracy, which is accomplished with an optimal threshold, constituting the most generalized metric and regards the maximum percentage of correct classification that can be achieved.

$$BCCR(\theta) = \frac{\left|\left\{x \mid x \in D_c \wedge \max_c P(c \mid x) \geq \theta\right\}\right|}{\left|D_c\right|} \tag{3}$$

Moreover, we focused on the partitioning of the data in a training and a test. The objective is to evaluate the models on a realistic and challenging test set. In this application we address the problem for which the model will be possibly deployed in a different data setting than it is trained on. The specified arbitrarily change in the data distribution is mentioned as distribution shift [39], [40] and the corresponding data, out-of-distribution data. Due to this effect, the model's generalization ability is disputed, and the evaluation of the model's performance becomes demanding. Attempting to face the challenge, we excluded the measurement conditions 3 and 5 from the training set and only included them in the test set. Furthermore, the test set included 8% of remaining precursors' samples and the negative challenge samples, while the training set contained the complementary 92% of the precursors' samples from the remaining six measurement conditions.

3. RESULTS AND DISCUSSION

3.1. Preprocessing

For every precursor, samples from 8 different measurement conditions were collected. *Figure 3*A, shows the unprocessed raw data from Urea, regrading MC 4 (38°C, 60%RH). In *Figure 3*B, the unprocessed raw data for all MCs are depicted, but for simplicity reasons, only five spectra per MC are taken. Each color indicates a different measurement condition. For simplicity reasons, only five spectra per measurement condition were plotted. Additive baseline shifts and further slight spectral dissimilarities can be noticed between spectra of a single measurement condition. We attributed this effect to three reasons. First, during sampling, we scanned the compound from different distances between 0.2 cm and 2 cm, which could be the case in the actual on-scene analysis. Smaller distances between the fiber probe and the sample induce higher signal intensity and on the contrary, bigger distances lead to lower signal intensity. Second, the substance was irradiated from slightly different angles, and at different spots, leading to various path lengths, due to the stochastic scattering of the light in the compound. Third, during the collection of the spectral data, changes in ambient temperature and relative humidity were taking place. For eliminating the effects above, we applied SNV to the data (*Figure 3*C), as well as SNV followed by SG 1st derivative with a 19-datapoint smoothing window (*Figure 3*D). We can observe the decrease in the intensity of the spectra after the application of the 1st derivative, which can be a problem for the models distinguishing the different classes, considering the already low spectral intensity of Potassium and Sodium Nitrate.

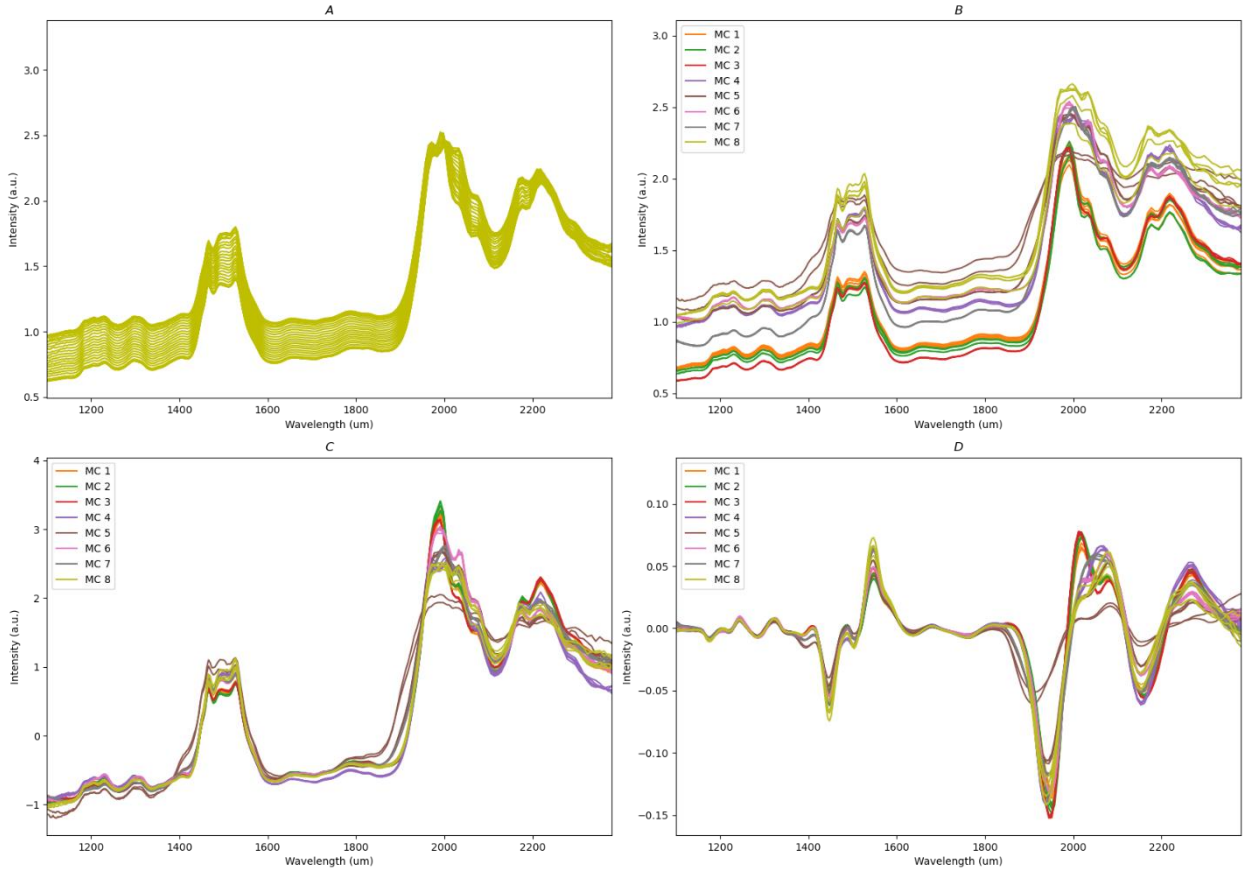

Figure 3. A) All spectra from Urea measured under measurement condition No.4. Effect of preprocessing in NIR spectral data of Urea, 5 spectra per measurement condition; B) raw, C) after SVN preprocessing, D) after SVN preprocessing, followed by Savitzky-Golay smoothing with a first-order derivative.

3.2. Model performance evaluation

Several experiments were performed with the aim of finding the optimal model for our dataset. We used RF and SVM algorithms, alone and in combination with PCA, where the 200 first components were used, as they explained more than 99% of the variability. PCA algorithm is commonly utilized against the curse of dimensionality [41], which means than high dimensional datasets often come with several problems, as for example overfitting, which could possibly be a threat in our small dataset. Furthermore, we evaluated the models separately for every preprocessing method. Regarding the training of the ML models, we used the Grid Search algorithm combined with 3-Fold Cross-Validation method for defining the optimal hyperparameters for each of the RF and SVM algorithms. *Table 1* presents the Correct Classification Rates and Binary Correct Classification Rate at False Positive Rates of 1%, 2% and 5%, as well as the maximum accuracy for the open-set problem and for all models including the HQI algorithm. RF and SVM models are built with a degree of stochasticity, therefore we computed 10 times each model and extracted the evaluation metrics, their mean value, as well as the standard deviation. We have emboldened the best performance for each evaluation metric.

3.3 Discussion

Supplementarily to Table 1 and for a better overview of the evaluations' results, *Figure 4* depicts Open-Set Classification Rate and Binary Open-Set Classification Rate curves.

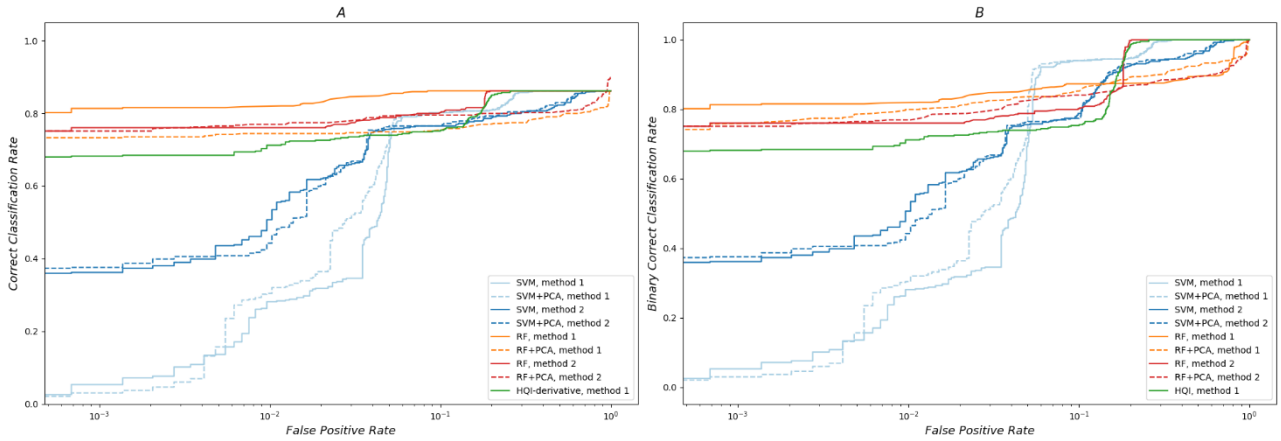

Figure 4. Evaluation curves, where method 1 corresponds to SVN preprocessing and method 2 to SVN followed by SG 1st derivative; A. Open-Set Classification Rate curve, B. Binary Open-Set Classification Rate curve

Table 1. Performance of multivariate classification models

Model	*Preprocessing Method*	*CCR (%)*			*BCCR (%)*			*Max. Accuracy*
		@1	@2	@5	@1	@2	@5	
RF	SNV	**83.5 ± 1.7**	**84.7 ± 1.5**	**85.9 ± 0.4**	**83.5 ±1.8**	**84.8 ± 1.6**	**86.2 ± 0.7**	**96.1 ± 0.4**
	SNV + SG	76.6 ± 1.2	77.2 ± 1.4	80.0 ± 0.6	76.6 ± 1.2	77.2 ± 1.4	80.0 ± 0.7	94.5 ± 0.2
RF + PCA	SNV	74.1 ± 0.2	74.3 ± 0.2	75.4 ± 0.9	77.7 ± 2.9	78.9 ± 2.7	81.8 ± 2.7	93.8 ± 0.1
	SNV + SG	76.5 ± 2.1	76.8 ± 2.1	78.3 ± 2.6	76.5 ± 2.1	77.0 ± 2.2	78.9 ± 3.1	94.4 ± 0.5
SVM	SNV	30.6 ± 1.9	37.9 ± 5.9	74.7 ± 1.5	30.6 ± 1.9	37.9 ± 5.9	79.4 ± 5.0	90.7 ± 0.1
	SNV + SG	54.0 ± 3.8	61.4 ± 1.8	75.8 ± 0.3	54.0 ± 3.8	61.4 ± 1.8	75.8 ± 0.4	91.2 ± 0.1
SVM + PCA	SNV	31.7 ± 0.5	36.7 ± 2.8	76.0 ± 0.1	31.7 ± 0.4	36.7 ± 2.8	82.8 ± 1.6	90.8 ± 0.1
	SNV + SG	45.1 ± 1.6	55.5 ± 2.5	75.8 ± 0.3	45.1 ± 1.6	55.0 ± 2.5	75.8 ± 0.4	91.3 ± 0.1
HQI	SNV	71.2	72.6	73.9	71.2	72.6	73.9	92.6

Generally, Random Forest combined with SNV preprocessing showed the best performance, achieving 96.1% maximum accuracy and over 83% Correct Classification Rate and Binary Correct Classification Rate even for the lowest False Positive Rates. On the other hand, SVM performed poorly at low FPR values, while having a rapid boost and high efficiency at higher FPRs. For 5% FPR, RF model was found the most suitable, SVM performed adequately and lastly HQI yielded the poorest score. We notice that in our case PCA does not boost the performance overall, and in the cases where it does, the enhancement is light. Furthermore, SG had different impacts on the different algorithms. A probable reason for SG performing poorly is the weak intensity of the Potassium Nitrate's and Sodium Nitrate's spectra. While we observe that for the models trained with SG preprocessed data, the performance is invariably altered in CCR and BCCR for low FPR values, from the observation of the OSCR curve, it becomes evident that RF + PCA model combined with SNV and SG display the highest accuracy in the close-set problem. As expected, the best performing ML model outperforms the HQI search method, while this finding is consistent with the well-known importance of ML methods on NIR applications.

Additionally, considering the different evaluation metrics, it becomes notable that in open-set tasks, OSCR and Binary OSCR curves describe the model's performance much more efficiently. Although maximum accuracy gives quantitively a good overview of the model's ability to make correct predictions, it can be misleading regarding the utility and application of the model. For example, we can observe from the graphs (*Figure 4*) and the table (*Table 1*), that SVM does reach a high accuracy and CCR for a threshold value, however, considering the demand for low FPR, SVM would

not be a prudent choice. Moreover, the closeness of CCR and BCCR indicates that the negative class is in fact the main challenge for all the models and spotlights the demandingness of the open-set classification problems.

4. CONCLUSION

This study demonstrates that the on-site detection of Ammonium Nitrate, Potassium Nitrate, Sodium Nitrate and Urea in the 1100-2500 nm region is possible, using a rather low-cost, hand-held spectroscopic system coupled with a suitable AI model. Additionally, this constitutes a precious tool, on the grounds that the method is non-invasive, thus no previous sample pretreatment is needed, intrinsically safer and FTIR technique provides rapid predictions. The systematic trial of different preprocessing methods (SG 1st derivative, SNV) and modelling with RF and SVM multivariate models as well as using the HQI search algorithm, indicated that RF model combined with SNV preprocessing method showed superiority with more than 80% correct classification rate for 1% of false positives. Therefore, the proposed sensor has the potential of a non-destructive, rapid and robust detection of the aforementioned precursors, adopting also an edge computing framework. Furthermore, the present study lays the foundations for some future work on precursors' detection in the NIR spectrum. Namely, we recommend the collection of additional samples and the utilization of a Neural Network to achieve similar or higher CCRs in lower FPRs. Moreover, we could see the potential of detecting more precursors, as for example sodium nitrite and hexamine, which contain O-H, N-H and C-H bonds. Lastly, since the evaluation of a model for an open-set recognition application is a demanding task, an interesting future study might involve testing the spectroscopic system in the real-world, with a view to conducting a thorough investigation regarding the final evaluation of the model.

ACKNOWLEDGMENTS

This work is a part of the S4AllCities project. This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 883522. Content reflects only the authors' view, and the Research Executive Agency (REA) / European Commission is not responsible for any use that may be made of the information it contains.

REFERENCES

[1] https://www.agilent.com/en/product/molecular-spectroscopy/ftir-spectroscopy/ftir-compact-portable-systems/4300-handheld-ftir#features

[2] https://www.agilent.com/en/product/molecular-spectroscopy/ftir-spectroscopy/ftir-compact-portable-systems/4100-exoscan-series-ftir-handheld

[3] https://www.thermofisher.com/order/catalog/product/TRUDEFENDERFTX

[4] https://www.bruker.com/en/products-and-solutions/cbrne-detectors/ims/roadrunner.html

[5] Ozaki, Y., Huck, C., Tsuchikawa, S., Engelsen, S. B., "Near-Infrared Spectroscopy," Springer Nature Singapore Pte Ltd. (2021).

[6] Luypaert, J., Massart, D. L., Vander Heyden, Y., "Near-infrared spectroscopy applications in pharmaceutical analysis," Talanta 72, 865-883 (2007).

[7] Liu, C., Yang, S. X., Deng, L., "Determination of internal qualities of Newhall navel oranges based on NIR spectroscopy using machine learning," Journal of Food Engineering 161, 16-23 (2015).

[8] Liu, J., Sun, S., Tan, Z., Liu, Y., "Nondestructive detection of sunset yellow in cream based on near-infrared spectroscopy and interval random forest," Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 242 (2020).

[9] Cruz-Tirado, J. P., da Silva Medeiros, M. L., Barbin, D. F., "On-line monitoring of egg freshness using a portable NIR spectrometer tandem with machine learning," Journal of Food Engineering 306 (2021).

[10] Teye, E., Amuah, C. L Y., McGrath, T., Elliott, C., "Innovative and rapid analysis for rice authenticity using hand-held NIR spectrometry and chemometrics," Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 217, 147-154 (2019).

[11] Liu, C. M., Han, Y., Min, S. G., Jia, W., Meng, X., Liu, P. P., "Rapid qualitative and quantitative analysis of methamphetamine, ketamine, heroin, and cocaine by near-infrared spectroscopy", Forensic Science International, 162-168 (2018).

[12] Zapata, F., Ferreiro-González, M., García-Ruiz, C., "Interpreting the near infrared region of explosives," Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 204, 81-87 (2018).

[13] Hotchkiss, P. J., "Explosive Threats: The Challenges they Present and Approaches to Countering Them," Springer Nature Switzerland AG (2018)

[14] Geng, C., Huang, S. J., Chen, S., "Recent Advances in Open Set Recognition: A Survey," IEEE TPAMI (2020).

[15] Dhamija, A. R., Günther, M., Boult, T.E., "Reducing Network Agnostophobia," 32nd Conference on Neural Information Processing Systems (2018).

[16] Xu, X., Xie, L., Ying, Y., "Factors influencing near infrared spectroscopy analysis of agro-products: a review," Front. Agr. Sci. Eng., 105–115 (2019).

[17] Büning-Pfaue, H., "Analysis of water in food by near infrared spectroscopy," Food Chemistry 82, 107-115 (2003).

[18] Maeda, H., Ozaki, Y., Tanaka, M., Hayashi, N., Kojima, T., "Near infrared spectroscopy and chemometrics studies of temperature- dependent spectral variations of water: relationship between spectral changes and hydrogen bonds," J. Near Infrared Spectrosc. 3, 191–201 (1995).

[19] https://www.hamamatsu.com/eu/en/product/type/C15511-01/index.html

[20] http://www.arcoptix.com/nir_ir_lamp.htm

[21] https://www.oceaninsight.com/products/fibers-and-probes/probes/reflectionbackscatter-probes/nanoq-rprobe-600-vis-nir/?qty=1

[22] https://ark.intel.com/content/www/us/en/ark/products/95061/intel-nuc-kit-nuc7i5bnk.html

[23] Maltezos, E., Karagiannidis, L., Dadoukis, A., Petousakis, K., Misichroni, F., Ouzounoglou, E., Gounaridis, L., Gounaridis, D., Kouloumentas, C., Amditis, A., "Public safety in smart cities under the edge computing concept," 2021 IEEE International Mediterranean Conference on Communications and Networking (MeditCom), 88-93 (2021).

[24] Maltezos, E., Lioupis, P., Dadoukis, A., Karagiannidis, L., Ouzounoglou, E., Krommyda M., Amditis, A., "A video analytics system for person detection combined with edge computing," Computation 10, 35, (2022).

[25] T. L. Foundation, https://www.edgexfoundry.org (2022).

[26] Jin S. et al., "Video Sensor Security System in IoT Based on Edge Computing," International Conference on Wireless Communications and Signal Processing (WCSP) (2020)

[27] Villali, V., Bijivemula, S., Narayanan, S. L., Mohana Venkata Prathusha, T., Krishna Sri, M. S., Khan, A., "Open-source Solutions for Edge Computing," 2nd International Conference on Smart Electronics and Communication (ICOSEC) (2021).

[28] Rinnan, Å., van den Berg, F., Engelsen, S. B., "Review of the most common pre-processing techniques for near-infrared spectra," Trends in Analytical Chemistry, 1201-1222 (2009).

[29] Barnes, R. J., Dhanoa, M. S., Lister, S. J., "Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra," Applied Spectroscopy, 772-777 (1989).

[30] https://www.whoi.edu/cms/files/derivative_spectroscopy_59633940_175744.pdf

[31] Blanco, M., Romero, M. A., "Near-infrared libraries in the pharmaceutical industry: a solution for identity confirmation," Analyst 126, 2212-2217 (2001).

[32] http://www.thermo.com.cn/resources/201211/141413246.pdf

[33] https://tools.thermofisher.com/content/sfs/brochures/TS-Pharma-pvalue-HQI.pdf

[34] Rodriguez, J. D., Westenberger, B. J., Buhse, L. F., Kauffman, J. F., "Standardization of Raman spectra for transfer of spectral libraries across different instruments," Analyst 136, 4232-4240 (2011)

[35] https://www.americanpharmaceuticalreview.com/1504-White-Papers-Application-Notes/147135-Pros-and-Cons-of-Using-Correlation-Versus-Multivariate-Algorithms-for-Material-Identification-via-Handheld-Spectroscopy/

[36] Noble, W. S., "What is a support vector machine?," Nature Biotechnology 24, 1565–1567 (2006)

[37] Cortes, C., Vapnik, V., "Support-Vector Networks," Machine Leaming 20, 273-297 (1995).

[38] Breiman, L., "Random Forests," Machine Learning 45, 5–32 (2001).

[39] Malinin, A., Band, N., Gal, Y., Gales, M. J. F., Ganshin, A., Chesnokov, G., Noskov et al., "Shifts: A Dataset of Real Distributional Shift Across Multiple Large-Scale Tasks," 35th Conference on Neural Information Processing Systems (2021).

[40] Bickel, S., Brückner, M., Scheffer, T., "Discriminative Learning Under Covariate Shift," Journal of Machine Learning Research 10 (2009).

[41] Verleysen, M., François, D., "The Curse of Dimensionality in Data Mining and Time Series Prediction," Springer-Verlag Berlin Heidelberg (2005).